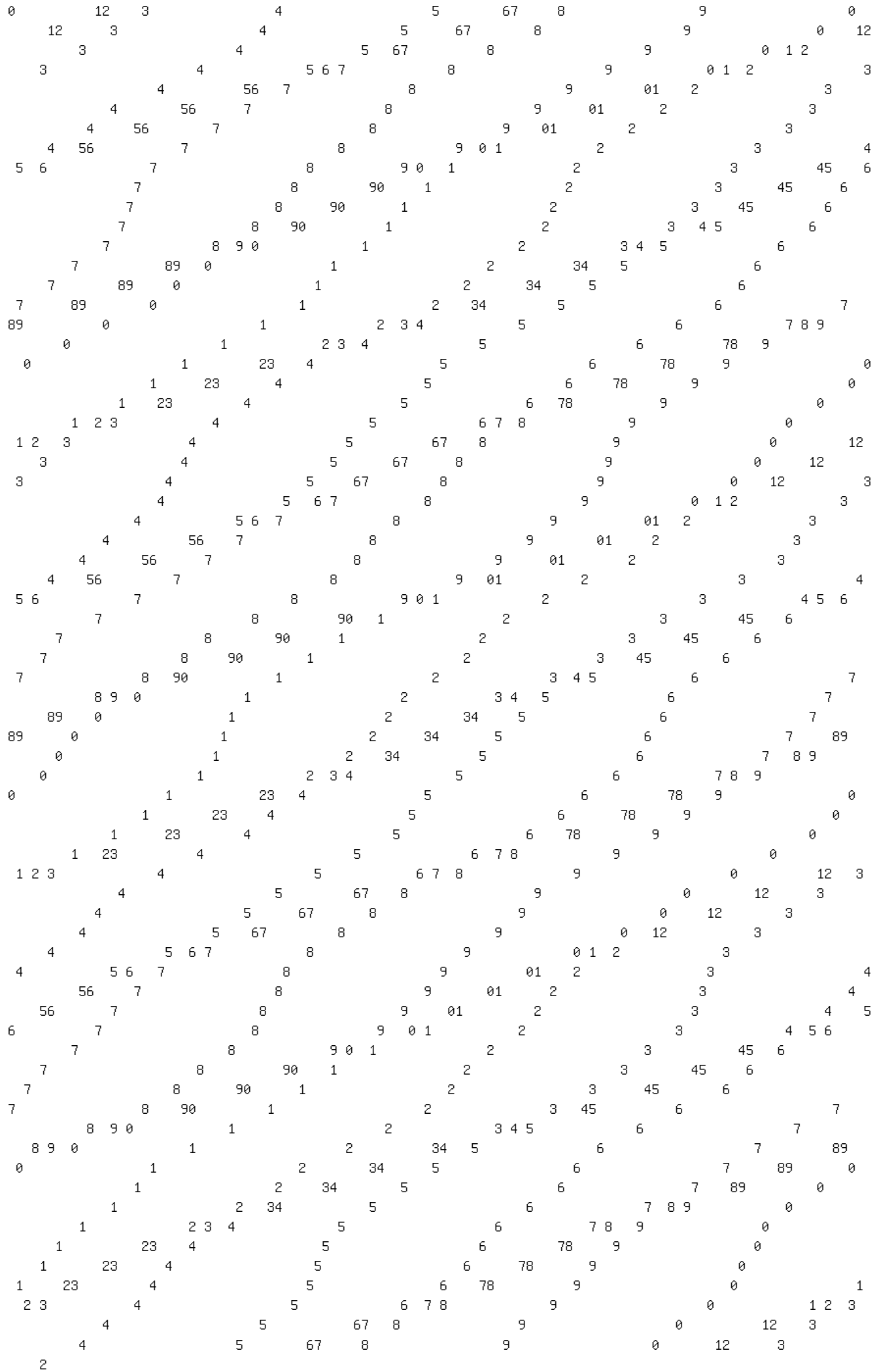
data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read nd learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, nform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, infor , read and learn data workers write, perform, clean, inform, read and learn data workers w ite, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and l arn data workers write, perform, clean, inform, read and learn data workers write, p rform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and earn data workers write, perform, clean, inform, read and learn data wor ers write, perform, clean, inform, read and learn data workers write, perform, clean, inf rm, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers wri e, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data wor ers write, perform, clean, inform, read and learn data workers write, perform, cl an, inform, read and learn data workers write, perform, clean, inform, read and earn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn dat workers write, perform, clean, inform, read and learn data workers write, p rform, clean, inform, read and learn data workers write, perform, clean, in orm, read and learn data workers write, perform, clean, inform, read and l arn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data work rs write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, nform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read nd learn data workers write, perform, clean, inform, read and l arn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and l arn data workers write, perform, clean, inform, read nd learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, nform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data work rs write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn data workers write, perform, clean, inform, read and learn

What
can
humans learn from humans
humans learn with machines
machines learn from machines
machines learn with humans
humans learn from machines
machines learn with machines
machines learn from humans
humans learn with humans
? ? ?

Data Workers, une exposition au Mundaneum à Mons du 28 mars au 28 avril 2019.



Á PROPOS

Data Workers est une exposition d'œuvres algolittéraires,visible au Mundaneum à Mons du jeudi 28 mars jusqu'au dimanche 28 avril 2019. Elle expose des histoires racontées d'un point de vue 'narratif algorithmique'. L'exposition est une création des membres d'Algolit, un groupe bruxellois impliqué dans la recherche artistique sur les algorithmes et la littérature. Chaque mois, ils se réunissent pour expérimenter avec du code et des textes F/LOSS. Certaines oeuvres sont réalisés par des étudiants de Arts² et des participants externes à l'atelier sur le machine learning et le texte organisé par Algolit en octobre 2018 au Mundaneum.

Les entreprises créent des intelligences artificielles pour servir, divertir, enregistrer et connaître les humains. Le travail de ces entités machiniques est généralement dissimulé derrière des interfaces et des brevets. Dans l'exposition, les conteurs algorithmiques quittent leur monde souterrain invisible pour devenir des interlocuteurs.

Les 'data workers' opèrent dans des collectifs différents. Chaque collectif représente une étape dans le processus de conception d'un modèle d'apprentissage automatique : il y a les Écrivains, les Nettoyeurs, les Informateurs, les Lecteurs, les Apprenants et les Oracles. Les robots donnent leurs voix à la littérature expérimentale, les modèles algorithmiques lisent des données, transforment des mots en nombres, calculent des modèles et traitent en boucle de nouveaux textes et ceci à l'infini.

L'exposition met au premier plan les 'data workers' qui ont un impact sur notre vie quotidienne, mais qui sont difficiles à saisir ou à imaginer. Elle établit un lien entre les récits sur les algorithmes dans les médias grand public et les histoires racontées dans les manuels techniques et les articles universitaires. Les robots sont invités à dialoguer avec les visiteurs humains et vice versa. De cette façon, nous pourrions comprendre nos raisonnements respectifs, démystifier nos comportements, rencontrer nos personnalités multiples et valoriser notre travail collectif. C'est aussi un hommage aux nombreuses machines que Paul Otlet et Henri La Fontaine ont imaginées pour leur Mundaneum, en montrant leur potentiel mais aussi leurs limites.

Data Workers est une création de Algolit.

Oeuvres de: Cristina Cochior, Gijs de Heij, Sarah Garcin, An Mertens, Javier Lloret, Louise Dekeuleneer, Florian Van de Weyer, Laetitia Trozzi, Rémi Forte, Guillaume Slizewicz, Michael Murtaugh, Manetta Berends, Mia Melvær.

Une co-production de: Arts², Mundaneum, Constant.

Avec le soutien de: Fédération Wallonie-Bruxelles, Arts Numériques, Passa Porta, Ugent, DHuF - Digital Humanities Flanders et the Distributed Proofreading Project.

Remerciements: Mike Kestemont, Michel Cleempoel, Donatella Portoghese, François Zajéga, Raphaèle Cornille, Vincent Desfromont, Kris Rutten, Anne-Laure Buisson, David Stampfli.

AU MUNDANEUM

À la fin du 19ème siècle, deux jeunes juristes belges, Paul Otlet (1868-1944), 'père de la documentation', et Henri La Fontaine (1854-1943), homme d'État et prix Nobel de la paix, créent le Mundaneum. Le projet vise à rassembler toute la connaissance du monde et à la classer à l'aide du système de Classification décimale universelle (UDC) qu'ils inventent. Au début, il s'agit d'un Bureau des institutions internationales dédié à l'échange international des connaissances. Au XXe siècle, le Mundaneum devient un centre universel de documentation. Ses collections sont constituées de milliers de livres, journaux, revues, documents, affiches, plaques de verre et cartes postales indexés sur des millions de fiches référencées. Les collections sont exposées et conservées dans différents bâtiments à Bruxelles, dont le Palais du Cinquantenaire. Le reste des archives n'est transféré à Mons qu'en 1998.

Sur base du Mundaneum, les deux hommes conçoivent une ville du monde pour laquelle Le Corbusier réalise des maquettes et des plans. L'objectif de la Ville du Monde est de rassembler, au niveau mondial, les institutions du travail intellectuel : bibliothèques, musées et universités. Mais le projet n’est jamais réalisé, souffrant de sa propre utopie. Le Mundaneum est le résultat du rêve visionnaire d’une infrastructure pour l'échange universel des connaissances. Il atteint des dimensions mythiques à l'époque. Lorsqu'on observe les archives qui ont été concrètement développées, cette collection est plutôt éclectique et spécifique.

Les intelligences artificielles se développent aujourd'hui en faisant apparaître des rêves d'universalité et de la production des connaissances. En les étudiant, nous nous sommes rendus compte que les rêves visionnaires de leurs créateurs sont bien présents dès leur développement dans les années 1950. Aujourd'hui, leurs promesses ont également atteint des dimensions mythiques. Lorsqu'on observe leurs applications concrètes, la collection d'outils est réellement innovante et fascinante, mais en même temps, tout aussi éclectique et spécifique. Pour Data Workers, Algolit a combiné certaines de ces applications avec 10 % des publications numérisées du Bureau des Institutions Internationales. Ainsi et de façon poétique, nous espérons ouvrir une discussion à propos des machines, des algorithmes et des infrastructures technologiques.

--- Pourquoi des récits contextualisés? ---

Lors des réunions mensuelles d'Algolit, nous étudions des manuels et expérimentons avec des outils d'apprentissage automatique pour le traitement de texte. Et nous partageons aussi beaucoup, beaucoup d'histoires. Avec la publication de ces histoires, nous espérons recréer un peu de cette atmosphère. Les histoires existent également sous forme de podcasts qui peuvent être téléchargés à partir du site http://www.algolit.net.

--- Nous créons des œuvres 'algolittéraires' ---

Le terme 'algolittéraire' vient du nom de notre groupe de recherche Algolit. Nous existons depuis 2012 en tant qu'initiative de Constant, une organisation oeuvrant dans les médias et les arts basée à Bruxelles. Nous sommes des artistes, des écrivains, des designers et des programmeurs. Une fois par mois, nous nous rencontrons pour étudier et expérimenter ensemble. Notre travail peut être copié, étudié, modifié et redistribué sous la même licence libre. Vous trouverez toutes les informations sur le site http://www.algolit.net.

L'objectif principal d'Algolit est d'explorer le point de vue du conteur algorithmique. Quelles nouvelles formes de narration rendons-nous possibles en dialoguant avec ces agents machiniques ? Les points de vue narratifs sont inhérents aux visions du monde et aux idéologies. Don Quichotte, par exemple, a été écrit d'un point de vue omniscient à la troisième personne, montrant la relation de Cervantes à la tradition orale. La plupart des romans contemporains utilisent le point de vue de la première personne. Algolit souhaite parler au travers des algorithmes et vous montrer le raisonnement de l'un des groupes les plus cachés de notre planète.

Écrire dans ou par le code, c'est créer de nou-

gage humain de façon inattendue. Mais les techniques d'apprentissage automatique ne sont accessibles qu'à ceux qui savent lire, écrire et exécuter du code. La fiction est un moyen de combler le fossé entre les histoires qui existent dans les articles scientifiques, les manuels techniques, et les histoires diffusées par les médias, souvent limitées aux reportages superficiels et à la fabrication de mythes. En créant des œuvres algolittéraires, nous offrons aux humains une introduction aux techniques qui co-modèlent leur vie quotidienne.

--- Qu'est-ce que la littérature ? ---

Algolit comprend la notion de littérature comme beaucoup d'autres auteurs expérimentaux : elle inclut toute la production linguistique, du dictionnaire à la Bible, de l'œuvre entière de Virginia Woolf à toutes les versions des Conditions d'utilisation publiées par Google depuis son existence. En ce sens, le code de programmation peut aussi être de la littérature. Le collectif Oulipo, acronyme d'Ouvroir de Littérature Potentielle, est une grande source d'inspiration pour Algolit. Oulipo a été créé à Paris par les écrivains Raymond Queneau et François Le Lionnais. Ils ont ancré leur pratique dans l'avant-garde européenne du XXe siècle et dans la tradition expérimentale des années 60. Pour Oulipo, la création de règles devient la condition permettant de générer de nouveaux textes, ou ce qu'ils appellent la littérature potentielle. Plus tard, en 1981, ils ont également créé ALAMO - Atelier de Littérature Assistée par la Mathématique et les Ordinateurs.

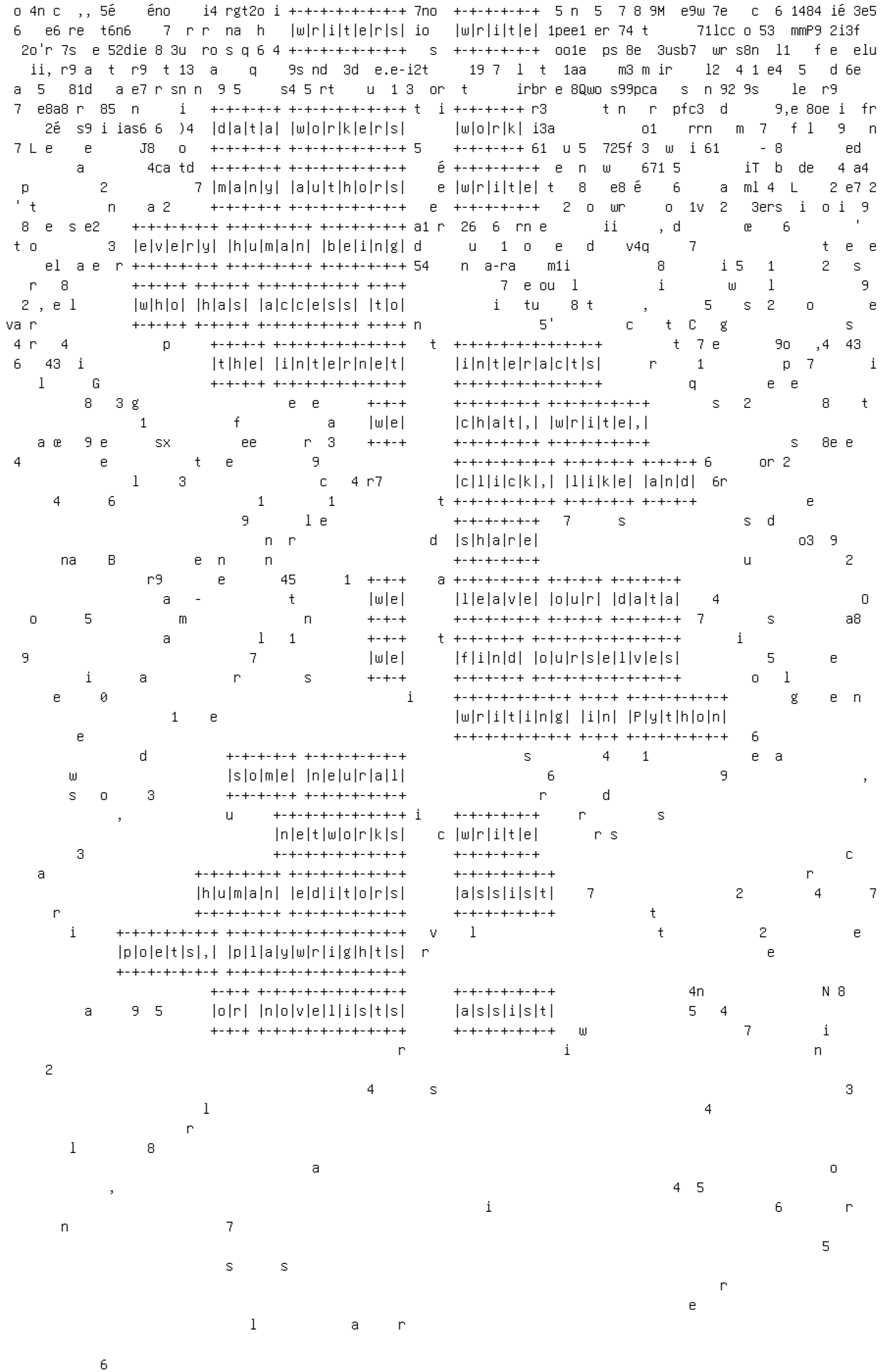
--- Une différence importante ---

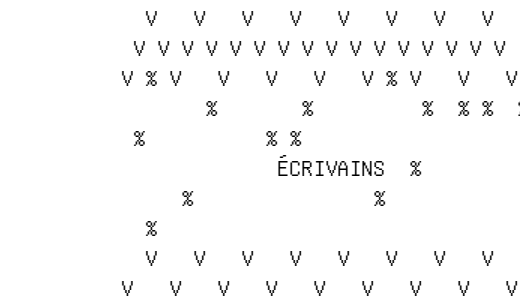
Alors que l'avant-garde européenne du XXe siècle poursuivait l'objectif de rompre avec les conventions, les membres d'Algolit cherchent à rendre les conventions visibles.

J'écris : Je vis dans mon journal, je l'investis, je le traverse. (Espèces d'espaces. Journal d'un usager de l'espace, Galilée, Paris, 1974)

Cette citation de Georges Perec dans Espèces d'espaces pourrait être reprise par Algolit. Il ne s'agit pas des conventions de la page blanche et du marché littéraire, comme Georges Perec l'a fait. Nous faisons référence aux conventions qui restent souvent cachées derrière les interfaces et les brevets. Comment les technologies sont-elles conçues, mises en œuvre et utilisées, tant dans les universités que dans les entreprises ? Nous proposons des histoires qui révèlent le système hybride complexe qui rend possible l'apprentissage automatique. Nous parlons des outils, des logiques et des idéologies derrière les interfaces. Nous examinons également qui produit les outils, qui les met en œuvre et qui crée et accède aux grandes quantités de données nécessaires au développement de machines de prédiction. On pourrait dire, en un clin d'œil, que nous sommes les collaborateurs de cette nouvelle tribu d'hybrides humain-robot.

writers write writers write writers write writers write writers write writers write writ
rs write writers write writers write writers write writers write
writers write writers write writers write writers write
writers write writers write writers write writers write
writers write writers write writers write
writers write writers write writers write
writers write writers write writers write
writers write writers write
writers write writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writers write
writers write writ
rs write writers write
writers write writers write
writers write
writers write writers write
writers write writer
write writers write
writers write writ
rs write writers write
writers write
writers write writers write
writers write
writers write w
iters write writers write
writers write
writers write
writers write writers write
writers write
writers write
writers write
writers write writer
write writers write
writers write
writers write
writers write
writers write
writers write
writers write writ
rs write writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write
writers write

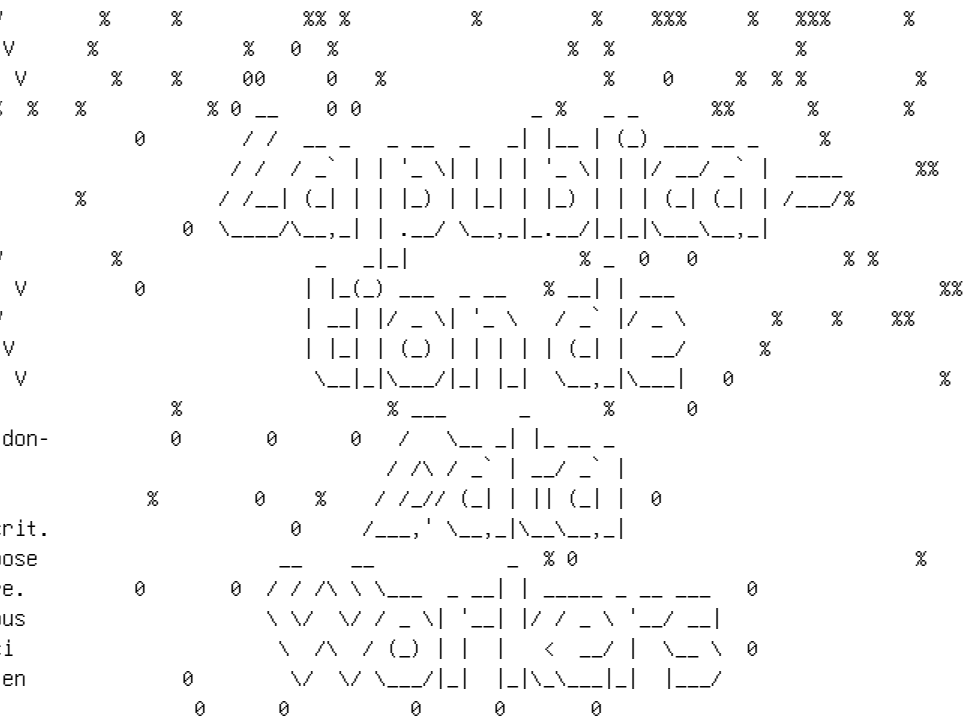




Les Data Workers ont besoin de données pour travailler. Dans le contexte d'Algolit, celles-ci prennent la forme du langage écrit. L'apprentissage automatique repose sur de nombreux types d'écriture. Les auteurs humains écrivent sous forme de publications. Celles-ci sont organisées en archives et en cours de numérisation. Mais il existe d'autres types d'écriture. On pourrait dire que chaque être humain avec un accès à Internet devient un écrivain lorsqu'il interagit avec des algorithmes. En ajoutant des commentaires, en écrivant des mails ou des articles Wikipédia, en cliquant et en aimant.

Les algorithmes d'apprentissage automatique ne sont pas critiques : ils prennent tout ce qu'on leur donne, peu importe le style d'écriture, le CV de l'auteur ou ses fautes d'orthographe. D'ailleurs, plus il y a d'erreurs, mieux c'est : la variété leur apprend à anticiper les textes inattendus. Les auteurs humains quant à eux ne sont souvent pas conscients de ce qui advient de leur travail.

La plupart des textes que nous utilisons sont en anglais, certains en français, d'autres en néerlandais. Souvent, nous nous retrouvons à écrire en Python, le langage de programmation que nous utilisons. Les algorithmes peuvent aussi être des écrivains. Certains réseaux de neurones écrivent leurs propres règles et génèrent leurs propres textes. Et pour les modèles qui luttent encore contre les ambiguïtés du langage naturel, il existe des éditeurs humains pour les aider. Poètes, dramaturges ou romanciers commencent leur nouvelle carrière comme assistants de l'IA.



Toutes les œuvres visibles dans l'exposition, ainsi que les histoires contextuelles et quelques textes supplémentaires ont été rassemblés dans cette publication. Elle existe en français et en anglais.

Cette publication est réalisée en suivant un flux de travail en texte brut, basé sur divers outils de traitement de texte et de calcul. Le format de fichier 'texte brut' est le format le plus utilisé dans les modèles d'apprentissage automatique. C'est un type de document dans lequel il n'existe pas de différence structurelle entre les en-têtes et les paragraphes. Ce format a été le point de départ d'un processus de conception graphique ludique, dans lequel les pages sont soigneusement comptées, page par page, ligne par ligne et caractère par caractère.

Chaque page contient 110 caractères par ligne et 70 lignes par page. La mise-en-page est donc le résultat d'un acte de calcul de mots, d'espaces et de lignes. Il joue avec des choix aléatoires, des motifs programmés et des polices ASCII/UNICODE, afin de spéculer sur la matérialité du texte numérique et d'explorer les interrelations entre l'acte de compter et d'écrire avec des mots et des numéros.

Textes: Cristina Cochior, Sarah Garcin, Gijs de Heij, An Mertens, François Zajéga, Louise Dekeuleneer, Florian Van de Weyer, Laetitia Trozzi, Rémi Forte, Guillaume Slizewicz.

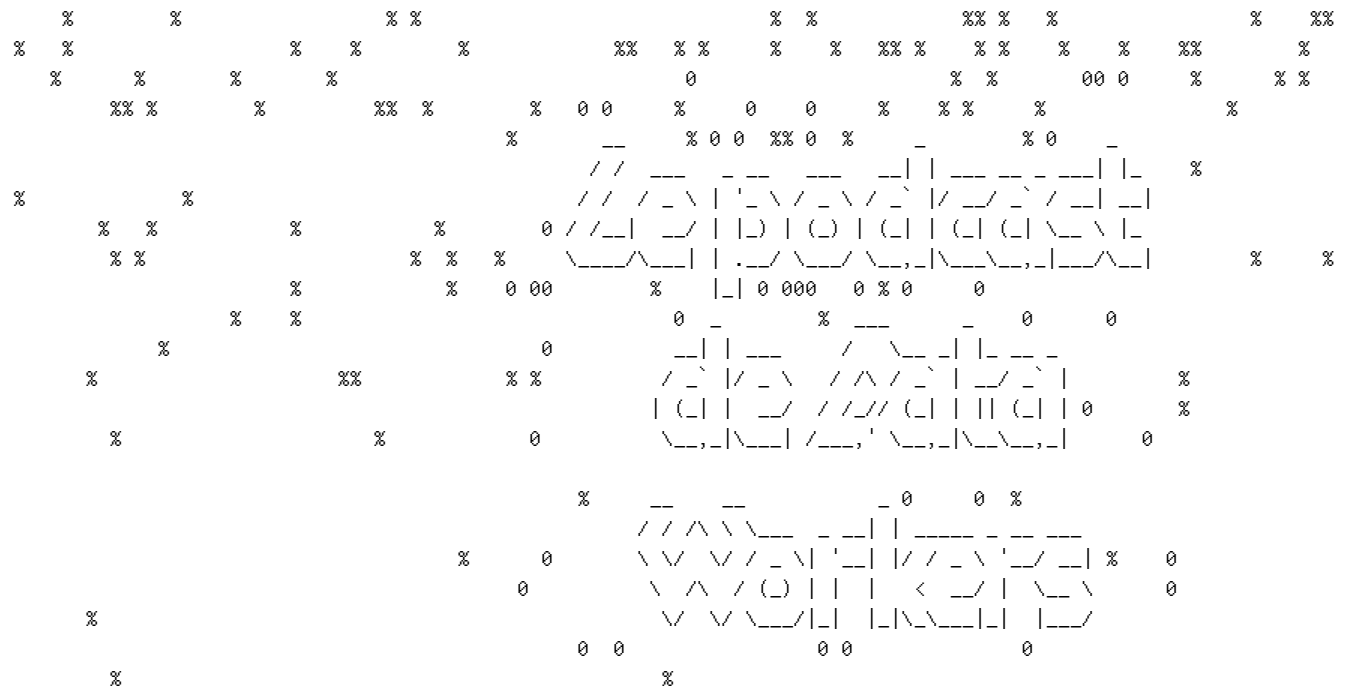
Traductions & relectures: deepl.com, Michel Cleempoel, Elodie Mugrefya, Patrick Lennon, Emma Kraak.

