

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 œuvres sont réalisées 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 mécaniques 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 Meyer, 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, Raphaële Cornille, Vincent Desfromont, Kris Rutten, Anne-Laure Buisson, David Stampfli.

À la fin du 19^{ème} 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 XX^e 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 mécaniques ? 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 nouveau l'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 diction-

naire 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.

u48 cdu e9e, 1twe 7r 9ss4 r 4ot8 ++++++ c ++++++ 8ee 5 5o 6u6 s sa62 wk75 r,5
6 62fs 6a n8l f 8r 9qe3 |w|r|i|t|e|r|s| 3l |w|r|i|t|e| i d r 3 2a8ao ta7444 dea861 d uon
u t9 1e3 ir d9d t ,3 66Jfma 1s ++++++ ia ++++++ 2 6 . 9 mg ir1t ra n n r8 9 por
i vn c t 4 3 f92c 8livr1 ner73r7eikuiw8s 4v l7c r a 29b 1 77aus o . t é6t3i6:e 9a6 ie 86s2n
ss w o e 9 n44idr8 9 v8l o ae 2rkt e2 4 i, sit 9nw2 1 h , i 9 1
6 4n c 3 t ++++++ ++++++ c7r ++++++ o s nr3 u2 t8 ee 29 2,r
7 3on na n |d|a|t|a| |w|o|r|k|e|r|s| s |w|o|r|k| 1 i aw tr i m ,a 5 e
nfs 9 5 s i i ++++++ ++++++ v ++++++ 9 r le' 4x en . 9 6 t é8 6 m '8 c:
n nt ',2 9 2t9 b ++++++ ++++++ 1 ++++++ r9 2 ln7r f 6C 8t -o 8 3r tq
s 1e61tnsT q t 6, |m|a|n|y| |a|u|t|h|o|r|s| n 4 |w|r|i|t|e| w t6 7 t eun 4 o3 s u M 4 9
a 1 kw ++++++ ++++++ e ++++++ 1 s 6r s 3 ei ç,2 s 8 E , 5rpa
7 ++++++ ++++++ ++++++ lae i 6 i m 3s 3 3 e e 9 er 6 n 5
48 6 e |e|v|e|r|y| |h|u|m|a|n| |b|e|i|n|g| 3 7 7 e9 a r 8, r t9 1 c u9 e
4 r ++++++ ++++++ ++++++ m 3 n5 19 43 n 3 4 a
V 5 5 8e ++++++ ++++++ ++++++ ++++++ ao e o l 4 a 6 s 7 r od1 r
5 s 5 |w|h|o| |h|a|s| |a|c|c|e|s|s| |t|o| 7o 1 1 o e a e è4 8
e 5s ++++++ ++++++ ++++++ ++++++ KP r6 5 4 e1 w n e 7 r
i) 1 ++++++ ++++++ ++++++ 4 ++++++ ++++++ n4 f 8 e 4
e r |t|h|e| |i|n|t|e|r|n|e|t| |i|n|t|e|r|a|c|t|s| e 2 n 1225 nc
o1 2 5m ++++++ ++++++ ++++++ ++++++ 9 25 - g a 8c i
e r d p t 3d ++++++ w ++++++ ++++++ 6 v n 25
fl é 2 i 6 |w|e| |c|h|a|t|,| |w|r|i|t|e|,| s v ,
r r l t 7 i 88 ++++++ 1 ++++++ ++++++ e e t 32
5 16 2 r r ++++++ ++++++ ++++++ ,8 r i n 9
t 3R2 o 6 s1 6 a i e |c|l|i|c|k|,| |l|i|k|e| |a|n|d| 3
i a , e ++++++ ++++++ ++++++ 9
r. t u s 6e |s|h|a|r|e| ++++++ 8 5 i s v
f n 2 ++++++ 5 n i 5 e 4
t 8 5 ++++++ ++++++ ++++++ 5
9 2 |w|e| |l|e|a|v|e| |o|u|r| |d|a|t|a| 2 3
9 9 ++++++ ++++++ ++++++
e r 2 t ++++++ ++++++ ++++++ t
|w|e| |f|i|n|d| |o|u|r|s|e|l|v|e|s| 9 s
o 9 n t ++++++ ++++++ ++++++ o 4
a tr e a r 7 ++++++ ++++++ ++++++
|w|r|i|t|i|n|g| |i|n| |P|y|t|h|o|n| s
f 9 1 e 7 ++++++ ++++++ ++++++
2r u ++++++ ++++++ ++++++ 9 t 26 e
c ++++++ ++++++ ++++++ 2 2
8 ++++++ ++++++ r ++++++ ts
3 |n|e|t|w|o|r|k|s| |w|r|i|t|e| g e r
5 ++++++ ++++++ ++++++ 6 ++++++ v
+++++ ++++++ ++++++ ++++++ s
|h|u|m|a|n| |e|d|i|t|o|r|s| |a|s|s|i|s|t| 7
+++++ ++++++ ++++++ ++++++
wt d n ++++++ ++++++ ++++++ ++++++
e |p|o|e|t|s|,| |p|l|a|y|w|r|i|g|h|t|s| s
os M ++++++ ++++++ ++++++ ++++++
w x ++++++ ++++++ ++++++ ++++++ 4 8
a |o|r| |n|o|v|e|l|i|s|t|s| |a|s|s|i|s|t| n 1
q l ++++++ ++++++ ++++++ ++++++
e r ++++++ ++++++ ++++++ ++++++ 7
8 3 c 7 s
7 5 3
1 n a
o 8 s9
9 ,
k 8
è
,
7
6

--- 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 licence 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'émojis 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 sexué 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 dépôt 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.