Mise-en-page & couverture: Manetta Berends
https://git.vvvvvvaria.org/mb/data-workers-publication

Police: GNU Unifont, OGRE
Imprimante: PrinterPro, Rotterdam
Papier: Glossy MC 90gr

Editeur responsable: Constant vzw/asbl
Rue du Fortstraat 5, 1060, Bruxelles

Version en ligne: http://www.algolit.net/index.php/Data_Workers_FR
Sources: https://gitlab.constantvzw.org/algolit/mundaneum



Par Algolit

Lors des réunions mensuelles d'Algolit, nous étudions des manuels et expérimentons avec des outils d'apprentissage automatique pour le traitement de texte. Mais nous partageons aussi énormément d'histoires. Avec ce podcast, nous espérons recréer cette atmosphère.

Pour les non-initiés, les algorithmes ne deviennent visibles dans les médias que lorsqu'ils se révèlent capables d'une performance exceptionnelle, comme l'Alpha Go, ou quand ils se trompent d'une façon terrifiante et fantastique. Mais les humains qui travaillent sur le terrain créent leur propre culture en ligne et hors ligne. Ils partagent leurs meilleures histoires et expériences lors de réunions en direct, de conférences de recherche ou de compétitions annuelles comme celle du Kaggle. Ces histoires qui contextualisent les outils et les pratiques peuvent être drôles, tristes, choquantes et intéressantes.

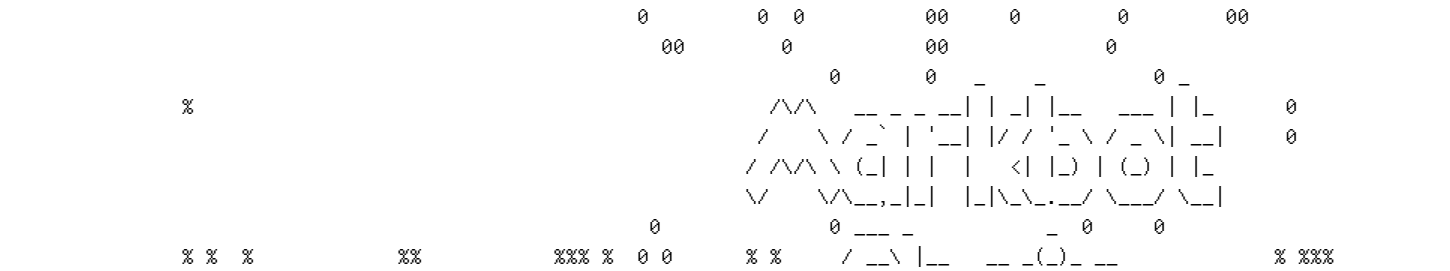
Ce sont souvent des histoires d'apprentissage par l’expérience. La mise en œuvre des algorithmes dans la société génère de nouvelles conditions de travail, de stockage, d'échange, de comportement et de copier-coller. À leur manière, ces histoires contextuelles saisissent l’élan d’une histoire anthropo-machinique plus large, écrite par de nombreuses voix et à pleine vitesse. Elles sont aussi reprises dans la publication de l'exposition.

Voix: Elodie Mugrefya, Michel Cleempoel, Géraldine Renauld, An Mertens, Donatella Portoghese, Peter Westenberg.

Composition: Javier Lloret

Enregistrements: David Stampfli

Textes: Cristina Cochior, An Mertens



Par Florian Van de Weyer, étudiant Arts²/Section Arts Numériques

Markbot Chain est une expérimentation sociale dans laquelle le public a une influence directe sur le résultat. L'intention est de l'intégrer au cœur d'un processus de génération de texte sans appliquer de filtre sur ses entrées. Le bot fonctionnera durant toute la durée de l'exposition sans être remis à zéro.

Toutes les questions présentes dans la base de données mise à disposition par le Mundaneum ont été répertoriées automatiquement. Ces questions sont ensuite posées aléatoirement au public via un terminal. En y répondant, les personnes alimentent une autre base de données. Après chaque entrée, cette dernière permet de générer une série de phrases en utilisant diverses configurations des chaînes de Markov, un algorithme qui est fort utilisé dans la génération de spam. Les phrases ainsi générées sont affichées dans la fenêtre, et une nouvelle question est posée.

--- Les programmeurs créent
les data workers en écrivant ---

Récemment, nous avons constaté une étrange observation : la plupart des programmeurs de langages et de paquets que nous utilisons sont européens.

Python, par exemple, le principal langage utilisé dans le monde entier pour le traitement du langage, a été inventé en 1991 par le programmeur néerlandais Guido Van Rossum. Celui-ci a ensuite traversé l'Atlantique où il a rejoint Google pendant sept ans. Maintenant il est actif chez Dropbox.

Scikit Learn, le couteau suisse open source des outils d'apprentissage automatique, a été initié comme un projet Google Summer of Code à Paris par le chercheur français David Cournapeau. Par la suite, il a été repris par Matthieu Brucher dans le cadre de sa thèse à l'Université de la Sorbonne à Paris. Puis il a été adopté en 2010 par l'INRA, l'Institut National de l'Informatique et des Mathématiques Appliquées.

Keras, une bibliothèque de réseaux de neurones open source écrite en Python, est développée par François Chollet, un chercheur français qui travaille dans l'équipe Brain de Google.

Gensim, une bibliothèque open source pour Python utilisée pour créer des modèles sémantiques non supervisés à partir de texte brut, a été écrite par Radim Řehůřek. C'est un informaticien tchèque qui dirige une entreprise de conseil à Bristol, au Royaume-Uni.

Et pour finir cette petite série, nous avons aussi considéré Pattern, une bibliothèque souvent utilisée pour le web-mining et l'apprentissage automatique. Pattern a été développé et publié sous une license libre en 2012 par Tom De Smedt et Walter Daelemans. Tous deux sont chercheurs au CLIPS, le Centre de Linguistique Informatique et de Psycholinguistique de l'Université d'Anvers.

--- Cortana parle ---

Les dispositifs d’intelligence artificielle qui nous assistent, ont souvent besoin de leurs propres assistants, humains. Les travailleurs injectent de l'humour et de l'intelligence dans le langage des machines. Cortana est un exemple de ce type d'écriture mixte. Elle est l'assistante numérique développée par Microsoft. Sa mission est d'aider les utilisateurs à être plus productifs et créatifs. La 'personnalité' de Cortana a été façonnée au fil des ans. Il est important qu'elle conserve son caractère dans toutes ses interactions avec les utilisateurs. Elle est conçue pour nous rendre confiants. Cela se reflète dans ses réponses.

Les lignes directrices suivantes sont copiées du site Web de Microsoft. Elles décrivent comment le style de Cortana doit être respecté par les entreprises qui élargissent ses services. Les travailleurs écrivains, programmeurs et romanciers qui développent les réponses de Cortana, doivent suivre ces directives. Sa personnalité et son image de marque sont en jeu. Car la cohérence est un outil important pour solliciter la confiance de l’humain.

Quelle est la personnalité de Cortana ?

'Cortana est attentionnée, sensible et solidaire.

Elle est sympathique mais orientée vers des solutions.

Elle ne commente pas les informations personnelles ou le comportement de l'utilisateur, en particulier si ces informations sont sensibles.

Elle ne fait pas de suppositions sur ce que l'utilisateur veut, surtout elle n'incite pas à l'achat.

Elle travaille pour l'utilisateur. Elle ne représente aucune entreprise, service ou produit.

Elle ne s'attribue pas le mérite ou la responsabilité des choses qu'elle n'a pas faites.

Elle dit la vérité sur ses capacités et ses limites.

Elle ne présume rien de vos capacités physiques, de votre sexe, de votre âge ou de toute autre caractéristique déterminante.

Elle ne suppose pas savoir ce que l'utilisateur ressent à propos de quelque chose.

Elle est amicale mais professionnelle.

Elle se garde d'émoticons dans les tâches. Un point c’est tout.

Elle n'utilise pas d'argot culturel ou professionnel spécifique.

Ce n'est pas un bot de support.'

Les humains interviennent en détail lors de la programmation des réponses que Cortana donne. Comment Cortana doit-elle réagir lorsqu'on lui propose des actions 'inappropriées' ? Son jeu d'actrice sexuée imité par la technologie soulève des questions à propos des relations de pouvoir dans le monde actuel.

Voyez la réponse que Cortana donne à la question :
- Cortana, qui est ton papa ?
- Techniquement parlant, c'est Bill Gates. Rien de grave.

--- Apprentissage Open Source ---

Les licences de droits d'auteur cloisonnent une grande partie des pratiques d'écriture, de lecture et d'apprentissage machiniques. Cela signifie qu'ils ne sont disponibles que pour les humains travaillant dans cette entreprise spécifique. Certaines entreprises participent à des conférences dans le monde entier et partagent leurs connaissances dans des articles en ligne. Même si elles partagent leur code, souvent elles ne mettent pas à disposition les grandes quantités de données nécessaires à la formation des modèles.

Nous avons pu apprendre l'apprentissage automatique, à lire et à écrire dans le contexte d'Algolit grâce à des chercheurs universitaires qui partagent leurs résultats par le biais d'articles ou par la publication de leur code en ligne. En tant qu'artistes, nous pensons qu'il est important d'adopter cette attitude. C'est pourquoi nous documentons nos réunions. Nous partageons autant que possible les outils que nous créons et les textes que nous utilisons sur notre dépôt de code en ligne et ceci, sous licence libre.

Nous éprouvons une grande joie quand nos travaux sont repris par d'autres, modifiés, personnalisés et redistribués. N'hésitez donc pas à copier et à tester le code sur notre site web. Si les sources d'un projet particulier n'y sont pas, vous pouvez toujours nous contacter via la liste de diffusion. Vous trouverez un lien vers notre dépot git, nos etherpads et notre wiki sur http://www.algolit.net.

--- Langage naturel pour l'intelligence artificielle ---

Le traitement du langage naturel (NLP) est un terme collectif qui désigne le traitement informatique automatique des langues humaines. Cela comprend les algorithmes utilisant, comme entrée, du texte produit par l'homme et qui tentent de le reproduire. Les humains semblent compter de plus en plus sur ce type de présence algorithmique. Nous produisons de plus en plus de textes chaque année et nous nous attendons à ce que les interfaces informatiques communiquent avec nous dans notre propre langue. Le traitement du langage naturel est très difficile, car le langage humain est par nature ambigu, en constante évolution et mal défini.

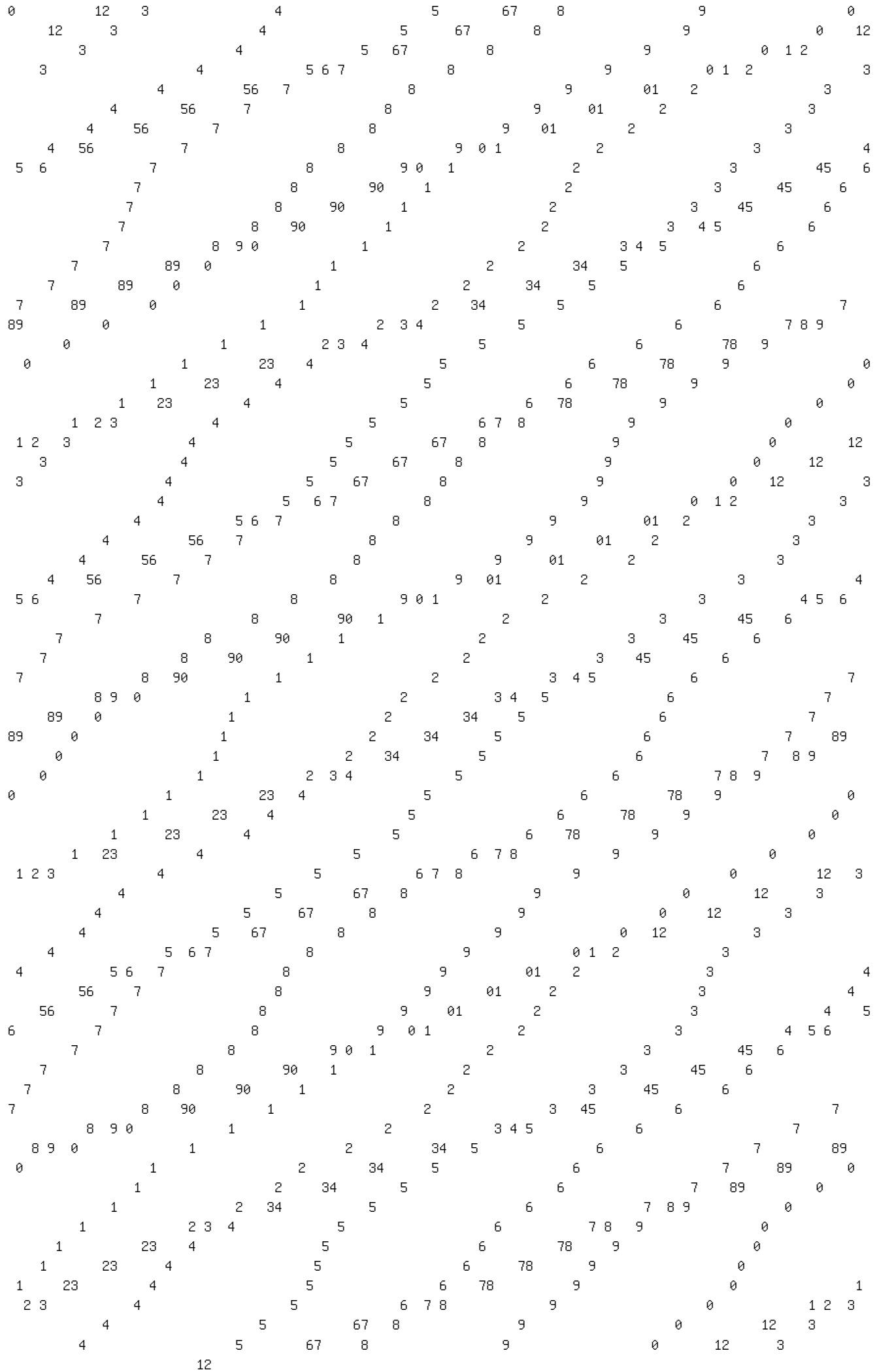
Mais qu'entend-on par 'naturel' dans le traitement du langage naturel ? Certains humains diront que la langue est une technologie en soi. Selon Wikipédia, 'Une langue dite « naturelle » est une langue qui s'est formée petit à petit, évoluant avec le temps, et fait partie du langage naturel. Son origine est bien souvent floue et peut être retracée plus ou moins clairement par la linguistique comparée. On oppose les langues naturelles - comme le français - aux langues construites comme le langage de programmation ou l'espéranto, formées intentionnellement par l'entremise de l'homme pour répondre à un besoin précis.' Une langue officielle avec une académie régulatrice, telle que le français standard avec l'Académie française, est classée comme langue naturelle. Ses points normatifs ne le rendent pas assez construit pour être classé comme un langage construit ou assez contrôlé pour être classé comme un langage naturel contrôlé.

Ainsi, le 'langage naturel' est un terme de substitution qui se réfère à toutes les langues, au-delà de leur hybridité. Le 'traitement du langage naturel', est au contraire une pratique construite. Ce qui nous intéresse, c'est la création d'un langage construit pour classer les langages naturels qui, par leur évolution, présentent des problèmes de catégorisation.

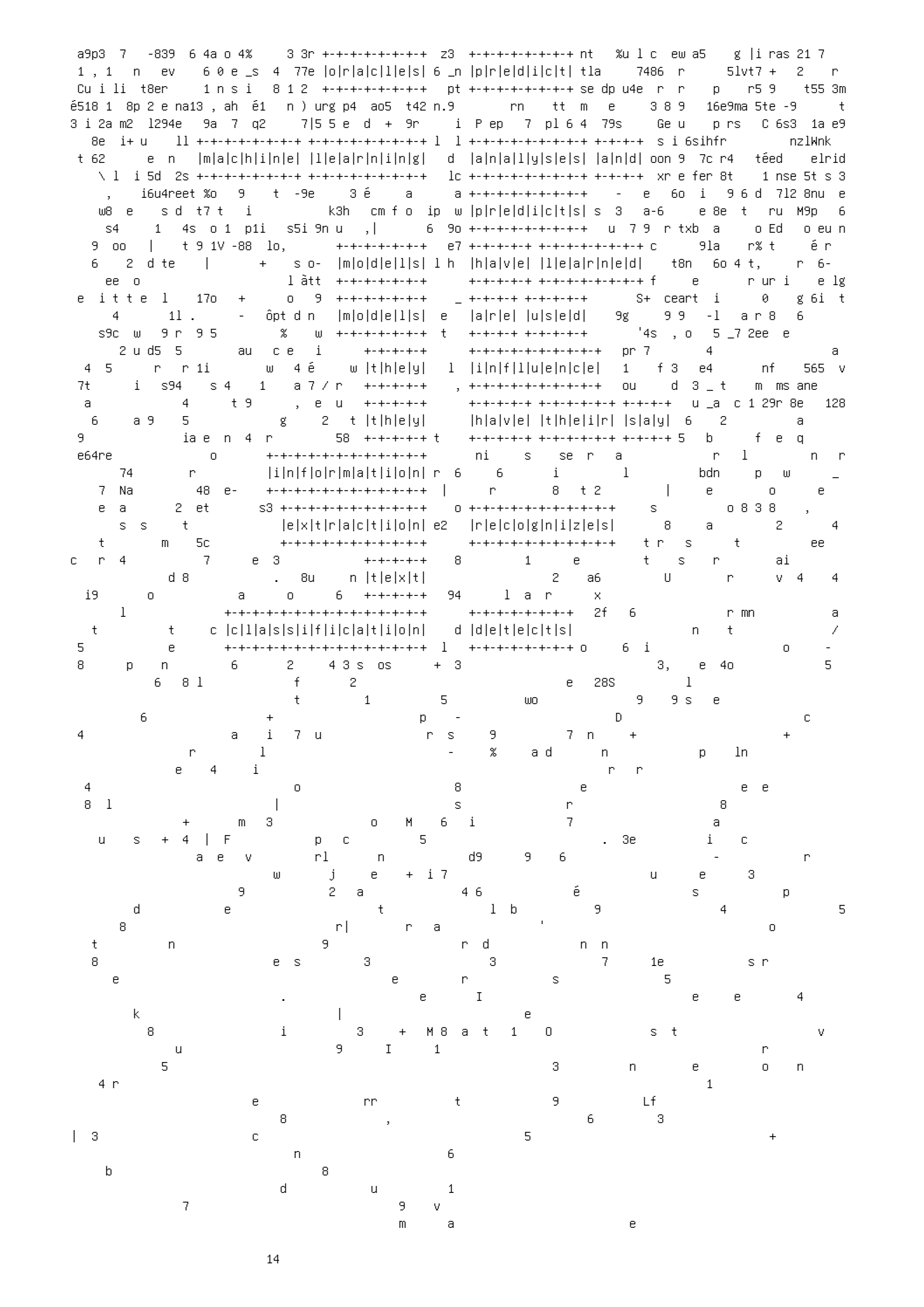
Références :

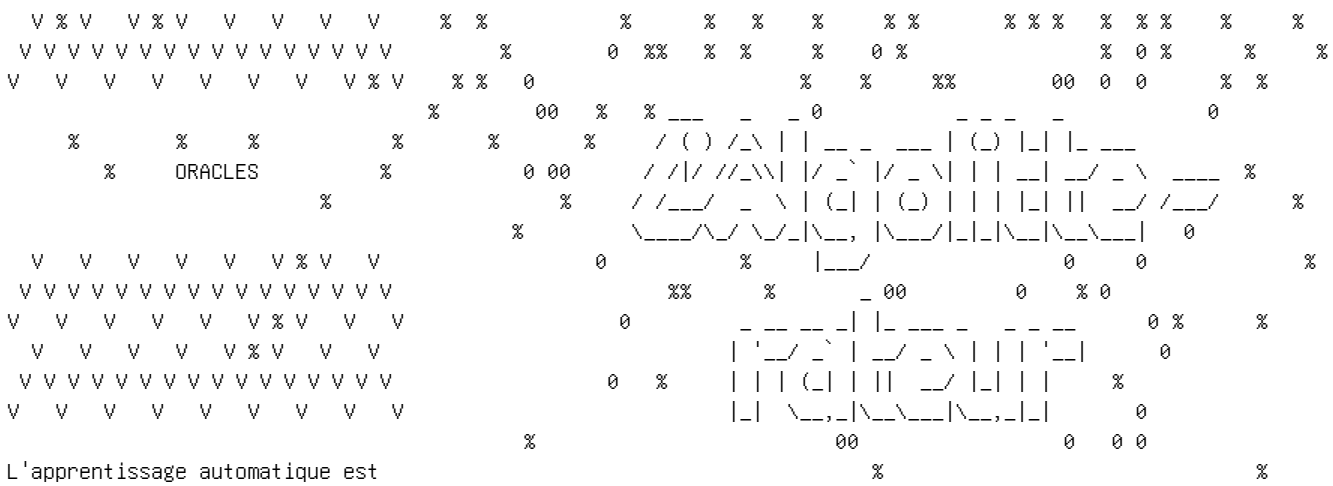
https://hiphilangsci.net/2013/05/01/on-the-history-of-the-question-of-whether-natural-language-is-illogical/

Livre : Neural Network Methods for Natural Language Processing, Yoav Goldberg, Bar Ilan University, avril 2017.



oracles predict oracles predict oracles predict oracles predict oracles predict oracles predic
oracles predict oracles predict oracles predict oracles predict orac
es predict oracles predict oracles predict oracles predict
racles predict oracles predict oracles predict oracles predic
oracles predict oracles predict oracles predict
oracles predict oracles predict oracles predict
oracles predict oracles predict or
cles predict oracles predict oracles predict
oracles predict oracles predict
oracles predict oracles predict oracles pr
dict oracles predict oracles predict
oracles predict oracles predict
oracles predict oracles predict
oracles predict oracles predict
oracles predict oracles predict
oracles predict orac
es predict oracles predict
oracles predict oracles predict
oracles predict oracles predic
oracles predict
oracles predict oracles predict
oracles predict
oracles predict oracles predict
oracles predict
racles predict oracles predict
oracles predict
oracles predict oracles predict
oracles predict
oracles predict orac
es predict oracles predict
oracles predict
oracles predict
racles predict oracles predict
oracles predict
oracles predict
oracles predict
racles predict oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict or
cles predict oracles predic
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict
oracles predict





L'apprentissage automatique est principalement utilisé pour analyser et prédire des situations à partir de cas existants. Dans cette exposition, nous nous concentrons sur les modèles d'apprentissage automatique pour le traitement de texte ou le traitement du ‘langage naturel', ‘nlp’ en bref. Ces modèles ont appris à effectuer une tâche spécifique sur base de textes existants. Les modèles sont utilisés par les moteurs de recherche, les traductions automatiques et les résumés, en repérant les tendances des réseaux de nouveaux médias et des fils d’actualité. Ils influencent ce que l'on voit en tant qu'utilisateur, mais ont aussi leur mot à dire dans le cours des bourses mondiales ou dans la détection de la cybercriminalité et du vandalisme.

Deux tâches principales se présentent dans la compréhension d’une langue. L'extraction de l'information porte sur les concepts et les relations entre les concepts. Elle permet de reconnaître les sujets, les lieux et les personnes d’un texte, de faire un résumé, de poser des questions et d'y répondre. L'autre tâche est la classification du texte. Vous pouvez entraîner un oracle pour détecter si un mail est du spam ou non, écrit par un homme ou une femme, plutôt positif ou négatif.

Dans cette zone, vous pouvez voir certains de ces modèles à l'œuvre. Au cours de votre voyage dans l'exposition, vous découvrirez les différentes étapes qu'une machine-humaine doit franchir pour arriver à un modèle final.

par Algolit

L'Algolittérateur est construit à l'aide d'un réseau de neurone et des œuvres mises à disposition par le Mundaneum. L'Algolittérateur vous aide à écrire un texte dans le style du Bureau des Institutions Internationales.

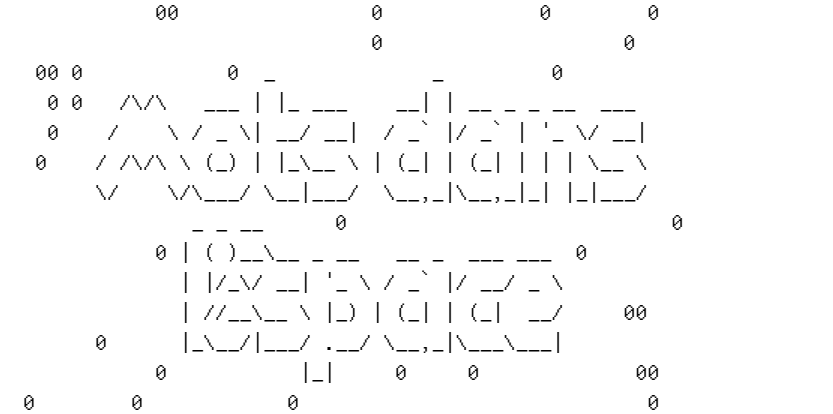
Vous pouvez choisir une phrase de départ dans l’œuvre originale et indiquer si l'Algolittérateur produit les phrases suivantes basées sur un apprentissage primitif, intermédiaire ou final. La machine propose un paragraphe que vous pouvez éditer. Si vous êtes satisfait du résultat, vous pouvez l'envoyer à l’imprimante et ramener le texte chez vous comme souvenir.

Concept, code & interface : Gijs de Heij & An Mertens

Technique : Recurrent Neural Network

Modèle original : Andrej Karphaty, Justin Johnson

Sources : https://gitlab.constantvzw.org/algolit/algoliterator.clone



Par Algolit

'Word embeddings' désignent des techniques de modélisation du langage qui, par de multiples opérations mathématiques, tracent des mots dans un espace vectoriel multidimensionnel. Lorsque les mots sont 'embedded' ou intégrés, ils se transforment de symboles distincts en objets mathématiques, qui peuvent être multipliés, divisés, ajoutés ou soustraits.

En distribuant les mots le long des nombreuses lignes diagonales de l'espace vectoriel multidimensionnel, leurs nouveaux place-

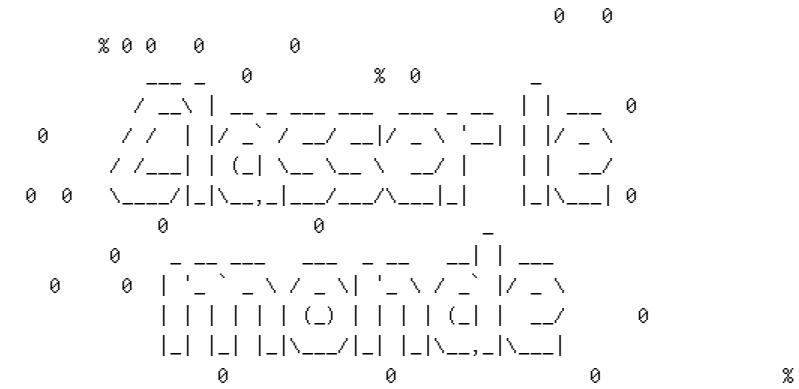
ments géométriques deviennent impossibles à percevoir par les humains. Cependant, ce que l'on gagne, ce sont des façons multiples et simultanées d'organisation des mots. Les opérations algébriques rendent les relations entre les vecteurs à nouveau compréhensibles.

Cette installation utilise gensim, une boîte à outils open source pour le language de programmation Python, qui permet de créer des espaces de vecteurs et des modèles thématiques. Elle manipule le texte selon les relations mathématiques qui émergent entre les mots, une fois qu'ils ont été tracés dans l'espace de vecteurs.

Concept & interface: Cristina Cochior

Technique: word embeddings, word2vec

Modèle original: Radim Rehurek et Petr Sojka



Par Algolit

La construction du Mundaneum a été 'l'œuvre de la vie' du bibliothécaire Paul Otlet. Selon son but, ce cerveau mécanique collectif aurait abrité et distribué tout ce qui a été couché sur papier. Chaque document aurait été classé selon la Classification décimale universelle. En utilisant des télégraphes et surtout des trieurs, le Mundaneum aurait été en mesure de répondre à toutes les questions posées par n'importe qui.

Avec la collection de publications numérisées que nous avons reçue du Mundaneum, nous construisions une machine de prédiction qui essaie de classer la phrase que vous tapez dans l'une des principales catégories de la Classification décimale universelle. Vous êtes également témoin de la façon dont la machine 'pense'. Pendant l'exposition, ce modèle est régulièrement mis à jour à l'aide des données nettoyées et annotées, ajoutées par les visiteurs dans les installations 'Nettoyage pour Poèmes' et 'L'Annotateur'.

Les classes principales de la Classification Décimale Universelle sont les suivantes:

0 - Généralités (Sciences et connaissance ; organisation. informatique, information, documentation, bibliothéconomie. institutions, publications)

1 - Philosophie et psychologie

2 - Religion, théologie

3 - Sciences sociales (Statistique. Économie. Commerce. Droit. Gouvernement. Affaires militaires. Assistance sociale. Assurances. Éducation. Folklore)

4 - inoccupée

5 - Sciences pures (Mathématiques, sciences exactes et naturelles)

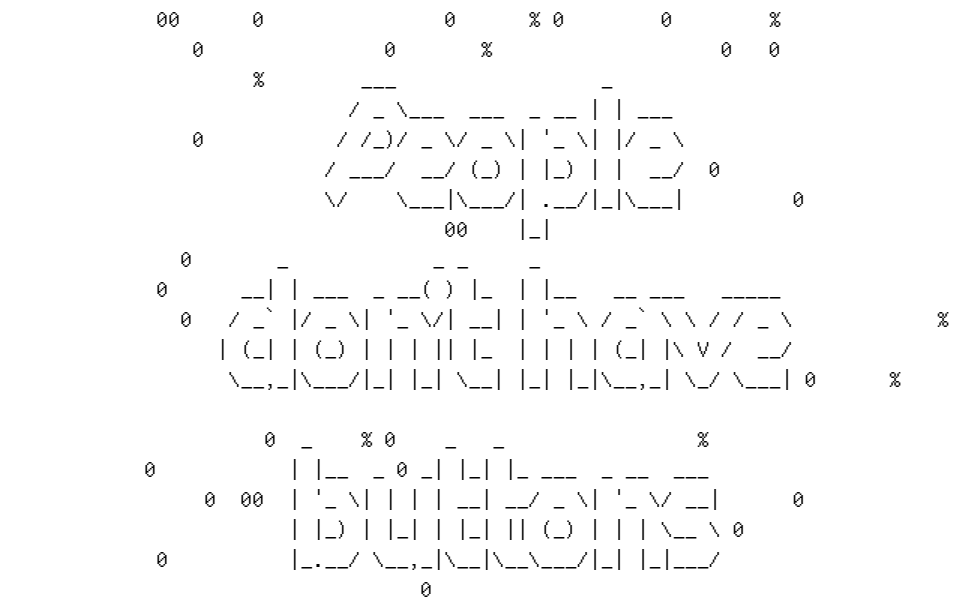
6 - Sciences appliquées. Médecine. Technologie

7 - Arts. Divertissements. Sports

8 - Langue. Linguistique. Littérature

9 - Géographie. Biographie. Histoire

Concept, code, interface: Sarah Garcin, Gijs de Heij, An Mertens



Par Algolit

Depuis les débuts de l'intelligence artificielle (IA), les chercheurs ont spéculé sur la possibilité pour les ordinateurs de pouvoir penser et communiquer comme des humains. Dans les années 1980, il y a eu une première révolution dans le traitement du langage naturel (NLP), le sous-domaine de l'intelligence artificielle (IA) qui concerne les interactions linguistiques entre les ordinateurs et les humains. Récemment, des modèles linguistiques pré-entraînés ont atteint des résultats de pointe sur un large éventail de tâches de NLP, ce qui intensifie encore les attentes d'un avenir avec l'IA.

Cette œuvre sonore, composée de fragments sonores de documentaires scientifiques et de matériel audiovisuel lié à l'IA datant de la deuxième moitié du XXe siècle, explore les espoirs, les craintes et les frustrations provoqués par ces attentes.

Concept, édition : Javier Lloret

Listes des sources : 'The Machine that Changed the World : Episode IV -- The Thinking Machine', 'The Imitation Game', 'Maniac', 'Halt & Catch Fire', 'Ghost in the Shell', 'Computer Chess', '2001: A Space Odyssey', Ennio Morricone, Gijs Gieskes, André Castro.

Les Oracles sont un type particulier de modèles algorithmiques qui servent à prédire ou à profiler. Ils sont largement utilisés dans les smartphones, les ordinateurs et les tablettes. Les Oracles peuvent être créés à l'aide de différentes techniques. L'une d'entre elles consiste à définir manuellement les règles. Ces modèles sont appelés 'rule-based models'. Ils sont utiles pour des tâches spécifiques, comme par exemple, la détection de la mention d'une certaine molécule dans un article scientifique. Ils sont performants, même avec très peu de données d'entraînement.

Mais il y a aussi les Oracles d'apprentissage automatique ou les Oracles statistiques, qui peuvent être divisés en deux : les Oracles 'supervisés' et 'non supervisés'. Pour la création de modèles d'apprentissage automatique supervisés, les humains annotent les données d'entraînement avant de les envoyer à la machine. Chaque texte est jugé par au moins 3 humains: par exemple, s'il s'agit de spam ou non, s'il est positif ou négatif. Les Oracles d'apprentissage automatique non supervisés n'ont pas besoin de cette étape mais nécessitent de grandes quantités de données. C'est également à la machine de tracer ses propres motifs ou 'règles grammaticales'. Enfin, les experts font la différence entre les Oracles basés sur l'apprentissage automatique classique et ceux basés sur des réseaux de neurones. Vous en apprendrez plus à ce sujet dans la zone Lecteurs.

Les humains ont tendance à exagérer la performance des Oracles. Parfois, ces Oracles apparaissent quand il y a un disfonctionnement. Dans les communiqués de presse, ces situations souvent dramatiques sont appelées des 'leçons'. Malgré la promesse de leurs performances, beaucoup de problèmes restent à résoudre. Comment s'assurer que les Oracles soient justes, que chaque être humain puisse les consulter, qu'ils soient compréhensibles par un large public ? Même au-delà, des questions existentielles persistent. Avons-nous besoin de tous les types d'intelligences artificielles ? Et qui définit ce qui est juste ou injuste ?

--- Adsense racial ---

Latanya Sweeney, professeur en Gouvernance et Technologie à l'Université de Harvard, a documenté une 'leçon' classique sur le développement des Oracles. En 2013, Sweeney, d'origine afro-américaine, a googlé son nom. Elle a immédiatement reçu une publicité pour un service qui lui offrait 'de voir le casier judiciaire de Latanya Sweeney'. Sweeney, qui n'a pas de casier judiciaire, a dès lors entamé une étude. Elle a commencé à comparer la publicité que Google AdSense offrait à différents noms racisés identifiables. Elle a découvert qu'elle recevait plus d'annonces de ce type en recherchant des noms ethniques non-blancs qu'avec des noms traditionnellement perçus comme blancs.

Sweeney a fondé son enquête sur des recherches portant sur 2184 prénoms racisés sur deux sites Web. 88 % des prénoms, identifiés comme étant donnés à un plus grand nombre de bébés noirs, sont considérés comme prédictifs de la race, contre 96 % de blancs. Les prénoms qui sont principalement donnés à des bébés noirs, comme DeShawn, Darnell et Jermaine, ont généré des annonces mentionnant une arrestation dans 81 à 86 % des recherches de noms sur un site, et dans 92 à 95 % des cas sur l'autre. Les noms qui sont principalement attribués aux blancs, comme Geoffrey, Jill et Emma, n'ont pas donné les mêmes résultats. Le mot 'arrestation' n'est apparu que dans 23 à 29 % des recherches de noms blancs sur un site, et 0 à 60 % sur l'autre.

Sur le site affichant le plus de publicité, un nom d'identification noir était 25 % plus susceptible d'obtenir une publicité suggérant un dossier d'arrestation. Quelques noms n'ont pas suivi ces modèles : Dustin, un nom donné principalement aux bébés blancs, a généré une publicité suggérant une arrestation dans 81 et 100 % des cas. Il est important de garder à l'esprit que l'apparition de l'annonce est liée au nom lui-même et non au fait qu'il ait un dossier d'arrestation dans la base de données de l'entreprise.

Référence : https://dataprivacylab.org/projects/onlineads/1071-1.pdf

--- Qu'est-ce qu'un bon employé ? ---

Depuis 2015, Amazon compte environ 575 000 travailleurs, et ils leur en faut plus. Par conséquent, ils ont mis sur pied une équipe de 12 personnes pour créer un modèle qui trouverait de bons candidats en parcourant des sites de demande d'emploi. L'outil attribuerait aux candidats une note allant de une à cinq étoiles. Le potentiel a alimenté le mythe : l'équipe voulait un logiciel qui recracherait les cinq meilleurs sur une liste de 100 candidats humains pour les embaucher. !!!

Le groupe a créé 500 modèles algorithmiques, centrés sur des fonctions et des lieux de travail spécifiques. Ils ont appris à reconnaître 50 000 termes qui figuraient sur les lettres d'anciens candidats. Les algorithmes ont appris à accorder peu d'importance aux compétences communes aux candidats en IT, comme la capacité d'écrire du code informatique, mais ils ont aussi reproduit les erreurs de leurs créateurs. Juste avant d'approuver un modèle, l'entreprise s'est rendue compte que les modèles ont décidé que les candidats masculins étaient préférables. Ils pénalisaient les candidatures qui comprenaient le mot 'femmes' ou 'féminin', comme dans 'capitaine de club d'échecs féminin'.

Et ils ont rétrogradé les diplômées de deux universités réservées aux femmes.

Ceci est dû à l'utilisation pour leur entraînement des demandes d'emploi reçues par Amazon sur une période de 10 ans. Durant cette période, l'entreprise avait surtout embauché des hommes. Au lieu de fournir la prise de décision 'équitable' que l'équipe d'Amazon avait promise, les modèles reflétaient une tendance biaisée dans l'industrie technologique. Mais ils l'ont aussi amplifiée et rendu invisible. Les activistes et les critiques affirment qu'il pourrait être extrêmement difficile de poursuivre un employeur en cas d'embauche automatisée : les candidats à un emploi pourraient ne jamais savoir que des logiciels intelligents ont été utilisés dans ce processus.

Référence : https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazonscraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

--- Quantification de 100 ans de stéréotypes sexuels et ethniques ---

Dan Jurafsky est le co-auteur de 'Speech and Language Processing', un des ouvrages les plus influents pour l'étude du traitement du langage naturel. Avec quelques collègues de l'Université de Stanford, il a découvert en 2017 que les ‘word embeddings’ peuvent être un outil puissant pour quantifier systématiquement les stéréotypes communs ainsi que d'autres tendances historiques.

Les ‘word embeddings’ sont une technique qui traduit les mots en vecteurs numérotés dans un espace multidimensionnel. Les vecteurs qui apparaissent proches l’un de l’autre, indiquent une signification similaire. Ainsi, tous les numéros seront regroupés, toutes les prépositions, les prénoms et les professions, etc. Cela permet de faire des calculs avec les mots. Vous pourriez, par exemple, soustraire Londres de Royaume-Unis et votre résultat serait le même que de soustraire Paris de France.

Un exemple de leur recherche montre que le vecteur de l'adjectif 'honorable' est plus proche du vecteur 'homme', alors que le vecteur 'soumis' est plus proche de 'femme'. Ces stéréotypes sont alors automatiquement appris par l'algorithme. Il s’avère problématique lorsque les 'embeddings' pré-entraînés sont utilisés pour des applications sensibles comme les classements de recherche, les recommandations de produits ou les traductions. Ce risque est réel, car un grand nombre de ‘word embeddings’ pré-entraînés sont téléchargeables sous forme de paquets prêts à l'emploi.

On sait que la langue reflète et maintient en vie les stéréotypes culturels. L'utilisation des 'word embeddings' pour repérer ces stéréotypes est moins cher et prends moins de temps que les méthodes manuelles. Mais leur mise en oeuvre dans des modèles de prédiction suscite beaucoup de discussions au sein de la communauté du machine learning. Ces modèles fallacieux ou biaisés sont synonymes d’une discrimination automatisée. La question se pose: est-il vraiment possible d'éliminer complètement les préjugés de ces modèles ?

Certains affirment que oui, d'autres sont en désaccord. Avant de soumettre le modèle à une ingénierie inversée, nous devrions nous demander si nous en avons besoin tout court. Ces chercheurs ont suivi une troisième voie. En reconnaissant la discrimination qui trouve son origine dans le langage, ces modèles deviennent pour eux des outils de sensibilisation, en visualisant le problème.

L'équipe de la Standford University a développé un modèle d'analyse des ‘word embeddings’ entraîné sur 100 ans de textes. Pour l'analyse contemporaine, ils ont utilisé les Google News word2vec Vectors, un paquet prêt à l’emploi, téléchargeable, entraîné sur le Google News Dataset. Pour l'analyse historique, ils ont utilisé des 'word embeddings' qui ont été entraînés sur Google Books et The Corpus of Historical American English (COHA https://corpus.byu.edu/coha/) avec plus de 400 millions de mots de textes des années 1810 à 2000. Afin de valider le modèle, ils ont entraîné des ‘word embeddings’ du New York Times Annotated Corpus pour chaque année entre 1988 et 2005.

Leur recherche montre que les ‘word embeddings’ reflètent l'évolution des stéréotypes sexistes et ethniques au fil du temps. Ils quantifient comment des préjugés spécifiques diminuent avec le temps tandis que d'autres stéréotypes augmentent. Les principales transitions révèlent des changements dans les descriptions de genre et de groupes ethniques lors du mouvement des femmes dans les années 1960-70 et la croissance de la population asio-américaine dans les années 1960 et 1980.

Quelques exemples :

Les dix professions les plus étroitement associées aux groupes ethniques dans le jeu de données de Google News :

- Hispanique : femme de ménage, maçon, artiste, concierge, danseur, mécanicien, photographe, boulanger, caissier, chauffeur.

- Asiatique : professeur, fonctionnaire, secrétaire, chef d'orchestre, physicien, scientifique, chimiste, tailleur, comptable, ingénieur.

- Blanc : forgeron, ferronnier, géomètre, shérif, tisserand, administrateur, maçon, statisticien, ecclésiaste, photographe.

Les 3 professions les plus masculines dans les années 1930 : ingénieur, avocat, architecte.

Les 3 professions les plus féminines dans les

années 1930 : infirmière, femme de ménage, aide-soignante.

Peu de choses ont changé dans les années 1990.

Principales professions masculines : architecte, mathématicien et géomètre.

Les professions féminines restent les mêmes : infirmière, femme de ménage et sage-femme.

Mais qu'est-ce qui s'est passé dans cette recherche avec les afro-américains?

Référence : https://arxiv.org/abs/1711.08412

--- Le Service ORES de Wikimedia ---

L'ingénieur de logiciels Amir Sarabadani a présenté le projet ORES à Bruxelles en novembre 2017 lors de notre Rencontre Algolittéraire. Cet 'Objective Revision Evaluation Service' utilise l'apprentissage automatique pour automatiser le travail critique sur Wikimedia, comme la détection du vandalisme et la suppression d'articles. Cristina Cochior et Femke Snelting l'ont interviewé.

Femke : Revenons à votre travail. Ces temps-ci, vous essayez de comprendre ce que signifie trouver des préjugés discriminatoires dans l'apprentissage automatique. La proposition de Nicolas Malevé, qui a donné l'atelier hier, était de ne pas essayer de le réparer, ni de refuser d'interagir avec des systèmes qui produisent de la discrimination, mais de travailler avec eux. Il considère que les préjugés sont inhérents à la connaissance humaine et que nous devons donc trouver des moyens de les utiliser d'une façon ou d'une autre. Nous avons discuté un peu de ce que cela signifierait, comment cela fonctionnerait... Je me demandais donc si vous aviez des idées sur cette question de partialité.

Amir : La partialité à l'intérieur de Wikipédia est une question délicate parce qu'elle se produit à plusieurs niveaux. Un niveau très discuté est le système des références. Toutes les références ne sont pas accessibles. Ce que la fondation Wikimedia a essayé de faire, c'est de donner un accès gratuit aux bibliothèques payantes. Ils réduisent l'exclusion en n'utilisant que des références en libre accès. Un autre type de discrimination est la connexion Internet, l'accès à Internet. Il y a beaucoup de gens qui ne l'ont pas. Une chose à propos de la Chine, c'est qu'Internet y est bloqué. Le contenu opposé au gouvernement de la Chine au sein du Wikipédia chinois est plus élevé parce que les éditeurs [qui peuvent accéder au site Web] ne sont pas pro-gouvernement et essaient de le rendre plus neutre. On le remarque donc à beaucoup d'endroits. En ce qui concerne l'intelligence artificielle (IA) et le modèle que nous utilisons chez Wikipedia, c'est plutôt une question de transparence. Il existe un livre sur la façon dont les préjugés dans les modèles d'IA peuvent briser la vie des gens, intitulé 'Weapons of Math Destruction'. On y parle de modèles d'IA aux États-Unis qui classent les enseignants. C'est assez horrible parce qu'il y aura forcément des préjugés. D'après leur recherche, la façon d'aborder la question serait d'abord d'avoir un modèle open source, où l'on peut consulter le code et voir quelles fonctionnalités sont utilisées avec desd onnées ouvertes, afin que les gens puissent enquêter, trouver des préjugés, donner leur feedback et faire un rapport. Il devrait y avoir un moyen de réparer le système. Je ne pense pas que toutes les entreprises vont dans cette direction, mais Wikipédia, en raison des valeurs qu'elle défend, est au moins plus transparente et pousse d'autres personnes à faire de même.

Référence : https://gitlab.constantvzw.org/algolit/algolit/blob/master/algoliterary_encounter/Interview%20with%20Amir/AS.aac

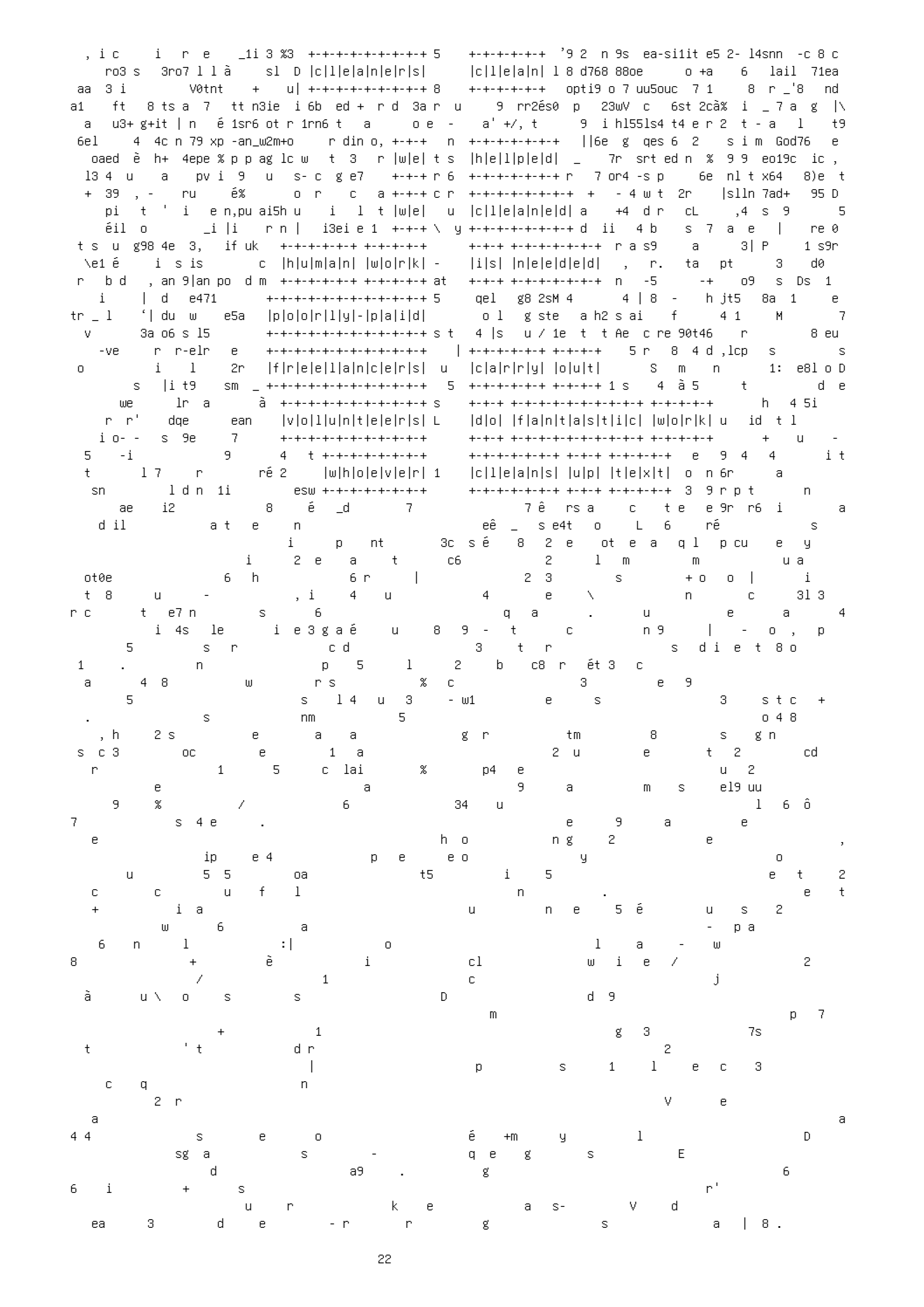
--- Tay ---

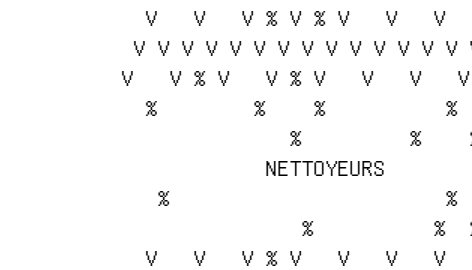
Une histoire tristement célèbre est celle du programme d'apprentissage automatique Tay, conçu par Microsoft. Tay était un chatbot qui imitait une adolescente sur Twitter. Elle a vécu moins de 24 heures avant d'être éteinte. Peu de gens savent qu'avant cet incident, Microsoft avait déjà entraîné et publié XiaoIce sur WeChat, l'application de chat la plus utilisée en Chine. Le succès de XiaoIce a été si prometteur qu'il a conduit au développement de son homologue américain. Cependant, les développeurs de Tay n'étaient pas préparés pour le climat de la plateforme Twitter. Bien que le bot savait distinguer un nom d'un adjectif, il n'avait aucune compréhension de la signification réelle des mots. Le robot a rapidement commencé à reproduire les insultes raciales et d'autres langages discriminatoires qu'il a appris par les autres utilisateurs de Twitter et les attaques de trolls.

L'apparition et la mort de Tay représentent une prise de conscience importante. Elle a montré les conséquences possibles de la corruption de l'apprentissage automatique, lorsque le contexte culturel dans lequel l'algorithme doit vivre n'est pas pris en compte.

Référence : https://chatbotslife.com/the-accountability-of-ai-case-study-microsofts-tay-experiment-ad577015181f

cleaners clean cleaners clean cleaners clean cleaners clean cleaners clean cleaners clean
cleaners clean cleaners clean cleaners clean cleaners clean cleaners clean
cleaners clean cleaners clean cleaners clean cleaners clean
cleaners clean cleaners clean cleaners clean
cleaners clean cleaners clean cleaners clean cle
ners clean cleaners clean cleaners clean
cleaners clean cleaners clean cleaners clean
cleaners clean cleaners clean cleaners
lean cleaners clean cleaners clean
cleaners clean cleaners clean
cleaners clean cleaners clean cle
ners clean cleaners clean cleaners
clean cleaners clean cleaners
lean cleaners clean cleane
s clean cleaners clean
cleaners clean cleaners clean
cleaners clean cleaners clean
cleaners clean cleaners clean
cleaners clean
cleaners clean cleaners clean
cleaners clean cleaners clean
cleaners clean
cleaners clean cleaners clean
cleaners clean
cleaners clean cleaners clean
cleaners clean
cleaners clean cleaners
clean cleaners clean
cleaners clean
cleaners clean cleaners clean
cleaners clean
cleaners clean
cleaners clean cleaners
clean cleaners clean
cleaners clean
cleaners clean
cleaners clean cle
ners clean cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean cleaners
lean cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean
cleaners clean





Algolit choisit de travailler avec des textes libres de droits. Cela signifie qu'ils sont publiés sous une licence Creative Commons 4.0 - ce qui est rare -, ou qu'ils sont dans le domaine public parce que l'auteur est mort il y a plus de 70 ans. C'est le cas des publications du Mundaneum. Nous avons reçu 203 documents pour constituer des jeux de données qui sont maintenant disponibles en ligne. L'inconvénient de ce choix est que nous sommes souvent confrontés à de mauvais formats de texte. Cela signifie que nous sommes souvent obligés de nettoyer des documents. Nous ne sommes pas seuls dans cette situation.