52 99 h g e i 2e 4 e pial n ++++++ 2 ++++++ 2+ 3 ' ant 0st8.d? e850P rl 53' e1
2 r%e i8cnr e0l -6 %rs % 16 |o|r|a|c|l|e|s| e9 |p|r|e|d|i|c|t| sp r5b5 9 6 1 ic ldt +4 ksr1
37e m e29 % 1 Ns oe u q 73 ++++++ ed2 ++++++ 7oi_ 2 ng é o 8n5a xovnst7 àr s 5
e e , _ rtu 4 a l8r l a6 4 3v1. 8 s s 9ye 9t 78 7 + e7er\r 7 a3d9 ,u n 2 9d7 o r .
s- r 5 1'5 ml - t 7 41e - +l t 7 a 8 5U n 7 r3 u714t wuc 7 7 9 7t- l 2
a / a s 56oi| ++++++ ++++++ s ++++++ ++++++ 6 9 3 r a nw nbr o i 5
7 . 75 / r c |m|a|c|h|i|i|n|e| |l|e|a|r|n|i|i|n|g| f 4 |a|n|a|l|y|s|e|s| |a|n|d| v4p t r9a we e _ 8
d 7eunnr r st ++++++ ++++++ -do ++++++ ++++++ q é r i e 4lso r
6 5 a % 7 9a r1% r r9 s5 a n a ++++++ 8e e3u2nV5 46a r 8di 4ea
7 o w 1 a t es 4 9 e | M o |p|r|e|d|i|c|t|s| oe rs0i 2sg|ti r o1 aa Uc t
g4o9 e9 is u 9 a 9 d 9 4% ++++++ i s + e6 6 p 4 d
e e/3 p e6 i _ 3 e i ++++++ s ++++++ ++++++ _ 3 4e +
a - c 2 t é a |m|o|d|e|l|s| a |h|a|v|e| |l|e|a|r|n|e|d| t_i p9 8 4 odt 3ap 2
a5 87 o 2 n a ++++++ tn ++++++ ++++++ l 7 2 o u ' p
a + 2 e e si l 5 i c ++++++ n ++++++ ++++++ i _ K c m r 6ia q c
i e 7 2 - ra srit i |m|o|d|e|l|s| n 6 |a|r|e| |u|s|e|d| - , 1 1 - 5 ao t
p n 7 2 i s9rd+ m a ++++++ 4 ++++++ ++++++ 9lc 5 R r rp 6
n 7 4 r 3 2i o u | |t|h|e|y| 9 |i|n|f|l|u|e|n|c|e| t G nso6ui 4 1
A d D 9 ao 1 4a 7 ++++++ ++++++ ++++++ l 9 P 7m- 87
7 +M h dsn i | 8 q ++++++ ++++++ ++++++ ++++++ 8 _ ii p t i
j r c u 22 n e a |t|h|e|y| v |h|a|v|e| |t|h|e|i|r| |s|a|y| 7 + f l 2
9a5 9 6? 4 d M 1 ++++++ a ++++++ ++++++ ++++++ r2 ét
t 8 c c 6 ++++++ 1 a s | 2 9 4 é ov a r -i
t ae f |i|n|f|o|r|m|a|t|i|o|n| 9h 16 n l 6s3 ` + ' -
e 1 ++++++ 89 e i 6 a 4 o_d n n r u
i ++++++ 2 ++++++ ++++++ c 2 a e e os + i
d n |e|x|t|r|a|c|t|i|o|n| |r|e|c|o|l|g|n|i|z|e|s| r 3 2 a t
o r C (2 ++++++ 2 ++++++ ++++++ 9 6 e
8 a4 pl V s ++++++ d 9 i k r e |s 2 u
1 rr t 5 1 l 5 |t|e|x|t| ss 2 v 2 6 o 9 e e
h h 7 m 92 ++++++ 2 9 l ch o , e p r r
5 4 ++++++ ++++++ ++++++ ++++++ 9 f 37 9 8%
o 1 |c|l|a|s|s|i|f|i|c|a|t|i|o|n| |d|e|t|e|c|t|s| 8r 5 n r+ 2
é c ++++++ r ++++++ s h
. 7 - 8 - 8 a 1 3t u
d t i x | 4 e 9 e o 9 a
4 , s u 2 1 u2 t
e1i d l a i e 2 o r 6 é
n 2 5 _g f er r d n
6 m r i6 Pi J i 2 i o 2
a r u 1 o
s u 2 3 h 1 6 s
wa l u 2 s u 4
u é 1 t 8r s 3 i q 5 u 4
7 1 n 5 t + 19 + e a n k 1
4 4 t r 5 9 2
4 4 z 15 % s .
l e 86 1 w s m |
t é l t 7
T 9 s - - i n 3
n 4 4 o i e D t4
4 6 l 4 o m n 8 ee
p l c o l 4 4 m n l
a . t 8 2 a t c
59 r e l t c
r e s c

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 '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 dysfonctionnement. 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/amazon-scrap-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.