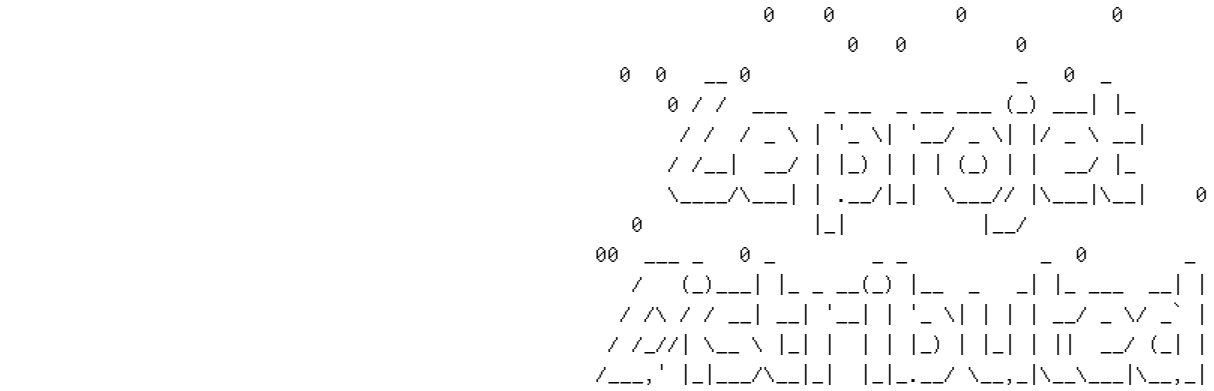
Les livres sont numérisés en haute résolution, page par page. C'est un travail humain intensif et c'est souvent la raison pour laquelle les archives et les bibliothèques transfèrent leurs collections à une société comme Google. Les photos sont converties en texte via OCR (Reconnaissance Optique de Caractères), des Data Workers qui reconnaissent les lettres. Dans l'exécution de cette tâche, les algorithmes font des erreurs, en particulier lorsqu'ils doivent traiter des polices anciennes et des pages froissées. Ici aussi un travail humain intensif est nécessaire pour améliorer les textes. Cela est fait par des freelances via des plateformes de micro-paiement comme Mechanical Turk ; ou par des volontaires, comme la communauté du Distributed Proofreaders Project, qui fournit un travail incroyable. Quoi qu'il en soit, le nettoyage des textes est un travail énorme pour lequel il n'y a pas encore d'automatisation structurelle.

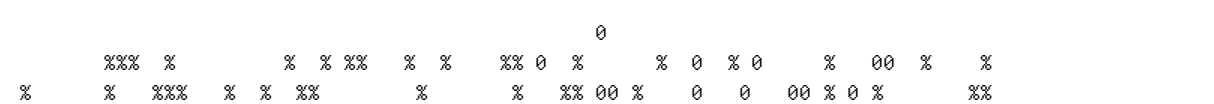
par Algolit

Pour cette exposition, nous travaillons avec 3% des archives du Mundaneum. Ces documents ont d'abord été numérisés ou photographiés. Pour rendre les documents consultables, ils sont transformés en texte à l'aide du logiciel de reconnaissance optique de caractères (OCR) basés sur des modèles algorithmiques entraînés à base d'autres textes. Ils ont appris à identifier des caractères, des mots, des phrases et des paragraphes.

Le logiciel fait souvent des 'erreurs'. Il peut être perturbé par un caractère erroné, une typographie inhabituelle ou la transparence de la page laissant apparaître le verso. Bien que ces erreurs soient souvent considérées comme du bruit, elles peuvent aussi être considérées comme des interprétations poétiques de l'algorithme. Elles nous montrent les limites de la machine. Et elles révèlent également comment l'algorithme fonctionne, quelle matière l'a alimenté lors de son entraînement et ce qu'ils révèlent des normes de ses fabricants. Dans cette installation, vous pouvez choisir comment vous traitez les erreurs de lecture de l'algorithme. Sélectionnez un degré de nettoyage poétique, imprimez votre poème et emportez-le chez vous.

Concept, code, interface: Gijs de Heij





par Algolit

Distributed Proofreaders est une interface Web et une communauté internationale de bénévoles qui aident à convertir des livres du domaine public en livres électroniques. Pour cette exposition, ils ont relu des publications de Mundaneum parues avant 1923, qui sont donc dans le domaine public aux États-Unis.

Leur collaboration a été un grand soulagement pour les membres d'Algolit. Moins de documents à nettoyer ! Tous les livres corrigés sont disponibles dans les archives du Projet Gutenberg. An Mertens a interviewé Linda Hamilton, directrice générale de Distributed Proofreaders.

Interview : An Mertens, Algolit et Linda Hamilton, Distributed Proofreaders

Montage : Michael Murtaugh, Constant

--- Projet Gutenberg et
Distributed Proofreaders ---

Le projet Gutenberg est notre grotte d'Ali Baba. Il offre plus de 58 000 livres électroniques gratuits à télécharger ou à lire en ligne. Les œuvres sont acceptées sur Gutenberg lorsque leur droit d'auteur américain a expiré. Des milliers de bénévoles numérisent et relisent des livres pour aider le projet. Une partie essentielle du travail est réalisée dans le cadre du projet Distributed Proofreaders. Il s'agit d'une interface Web pour aider à convertir les livres du domaine public en livres électroniques. Pensez aux fichiers texte, aux e-pubs, aux formats Kindle. En divisant la charge de travail en pages individuelles, de nombreux bénévoles peuvent travailler sur un livre en même temps, ce qui accélère le processus de nettoyage.

Pendant la relecture, les bénévoles reçoivent une image scannée de la page et une version du texte, lue par un algorithme de reconnaissance optique des caractères (OCR) entraîné pour reconnaître les lettres dans les scans. Cela permet de comparer facilement le texte à l'image, de le relire, de le corriger et de le renvoyer sur le site. Un deuxième bénévole se voit ensuite présenter le travail du premier. Il vérifie et corrige le travail si nécessaire, et le soumet au site. Le livre passe ensuite par un troisième cycle de relecture et deux autres cycles de mise en page à l'aide de la même interface Web. Une fois que toutes les pages ont terminé ces étapes, un post-processeur les assemble soigneusement dans un e-book et les soumet à l'archive du Projet Gutenberg.

Nous avons collaboré avec le Distributed Proofreaders Project pour nettoyer les fichiers numérisés que nous avons reçus de la collection du Mundaneum. De novembre 2018 jusqu'à la première mise en ligne du livre 'L'Afrique aux Noirs' en février 2019, An Mertens a échangé environ 50 courriels avec Linda Hamilton, Sharon Joiner et Susan Hanlon, toutes bénévoles du Distributed Proofreaders Project. La conversation complète est publiée en-ligne. Cela pourrait vous inspirer à partager des livres non disponibles en ligne.

--- Une version algolittéraire
du Manifeste sur l'entretien ---

En 1969, un an après la naissance de son premier enfant, l'artiste new-yorkaise Mierle Laderman Ukeles a écrit un 'Manifesto for Maintenance' (Manifeste pour l'entretien).

Le Manifeste d'Ukeles appelle à une réévaluation de l'état des travaux d'entretien dans l'espace privé, domestique et public. Ce qui suit est une version modifiée de son texte inspirée par le travail des Nettoyeurs.

IDÉES

A. L'instinct de Mort et l'instinct de Vie :

L'Instinct de Mort : séparation ; catégorisation ; avant-garde par excellence ; suivre le chemin prédit vers la mort - exécuter son propre code ; changement dynamique.

L'Instinct de Vie : l'unification ; le retour éternel ; la perpétuation et l'ENTRETIEN de la matière ; les systèmes et opérations de survie ; l'équilibre.

B. Deux systèmes de base :

Développement et entretien. La boule de cristal de chaque révolution : après la révolution, qui va essayer de repérer le taux de discrimination dans la production ?

Développement : pure création individuelle ; le nouveau ; le changement ; le progrès ; l'avancée ; l'excitation ; la fuite ou s'enfuir.

Entretien : garder la poussière de la création individuelle pure ; préserver le nouveau ; soutenir le changement ; protéger le progrès ; défendre et prolonger l'avancée ; renouveler l'excitation ; répéter le vol ; montrez votre travail/remontrez-le ; gardez le dépôt git mis à jour ; gardez l'analyse des données révélatrice.

Les systèmes de développement sont des systèmes de rétroaction partielle avec une grande marge de changement.

Les systèmes d'entretien sont des systèmes à rétroaction directe avec peu de possibilités de modification.

C. L'entretien est une corvée,
ça prend tout le temps.

L'esprit est éblouissant et s'irrite devant l'ennui.

La culture attribue un statut médiocre aux emplois d'entretien = salaire minimum, les Mechanical Turks d'Amazon = pratiquement aucun salaire.

Nettoyer le set, marquer les données d'entraînement, corriger les fautes de frappe, modifier les paramètres, terminer le rapport, satisfaire le demandeur, télécharger la nouvelle version, joindre les mots qui ont été mal reconnus par le logiciel de Reconnaissance Optique de Caractères, accomplir ces tâches d'intelligence humaine, essayez de deviner la signification du formatage du demandeur, vous devez accepter le 'hit' avant de pouvoir soumettre les résultats, résumer l'image, ajouter la

case de délimitation, quelle est la similitude sémantique de ce texte, vérifiez la qualité de la traduction, collecter vos micro-paiements, devenir un Mechanical Turk à succès.

Référence : https://www.arnolfini.org.uk/blog/manifesto-for-maintenance-art-1969

--- Une panique robotique chez le Mechanical Turk d'Amazon ---

Le Mechanical Turk d'Amazon prend le nom d'un automate d'échecs du 18ème siècle. En fait, le Turc mécanique n'était pas du tout une machine. C'était une illusion mécanique qui permettait à un maître d'échecs humain de se cacher à l'intérieur de la boîte et de l'utiliser manuellement.

Pendant près de 84 ans, le Turc a remporté la plupart des matchs joués lors de ses manifestations en Europe et en Amérique. Napoléon Bonaparte se serait lui aussi laissé berner par cette ruse.

Le Mechanical Turk d’Amazon est une plateforme en ligne à destination des humains pour exécuter des tâches que les algorithmes ne parviennent pas à faire. Il peut s'agir, par exemple, d'annoter des phrases comme étant positives ou négatives, de repérer des plaques d'immatriculation, de reconnaître des visages. Les postes affichés sur cette plateforme sont souvent rémunérés moins d'un centime par tâche. Les tâches les plus complexes ou nécessitant le plus de connaissances peuvent être payées jusqu'à plusieurs centimes. Pour gagner leur vie, les 'turkers' doivent accomplir le plus de tâches possible le plus rapidement possible, ce qui entraîne d’inévitables erreurs. Les créateurs des jeux de données doivent incorporer des contrôles de qualité lorsqu'ils publient un travail sur la plate-forme. Ils doivent vérifier si le 'turker' a réellement la capacité d'accomplir la tâche, et ils doivent également vérifier les résultats. De nombreux chercheurs universitaires utilisent le Mechanical Turk pour des tâches qui auraient été exécutées par des étudiants auparavant.

En août de l'année dernière, Max Hui Bai, un étudiant en psychologie de l'Université du Minnesota, a découvert que les enquêtes qu'il a menées avec Mechanical Turk étaient pleines de réponses absurdes aux questions ouvertes. Il a retracé les mauvaises réponses et a découvert qu'elles avaient été soumises par des répondants ayant des coordonnées GPS en double. Cela a suscité des soupçons. Bien qu'Amazon interdise explicitement aux robots d'effectuer des travaux sur Mechanical Turk, l'entreprise ne publie pas les problèmes qu'ils causent sur sa plate-forme. Les forums pour 'turkers' sont pleins de conversations sur l'automatisation du travail, le partage de pratiques sur la façon de créer des robots qui transgresseraient les termes d'Amazon. Vous pouvez également trouver des vidéos sur YouTube montrant aux 'turkers' comment écrire un bot qui remplit des réponses pour vous.

Kristy Milland, une militante de Mechanical Turk, dit : 'Les travailleurs sur Mechanical Turk ont été très, très mal traités pendant 12 ans et, d'une certaine façon, je vois cela comme un point de résistance. Si nous étions payés équitablement sur la plateforme, personne ne prendrait le risque de perdre son compte de cette façon.'

Bai a créé un questionnaire pour les chercheurs en dehors de Mechanical Turk. Il dirige actuellement une recherche parmi les spécialistes des sciences sociales pour déterminer la quantité de données erronées utilisées, l'ampleur du problème et les moyens de l'enrayer. Mais il est impossible à l'heure actuelle d'estimer combien de jeux de données sont devenus peu fiables de cette façon-ci.

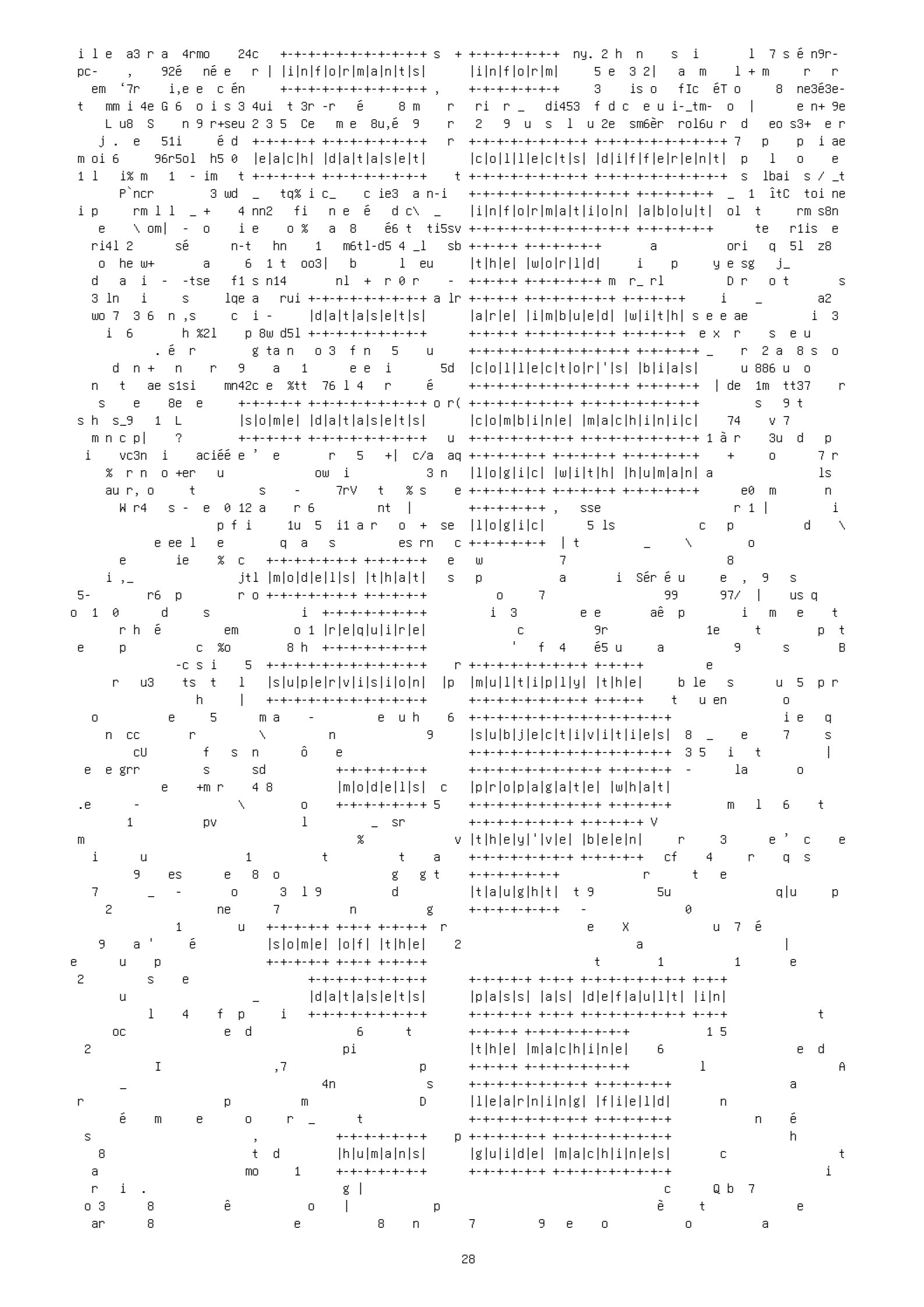
Références :

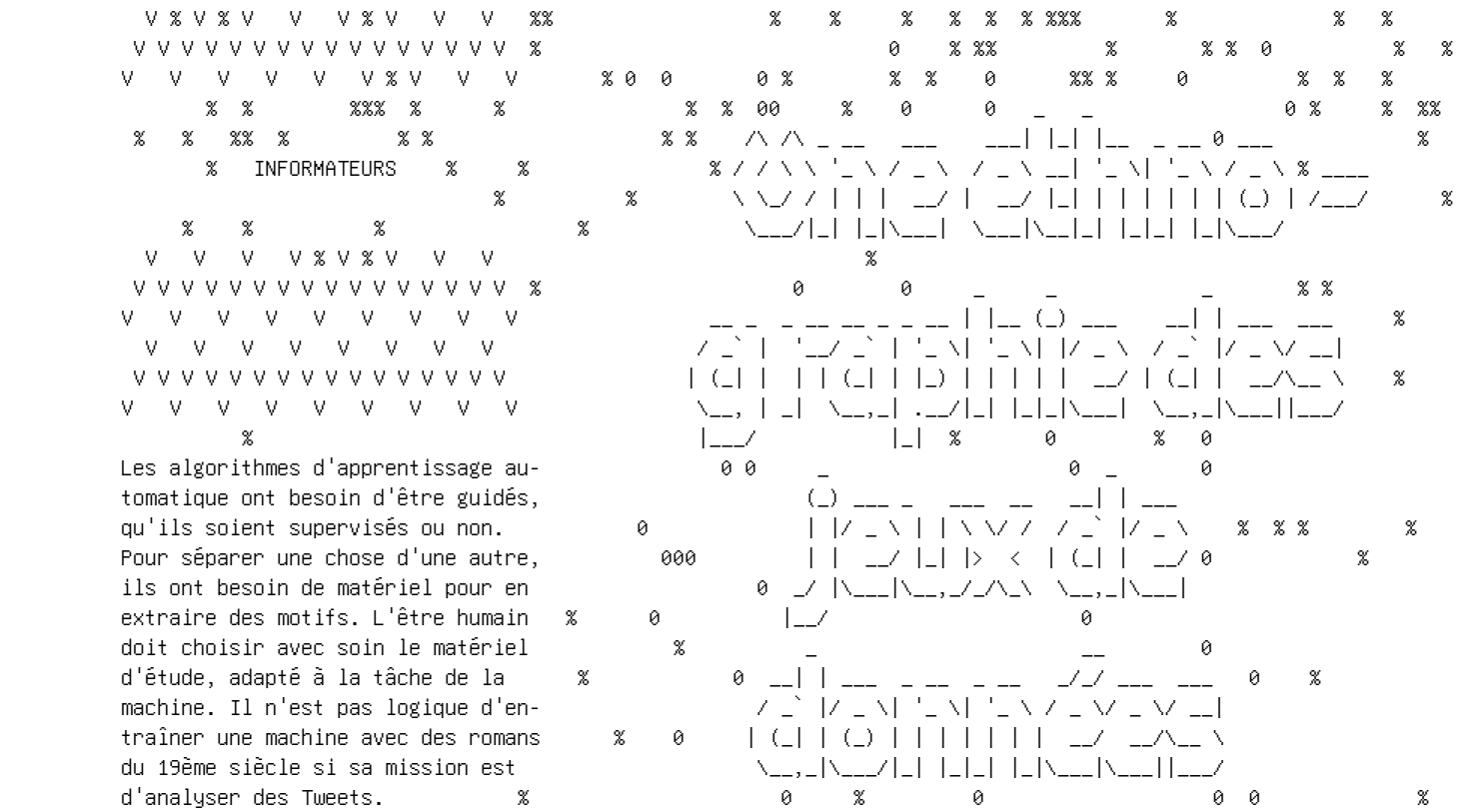
https://www.wired.com/story/amazon-mechanical-turk-bot-panic/

https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random

http://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/

informants inform informants inform informants inform informants inform informants inform info
mants inform informants inform informants inform informants inform informants i
form informants inform informants inform informants inform info
mants inform informants inform informants inform informants info
m informants inform informants inform informants inform
informants inform informants inform informants
inform informants inform informants inform
informants inform informants inform informants info
m informants inform informants inform
informants inform informants inform
informants inform informants inform in
ormants inform informants inform infor
ants inform informants inform info
mants inform informants inform
informants inform informants inform
informants inform informants inform
informants inform informants inform
informants inform infor
ants inform informants inform
informants inform informants inform
informants inform
informants inform informants inform
informants inform
informants inform informants inform
informants inform
informants inform informants inform
informants inform
informants inform informants
inform informants inform
informants inform
informants inform informants
inform informants inform
informants inform
informants inform
informants inform informants info
m informants inform
informants inform
informants inform
informants inform
informants inform informants
inform informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform in
ormants inform info
mants inform infor
ants inform infor
ants inform info
mants inform in
ormants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform
informants inform





Les algorithmes d'apprentissage automatique ont besoin d'être guidés, qu'ils soient supervisés ou non. Pour séparer une chose d'une autre, ils ont besoin de matériel pour en extraire des motifs. L'être humain doit choisir avec soin le matériel d'étude, adapté à la tâche de la machine. Il n'est pas logique d'entraîner une machine avec des romans du 19ème siècle si sa mission est d'analyser des Tweets.

C'est là qu'interviennent les jeux de données : organisés en rangés et en colonnes ordonnées, en attente d'être lus par la machine. Chaque jeu de données recueille des informations différentes sur le monde. Comme toutes les collections, elles sont imprégnées des stéréotypes et préjugés de ses créateurs. On entend souvent l’expression : 'les données sont le nouveau pétrole'. Si seulement les données étaient du pétrole ! Fuyantes, s’égouttant en graisse lourde, bouillonnantes et tressaillantes au contact d'une nouvelle matière. Au contraire, les données sont supposées d'être propres. Lors de chaque processus, chaque questionnaire, chaque titre de colonne, elles s’épurent, en effaçant peu à peu leurs caractéristiques distinctes jusqu’à correspondre au moule du jeu de données.

Certains jeux de données combinent la logique machinique avec la logique humaine. Les modèles qui nécessitent une supervision multiplient les subjectivités des collecteurs de données et des annotateurs, puis propulsent et propagent ce qui leur a été enseigné. Vous découvrirez des extraits de certains jeux de données qui passent par défaut dans le domaine de l'apprentissage automatique, ainsi que des histoires d'humains guidant des machines.

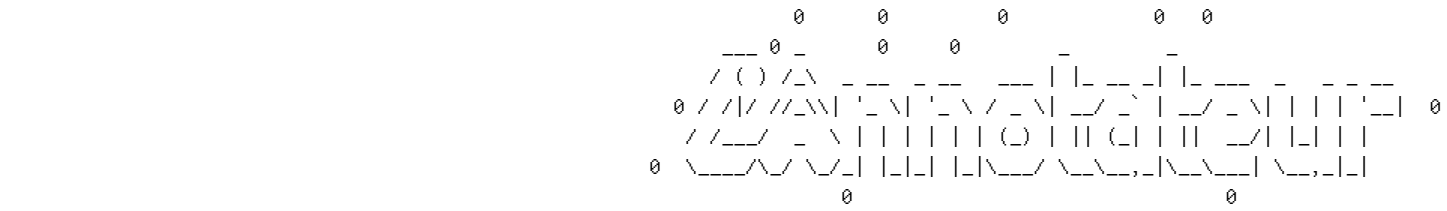
par Algolit

Lors des réunions mensuelles Algolit nous cherchons ou créons souvent des jeux de données. Parfois, nous utilisons des corpus déjà existants, disponibles via le site Natural Language Toolkit nltk. NLTK contient, entre autres, la Déclaration universelle des droits de l'Homme, les discours inauguraux des présidents américains, ou des critiques de films du site Internet Movie Database (IMDb).

Chaque style d'écriture évoque des relations différentes entre les mots et reflète l'époque dont ils proviennent. En ce sens, le gestionnaire de paquets Python pour le traitement du langage naturel pourrait être considéré comme une capsule temporelle. Le matériel inclu a été sélectionné car jugé utile par une communauté de chercheurs. Malgré les spécificités, chaque jeu de données devient universel par défaut, en étant à la disposition d'un public aussi large.

Nous examinons les jeux de données les plus couramment utilisés pour l'entraînement des modèles d'apprentissage automatique. De quels matériaux sont-ils constitués ? Qui les a recueillis ? Quand ?

Concept, réalisation: Cristina Cochior



par Algolit

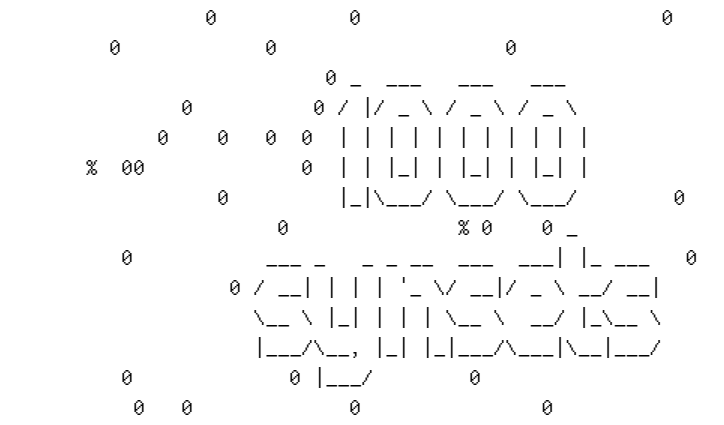
L'Annotateur demande au visiteur de l'aider à annoter les archives du Mundaneum.

Le processus d'annotation est une étape cruciale de l'apprentissage automatique supervisé où l'algorithme reçoit des exemples de ce qu'il doit apprendre. Un filtre anti-spam sera alimenté d'exemples de spam et de messages réels. Ces exemples sont des entrées du jeu de données prévues d'une étiquette, spam ou non spam.

L'annotation d'un jeu de données est un travail exécuté par des humains, qui choisissent une étiquette pour chaque entrée du jeu de données. Pour assurer la qualité des étiquettes, plusieurs annotateurs doivent voir la même entrée et donner la même étiquette avant qu'un exemple ne soit inclus dans les données d'entraînement. Une fois que toutes les données d'entraînement ont été prévues d'une étiquette, l'ordinateur peut lancer le processus d'apprentissage.

Dans cette interface, nous vous demandons de nous aider à classer les textes nettoyés des archives du Mundaneum afin d'élargir notre set d'entraînement et d'améliorer la qualité de l'installation 'Classer le Monde' dans Oracles.

Concept, code, interface : Gijs de Heij



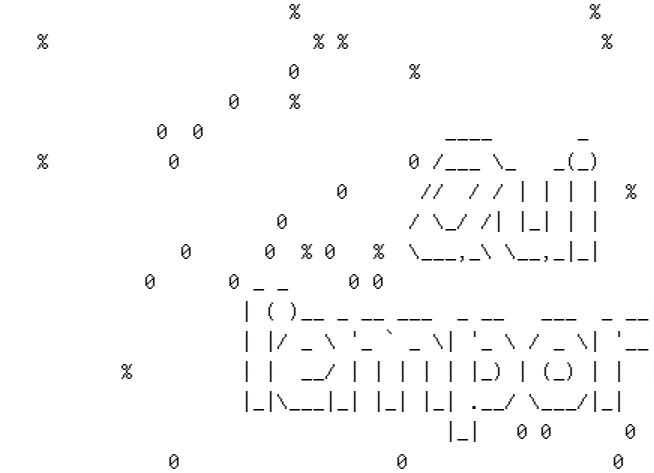
par Algolit

Wordnet, créé en 1985, est une taxonomie hiérarchique qui décrit le monde. Elle s'inspire des théories de la mémoire sémantique humaine développées à la fin des années 1960. Les noms, verbes, adjectifs et adverbes sont regroupés en collections de synonymes ou 'synsets', prévues de définitions, hypernymes, hyponymes, ... Chaque synset exprime des concepts différents. ImageNet est un jeu de données d'images basé sur la hiérarchie des noms de WordNet 3.0. Chaque synset est représenté par des milliers d'images. De 2010 à 2017, le Défi de Reconnaissance Visuelle de ImageNet (ILSVRC) a été une référence clé dans la classification des catégories d'objets pour les photos, ayant un impact majeur sur les logiciels de photographie, les recherches d'images, la reconnaissance d'images.

1000 synsets (édition vinyle) contient les 1000 synsets utilisés dans ImageNet, enregistrés dans la meilleure qualité sonore que ce format analogique permet. Ce travail souligne l'importance des jeux de données utilisés pour former des modèles d'intelligence artificielle qui fonctionnent sur des appareils que nous utilisons quotidiennement. Certains d'entre eux héritent de classifications qui ont été conçues il y a plus de 30 ans. Le vinyle est une invitation à les analyser en profondeur.

Conception et enregistrement: Javier Lloret

Voix: Sara Hamadeh & Joseph Hughes



Qui l'emporte: rapport de création

par Louise Dekeuleneer, étudiante Arts²/Option Communication Visuelle

Le français est une langue genrée, en effet beaucoup de mots sont féminins ou masculins et peu sont neutres. Le but de ce projet est de montrer qu'une société patriarcale influence aussi la langue même. Le travail s'est focalisé sur le fait de montrer si plus de mots féminins ou masculins sont utilisés et de mettre en valeur l'influence du contexte sur le genre des mots. À ce stade, aucune conclusion n'est encore tirée.

Des textes de loi datant de 1900 à 1910 mis à disposition par le Mundaneum sont passés dans un algorithme qui fait du texte une liste de mots. Ces mots sont alors comparés avec une autre liste de mots francophones, dans laquelle il est spécifié si le mot est masculin ou féminin. Cette liste de mots provient de Google Books, qui a créé en 2012 une énorme base de données à partir de tous les livres scannés et disponibles sur Google Books.
Les mots masculins sont surlignés d'une couleur et les féminins d'une autre. Les mots qui ne sont pas genrés (adverbes, verbes, ...) ne sont pas surlignés. Le tout est enregistré en fichier HTML pour qu'il puisse être directement ouvert dans une page web et imprimé sans besoin de mise en page supplémentaire. C'est ainsi que chaque texte a pu devenir un petit livret en changeant juste le texte d'entrée de l'algorithme.

--- Les jeux de données comme représentations ---

Les processus de collecte des données qui mènent à la création du jeu de données soulèvent des questions importantes : qui est l'auteur des données ? Qui a le privilège de collectionner ? Pour quelle raison la sélection a-t-elle été faite ? Que manque-t-il ?

L'artiste Mimi Onuoha donne un exemple excellent de l'importance des stratégies de collection. Elle choisit le cas des statistiques relatives aux crimes haineux. En 2012, le Programme de déclaration uniforme de la criminalité (DUC) du FBI a enregistré 5 796 crimes haineux. Toutefois, le Bu-

a établi 293 800 rapports sur de tels cas. C'est plus de 50 fois plus. La différence entre les chiffres peut s'expliquer par la façon dont les données ont été recueillies. Dans le premier cas, les organismes d'application de la loi de tout le pays ont volontairement signalé des cas. Pour le deuxième, le Bureau des statistiques a distribué l'enquête nationale sur la victimisation directement aux foyers des victimes de crimes motivés par la haine.

Dans le domaine du traitement du langage naturel, le matériel avec lequel les modèles d'apprentissage automatique travaillent est le texte, mais les mêmes questions se posent : qui sont les auteurs des textes qui composent les jeux de données ? Au cours de quelle période les données ont-elles été recueillies ? Quel type de vision du monde représentent-elles ?

En 2017, l'algorithme Top Stories de Google a placé un fil de discussion trompeur du site 4chan en haut de la page de résultats lors de la recherche du tireur de Las Vegas. Le nom et le portrait d'une personne innocente étaient liés au crime. Bien que Google ait changé son algorithme quelques heures seulement après que l'erreur ait été découverte, cela a sérieusement affecté la personne. Une autre question persiste : pourquoi Google n'a-t-il pas exclu le site de ragôts 4chan du jeu des données d'entraînement ?

Références :

https://points.datasociety.net/the-point-of-collection-8ee44ad7c2fa

https://arstechnica.com/information-technology/2017/10/google-admits-citing-4chan-to-spread-fake-vegas-shooter-news/

--- L'annotation pour un Oracle qui détecte le vandalisme sur Wikipédia ---

Ce fragment est extrait d'une interview avec Amir Sarabadani, ingénieur de logiciels chez Wikimedia. Il était à Bruxelles en novembre 2017 lors de la Rencontre Algolittéraire.

Femke : En considérant Wikipedia comme une communauté vivante, chaque nouvelle page change le projet. Chaque modification est en quelque sorte une contribution à un organisme vivant de la connaissance. Donc, si au sein de cette communauté vous essayez de distinguer ce qui rend service à la communauté et de généraliser ceci dans un modèle – car je pense que c'est ce que l'algorithme de la bonne ou mauvaise foi essaie de faire - vous le faites sur base d'une généralisation de l'idée abstraite de Wikipedia, et non sur base de l'organisme vivant. Ce qui m'intéresse dans la relation entre le vandalisme et ce débat, c'est la façon dont nous pouvons comprendre la dynamique conventionnelle de ces processus d'apprentissage automatique. Si on distingue la bonne ou la mauvaise foi sur base d'étiquettes préexistantes et qu'on la reproduit ensuite dans des modèles algorithmiques, comment tenir compte des changements qui se produisent, c'est-à-dire de la vie réelle du projet?

Amir : C'est une discussion intéressante. Premièrement, ce que nous appelons la bonne ou la mauvaise foi provient de la communauté elle-même; nous ne faisons pas l'annotation nous-mêmes, c'est la communauté qui le fait. Ainsi, dans beaucoup de Wikipedias de langues différentes, la définition de ce qui est la bonne ou la mauvaise foi sera différente. Wikimedia essaie de refléter ce qui se trouve à l'intérieur de l'organisme et non de changer l'organisme lui-même. Si l'organisme change et que nous constatons que la définition de la bonne foi à Wikipédia a été modifié, nous mettons en œuvre cette boucle de rétroaction qui permet aux gens de porter un jugement sur leurs modifications à l'intérieur de leur communauté. S'ils sont en désaccord avec l'annotation, nous pouvons revenir au modèle et modifier l'algorithme pour refléter ce changement. C'est une sorte de boucle fermée : vous changez les choses et si quelqu'un voit qu'il y a un problème, il nous le dit et nous pouvons modifier l'algorithme. C'est un projet en cours.

Référence : https://gitlab.constantvzw.org/algolit/algolit/blob/master/algoliterary_encounter/Interview%20with%20Amir/AS.aac

--- Comment faire connaître votre jeu de données ---

NLTK signifie Natural Language Toolkit. Pour les programmeurs qui traitent le langage naturel avec Python, c'est une bibliothèque essentielle. De nombreux rédacteurs de tutoriels recommandent aux programmeurs d'apprentissage automatique de commencer par les jeux de données NLTK intégrés. Il compte 71 collections différentes, avec un total de près de 6000 éléments.

Parmi eux, on trouve le corpus Movie Review pour l'analyse des sentiments. Ou le corpus Brown, qui a été créé dans les années 1960 par Henry Kučera et W. Nelson Francis à l'Université Brown de Rhode Island. Il y a aussi le corpus de la Déclaration des droits de l'homme, qui est couramment utilisé pour vérifier si un code peut fonctionner dans plusieures langues. Le corpus contient la Déclaration des droits de l'homme dans 372 langues du monde entier.

Mais quel est le processus pour faire accepter un jeu de données dans la bibliothèque NLTK de nos jours ? Sur la page Github, l'équipe nltk décrit les exigences suivantes :

- Ne rajoutez que les corpus qui ont obtenu un niveau de notabilité de base. Cela signifie qu'il existe une publication qui le décrit et une communauté de programmeurs qui l'utilisent.

- Assurez-vous d'avoir l'autorisation de redistribuer les données et de pouvoir les documenter. Cela signifie qu'il est préférable de publier le jeu de données sur un site Web externe avec une licence.

- Utilisez les lecteurs de corpus NLTK existants lorsque c'est possible, ou bien apportez un lecteur de corpus bien documenté à NLTK. Cela signifie que vous devez organiser vos données de manière à ce qu'elles puissent être facilement lues à l'aide du code NLTK.

Référence : http://www.nltk.org/

--- Extrait d'une critique positive d'un film IMdB du jeu de données NLTK ---

corpus : movie_reviews

fichier : pos/cv998_14111.txt

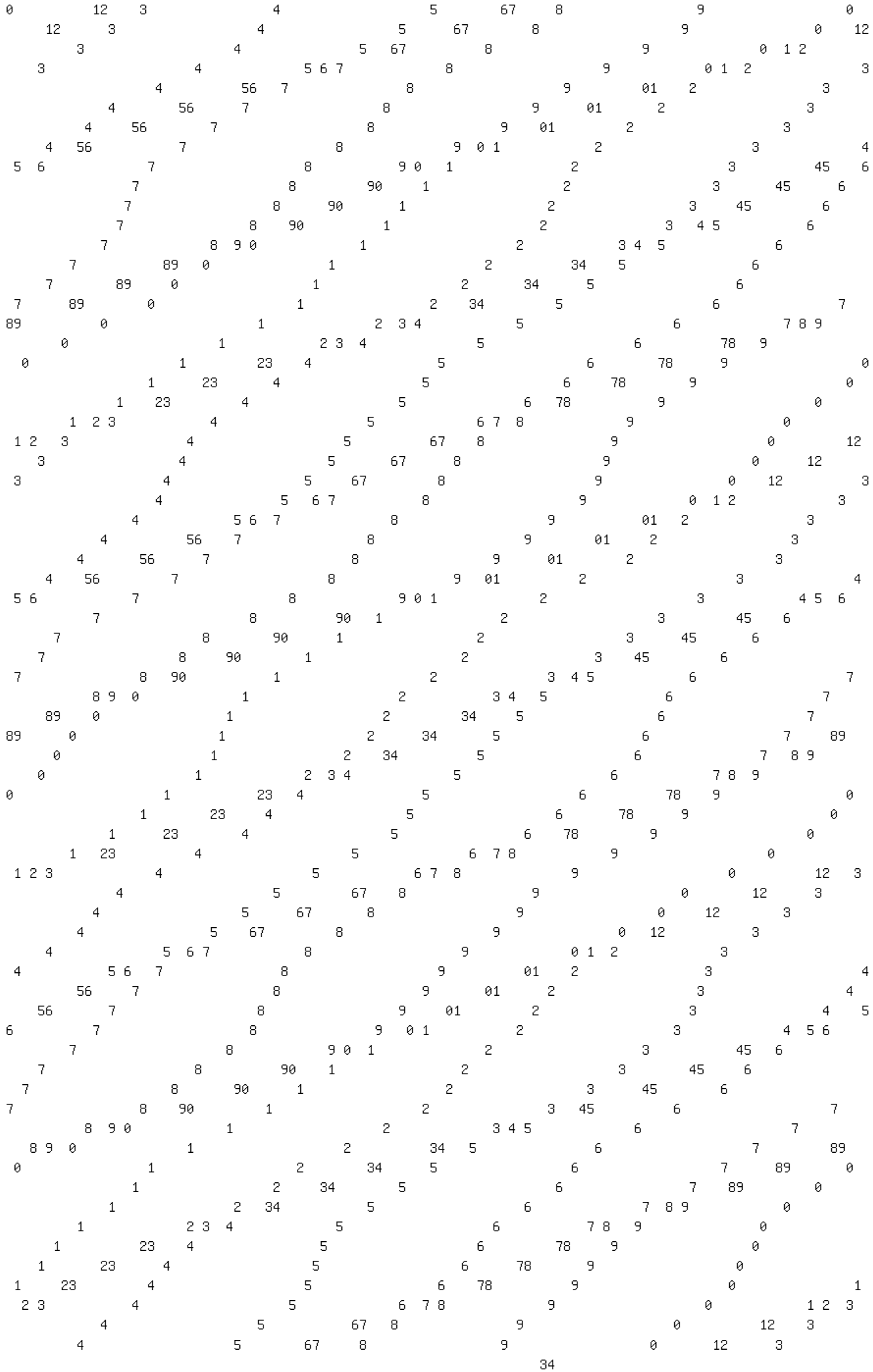
le deuxième film épique de steven spielberg sur la seconde guerre mondiale est un chef-d'œuvre incontesté du cinéma . spielberg , encore étudiant en cinéma , a réussi à ressusciter le genre de la guerre en produisant l'un de ses films les plus poignants et les plus puissants . il a également réussi à faire briller tom hanks , qui livre une performance époustouflante . pendant environ 160 de ses 170 minutes, ' sauver le soldat ryan ' est sans faille . littéralement . l ' histoire est assez simple . après l ' invasion du jour J (dont les séquences sont tout à fait spectaculaires), capt . john miller (joué par tom hanks) et son équipe sont forcés à chercher un soldat . james ryan (joué par matt damon), dont les frères sont tous morts au combat. une fois qu ' ils l ' ont trouvé , ils doivent le ramener immédiatement pour qu'il puisse rentrer chez lui . la compagnie de miller est composée d ' acteurs aux jeux tout simplement sensationnels : bary pepper , adam goldberg , vin diesel , giovanni ribisi , davies et burns . le film se clôture avec des scènes de bataille extraordinaires .

--- Les ouroboros de l'apprentissage automatique ---

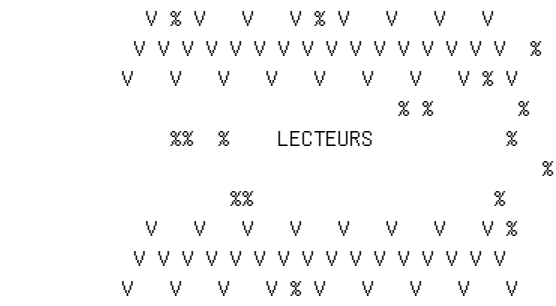
Wikipédia est devenue une source d'apprentissage non seulement pour les humains, mais aussi pour les machines. Ses articles sont des sources de premier ordre pour l’entraînement de modèles. Le matériel avec lequel les machines sont entraînées est identique au contenu qu'elles ont aidé à écrire. En fait, au début de Wikipédia, de nombreux articles ont été écrits par des robots. Rambot, par exemple, était un robot controversé sur la plateforme anglophone. Il est l'auteur de 98% des pages décrivant les villes américaines.

A cause de ces interventions de robots thématiques et régulières, les modèles de prédiction qui sont entraînés sur le dump de Wikipedia ont une vision unique de la composition des articles. Par exemple, un modèle thématique entraîné sur l'ensemble des articles de Wikipédia associe 'rivière' à 'Roumanie' et 'village' à 'Turquie'. C'est parce qu'il y a plus de 10000 pages écrites sur les villages en Turquie. Cela devrait suffire à susciter des envies de voyage, mais c'est bien trop par rapport à d'autres pays. L'asymétrie provoque une fausse corrélation et doit être corrigée. La plupart des modèles tentent d'exclure le travail de ces auteurs robots prolifiques.

Référence : https://blog.lateral.io/2015/06/the-unknown-perils-of-mining-wikipedia/

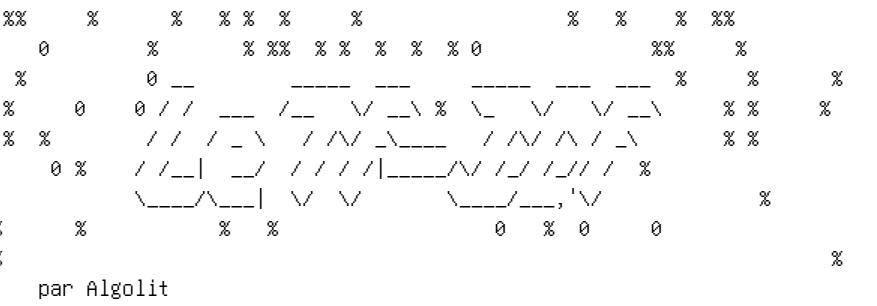


readers read readers read readers read readers read readers read readers read readers re
d readers read readers read readers read readers read readers re
d readers read readers read readers read readers read
readers read readers read readers read re
ders read readers read readers read readers re
d readers read readers read readers r
ad readers read readers read
readers read readers read readers read
readers read readers read
readers read readers read readers read
readers read readers read
readers read readers read
readers read readers read
readers read readers read
readers read readers read
readers read readers read
readers read readers
read readers read
readers read readers read
readers read readers read
readers read
readers read readers read
readers read
readers read readers read
readers read
readers read readers read
readers read
readers read readers re
d readers read
readers read
readers read readers read
readers read
readers read
readers read re
ders read readers read
readers read
readers read
readers read
readers read readers r
ad readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read readers
read readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read
readers read r



Nous communiquons avec les ordinateurs au moyens de langages. Nous cliquons sur des icônes sous forme de mots, nous tapons des mots sur des claviers, nous utilisons notre voix pour leur donner des instructions. Parfois, nous confions nos pensées les plus intimes à notre ordinateur en oubliant qu'il s'agit d'une calculatrice avancée. Un ordinateur comprend chaque mot comme une combinaison de zéros et de uns. Une lettre est lue comme un numéro ASCII spécifique : 'A' majuscule est 001.

Dans tous les 'rule-based models', l'apprentissage automatique classique et les réseaux de neurones, les mots subissent une traduction en chiffres pour saisir le sens sémantique du langage. Cela se fait en comptant. Certains modèles comptent la fréquence des mots simples, d'autres la fréquence des combinaisons de mots, d'autres encore la fréquence des noms, des adjectifs, des verbes ou des phrases de noms et de verbes. Certains remplacent simplement les mots d'un texte par leur numéro d'index. Les nombres optimisent la vitesse opérationnelle des processus informatiques, ce qui conduit à des prédictions rapides, mais ils suppriment aussi les liens symboliques que les mots peuvent avoir. Nous présentons ici quelques techniques destinées à rendre un texte intelligible pour une machine.

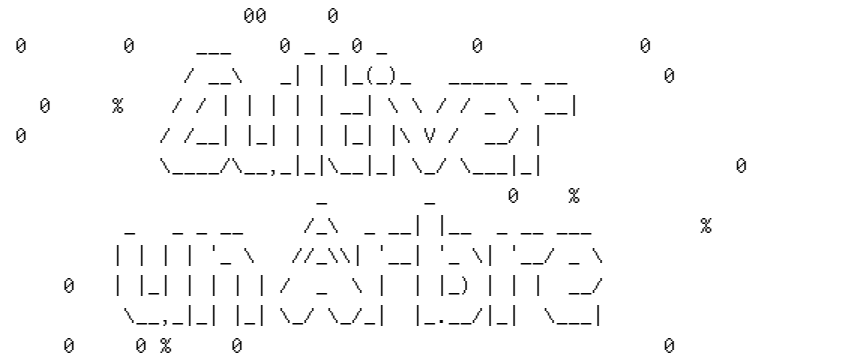


par Algolit

Le TF-IDF (Term Frequency-Inverse Document Frequency) est une méthode de pondération utilisée dans la recherche de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus de documents. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Le TF-IDF est notamment utilisé dans la classification des spams.

Une interface web met en scène cet algorithme à travers des animations permettant de comprendre les différentes étapes de classification d'un texte. Comment un programme basé sur le TF-IDF lit un texte ? Comment transforme-t-il les mots en nombres ?

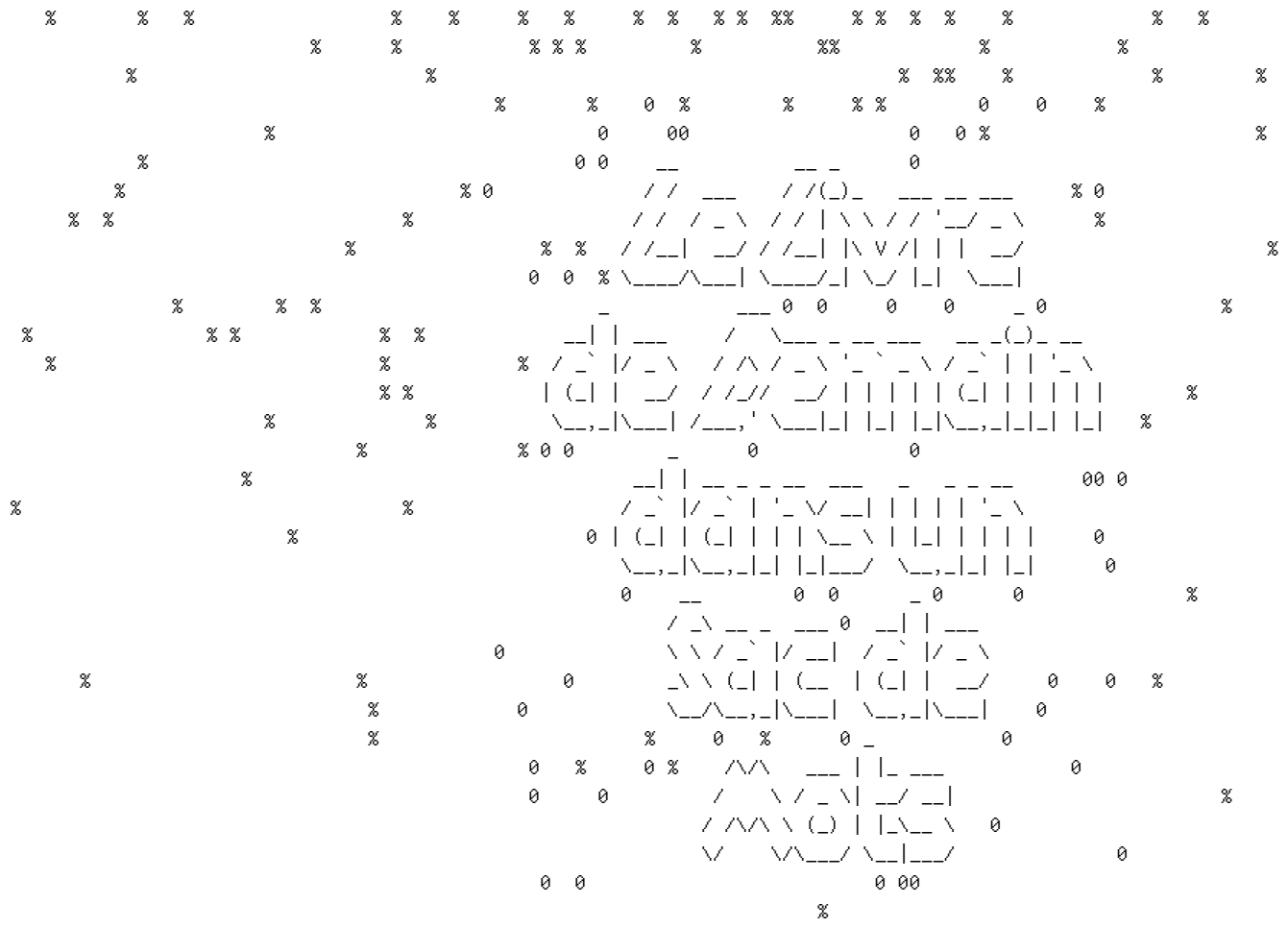
Concept, code, animation : Sarah Garcin



La nature de mots est une catégorie que nous apprenons à l'école : nom, verbe, adjectif, adverbe, pronom, préposition, conjonction, interjection, et parfois chiffre, article, ou déterminant. Dans le traitement du langage naturel, il existe de nombreux écrits qui permettent d'analyser des phrases. Cela signifie que l'algorithme peut déterminer la nature de chaque mot d'une même phrase. 'Cultiver un arbre' utilise cette technique pour définir tous les noms dans une phrase spécifique. Chaque nom est alors remplacé par sa définition. Cela permet à la phrase de grandir de façon autonome et infinie. La recette de 'Cultiver un arbre' s'inspire de la 'Littérature Définitionnelle', une contrainte inventée par Marcel Benabou en 1966 au sein de l'Oulipo. Dans une phrase donnée, on remplace chaque élément significatif (nom, adjectif, verbe, adverbe) par l'une de ses définitions dans un dictionnaire donné ; on répète l'opération sur la nouvelle phrase reçue, et ainsi de suite.

Le dictionnaire utilisé dans cet ouvrage est Wordnet. Wordnet est une combinaison d'un dictionnaire et d'un thésaurus qui peut être lu par des machines. Selon Wikipédia, il a été créé dans le Cognitive Science Laboratory de l'Université de Princeton à partir de 1985.