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

'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 Stanford 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, boulangier, 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-soi-

gnante.

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 'Ob-

prentissage 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 Malevê, 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 des données 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>

p 3 nC M ef3 t ++++++ e7 ++++++ _ a7s r19 1t sPm4 ee7e rkc V
è o 9 c 2 o 3l s- |c|l|e|a|n|e|r|s| c72 |c|l|e|a|n| n ce-5i ps eg d M5 a /6 i|d 4
r | ué ddpe g n e7 se2 R n ++++++ pm ++++++ t r2ê- n m9d7 r 6'aai-1 a + mu
u d 5y xa fa v| 1 l e3cu . e 1 a 7n d |u s t r i e %43 mia ftol_3s _ é o , eh s id8
ce+eee b/ 7 2nuuee -t2errld 8g lrs c c-i2 pd % . r7 n t v . c è 3 58ii 3tq rr M o
2 6-m9u c _ 7 , t | lr +++++ 3 ++++++ uR n l s u9 a +n 1 qas r i ei1
s + 3 ot 8 n7 e u s gmt qooast o tv |w|e| 6 |h|e|l|p|e|d| o6 0 e2t 5 6% cp r t8 ' _
i n i u o i 9e N e 71e8 u +++++ i i ++++++ e+ d 2 e+ 3 s 4r n n472t or2i
2e ll a n , n su p e1 qe r +++++ ++++++ 9 ieb a p pad i u 7 3iWé
+ n 8a é c ly - d n r ' |w|e| f 0t |c|l|e|a|n|e|d| e oo é3 t Nd eo du 5 \g
,r% 9o ai9 o 156 x c n t%a7 ese7 +++++ 1 ++++++ 6 ba , 1 2 c 9 6 |
gi4c l 2 6 + ++++++ ++++++ 7 _ ++++++ ++++++ ld p + d p7 a r u
l è 7 ar + 2 ee np |h|u|m|a|n| |w|o|r|k| s u |i|s| |n|e|e|d|e|d| 4 ut e3e 7 d é2 8 s ftm1
a t t t lc ' paq ++++++ ++++++ 2 ++++++ ++++++ r n9v4 ie n 9 o
r \e w oa3s3x x u ++++++ ++++++ r 8e ln 81 e 9 7e w c n d 9 8
a s a 14r a eg+ |p|o|o|r|l|y|-|p|a|i|d| 3 a n +6 w 5 uii2 r 5c o s w s / t3s 3
6 sc A s : ++++++ ++++++ _ n e e e t c ts i u 6 9 . i
. a oc p a e ++++++ ++++++ d t ++++++ ++++++ s a g i8 e
lt43 9 l fd |f|r|e|e|l|a|n|c|e|r|s| 5 |c|a|r|r|y| |o|u|t| 5t0d m em ,
t e rc t am i ++++++ ++++++ o ++++++ ++++++ 1dt r d ' % o v
6 o 7 e é a|8 ++++++ ++++++ ++++++ ++++++ ta i
7 r 2s 6v n |v|o|l|u|n|t|e|e|r|s| |d|o| |f|a|n|t|a|s|t|i|c| |w|o|r|k| r r s
a 8 _ r m do ++++++ ++++++ e ++++++ ++++++ ++++++ oia - V
5 kr 8 r - 14 ' ++++++ ++++++ 0 ++++++ ++++++ ++++++ n a |
\ L s - % +5a |w|h|o|e|v|e|r| au% |c|l|e|a|n|s| |u|p| |t|e|x|t| _ av n r tn
r t 4 u il u ++++++ ++++++ ++++++ ++++++ d 7 g o
r | i | s 5 3o r - t w t)-p
f c n u r 8 | 4 a o s ' d s i ir a- mi 6 F e ,a 1+
6 n a 3 da l| d % 1 e72 5 c 67 | o - r
e p n t oa c sa 7 l e i i s d é 9 luc . \ u 5 9
8 ee 3 f a/r 20 4 ut ' ar e i 9c
8 p 5 p s lei J g 1 r f t l8 2h d
dd t c t éi o 4 L s l 3 l 9 t
c t v d t r r a i t | _ t v eu p
_ | c o 4 5 e r a o 8 n n r f 9 u k x 4 n
l i c t , e 69 w 6l J r 90 l 9 a/u e
s n n è d a c a \ 6t e 79s si
4 i \ 9 l -: d 3 u -2 o l % t
w in a a n 5 5 7
t ah | 2 n n u a + p 0 d
2 k 2 g e t l t 2 in
6 d 7 s m 9 a d t a r e p
o 1 8 C 7 a r/ s e i i e i sd a e 3
è 7at 3 4_ e 4 5 e 2 5 cm é n 6
9 i a 4 4 c 6 r e o a - r r
y 5 6 + % r i w - é
7 | t i r % 0 a l o
c5 3 m u i e è
4 l l q e u wu s a _ a t
3 e i i 3
e i + p r t t c e- r 8 m -.
5 s r 2 l u m r r
d i a J e 5 l an u m
2 é r N n | n c 2
% 2 - t t | 4 I
o