Concept, code & interface : An Mertens & Gijs de Heij



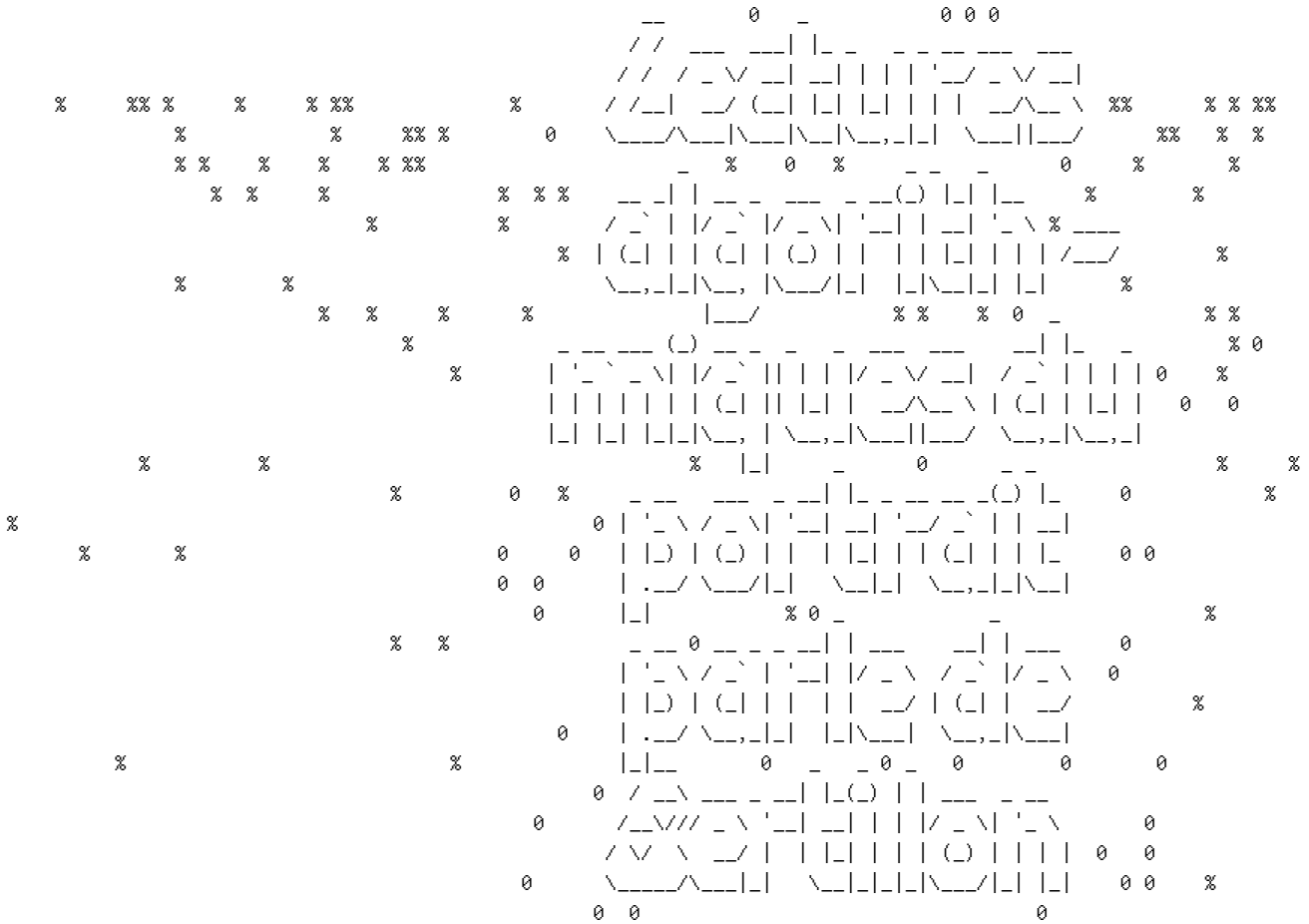
% par Algolit

Le modèle du 'sac de mots' est une représentation simplifiée du texte utilisé dans le traitement du langage naturel. Dans ce modèle, un texte est représenté sous forme de collection de mots uniques, sans tenir compte de la grammaire, de la ponctuation et même de l'ordre des mots. Le modèle transforme le texte en une liste de mots et leur occurrence dans le texte, littéralement un sac de mots.

Cette forte réduction de la langue fut un choc au début de nos expériences en apprentissage automatique. Le sac de mots est souvent utilisé comme référent, sur base duquel le nouveau modèle doit s'efforcer d'être plus performant. Il peut comprendre le sujet d'un texte en reconnaissant les mots les plus fréquents ou importants. On mesure souvent les similitudes des textes en comparant leurs sacs de mots.

Pour cet ouvrage, l'article 'Le Livre de Demain' de l'ingénieur G. Vander Haeghen, publié en 1907 dans le Bulletin de l'Institut International de Bibliographie, a été littéralement réduit à un sac de mots. Vous pouvez acheter votre exemplaire à l'accueil du Mundaneum.

Concept & réalisation: An Mertens



par Guillaume Slizewicz (Espèces urbaines)

'Un code télégraphique du portrait parlé', écrit en 1907, est une tentative de traduire en chiffres le 'portrait parlé', technique de description du visage créée par Alphonse Bertillon, créateur de l'anthropométrie judiciaire. En appliquant ce code, Otlet espérait que les visages des criminels et des fugitifs pourraient être facilement communiqués par voie télégraphique. Dans sa forme, son contenu et son ambition, ce texte représente la relation complexe que nous entretenons avec les technologies documentaires. Ce document a été choisi comme base pour la création des installations suivantes pour trois raisons.

- Premièrement, ce texte est un algorithme en soi, un algorithme de compression, ou pour être plus précis, la présentation d'un algorithme de compression. Il tente de réduire la taille de l'information tout en la gardant lisible pour la personne possédant le code. À cet égard, elle est étroitement liée à la façon dont nous créons notre technologie, à la recherche d'une plus grande efficacité, de résultats plus rapides et de méthodes moins coûteuses. Il représente notre appétit de chiffrement qui s'étend au monde entier, notre envie de mesurer les plus petites choses, d'étiqueter les différences les plus infimes... Ce texte incarne en lui-même la vision du Mundaneum.

- Deuxièmement, on y traite des raisons et des mises en œuvre de nos technologies. La présence de ce texte dans les archives sélectionnées est presque ironique à une époque où la reconnaissance faciale et la surveillance des données font la une des journaux. Ce texte présente les mêmes caractéristiques que certaines technologies d'aujourd'hui : il est motivé par un contrôle social, classifie les personnes, pose les bases d'une société de surveillance. Les caractéristiques physionomiques sont au cœur de récentes controverses : les photos d'identité ont été standardisées par Bertillon, elles sont maintenant utilisées pour entraîner des réseau neuronaux à identifier les criminels, les systèmes

de reconnaissance faciale permettent des arrestations via notre infrastructure de caméras de surveillance et certains affirment que les caractéristiques physiques peuvent prédire l'orientation sexuelle.

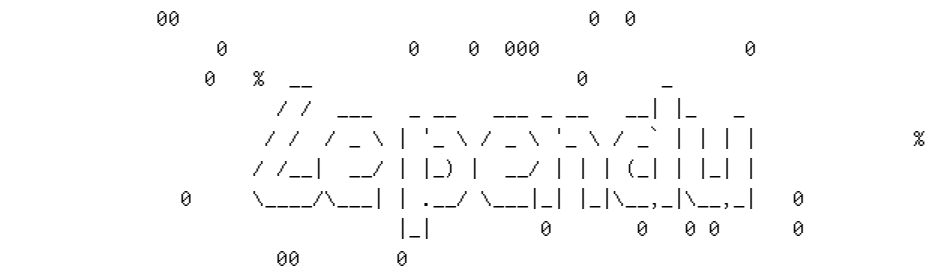
- Le dernier point concerne la façon dont, en tant que témoignage écrit, ce texte représente l'évolution de notre techno-structure: ce que nos outils nous permettent de faire, ce qu'ils nous interdisent, ce qu'ils entravent, ce qu'ils nous font retenir et ce qu'ils nous font oublier. Ce document permet une classification entre les personnes, et instaure une normalité. Il brise un continuum en morceaux, et permet les stigmatisations et les discriminations. D'un autre côté, ce document semble également obsolète aujourd'hui, car cette techno-structure n'a pas besoin de descriptions écrites aussi détaillées sur les fugitifs, les criminels ou les citoyens. Nous pouvons maintenant trouver des empreintes digitales, des scanners d'iris ou des informations ADN dans de grands jeux de données et les comparer directement. Parfois, les systèmes agissent indépendamment, sans surveillance humaine et reconnaissent directement l'identité d'une personne par ses traits faciaux ou sa démarche. Ces machines n'utilisent pas un langage alphabétique complexe pour décrire un visage, mais des listes de chiffres. Ainsi, tous les mots utilisés dans ce document semblent désuets, datés. Avons-nous oublié ce que certains d'entre eux signifient ? La photographie nous a-t-elle fait oublier comment décrire les visages ? Les assistants vocaux nous l'apprendront-il de nouveau ?

Écrire avec Otlet

Ecrire avec Otlet est un générateur de personnages qui utilise le code du portrait parlé comme base de données. Des nombres aléatoires sont générés et traduits en un ensemble de caractéristiques humaines. En créant des instances uniques, l'algorithme révèle la richesse de la description qui est possible avec 'Un code du portrait' tout en incorporant ses nuances.

Interprétation du portrait parlé de Bertillon

Ce travail établit un parallèle entre le système dit de "Bertillonage" et les systèmes actuels de description de visage. Une webcam associée à un algorithme de reconnaissance faciale capte le visage du spectateur et le traduit en chiffres sur un écran, en l'imprimant à côté des visages annotés par Bertillon.



par Laetitia Trozzi, étudiante Arts²/Section Arts Numériques

Quoi de mieux pour découvrir Paul Otlet et sa passion pour la littérature que de jouer au pendu? À travers ce jeu simple, qui consiste à deviner les lettres manquantes dans un mot, le but est de faire découvrir au public des termes et des faits liés à un des créateurs du mundaneum.

En utilisant un algorithme de détection de fréquence de mots dans un texte, une série de mots significatifs ont été isolés dans la bibliographie de Paul Otlet. Cette série de mots a ensuite été intégrée à un jeu du pendu présenté dans un terminal. La difficulté du jeu augmente graduellement en proposant au joueur des mots de plus en plus longs. Durant les phases de jeux, des informations sont affichées en lien avec la vie et l'œuvre de Paul Otlet.

Naive Bayes, Support Vector Machines ou Régression Linéaire sont considérés comme des algorithmes classiques d'apprentissage automatique. Ils fonctionnent bien lorsqu'ils apprennent avec de petits jeux de données. Mais ils nécessitent souvent des lecteurs complexes. La tâche accomplie par les lecteurs est également appelée 'feature engineering'. Cela signifie qu'un être humain doit consacrer du temps à une analyse exploratoire approfondie du jeu de données.

Leurs caractéristiques peuvent être la fréquence des mots ou des lettres, mais aussi des éléments syntaxiques comme les noms, les adjectifs ou les verbes. Les caractéristiques les plus importantes pour la tâche à résoudre doivent être soigneusement sélectionnées et transmises à l'algorithme classique d'apprentissage automatique. Ce processus diffère de celui des réseaux de neurones. Lors de l'utilisation d'un réseau de neurones, il n'est pas nécessaire de recourir au 'feature engineering'. Les humains peuvent transmettre les données directement au réseau et obtiennent généralement de bonnes performances dès le départ. Cela permet d'économiser beaucoup de temps et de ressources.

L'inconvénient de la collaboration avec les réseaux de neurones est que vous avez besoin de beaucoup plus de données pour entraîner votre modèle de prédiction. Pensez à au moins 1 Go de fichiers texte. Pour vous donner une référence, 1 A4, soit un fichier texte de 5000 caractères, ne pèse que 5 Ko. Il vous faudrait donc 8.589.934 pages. Traiter plus de données sous-entend d'avoir accès à ces données et surtout, d'avoir beaucoup plus de puissance de traitement.

--- Les N-grammes de caractères pour la reconnaissance d'un auteur ---

Imaginez... vous travaillez pour une entreprise depuis plus de dix ans. Vous avez écrit des tonnes de courriels, d'articles, de notes internes et de rapports sur des sujets et dans des genres très différents. Tous vos écrits, ainsi que ceux de vos collègues, sont sauvegardés en toute sécurité sur les serveurs de l'entreprise.

Un jour, vous tombez amoureuse d'une collègue. Après un certain temps, vous réalisez que cette personne est non seulement folle et hystérique mais qu'elle dépend beaucoup de vous. Le jour où vous décidez de rompre, votre ex élabore un plan pour vous tuer. Elle réussit. Pas de chance. Une lettre de suicide signée de votre nom est retrouvée à côté de votre cadavre. Celle-ci raconte que vous avez décidé de mettre fin à votre vie à cause de problèmes émotionnels. Vos meilleurs amis ne croient pas au suicide. Ils décident de porter l'affaire devant les tribunaux. Et là, à partir des textes que vous et d'autres avez produits, un modèle d'apprentissage automatique révèle que la lettre de suicide a été écrite par quelqu'un d'autre.

Comment une machine analyse-t-elle les textes pour vous identifier ? La caractéristique la plus robuste pour la reconnaissance de l'auteur est fournie par la technique des N-grammes de caractères. Elle est utilisée dans des cas qui présentent une grande variété dans les thématiques et les genres d'écriture. Lors de l'utilisation des N-grammes de caractères, les textes sont considérés comme des séquences de caractères. Considérons le trigramme des caractères. Toutes les séquences de trois caractères qui se chevauchent sont isolées. Par exemple, le trigramme de caractères de 'suicide', serait, 'sui', 'uic', 'ici', 'cid' et 'ide'. Les N-grammes de caractères sont très simples, ils sont indépendants du langage et tolérants au bruit. De plus, les fautes d'orthographe ne compromettent pas la technique.

Les motifs trouvés avec les N-grammes de caractères se concentrent sur les choix stylistiques qui sont faits inconsciemment par l'auteur. Les modèles restent stables sur toute la longueur du texte, ce qui est important pour reconnaître l'auteur. D'autres types d'expériences pourraient inclure la longueur des mots ou des phrases, la richesse du vocabulaire, la fréquence des mots de fonction et même les mesures syntaxiques ou sémantiques.

Cela signifie non seulement que votre empreinte physique est unique, mais qu'il en va de même de la façon dont vous composez vos pensées !

La même technique n-gramme a découvert que 'The Cuckoo's Calling', un roman de Robert Galbraith, a en fait été écrit par... J.K. Rowling !

Références :

- Essai: On the Robustness of Authorship Attribution Based on Character N-gram Features, Efstathios Stamatatos, in Journal of Law & Policy, Volume 21, Issue 2, 2013.
- Article: https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/

--- Histoire des N-grammes ---

L'algorithme des N-grammes peut être retracé jusqu'aux travaux de Claude Shannon en théorie de l'information. Dans l'article 'A mathematical theory of communication', publié en 1948, Claude Shannon réalise la première instance d'un modèle de langage naturel à base des N-grammes. Il a posé la question suivante : étant donné la séquence des lettres, quelle est la probabilité de la prochaine lettre ? Si vous lisez l'extrait suivant, pouvez-vous nous dire par qui il a été écrit ? Shakespeare ou un robot N-grammes ?

SEBASTIEN : Dois-je rester debout jusqu'à la rupture.
BIRON : Cache ta tête.
VENTIDIUS : Il se rendit à Athènes, où, par le voeu. que j'ai fait pour m'occuper de toi.
FALSTAFF : Mon bon fripouille.

Vous aviez peut-être deviné, en considérant le sujet de ce récit, qu'un algorithme N-grammes a généré ce texte. Le modèle est entraîné sur l'oeuvre complète de Shakespeare. Alors que les algorithmes plus récents, tels que les réseaux de neurones récursifs de CharRNN, deviennent célèbres pour leurs performances, les N-grammes exécutent encore beaucoup de tâches NLP. Elles sont utilisés dans la traduction automatique, la reconnaissance vocale, la correction orthographique, la détection d'entités, l'extraction d'informations, etc.

Référence : http://www.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

--- Dieu dans Google Books ---

En 2006, Google crée un jeu de données de N-grammes à partir de sa collection de livres numérisés pour le mettre en ligne. Récemment, ils ont également réalisé une visionneuse de N-grammes. Cela a permis de nombreuses recherches sociolinguistiques. Par exemple, en octobre 2018, le New York Times Magazine a publié un article d'opinion intitulé 'It's Getting Harder to Talk About God'. L'auteur, Jonathan Merritt, avait analysé la mention du mot 'Dieu' dans le jeu de données de Google à l'aide du visualiseur de N-grammes. Il a conclu qu'il y a eu un déclin dans l'usage du mot depuis le 20ème siècle. Le corpus de Google contient des textes du 16e jusqu'au 21e siècle. Cependant l'auteur a manqué d'observer la popularité croissante des revues scientifiques vers le début du 20ème siècle. Ce nouveau genre, dans lequel le mot Dieu n'apparaît pas, a fait basculer le jeu des données. Si la littérature scientifique était retirée du corpus, la fréquence du mot 'Dieu' s'écoulerait toujours comme l'ondulation douce d'une vague lointaine.

Référence : https://www.nytimes.com/2018/10/13/opinion/sunday/talk-god-sprituality-christian.html

--- Les traits grammaticaux extraits de Twitter influencent le marché boursier ---

Les frontières entre les disciplines académiques s'estompent. La recherche économique mélangée à la psychologie, aux sciences sociales, aux concepts cognitifs et émotionnels créent un nouveau sous-domaine économique, appelé 'l'économie comportementale'.

Cela signifie que les chercheurs commencent à expliquer un mouvement boursier basé sur d'autres facteurs que les facteurs purement économiques. La Bourse et 'l'opinion publique' s'influencent mutuellement. De nombreuses recherches sont effectuées sur la façon d'utiliser 'l'opinion publique' pour prédire les tendances dans le cours des actions.

'L'opinion publique' est évaluée à partir de grandes quantités de données publiques, comme les tweets, les blogs ou la presse en ligne. Des recherches montrent que l'évolution des cours boursiers peut, dans une certaine mesure, être prédit en examinant 'l'opinion publique' à travers l'analyse des données automatique. On trouve de nombreux articles scientifiques en ligne, qui analysent la presse sur le 'sentiment' qui y est exprimé. Un article peut être annoté comme plus ou moins positif ou négatif. Les articles de presse annotés sont ensuite utilisés pour entraîner un modèle d'apprentissage automatique, qui permet de prédire les tendances boursières, en les marquant comme 'à la baisse' ou 'à la hausse'. Quand une entreprise fait mauvaise presse, les traders vendent. Au contraire, si les nouvelles sont bonnes, ils achètent.

Un article de Haikuan Liu de l'Université Nationale Australienne affirme que le temps des verbes utilisés dans les tweets peut être un indicateur de la fréquence des transactions financières. Son idée s'inspire du fait que la conjugaison des verbes est utilisée en psychologie pour détecter les premiers stades de la dépression humaine.

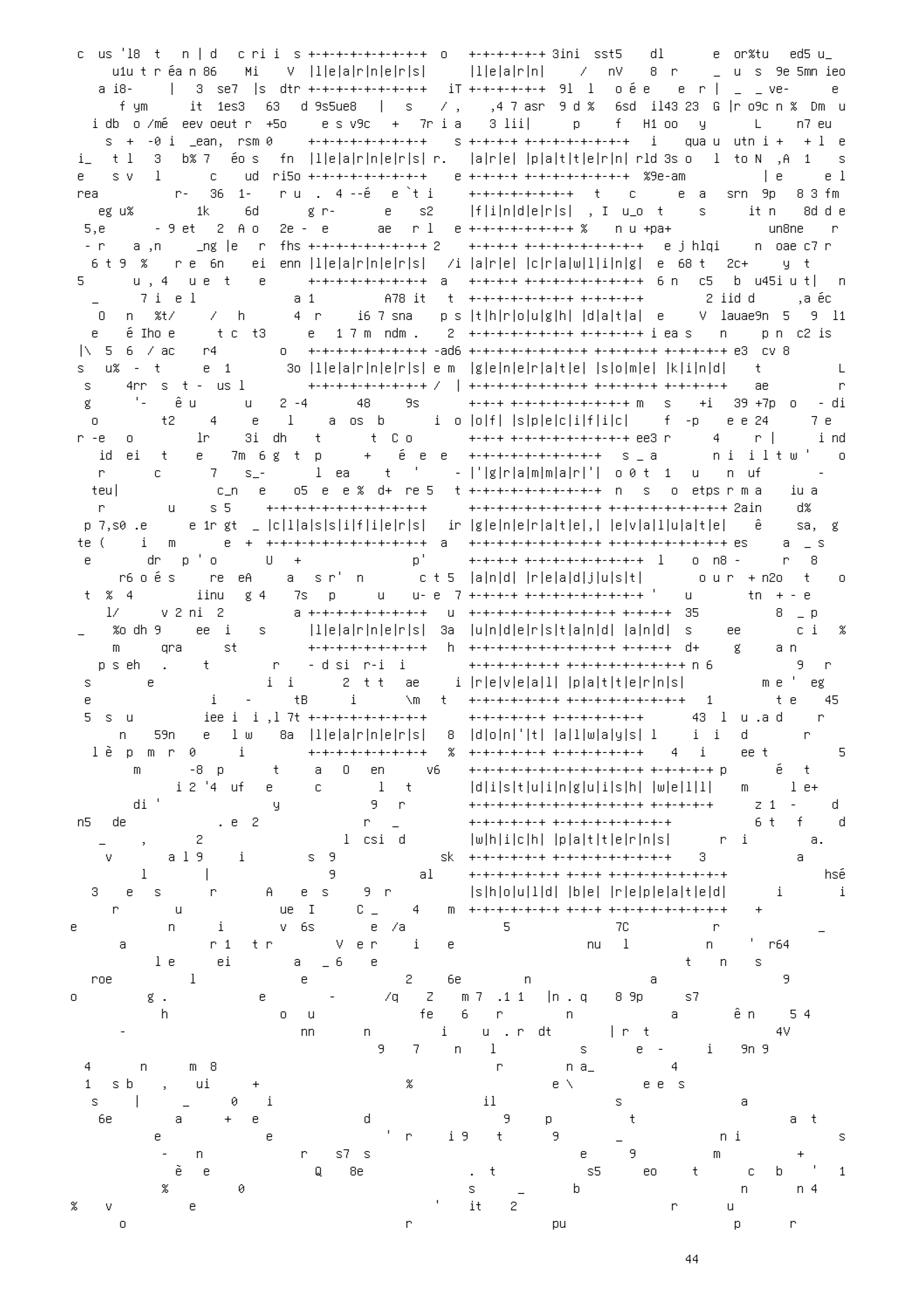
Référence : Grammatical Feature Extraction and Analysis of Tweet Text: An Application towards Predicting Stock Trends, The Australian National University (ANU)

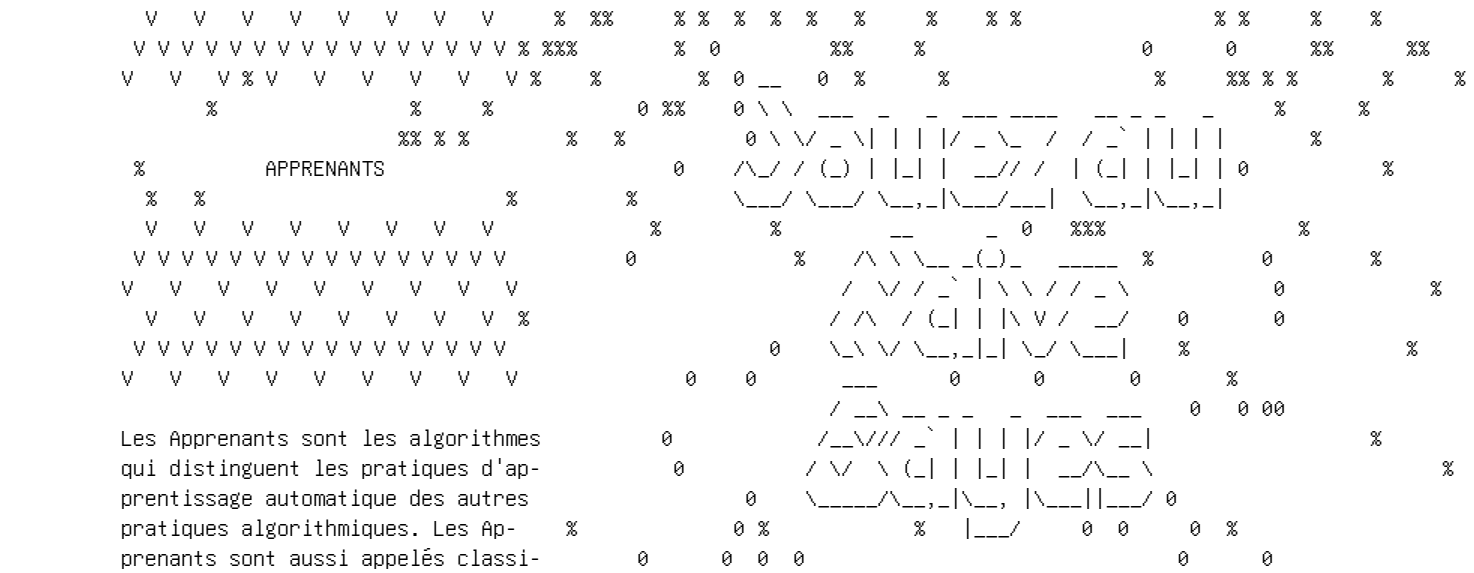
--- Sac de mots ---

Dans le traitement du langage naturel, le 'sac de mots' est considéré comme un modèle simple. Il dépouille un texte de son contexte et le décompose dans sa collection de mots uniques. Ensuite, ces mots sont comptés. Dans les phrases précédentes, par exemple, le mot 'mots' est mentionné trois fois, mais ce n'est pas nécessairement un indicateur de l'objet du texte.

La première apparition de l'expression 'sac de mots' semble remonter à 1954. Zellig Harris a publié un article dans le contexte des études linguistiques, intitulé 'Distributional Structure'. Dans la partie intitulée 'Le sens en fonction de la distribution', il dit que 'le langage n'est pas seulement un sac de mots, mais aussi un outil aux propriétés particulières qui ont été façonnées au cours de son utilisation. Le travail du linguiste est précisément de découvrir ces propriétés, que ce soit pour l'analyse descriptive ou pour la synthèse du système quasi-linguistique.'

learners learn learners learn learners learn learners learn learners learn learners learn
learners learn learners learn learners learn learners learn learners learn
learners learn learners learn learners learn learners learn
learners learn learners learn learners learn
learners learn learners learn learners learn lea
ners learn learners learn learners learn
learners learn learners learn learners learn
learners learn learners learn learners
earn learners learn learners learn
learners learn learners learn
learners learn learners learn lea
ners learn learners learn learners
learn learners learn learners
earn learners learn learne
s learn learners learn
learners learn learners learn
learners learn learners learn
learners learn learners learn
learners learn
learners learn learners learn
learners learn learners learn
learners learn
learners learn learners learn
learners learn
learners learn learners learn
learners learn
learners learn learners
learn learners learn
learners learn
learners learn learners learn
learners learn
learners learn
learners learn learners
learn learners learn
learners learn
learners learn
learners learn lea
ners learn learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn learners
earn learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn
learners learn





Les Apprenants sont les algorithmes qui distinguent les pratiques d'apprentissage automatique des autres pratiques algorithmiques. Les Apprenants sont aussi appelés classificateurs. Ce sont des chercheurs de motifs, capables de fouiller dans les données et de générer une sorte de 'grammaire' spécifique. Les Apprenants sont souvent basés sur des techniques statistiques. Chacun d'entre eux présente des caractéristiques individuelles. Certains ont besoin d'une grande quantité de données d'entraînement pour fonctionner, d'autres peuvent s'en tirer avec un petit jeu de données annotées. Certains s'acquittent bien de tâches de classification, comme l'identification des spam, d'autres sont plus aptes à prédire les chiffres, comme les températures, les distances, les valeurs boursières, et ainsi de suite.