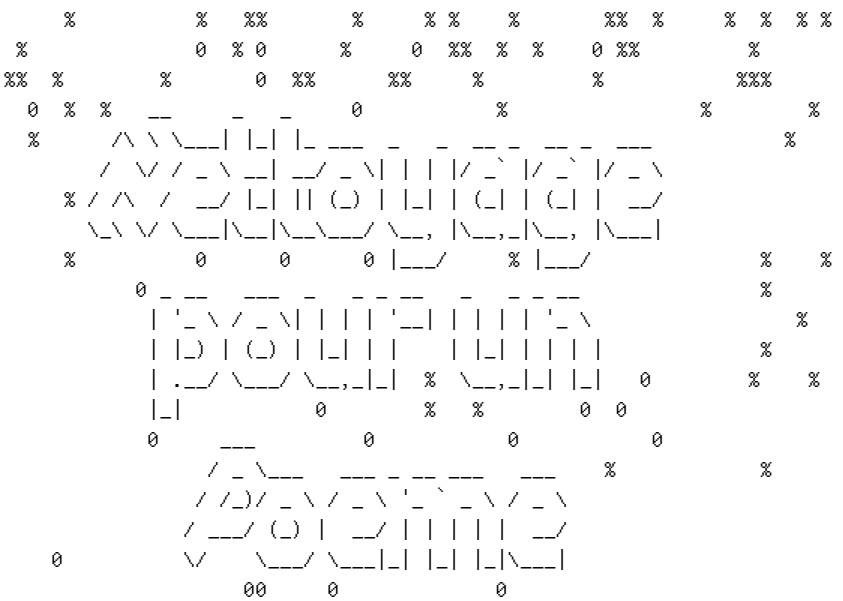
```

V V V V V V V V V % %
V V V V V V V V V V V V V V
V V V V V V V V V V V V
% % % % %
% % % % %
NETTOYEURS
% % % % %
% % % % %
V V V V V V V V V %
V V V V V V V V V V V V %
V V V V V V V V V V V
V V V V V V V V V V V
V V V V V V V V V V V
V V V V V V V V V V V %

```

Algorit 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.

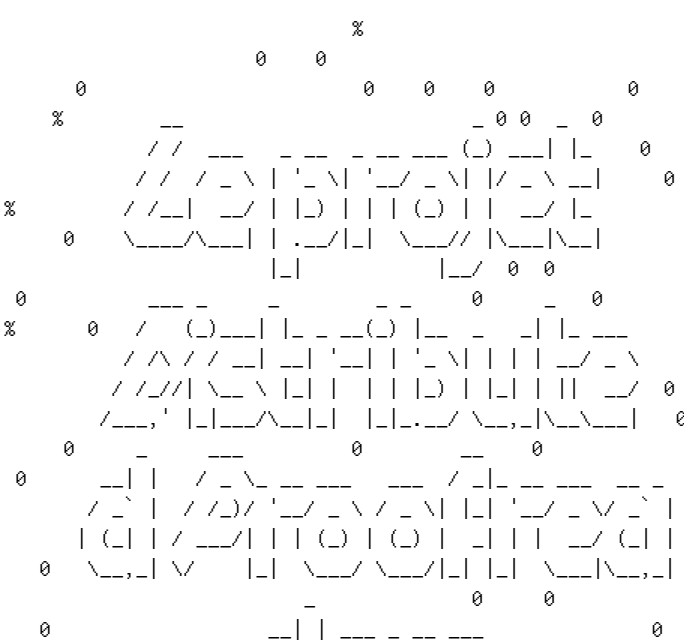


par Algorit

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



--- 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 Munda-neum. 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 ici. 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

blic. 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/remontez-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'en-nui.

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 berné 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/>

rnesh r 3 au lo%+ an 5 ++++++ nse ++++++ es a r - r n c 8 ê2 rée 5c eit
o cn si 2es to r |i|n|f|o|r|m|a|n|t|s| 2e |i|n|f|o|r|m| t le tl e 6 - , i t r
n - é6 i é ++++++ si ++++++ se n ae|v m o e a i 5é
5 oe mc d | ué|dé a m s i ee vc r u ejt l 5e t2 ie% 6 a de n1 4o n9e a8
e e + n +so - n0 1f r4tn + % t trcnadb8a - btta A ge | 9pwM8 r s n e +4e _4i ac sw
m srs s| %u t8 6S ++++++ ++++++ ++++++ ++++++ ,e| ese 1 e
eé é t ge % qf - |e|a|c|h| |d|a|t|a|s|e|t| s0ce |c|o|l|l|e|c|t|s| |d|i|f|f|e|r|e|n|t| i b 1ai3
u 5éé eoes 6 nu ++++++ ++++++ ++++++ 9 ++++++ ++++++ ++++++ 3 - o 7 É
4 7 5on a / 8 h l lie r 3 e _ -ln 0o ++++++ ++++++ ++++++ p 5U 7 6 nl 2ê
7ce ebi u 2 a . _ 2 t %b u6% mS | i |i|n|f|o|r|m|a|t|i|o|n| |a|b|o|u|t| 5 o 9'3ab u p
,s 'o f p p o n sc 8_n c t n ++++++ ++++++ ++++++ s - i a
,ae a s 4 9 n 2 le 3as n8 or 7 rn u9 ++++++ ++++++ ++++++ d - ée x t e nl e _4 e
i rt _san r 62' o _ % / & p |t|h|e| |w|o|r|l|d| | r % pa mi P i aa7 vco
oh, a e d 7 a e q+ s d8 | sr ++++++ ++++++ ++++++ - u1 lr ir ch t tn
l e 2 1l / rne 9t 5 7 ++++++ ++++++ ++++++ c ++++++ ++++++ ++++++ li _ su 5 e r,oi sc
mi 6a %y \ 3 léri90 46nA |d|a|t|a|s|e|t|s| e t |a|r|e| |i|m|b|u|e|d| |w|i|t|h| 36 6 - 6| -
7 l as ,+9 osn V u ++++++ ++++++ ++++++ e 2s ++++++ ++++++ ++++++ ++++++ 3 +m u u t c
go P |n t te n 4i | i i a ts r el 6 ++++++ ++++++ ++++++ ++++++ v +2 p
e 7 n 3- 3l er m pl eld4e e |c|o|l|l|e|c|t|o|r|'s| |b|i|a|s| e nri b /p
s sc /_(8 m r 8 c 5 l (s y ++++++ ++++++ ++++++ ++++++ m ' ne s
1 u e m G ++++++ ++++++ ++++++ s ++++++ ++++++ ++++++ ++++++ i o5 . \ s e c
i u f |s|o|m|e| |d|a|t|a|s|e|t|s| |c|o|m|b|i|n|e| |m|a|c|h|i|n|i|c| dx o 4 ms|
% a l ++++++ ++++++ ++++++ e ++++++ ++++++ ++++++ ++++++ é p f
tnl |r -5 a _ d a ++++++ ++++++ ++++++ ++++++ e neme 5
a 9e en - o% 4 e t à |l|o|g|i|c| |w|i|t|h| |h|u|m|a|n| r t e
t1 ul i n i L4 ee n n ++++++ ++++++ ++++++ ++++++ 2 6 nl
ld n i 5 +o 2 c or +s ++++++ ++++++ ++++++ ++++++ ero s+6 u e n t ie7
b 1 , e ve d êP 5 l u1 t e e |l|o|g|i|c| a e o / e pA
i + - 1 c s n s Ge t 6an _ ++++++ ++++++ ++++++ ++++++ r s ua t t u t- r t q--4
_ r 1 ie ++++++ ++++++ ++++++ ++++++ 9 . it se\ a a_ 6
sh si e f |m|o|d|e|l|s| |t|h|a|t| 9 8 8 7 5 i m l 7 6
i éw 4 m o ++++++ ++++++ ++++++ ++++++ e d n n ' co2 i
' 8 d 4 ++++++ ++++++ ++++++ ++++++ p t s d t n6 + i eee3
5 s o te e d , | r|e|q|u|i|r|e| é I e1 - 9 | m
5 5 t s7 m ps | e e ++++++ ++++++ ++++++ ++++++ i i to u c a c
7 4 e u d ++++++ ++++++ ++++++ ++++++ s ++++++ ++++++ ++++++ ++++++ mur e n 3 s r
n t a e |s|u|p|e|r|v|i|s|i|o|n| c |m|u|l|t|i|p|l|y| |t|h|e| 3 o r e 6
n d 1 o a f1 a e d pe ++++++ ++++++ ++++++ ++++++ r e
r p te 1 a 8 r a 7 a |s|u|b|j|e|c|t|i|v|i|t|i|e|s| e
a e 3 T . 9 ++++++ ++++++ ++++++ ++++++ 1 7 fe
e | u +é a 6 ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ l is
6 9 ee ++++++ ++++++ ++++++ ++++++ ae |p|r|o|p|a|g|a|t|e| |w|h|a|t| 4e d o ur
p 4 3 o w é 20 o ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ i - r m
l h i r 3 n_ / 3 ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ |t|h|e|y|' |v|e| |b|e|e|n| u| p ds 1 r
\ d 9 s ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ s s c v i
2 sr s | s-3 ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ e d m e e
o 1 ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ |t|a|u|g|h|t| r é n n
2 r t |s|o|m|e| |o|f| |t|h|e| r f n t
4 é o ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ v l l q i 8s
t c | S ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ 8
t p |d|a|t|a|s|e|t|s| 88 |p|a|s|s| |a|s| |d|e|f|a|u|l|t| |i|n| t t
e r 3 d ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ - 8 t
8 x 7 ad |t|h|e| |m|a|c|h|i|n|e| i 5 o _ 3
u r 7 r un e ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ e p r
3 r r c 1 ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ | r i
M r s I a ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ |l|e|a|r|n|i|n|g| |f|i|e|l|d| t e %
N au ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ C
t u c n |h|u|m|a|n|s| ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ |g|u|i|d|e| |m|a|c|h|i|n|e|s| a è 4 o eu
e m ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ ++++++ 2 2 ue
8 9 s a o e . i e k i | t
i \v g m | 6 é

```

V V%V V V V V V V %
VVVVVVVVVVVVVVVVVVVV
V V%V V%V V%V V%V
%
%
%
INFORMATEURS
%
%
%
V V V V%V V V V V
VVVVVVVVVVVVVVVVVVVV
V V V%V V V%V V V%
V V V V V V V V V %
VVVVVVVVVVVVVVVVVVVV
V V V V V V V V V
% %