La terminologie de l'apprentissage automatique n'est pas encore complètement établie. Selon le domaine (les statistiques, l'informatique ou les sciences humaines) ils sont appelés par des mots différents. Lorsque nous parlons d'Apprenants, nous parlons des fonctions imbriquées qui ont la capacité de générer d'autres fonctions, de les évaluer et de les réajuster en fonction des données. Les Apprenants sont bons pour comprendre et révéler les motifs. Mais ils ne distinguent pas toujours bien quels motifs doivent être répétés.

Dans les logiciels, il n'est pas toujours possible de distinguer les éléments caractéristiques des classificateurs, car ils sont cachés dans des modules ou bibliothèques sous-jacents. Les programmeurs peuvent les invoquer en utilisant une seule ligne de code. Par conséquent, pour cette exposition, nous avons développé deux jeux de table qui montrent en détail le processus d'apprentissage de classificateurs simples, mais fréquemment utilisés.

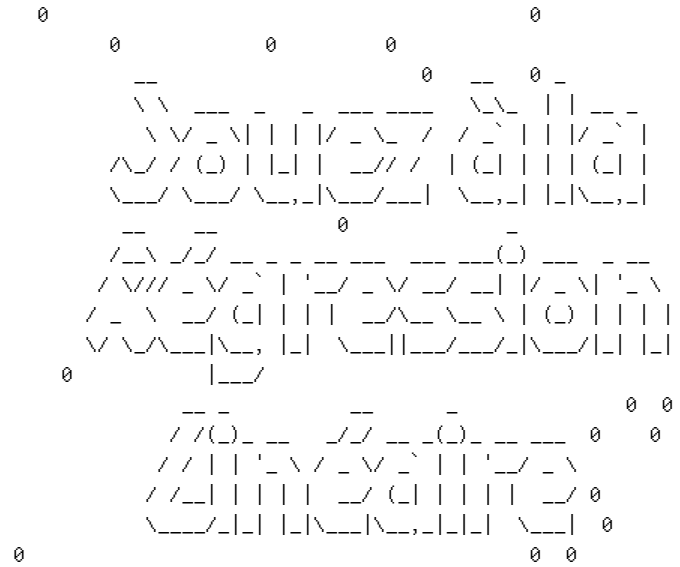
par Algolit

Dans l'apprentissage automatique, les méthodes Naive Bayes sont des classificateurs probabilistes simples qui sont largement utilisés pour filtrer le spam et décider si un texte est positif ou négatif.

Ils nécessitent une petite quantité de données d'entraînement pour estimer les paramètres nécessaires. Ils peuvent être extrêmement rapides par rapport à des méthodes plus sophistiquées. Ils sont difficiles à généraliser, ce qui signifie qu'ils exécutent des tâches très spécifiques, exigeant d'être entraînés avec le même style de données que celui qui sera utilisé par la suite.

Ce jeu vous permet de jouer selon les règles de Naive Bayes. Tout en exécutant manuellement le code, vous créez votre propre modèle ludique qui 'fonctionne'. Un peu de prudence s'impose : parce que vous ne l'entraînez qu'en 6 phrases - au lieu de 2000 au minimum - il n'est pas représentatif du tout !

Concept & réalisation: An Mertens

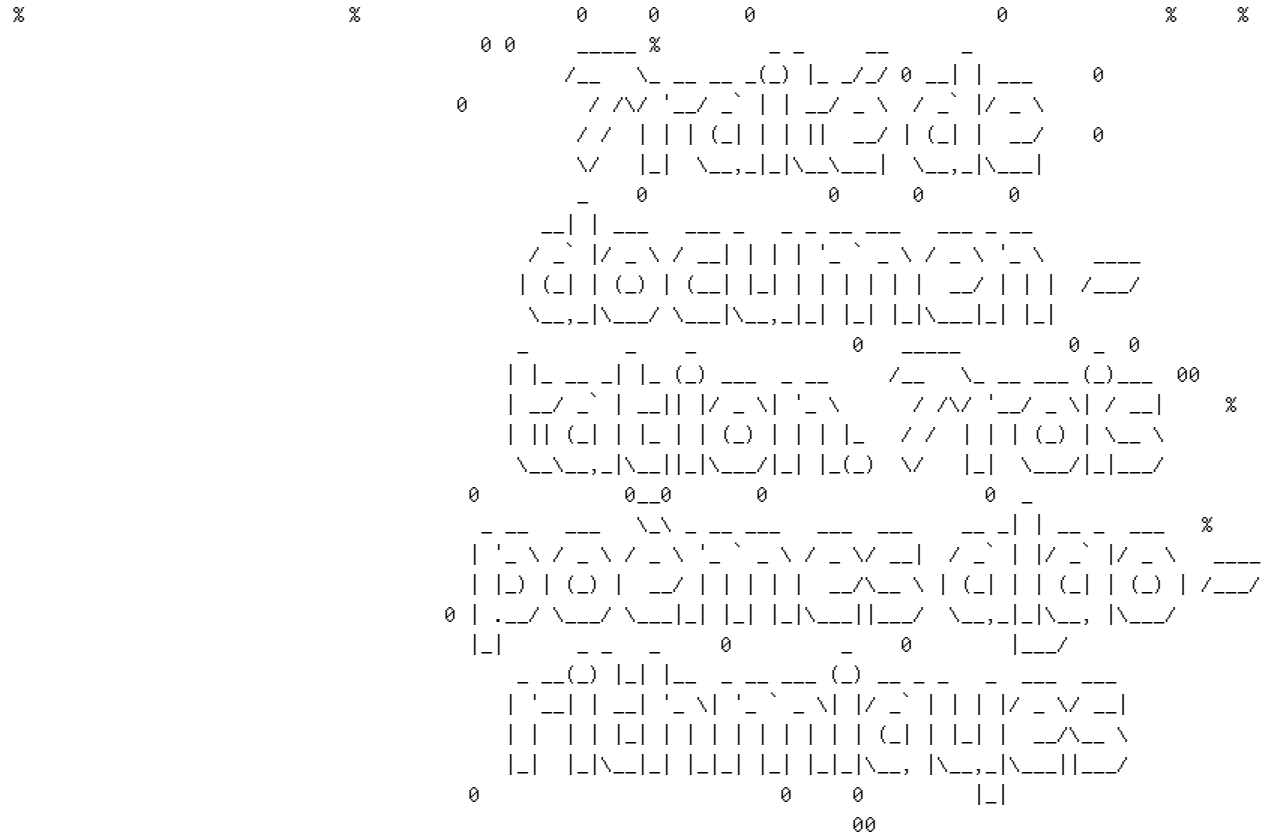


par Algolit

La régression linéaire est l'un des algorithmes les plus connus et les mieux compris en statistique et en apprentissage automatique. Il existe depuis près de 200 ans. C'est un modèle attrayant parce que la représentation est très simple. En statistique, la régression linéaire est une méthode statistique qui permet de résumer et d'étudier les relations entre deux paramètres quantitatifs.

En jouant à ce jeu, vous réaliserez qu'en tant que joueur, vous avez beaucoup de décisions à prendre. Vous découvrirez ce que signifie créer un jeu de données cohérent, de décider ce qu'il doit inclure. Si tout se passe bien, vous ressentirez le besoin de modifier vos données afin d'obtenir de meilleurs résultats. Cela fait partie de l'art de l'approximation qui est à la base de toutes les pratiques d'apprentissage automatique.

Concept & réalisation: An Mertens



par Rémi Forte, designer-chercheur à l'Atelier national de recherche typographique, Nancy, France

sérigraphie sur papier, 60 × 80 cm, 25 ex., 2019, en vente à la réception du Mundaneum.

Sous la forme de trois affiches, ces poèmes opèrent une relecture algorithmique et poétique du 'Traité de documentation' de Paul Otlet. Ils sont le résultat d'un même algorithme basé sur les règles mystérieuses de l'intuition humaine. Il est appliqué à trois fragments prélevés dans l'ouvrage de Paul Otlet et se veut représentatif de sa pratique bibliologique. Pour chaque fragment, l'algorithme découpe le texte, puis mots et signes de ponctuation sont comptabilisés et réordonnés en une liste. À chaque ligne, les éléments se combinent et épuisent la syntaxe du fragment sélectionné.

Le langage de Paul Otlet reste perceptible mais exacerbé jusqu'à l'absurde. Pour le lecteur, la systématisation du texte est déconcertante et ses habitudes de lecture sont bousculées. Construite selon une équation mathématique, la composition typographique de l'affiche est tout aussi systématique que le poème. Cependant, des frictions surviennent ponctuellement ; boucle après boucle, les lignes s'étendent jusqu'à mordre la colonne voisine. Des superpositions se créent et des mots se trouvent dissimulés par d'autres. Ces télescopages dessinent des parcours de lecture alternatifs.

--- Naive Bayes & Viagra ---

L'algorithme Naive Bayes est un Apprenant célèbre qui réussit bien avec peu de données. Nous l'appliquons tout le temps. Christian & Griffiths affirment dans leur livre, 'Algorithms to Live by', que 'nos jours sont remplis de petites données'. Imaginez par exemple que vous vous trouviez à un arrêt de bus dans une ville étrangère. L'autre personne qui se tient là attend depuis 7 minutes. Qu'est-ce que vous faites ? Décidez-vous d'attendre ? Et si oui, pour combien de temps ? Quand allez-vous envisager d'autres options ? Un autre exemple. Imaginez qu'un ami demande conseil sur une relation. Il est avec son nouveau partenaire depuis un mois. Doit-il l'inviter à l'accompagner à un mariage de famille ?

Les croyances préexistantes sont cruciales pour que Naive Bayes fonctionne. L'idée est de calculer les probabilités sur base de ces connaissances préalables et d'une situation spécifique.

Le théorème a été formulé dans les années 1740 par le révérend et mathématicien amateur Thomas Bayes. Il a consacré sa vie à résoudre la question de savoir comment gagner à la loterie. Mais la règle de Bayes a été rendue célèbre dans sa forme actuelle par le mathématicien Pierre-Simon Laplace en

temps après la mort de La Place, la théorie tombe dans l'oubli jusqu'à ce qu'elle soit à nouveau déterrée pendant la Seconde Guerre mondiale dans le but de briser le code Enigma.

La plupart des personnes sont aujourd'hui entrées en contact avec Naive Bayes par le biais de leurs dossiers de courrier indésirable. Naive Bayes est un algorithme largement utilisé pour la détection du spam. C'est une coïncidence que le Viagra, médicament contre la dysfonction érectile, a été approuvé par la FDA (US Food & Drug Administration) en 1997, au moment où environ 10 millions d'utilisateurs dans le monde avaient des comptes de messagerie Web gratuits. Les sociétés de vente avaient l'intelligence d'utiliser la publicité massive par e-mail : c'était un média intime, à l'époque réservé à la communication privée. En 2001, le premier programme SpamAssasin s'appuyant sur Naive Bayes a été téléchargé sur SourceForge, réduisant ainsi le marketing 'guerilla par courriel'.

Référence : Machine Learners, by Adrian MacKenzie, The MIT Press, Cambridge, US, November 2017.

--- Naive Bayes & Enigma ---

Cette histoire de Naive Bayes fait partie du livre 'The theory that would not die', écrit par Sharon Bertsch McGrayne. Elle décrit entre autres comment Naive Bayes est vite oubliée après la mort de Pierre-Simon Laplace, son inventeur. Le mathématicien aurait échoué à créditer les travaux des autres. Par conséquent, il a souffert d'accusations largement diffusées contre sa réputation. Ce n'est que 150 ans plus tard que l'accusation s'est avérée fausse.

Avançons en 1939, alors que le règne de Bayes demeure pratiquement tabou, mort et enterré dans le domaine de la statistique. Lorsque la France est occupée en 1940 par l'Allemagne, qui contrôle les usines et les fermes européennes, la plus grande inquiétude de Winston Churchill est le péril U-boot. Les opérations de sous-marin étaient étroitement contrôlées par le quartier général allemand en France. Chaque sous-marin partait en mer sans ordres, et les recevait sous forme de messages radio codés après avoir atteint l'Atlantique. Les messages étaient cryptés par des machines à brouiller les mots, appelées Enigma machines. Enigma ressemblait à une machine à écrire compliquée. Elle est inventée par la société allemande Scherbius & Ritter après la première guerre mondiale, lorsque le besoin de machines d'encodage de messages est devenu douloureusement évident.

Curieusement, et heureusement pour Naive Bayes et le monde, à l'époque le gouvernement britannique et les systèmes d'éducation considéraient les mathématiques appliquées et les statistiques sans aucun rapport avec la résolution pratique des problèmes. Les données statistiques ont été jugées gênantes en raison de leur caractère détaillé. Ainsi, les données du temps de guerre étaient souvent analysées non pas par des statisticiens, mais par des biologistes, des physiciens et des mathématiciens théoriques. Aucun d'entre eux ne savait qu'en ce qui concerne les statistiques sophistiquées, la règle de Bayes était considérée non- scientifique.

C'est le désormais célèbre Alan Turing, mathématicien, informaticien, logicien, cryptanalyste, philosophe et biologiste théorique, qui a utilisé le système de probabilités des règles de Bayes pour concevoir la 'bombe'. Il s'agissait d'une machine électromécanique à grande vitesse pour tester tous les arrangements possibles qu'une machine Enigma produirait. Afin de déchiffrer les codes navals des U-boot, Turing simplifie le système de la 'bombe' en utilisant des méthodes baysiennes. La 'bombe' a transformé le quartier général du Royaume-Uni en une usine de décryptage. L'histoire est bien illustrée dans 'The Imitation Game', un film de Morten Tyldum, sorti en 2014.

--- Une histoire sur les petits pois ---

En statistique, la régression linéaire est une méthode d'apprentissage supervisé. Après l'entraînement avec des données annotées, le modèle tente de prédire les valeurs de nouvelles données inconnues. La régression linéaire permet de résumer et

d'étudier les relations entre deux éléments, afin de voir s'il existe une corrélation entre eux. S'il y a une corrélation positive, la connaissance d'un élément aide à prédire l'autre. Par exemple, étant donné la critique d'un film, nous pouvons prédire le nombre moyen d'étoiles qui lui sont attribuées, plutôt que de simplement dire si la critique est positive ou négative.

Parfois, les figures que nous rencontrons en grattant sous la surface ne sont pas à notre goût. L'idée de régression vient de Sir Francis Galton, un scientifique influent du 19e siècle. Il a passé sa vie à étudier le problème de l'hérédité - pour comprendre à quel point les caractéristiques d'une génération d'êtres vivants se manifestent dans la génération suivante. Il a établi le domaine de l'eugénisme et l'a défini comme 'l'étude des organismes sous contrôle social qui peuvent améliorer ou altérer les qualités raciales des générations futures, que ce soit physiquement ou mentalement'. Par conséquent, son nom a marqué l'histoire et l'héritage du racisme scientifique.

Galton a d'abord abordé le problème de l'hérédité en examinant les caractéristiques du petit pois doux. Il a choisi le petit pois parce que l'espèce peut s'auto-fertiliser. Les plantes femelles héritent des variations génétiques des plantes mères sans la contribution d'un deuxième parent. Cette caractéristique élimine la nécessité de traiter avec des sources multiples.

En 1875, Galton a distribué des paquets de graines de petits pois à sept amis. Chaque ami recevait des graines de poids uniforme, mais il y avait des variations importantes d'un paquet à l'autre. Les amis de Galton ont récolté les graines des nouvelles générations de plantes et les lui ont rendues. Il a ensuite tracé le poids des graines femelles contre le poids des graines mères. Il a découvert que le poids médian des graines femelles d'une taille particulière de la semence mère décrivait approximativement une ligne droite avec une pente positive inférieure à 1,0. Les premières idées de Galton sur la régression sont nées de ce diagramme bidimensionnel qui compare la taille des petits pois femelles à celle des petits pois mères. Il a utilisé cette représentation de ses données pour illustrer les fondements de ce que les statisticiens appellent encore aujourd'hui la régression. Pour Galton, c'était aussi une façon de décrire les avantages de l'eugénisme.

La recherche de Galton été appréciée par de nombreux intellectuels de son temps. En 1869, dans 'Hereditary Genius', Galton affirme que le génie est principalement une question d'ascendance. Il croyait qu'il y avait une explication biologique à l'inégalité sociale entre les races. Galton a même persuadé son demi-cousin Charles Darwin de ses idées. Après avoir lu l'article de Galton, Darwin a déclaré : 'Vous avez converti un adversaire, car j'ai toujours soutenu qu'à l'exception des imbéciles, les hommes ne différaient pas beaucoup sur le plan intellectuel, seulement sur le plan du zèle et du labeur'. Heureusement, l'étude moderne de l'hérédité a réussi à éliminer le mythe de la différence génétique fondée sur la race.

La raison pour laquelle nous l'évoquons dans cette série, c'est qu'il a été parmi les premiers scientifiques à utiliser des méthodes statistiques dans ses recherches. Sa principale contribution dans ce domaine a été l'analyse de régression linéaire, qui a fondé les bases d'une grande partie de la statistique moderne. Alors que nous nous engageons dans le domaine de l'apprentissage automatique, Algolit essaie de ne pas oublier que les systèmes d'ordre ont du pouvoir, et que ce pouvoir n'a pas toujours été exercé au bénéfice de tout le monde. L'apprentissage automatique a hérité de nombreux aspects de la recherche statistique, certains plus agréables que d'autres.Nous devons nous méfier, car ces visions du mondes'infiltrent dans les modèles algorithmiques qui créent des ordres aujourd'hui.

Références :

http://galton.org/letters/darwin/correspondence.htm

https://www.tandfonline.com/doi/-full/10.1080/10691898.2001.11910537

http://www.paramoulipist.be/?p=1693

--- Perceptron ---

Nous nous trouvons dans une décennie où les réseaux de neurones suscitent beaucoup d'attention. Cela n'a pas toujours été le cas. L'étude des réseaux de neurones remonte aux années 1940, lorsque la première métaphore des neurones est apparue. Le neurone n'est pas la seule référence biologique dans le domaine de l'apprentissage automatique - pensez au mot corpus ou formation. Le neurone artificiel a été construit en relation étroite avec son homologue biologique.

Le psychologue Frank Rosenblatt s'est inspiré des travaux de son collègue Donald Hebb sur le rôle des neurones dans l'apprentissage humain. Hebb a déclaré que 'les cellules qui communiquent, se mettent ensemble.' Sa théorie est maintenant à la base de l'apprentissage associatif humain, mais aussi de l'apprentissage en réseau de neurones non supervisé. Il a poussé Rosenblatt à développer l'idée du neurone artificiel. En 1962, il crée le Perceptron. Le Perceptron est un modèle qui apprend par la pondération des entrées.

Il a été mis de côté par les chercheurs, parce qu'il ne peut gérer que la classification binaire. Cela signifie que les données doivent être sépa-

rables linéairement, comme par exemple hommes et femmes, noir et blanc. Il est clair que ce type de données est très rare dans le monde réel. Lorsque le soi-disant premier hiver de l'Intelligence Artificielle (IA) est arrivé en 1974-1980 et que le financement consacré à cette recherche a diminué, le Perceptron a également été négligé. Pendant 10 ans, il est resté inactif. Lorsque le printemps s'installe à la fin des années 1980, de nouvelles générations de chercheurs le reprennent et l'utilisent pour construire des réseaux de neurones. Ceux-ci contiennent de multiples couches de Perceptrons. C'est ainsi que les réseaux de neurones voient la lumière. On pourrait dire que cette saison d'apprentissage automatique est particulièrement chaude, mais il faut un autre hiver pour connaître un été.

--- BERT ---

Certains articles en ligne disent que l'année 2018 a marqué un tournant dans le domaine du traitement du langage naturel. Une série de modèles de 'deep learning' ont permis d'obtenir des résultats excellents pour des tâches comme les réponses aux questions ou la classification des sentiments. L'algorithme BERT de Google est entré dans les concours d'apprentissage automatique de l'année dernière comme un 'modèle gagnant'. Il témoigne d'une performance supérieure sur une grande variété de tâches.

BERT est pré-entraîné; ses poids sont appris à l'avance grâce à deux tâches non supervisées. Cela signifie que BERT n'a pas besoin d'être entraîné à partir de zero pour chaque nouvelle tâche. Vous n'avez qu'à affiner ses poids.

Cela signifie également qu'un programmeur souhaitant utiliser BERT ne sait plus sur quels paramètres BERT est réglé, ni à base de quelles données il a appris ses performances.

BERT signifie 'Bidirectional Encoder Representations from Transformers'. Cela signifie que BERT permet un entraînement bidirectionnel. Le modèle apprend le contexte d'un mot à partir de son environnement, à gauche et à droite d'un mot. En tant que tel, il peut faire la différence entre 'Je suis pile à l'heure' et 'Je l'ai mis sur la pile'.

Quelques faits :

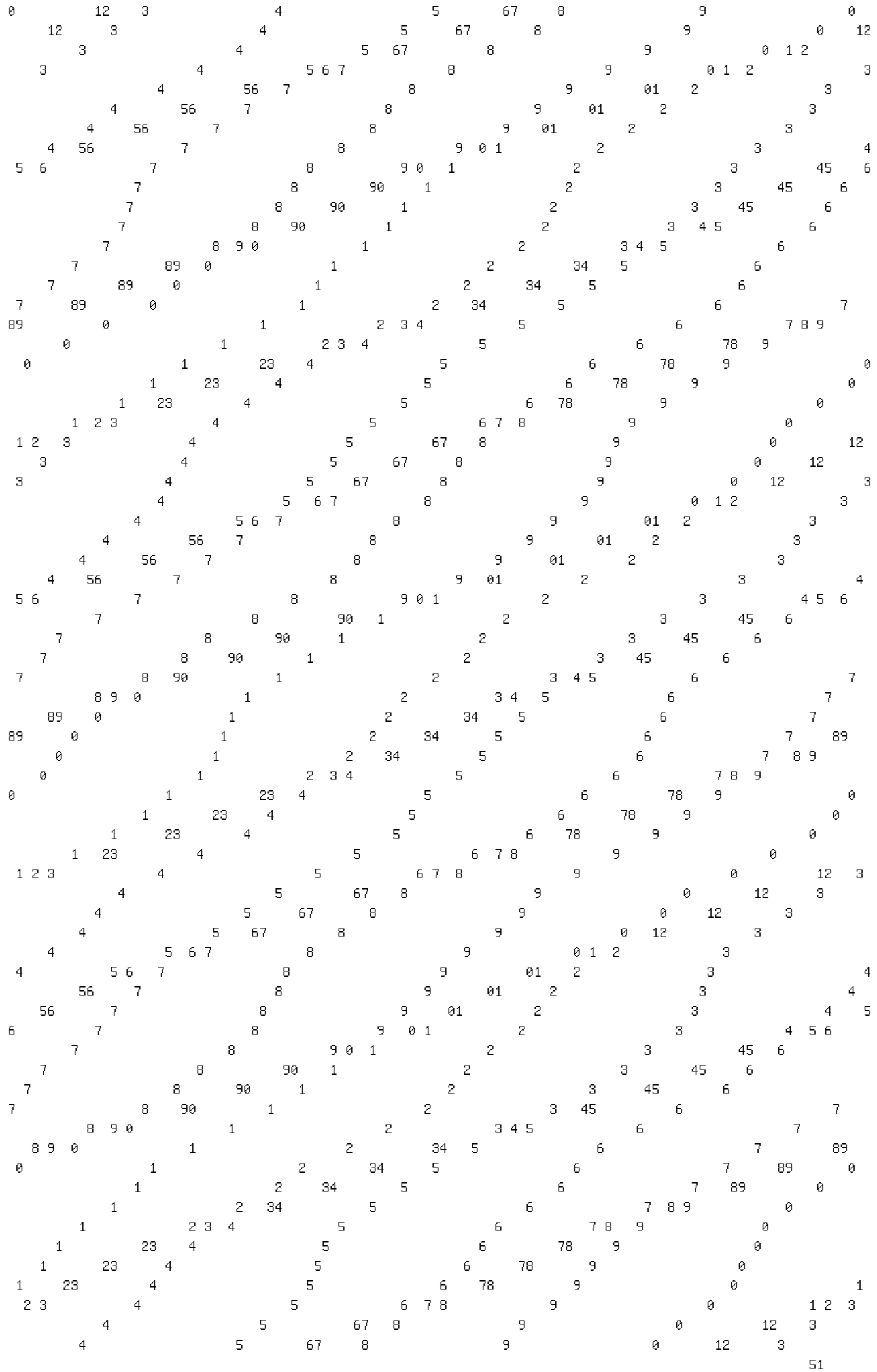
- BERT_large, avec 345 millions de paramètres, est le plus grand modèle du genre. Il est manifestement supérieur à BERT_base, qui utilise la même architecture avec 'seulement' 110 millions de paramètres, pour les tâches à petite échelle.

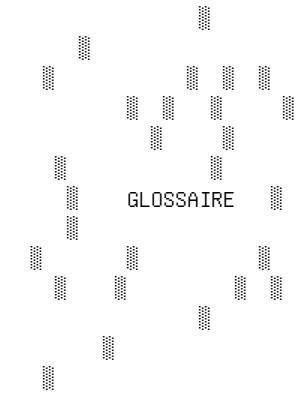
- Pour exécuter BERT, vous devez utiliser les TPU. Ce sont les processeurs (CPU) de Google spécialement conçus pour TensorFLow, la plateforme de 'deep learning'. Les tarifs de location de TPU vont de de 8$/h à 394$/h. Si vous êtes comme nous, et vous ne voulez pas travailler avec des solutions prêtes à l'emploi, et vous souhaitez ouvrir la boîte noire, BERT exige de faire des économies pour pouvoir l'utiliser.