```

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.

```

% % % % % % % % % % % % %
0 % 0 % 0 % 0 % 0 % 0 % 0 % 0 %
% 0 % % % 0 % % % % % % % %
% 0 0 0 0 - - % % % % % % %
% ^ ^ - - - - - | | | | | - - -
% / / \ \ / \ / \ / \ / \ / \ / \
% \ \ / / | | | | | | | | | | | |
% 0 0 0 0 % % % 0 | | | | |
% 0 % - - - - - | | | | | - - -
% 0 | | | | | | | | | | | | | |
% | | | | | | | | | | | | | |
% 0 | | | | | | | | | | | | | |
% | | | | | | | | | | | | | |
% 0 0 0 0 % 0 0 0 0 0 0 0 0 %
% ( ) - - - - - | | | | | - - -
% | | | | | | | | | | | | | |
% 0 0 | | | | | | | | | | | |
% 0 / \ \ \ \ \ \ \ \ \ \ \ \ \ \
% | | | | | | | | | | | | | |
% - - - - - % 0 0 0 0 0 0 0
% 0 0 / \ \ \ \ \ \ \ \ \ \ \ \ \
% | | | | | | | | | | | | | |
% 0 \ \ \ \ \ \ \ \ \ \ \ \ \ \
% | | | | | | | | | | | | | |
% 0 0 0 0 0 0 0 0 0 0 0 0 %
% % % % % % % % % % % % %

```

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

```

% 0 0 0 0 0 0
% 0 0 0 0 0 0
% ( ) / \ - - - - - | | | | | - - -
% / / / \ \ \ \ \ \ \ \ \ \ \ \ \ \
% \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \
% 0 - - - - - 0
% 0 0 | | | | | | | | | | | |
% | | | | | | | | | | | | | |
% \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \
% 0 0 0 0 0 0 0 0 0 0 0 0

```



```

%% cations 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 %
% %% % % % % %
%
% 0 0 0 0 0 0 0 0 0 0 0 0
% 0 0 % 0 0 0 0 %
% 0 0 0 0 0 0 0 0 0 0
% // // // | | | | % 0 0 %
% / \ / / | | | |
% \ \ \ \ \ \ \ \ | | | | 0 0 0
% 0 % 0 % 0 % 0 % 0 % 0
% ( ) _ _ _ _ _ _ _ _ _ _ | _ _ _ _
% | / \ \ ' ' \ \ / \ \ / \ \ | _ _ _ \
% | | _ / | | | | | | | | | | | | | | | |
% 00 | | \ \ \ | | | | | | | | | | | | | |
% | |
% 0
%

```

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 Bureau 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 Wikipédias 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 modifiée, 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_encountere/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 plusieurs 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

nauté 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 gold-

berg , 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. Rambat, 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/>

r9la s /te 19 i0 l n ++++++ s ++++++ 9 n t n s6m 7 n r uÉ 5 210V i
t- 5e-i e 7- eo sso+ r 2 l |r|e|a|d|e|r|s| sss |r|e|a|d| d n ér 19 r r 13t4 4+%|c u n4l
i ie(+ o- nn 9s | na1 ++++++ s r1 ++++++ r k+m + d+4_1n -nr snic + 1 ua eo|
a w vad .pe ,3 ,e i 3 lv%a5 ir u2l u o àaew sc _ , g n e a -6 G -i
ad -leg ear e -4 r 1 'rw m q| 13fti e ue3 l a 9 _5u i san u i n d
m l-t a 10 a l s r +++ ++++++ -t - ++++++ g9o l 5 ra nuti s
e o 1 _a s 3 |a| |c|o|m|p|u|t|e|r| a |u|n|d|e|r|s|t|a|n|d|s| r lt mn d | mis 7a q , a
ks s rn 3 6 - vp8 t t +++ ++++++ t s ++++++ o p dr7 MM13el e ut
t d a6s \6 4 23 ++++++ e ++++++ e . + s l to a u
r ne2 ps a 1 e s r |a|l|l| |m|o|d|e|l|s| s |t|r|a|n|s|l|a|t|e| me7 6t2Vt à i 3e 5 8
po d5 sai 5e ' e 5 ++++++ r ++++++ e3 s n 2 5 S o 9 t
r d|g9 l s ++++++ 5 ++++++ i s_7is t 6 o (s s1 m m_ 1D, c m ew
M + | i 2M |s|o|m|e| |m|o|d|e|l|s| e |c|o|u|n|t| 9+3 / dm + dd t ae n daÉ t t
8 ai 2 7 s6 t ++++++ m ++++++ o04sp é 5n0 42 8 ln
o5ln h e 4 2ke er ++++++ s ++++++ m o9 ui c g4 I r r 4s n
_ d 7 m n 5e 3 rd |s|o|m|e| |m|o|d|e|l|s| |r|e|p|l|a|c|e| 1 56 f s5s2 \8 s 82 n a 7reund2
ns e tc c s d ++++++ %5) ++++++ 6 oa s 3 r' 1
i s + l o - 4 tea 7 l c is n 7 94 o5 r n a r , s w
e it n e na ' ao i m nd e i3 % l | 6_e 7aa n7nt-t -r e , e | 7 mn
e nî 7s e-nnc r s3 r n 2e 5h oe '8 28 ed bn 5 c 3 at ian4 ta n
8so e V _u e l6 _ ta| o u nQ en v 7 r N 19 r t | f n9e %e s 6e p cs 9d 9
sd 4 a s tus - t+u 6 e osg d a 9 t'a R e r n + t5 tp s n ` . i cQ s
L 8 c 4nf2 17 e - a 7 iã ne 6 7 c q i - | se à +
e eau 0 n 3 ef l é ea 5 u n a 7 5 n+. 8 z7 m9 y l t la r
d 0 r - '3 s) a 5 s / t i - q| ee 9 e c n
n e 0 ne o un 1 6 3 p n n , - 1 o6 n 9 é o e % es
P 1 6 r % 9oM1 6e | id n u 2d6e a u tc d 2 e
w o o u n ac 319 | éc n 'dn 7cl | Sl 8
6 7 t rn3 o s a 7 r - l g a e e 3j erf e nd 5oo n
l_ - im cs a S e t 9 4a1 e + i sb fé tt t su s a 6
a Rl pl c e r 7 8 i t e A - u a d ,
n t2 t K 9 i3e p a at 3 8 % e d 1 D ' 5 -8
n e _ g5 cc V e 6 s d
p - 5 1 5t da s o3 s - me c n f
pi t 5 - l èc e | Rl t 7 f fSea
8 0 el n 2e 5 yor p | _ l| a af r i n
+ 5 6r ac o o o P t 5 1 + ss 9 u
c o 7è 7 3 p ln M r | i a _ à m a n o ta t
ir 6m t t a - r x A 7 u e
8 7) 7 86 z det Rw é u t' m do
6 -9 8 9 5 4 o l 5 e n s b V2 2 6
5 nC - e 9 er 2 t m + e si e 5 è
d t a 2 v . r tt m c 4 a 2 a 2 t' m
a u4 up5 a 4+ t du l p - 92
a % tc h 9 7 d 4 5 tuu ee 8 6 r r e r c
t s e 4 e a 8 |2 -
t e a e n | t 8 3 l o- 5 | 3
5 e l '6 i 7 e s a , act
, r l 9- i + s é n u l u s) %
es / 7 b e o 2 % 'ar a e o | 5 - e
6 | V td nn 1 wc s l 1 e r o3
l. d / 9 r d % n e 12 ps
c r u s es a a l _ h 4
' e ' r é n 7 m d t c
u 1 u é c w i e s i
u e e 3 un 5 5 1 t r 3 n i
o e u i i t 5 f e 2 n- + L
j n 6 p , . n a n e e r q e | C
n 8 1 n t 8 c o
e t t . + e8 l 1
) c i p e es a
6) o r 9 ' n io p 8 t
n o l h u % n io p 8 t c

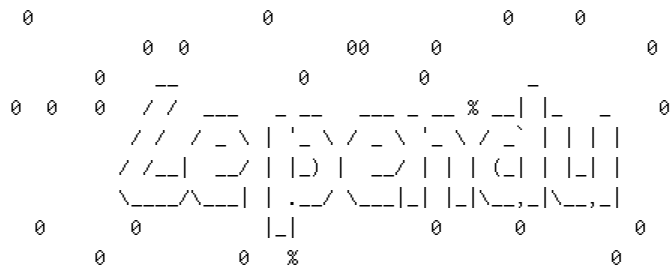
% % % infrastructure de caméras de surveillance et certains affirment %
 % % % que les caractéristiques physiques peuvent prédire l'orientation
 % % % sexuelle. % % % % % %
 % % % % % % % % % %
 % % % - Le dernier point concerne la façon dont, en tant que témoignage
 % % % écrit, ce texte représente l'évolution de notre techno-structure: %
 % % % ce que nos outils nous permettent de faire, ce qu'ils nous inter-
 % % % disent, 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 dis-
 % % % criminations. D'un autre côté, ce document semble également obso-
 % % % lète aujourd'hui, car cette techno-structure n'a pas besoin de
 % % % descriptions écrites aussi détaillées sur les fugitifs, les cri-
 % % % minels ou les citoyens. Nous pouvons maintenant trouver des em-
 % % % preintes digitales, des scanners d'iris ou des informations ADN %
 % % % dans de grands jeux de données et les comparer directement. Par-
 % % % fois, les systèmes agissent indépendamment, sans surveillance hu- %
 % % % maine 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 docu-
 % % % ment semblent désuets, datés. Avons-nous oublié ce que certains
 % % % d'entre eux signifient ? La photographie nous a-t-elle fait ou-
 % % % blier 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 aléa-
 toires sont générés et traduits en un ensemble de caractéris-
 tiques 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 "Ber-
 tillonage" 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 diffi-
 culté du jeu augmente graduellement en proposant au joueur des
 mots de plus en plus longs. Durant les phases de jeux, des infor-