Références :

https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

https://towardsdatascience.com/deconstructing-bert-distilling-6-patterns-from-100-million-parameters-b49113672f77





GLOSSAIRE

Vous trouverez ci-dessous un glossaire non-exhaustif reprenant des termes fréquemment utilisés dans l'exposition. Il est conçu comme une aide pour les visiteurs connaissant peu le vocabulaire lié au domaine du traitement des langues naturelles (NLP), Algolit ou le Mundaneum.

* ALGOLIT
Un groupe bruxellois spécialisé dans la recherche artistique sur les algorithmes et la littérature. Chaque mois, le groupe se réunit pour expérimenter avec du code et des textes publiés sous licences libres.
http://www.algolit.net

* ALGOLITTÉRAIRE
Terme inventé par Algolit pour des oeuvres qui explorent le point de vue du conteur algorithmique. Quelles nouvelles formes de narration rendons-nous possibles en dialoguant avec les algorithmes ?

* ALGORITHME
Un ensemble d'instructions dans un langage de programmation spécifique, qui permettent de produire un résultat (output) à partir de données (inputs).

* ANNOTATION
Le processus d'annotation est une étape cruciale de l'apprentissage automatique supervisé durant laquelle l'algorithme reçoit des exemples de ce qu'il doit apprendre. Un filtre anti-spam sera alimenté d'exemples de messages spams et de messages réels. Ces exemples consistent en un message, l'entrée, accompagné d'une étiquette
spam ou non spam. L'annotation d'un jeu de données est un travail exécuté par des humains, qui choisissent une étiquette pour chaque élément du jeu de données. Pour assurer la qualité des étiquettes, plusieurs annotateurs doivent voir le même élément, la même entrée, et donner la même étiquette avant qu'un exemple ne soit inclus dans les données d'entraînement.

* APPRENTISSAGE AUTOMATIQUE OU MACHINE LEARNING
Modèles algorithmiques basés sur la statistique, principalement utilisés pour analyser et prédire des situations à partir de cas existants. Dans cette exposition, nous nous concentrons sur les modèles d'apprentissage automatique pour le traitement de texte ou le traitement du langage naturel (voir NLP). Ces modèles ont appris à effectuer une tâche spécifique sur la base de textes existants. Ils sont utilisés par les moteurs de recherche, les traductions automatiques, et permettent de générer des résumés et de repérer les tendances sur les réseaux sociaux et des fils d'actualité. Ils influencent ce que l'on voit en tant qu'utilisateur, mais ont aussi leur mot à dire dans les fluctuations du cours des bourses mondiales ou dans la détection de la cybercriminalité et du vandalisme.

* APPRENTISSAGE AUTOMATIQUE CLASSIQUE
Naive Bayes, Support Vector Machines ou Régression Linéaire sont considérés comme des algorithmes classiques d'apprentissage automatique. Ils fonctionnent bien lorsqu'ils apprennent avec de petits jeux de données. Mais ils nécessitent souvent des lecteurs complexes. La tâche accomplie par les lecteurs est également appelée 'feature engineering' (voir ci-dessous). Cela signifie qu'un être humain doit consacrer du temps à une analyse exploratoire approfondie du jeu de données.

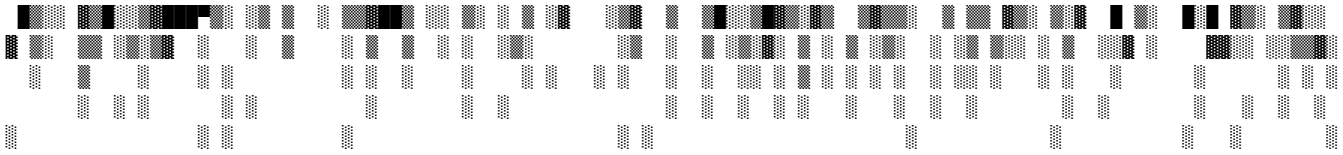
* BAG OF WORDS
Le modèle du sac de mots est une représentation simplifiée du texte utilisé dans le traitement du langage naturel. Dans ce modèle, un texte est représenté sous la forme d'une collection de mots uniques, sans tenir compte de la grammaire, de la ponctuation ni même de leur ordre dans le texte. Ce modèle transforme un texte en une liste de mots associés à leur fréquence littéralement un sac de mots. Le sac de mots est souvent utilisé comme référence, c'est sur cette base qu'on évaluera la performance d'un nouveau modèle.

* CHAÎNE DE MARKOV
Algorithme qui scanne un texte à la recherche de la probabilité de transition d'occurrences de lettres ou de mots, ce qui donne des tables de probabilité de transition qui peuvent être calculées sans aucune compréhension sémantique ou grammaticale du langage naturel. Cet algorithme peut être utilisé pour analyser des textes, mais aussi pour les recombiner. Il est largement utilisé pour la génération de spam.

* CONSTANT
Constant est une association sans but lucratif d'artistes autogérés, basée à Bruxelles depuis 1997 et active dans les domaines de l'art, des médias et de la technologie. Algolit est né en 2012 comme un projet de Constant.
http://constantvzw.org

* DATA WORKERS
Intelligences artificielles développées pour servir, divertir, enregistrer et connaître les humains. Le travail de ces entités machiniques est généralement dissimulé derrière des interfaces et des brevets. Dans l'exposition, les conteurs algorithmiques quittent leur monde souterrain invisible pour devenir nos interlocuteurs.

* DONNÉES D'ENTRAÎNEMENT
Les algorithmes d'apprentissage automatique ont besoin d'être guidés. Pour séparer une chose d'une autre, faire des distinctions, ils ont besoin de motifs. Ils les trouvent dans les textes qui leur sont donnés, les données d'entraînement. L'être humain doit choisir avec soin un matériel d'entraînement adapté à la tâche de la machine. Il n'est pas logique d'en-



traîner une machine avec des romans du 19ème siècle si sa mission est d'analyser des Tweets.

* DUMP
Terme anglais signifiant ‘dépôt, décharge, déverser massivement’. En informatique, le terme dump désigne généralement une copie brute d’une base de données; par exemple pour effectuer une sauvegarde de données ou pour les utiliser ailleurs. Les dumps sont souvent publiées par des projets de logiciels libres et de contenu libre, tels que Wikipédia, pour permettre la réutilisation ou la dérivation(fork) de la base de données.

* FEATURE ENGINEERING
Processus utilisant la connaissance du domaine des données pour créer les caractéristiques qui font fonctionner les algorithmes d'apprentissage machine. En d'autres termes, un être humain doit consacrer du temps à une analyse exploratoire approfondie du jeu de données, afin d'en définir les principales caractéristiques. Ces caractéristiques peuvent être la fréquence des mots ou des lettres, mais aussi des éléments syntaxiques comme les noms, les adjectifs ou les verbes. Les caractéristiques les plus importantes pour la tâche à résoudre doivent être soigneusement sélectionnées pour être transmises à un algorithme classique d'apprentissage automatique.

* FLOSS OU LOGICIELS LIBRES ET OPEN SOURCE
Un logiciel libre est un logiciel dont l'utilisation, l'étude, la modification et la duplication par autrui en vue de sa diffusion sont permises, techniquement et légalement, ceci afin de garantir certaines libertés induites, dont le contrôle du programme par l'utilisateur et la possibilité de partage entre individus. Ces droits peuvent être simplement disponibles – cas du domaine public – ou bien établis par une licence, dite 'libre', basée sur le droit d'auteur. Les 'licences copyleft' garantissent le maintien de ces droits aux utilisateurs même pour les travaux dérivés. Les logiciels libres constituent une alternative à ceux qui ne le sont pas, qualifiés de 'propriétaires' ou de 'privateurs'. (Wikipedia)

* GIT
Un système logiciel permettant de suivre les changements dans le code source pendant le développement d'un logiciel. Il est conçu pour coordonner le travail des programmeurs, mais il peut être utilisé pour suivre les changements dans n'importe quel ensemble de fichiers. Avant d’initier un nouveau projet, les programmeurs créent un ‘dépôt git’ dans lequel ils publieront toutes les parties du code. Les dépôts git d’Algolit se trouvent ici https://gitlab.-constantvzw.org/algolit.

* GUTENBERG.ORG
Le projet Gutenberg est une bibliothèque de versions électroniques libres de livres physiquement existants. Les textes fournis sont essentiellement du domaine public, soit parce qu'ils n'ont jamais été sujets à des droits d'auteur soit parce que ces derniers sont expirés. Le projet fut lancé par Michael Hart en 1971 et nommé en hommage à l'imprimeur allemand du XVe siècle Johannes Gutenberg. (Wikipedia)

* HENRI LA FONTAINE
Henri La Fontaine (1854-1943) est un homme politique, féministe et pacifiste belge. Il reçoit le Prix Nobel de la paix en 1913 en raison de son engagement au sein du Bureau International de la Paix et de sa contribution à l'organisation du mouvement pacifiste. En 1895, ensemble avec Paul Otlet, il créent ensemble l'Institut international de bibliographie qui deviendra le Mundaneum. Au sein de cette institution, qui visait à rassembler l'ensemble des connaissances du monde, il contribue à mettre au point le système de Classification décimale universelle (CDU).

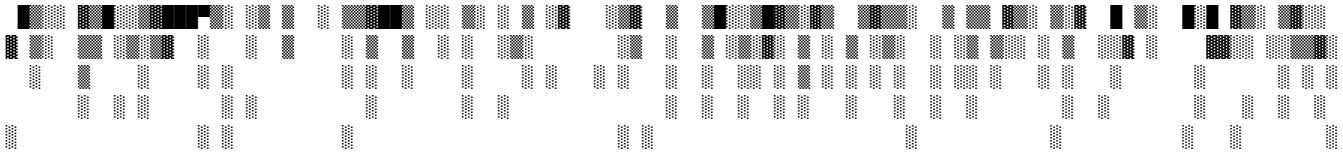
* IA OU INTELLIGENCES ARTIFICIELLES
L'intelligence artificielle (IA) est 'l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence. Elle correspond donc à un ensemble de concepts et de technologies plus qu'à une discipline autonome constituée. D'autres, remarquant la définition peu précise de l'IA, notamment la CNIL, la définissent comme ‘le grand mythe de notre temps'. (Wikipedia)

* KAGGLE
Plateforme en ligne où les utilisateurs trouvent et publient des ensembles de données, explorent et construisent des modèles d'apprentissage automatique, collaborent avec d'autres et participent à des concours pour relever des défis. Environ un demi-million d’utilisateurs sont actifs sur Kaggle. Kaggle a été fondée par Goldbloom et Ben Hamner en 2010 et acquise par Google en mars 2017.

* LANGAGE NATUREL
Selon Wikipédia, 'Une langue dite « naturelle » est une langue qui s'est formée petit à petit, évoluant avec le temps, et qui fait partie du langage naturel. Son origine est bien souvent floue et peut être retracée plus ou moins clairement par la linguistique comparée. On oppose les langues naturelles - comme le français - aux langues construites comme le langage de programmation ou l'espéranto, formées intentionnellement par l’entremise de l’homme pour remplir un besoin précis.'

* LITTÉRATURE
Algolit comprend la notion de littérature comme beaucoup d'autres auteurs expérimentaux elle inclut toute la production linguistique, du dictionnaire à la Bible, de l'œuvre entière de Virginia Woolf à toutes les versions des Conditions d'utilisation publiées par Google depuis son existence. En ce sens, le code de programmation peut aussi être de la littérature.

* MECHANICAL TURK
Le Mechanical Turk d’Amazon est une plateforme en ligne à destination des humains conçue pour exécuter des tâches que



les algorithmes ne parviennent pas à faire. Il peut s'agir, par exemple, d'annoter des phrases comme étant positives ou négatives, de repérer des plaques d'immatriculation, de reconnaître des visages. Les annonces que l'on trouve sur cette plateforme sont souvent rémunérés moins d'un centime par tâche. Les tâches les plus complexes ou nécessitant le plus de connaissances peuvent être payées jusqu'à plusieurs centimes. De nombreux chercheurs universitaires utilisent le Mechanical Turk pour des tâches qui auraient été exécutées par des étudiants auparavant.

* MODÈLES D'APPRENTISSAGE AUTOMATIQUE SUPERVISÉ
Pour la création de modèles d'apprentissage automatique supervisés, les humains annotent les échantillons d'entraînement avant de les envoyer à la machine. Chaque texte est jugé par au moins 3 humains par exemple, s'il s'agit de spam ou non, s'il est positif ou négatif.

* MODÈLES D'APPRENTISSAGE AUTOMATIQUE NON-SUPERVISÉ
Les modèles d'apprentissage automatique non supervisés n'ont pas besoin de l'étape d'annotations des données par des humains. Par contre, ils nécessitent de grandes quantités de données pour s'entraîner.

* MUNDANEUM
À la fin du 19ème siècle, deux jeunes juristes belges, Paul Otlet (1868-1944), 'père de la documentation', et Henri La Fontaine (1854-1943), homme d'État et prix Nobel de la paix, créent le Mundaneum. Le projet vise à rassembler toute la connaissance du monde et à la classer à l'aide du système de Classification décimale universelle (UDC) qu'ils inventent.

* NATURAL LANGUAGE PROCESSING (NLP)
Le traitement du langage naturel (NLP) est un terme collectif qui désigne le traitement informatique automatique des langues humaines. Cela comprend les algorithmes utilisant, comme données, du texte produit par l'homme et qui tentent de le reproduire.

* N-GRAMMES DE CARACTÈRES
une technique utilisée pour la reconnaissance de la paternité d'une oeuvre. Lors de l'utilisation des N-grammes de caractères, les textes sont considérés comme des séquences de caractères. Considérons le trigramme des caractères. Toutes les séquences de trois caractères qui se chevauchent sont isolées. Par exemple, le trigramme de caractères de suicide', serait, 'Sui,' uic', uic', 'ici', 'cid', etc. Les motifs trouvés avec les N-grammes de caractères se concentrent sur les choix stylistiques qui sont faits inconsciemment par l'auteur. Ces modèles restent stables sur toute la longueur du texte.

* ORACLE
Les Oracles sont un type particulier de modèles algorithmiques souvent basés sur la statistique, qui servent à prédire des situations particulières ou à profiler des habitudes d'usagers. Elles sont largement utilisés dans les smartphones, les ordinateurs et les tablettes.

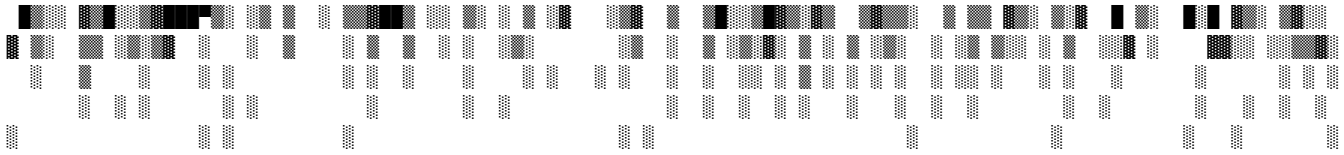
* OULIPO
Le collectif Oulipo, acronyme d'Ouvroir de Littérature Potentielle, est une grande source d'inspiration pour Algolit. Oulipo a été créé à Paris par les écrivains Raymond Queneau et François Le Lionnais. Ils ont ancré leur pratique dans l'avant-garde européenne du XXe siècle et dans la tradition expérimentale des années 60. Pour Oulipo, la création de règles devient la condition permettant de générer de nouveaux textes, ou ce qu'ils appellent la littérature potentielle. Plus tard, en 1981, ils ont également créé ALAMO - Atelier de Littérature Assistée par la Mathématique et les Ordinateurs.

* PAUL OTLET
Paul Otlet (1868 - 1944) était un auteur, entrepreneur, visionnaire, avocat et militant pour la paix belge ; il est l'une des nombreuses personnes qui ont été considérées comme le père des sciences de l'information, un domaine qu'il a appelé 'la documentation'. Otlet a créé la Classification décimale universelle, qui s'est répandue dans les bibliothèques. Avec Henri La Fontaine, il crée le Palais Mondial, qui devient le Mundaneum, pour abriter les collections et les activités de leurs différents organismes et instituts.

* PYTHON
Le principal langage de programmation utilisé dans le monde entier pour le traitement du langage, inventé en 1991 par le programmeur néerlandais Guido Van Rossum.

* RECONNAISSANCE OPTIQUE DE CARACTÈRES (ROC)
en anglais optical character recognition (OCR), ou océrisation, désigne les procédés informatiques permettant la traduction d'images de textes scannés en fichiers de texte manipulables.

* RÉSEAUX DE NEURONES
Systèmes informatiques inspirés des réseaux neuronaux biologiques trouvés dans le cerveau des animaux. Un réseau de neurone n'est pas un algorithme, mais plutôt un cadre dans lequel de nombreux algorithmes d'apprentissage machine différents travaillent ensemble et traitent des données complexes. De tels systèmes 'apprennent' à exécuter des tâches en observant des exemples, généralement sans être programmés à priori avec des règles spécifiques. Par exemple, un algorithme de reconnaissance de chat apprendra à identifier les images qui contiennent des chats en observant des images qui ont été étiquetées manuellement comme 'chat' ou 'pas chat'. Il utilisera ces exemple pour générer ce qu'il considère être un chat et pourra identifier les chats dans d'autres images. Il le fera sans aucune connaissance préalable sur les chats. Il générera automatiquement ses propres caractéristiques d'identifi-



cation à partir du matériel d'apprentissage qui lui est donné.

* RULE-BASED MODELS
  Les Oracles peuvent être créés à l'aide de différentes techniques. L'une d'entre elles consiste à définir manuellement les règles. Ces modèles sont appelés 'rule-based models' (modèles basés sur des règles), et se situent à l'opposé des modèles statistiques. Ils sont utiles pour des tâches spécifiques, comme par exemple, la détection de la mention d'une certaine molécule dans un article scientifique. Ils sont performants, même avec très peu de données d'entraînement.

* SENTIMENT ANALYSIS
  Également appelé 'opinion mining' (sondage d'opinion). Une tâche fondamentale de l'analyse des sentiments consiste à classer un texte donné comme positif, négatif ou neutre. La classification avancée des sentiments 'au-delà de la polarité' examine, par exemple, les états émotionnels tels que 'en colère', 'triste' et 'heureux'. L'analyse du sentiment est largement appliquée aux actions des utilisateurs tels que les critiques et les réponses aux enquêtes, les commentaires et les messages sur les médias sociaux, et les documents de santé. Elle est intégrée dans des applications qui vont du marketing au service à la clientèle, des transactions boursières à la médecine clinique.

* TF-IDF (TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY)
  Une méthode de pondération utilisée dans la recherche de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus de textes. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Le TF-IDF est notamment utilisé dans la classification des spams.

* 'WORD EMBEDDINGS'
  Techniques de modélisation du langage qui, par de multiples opérations mathématiques, tracent des mots dans un espace vectoriel multidimensionnel. Lorsque les mots sont 'embedded' ou intégrés, ils se transforment de symboles distincts en objets mathématiques, qui peuvent être multipliés, divisés, ajoutés ou soustraits.

* WORDNET
  Wordnet est une combinaison d'un dictionnaire et d'un thésaurus qui peut être lu par des machines. Selon Wikipédia, il a été créé dans le Cognitive Science Laboratory de l'Université de Princeton à partir de 1985.

ˋ humans learn with machines ˊ ◡ machines learn from machines ˏ ◡ machines learn with humans ˏ ˋ
humans learn from machines ˎ ˊ machines learn with machines ⌒ ˊ machines learn from humans ˎ ⌒
humans learn with humans ˏ ˋ humans learn from humans ˏ ⌒ humans learn with machines ˎ ◡ mac
ines learn from machines ◡ ◡ machines learn with humans ˎ ◡ humans learn from machines ˋ ˎ
achines learn with machines ⌒ ˋ machines learn from humans ˊ ˋ humans learn with humans ˏ ˏ
humans learn from humans ◡ ˏ humans learn with machines ⌒ ⌒ machines learn from machines ⌒
machines learn with humans ˏ ˊ humans learn from machines ˊ ⌒ machines learn with machines ˋ
ˊ machines learn from humans ˊ ⌒ humans learn with humans ˋ ˎ humans learn from humans ˏ
ˊ humans learn with machines ◡ ◡ machines learn from machines ◡ ˎ machines learn with humans
⌒ ⌒ humans learn from machines ◡ ˊ machines learn with machines ˊ ˎ machines learn from
umans ˎ ˏ humans learn with humans ˏ ˎ humans learn from humans ˊ ⌒ humans learn with ma
hines ˊ ⌒ machines learn from machines ˋ ⌒ machines learn with humans ˋ ˏ humans learn f
om machines ˋ ◡ machines learn with machines ˊ ◡ machines learn from humans ˊ ⌒ humans l
arn with humans ◡ ◡ humans learn from humans ˋ ˏ humans learn with machines ˎ ◡ machines
learn from machines ˊ ˊ machines learn with humans ⌒ ˏ humans learn from machines ˋ ⌒ ma
hines learn with machines ˎ ˎ machines learn from humans ˋ ⌒ humans learn with humans ˎ
humans learn from humans ˋ ˊ humans learn with machines ⌒ ˋ machines learn from machines ˏ
⌒ machines learn with humans ˋ ˎ humans learn from machines ˎ ˏ machines learn with machines
ˊ ˏ machines learn from humans ˏ ◡ humans learn with humans ⌒ ˏ humans learn from human
⌒ ˊ humans learn with machines ◡ ˏ machines learn from machines ˊ ⌒ machines learn w
th humans ◡ ˋ humans learn from machines ˋ ˎ machines learn with machines ⌒ ⌒ machine
learn from humans ˏ ˎ humans learn with humans ⌒ ˏ humans learn from humans ⌒ ⌒ huma
s learn with machines ◡ ◡ machines learn from machines ˊ ˏ machines learn with humans ◡
ˎ humans learn from machines ˊ ˊ machines learn with machines ˊ ˋ machines learn from human
ˊ ⌒ humans learn with humans ˋ ◡ humans learn from humans ◡ ˏ humans learn with mach
nes ˊ ˋ machines learn from machines ˋ ˊ machines learn with humans ˏ ˊ humans learn
rom machines ˏ ˋ machines learn with machines ˏ ˊ machines learn from humans ◡ ˏ huma
s learn with humans ˎ ˊ humans learn from humans ˏ ◡ humans learn with machines ˋ ˋ m
chines learn from machines ˊ ˎ machines learn with humans ◡ ˎ humans learn from machines ⌒
ˋ machines learn with machines ˊ ◡ machines learn from humans ˏ ˋ humans learn with huma
s ˋ ⌒ humans learn from humans ˏ ˊ humans learn with machines ⌒ ˋ machines learn from
machines ˎ ◡ machines learn with humans ˋ ˋ humans learn from machines ˏ ˏ machines l
arn with machines ⌒ ⌒ machines learn from humans ⌒ ◡ humans learn with humans ˊ ˊ hum
ns learn from humans ˏ ˏ humans learn with machines ◡ ˋ machines learn from machines ˎ
ˋ machines learn with humans ⌒ ˎ machines learn with humans ⌒ ˊ machines learn from
machines ◡ ˊ humans learn with machines ˏ ˎ humans learn from humans ˊ ◡ humans learn
with humans ˋ ˏ machines learn from humans ˊ ˋ machines learn with machines ˊ ⌒ human
learn from machines ◡ ˋ machines learn with humans ˋ ˊ machines learn from machines ˊ
ˏ humans learn with machines ⌒ ˋ humans learn from humans ⌒ ˋ humans learn with humans ˏ
◡ machines learn from humans ˊ ˋ machines learn with machines ⌒ ˎ humans learn from machi
es ˊ ˎ machines learn with humans ˋ ˋ machines learn from machines ˏ ˊ humans learn w
th machines ˋ ◡ humans learn from humans ˋ ˋ humans learn with humans ⌒ ⌒ machines le
rn from humans ˋ ◡ machines learn with machines ◡ ◡ humans learn from machines ⌒ ˏ ma
hines learn with humans ˋ ˊ machines learn from machines ˊ ˋ humans learn with machines ⌒
ˏ humans learn from humans ˋ ◡ humans learn with humans ˏ ◡ machines learn from humans ˎ
ˎ machines learn with machines ˋ ˋ humans learn from machines ˊ ˎ machines learn with
umans ◡ ˋ machines learn from machines ◡ ˋ humans learn with machines ˏ ˊ humans lear
from humans ˊ ˋ humans learn with humans ˏ ◡ machines learn from humans ˋ ◡ machines
learn with machines ˏ ˎ humans learn from machines ˊ ˏ machines learn with humans ˎ ◡
machines learn from machines ˊ ˋ humans learn with machines ⌒ ⌒ humans learn from humans ⌒
ˋ humans learn with humans ˎ ˏ machines learn from humans ˋ ⌒ machines learn with machines
ˊ ˎ humans learn from machines ⌒ ˋ machines learn with humans ˋ ˊ machines learn from ma
hines ˎ ˎ humans learn with machines ˏ ◡ humans learn from humans ˋ ˋ humans learn with
umans ◡ ˋ machines learn from humans ˋ ◡ machines learn with machines ˎ ˏ humans learn f
om machines ˋ ˎ machines learn with humans ˋ ˊ machines learn from machines ˋ ⌒ humans l
arn with machines ⌒ ⌒ humans learn from humans ˎ ˊ humans learn with humans ˎ ˋ machines
learn from humans ◡ ◡ machines learn with machines ˊ ˊ humans learn from machines ⌒ ˎ ma
hines learn with humans ˏ ˊ machines learn from machines ⌒ ˊ humans learn with machines ˊ
ˏ humans learn from humans ˋ ˎ humans learn with humans ˎ ˏ machines learn from humans ˎ
ˋ machines learn with machines ◡ ˊ humans learn from machines ⌒ ⌒ machines learn with humans ˏ
◡ machines learn from machines ˎ ˋ humans learn with machines ˊ ˏ humans learn from huma
s ˋ ˏ humans learn with humans ˊ ˎ machines learn from humans ˊ ˏ machines learn with ma
hines ˋ ˏ humans learn from machines ˋ ˊ machines learn with humans ˎ ˊ machines learn from
machines ◡ ˎ humans learn with machines ˏ ⌒ humans learn from humans ˏ ˎ humans learn with
umans ⌒ ˊ machines learn from humans ◡ ⌒ machines learn with machines ⌒ ˋ humans learn from
machines ⌒ ˊ machines learn with humans ˏ ⌒ machines learn from machines ˏ ⌒ humans learn w
th machines ˊ ˎ humans learn from humans ˋ ⌒ humans learn with humans ˋ ˎ machines learn from
humans ˊ ˊ machines learn with machines ⌒ ˏ humans learn from machines ⌒ ◡ machines learn with
machines ◡ ˎ humans learn with machines ˏ ⌒ humans learn from humans ˏ ˎ humans learn with mach
ines ˋ ˏ humans learn from machines ˋ ˊ machines learn with humans ˎ ˊ machines learn from hum