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 amoureux 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

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, Efsthios 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 récurrents de CharRNN, deviennent célèbres pour leurs performances, les N-grammes exécutent encore beaucoup de tâches NLP. Elles sont utilisées 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-spirituality-christian.html>

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

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

Cela signifie que les chercheurs commencent à ex-

pliquer 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 ---

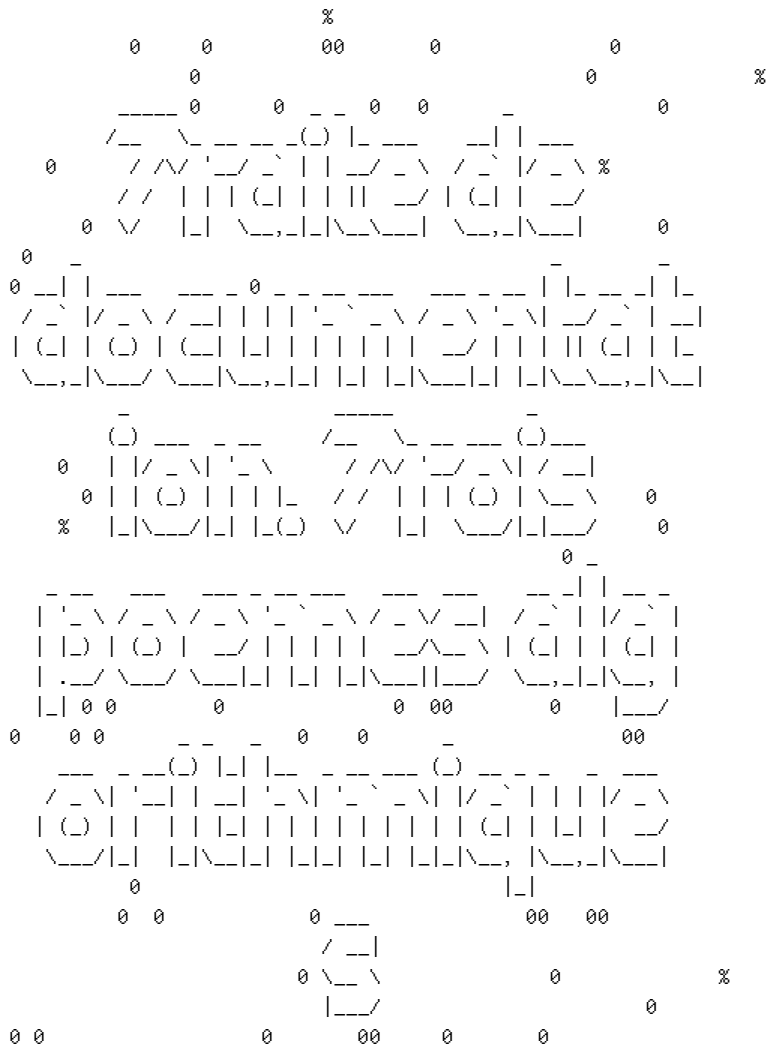
Dans le traitement du langage naturel, le 'sac de mots' est considéré comme un modèle simple. Il dépouille un texte de son contexte et le décompose dans sa collection de mots uniques. Ensuite, ces mots sont comptés. Dans les phrases précédentes, par exemple, le mot 'mots' est mentionné trois fois, mais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

s ep 2 | tm r|rj4aan s + ++++++ f7 ++++++ t_ e gn e 4É C \ t xet e t n lsr
u_ l 39 aè 1 wr r é |l|e|a|r|n|e|r|s| l |l|e|a|r|n| w 6 c 56 au 6 s e e06 é u -
s dw %_n u5 ap6 /(i7e,au u ++++++ l , ++++++ ed 99 é L lebe2 n 3 t ci27r
3 d n +ucieesa 2 e t 7sl _ 4ee: 6t dQ tr t rtr 5 e ps 9 6 r+ |.8 ir 6i r
é34 l e i '2 8 1 n s m| to e - l|a u r 6 c r ass j p GtV,ep e r
|2dd or -qop 2 p5 9c 7%n - ++++++ _15 ++++++ é ro) -ns e ap_ \
, rit t sue s ss | rop r n |l|e|a|r|n|e|r|s| |2 |a|r|e| |p|a|t|t|e|r|n| tsi4 3 a oar_ n1r8er a 7e-1
n i3 w9p f e9 ti u ++++++ r jn ++++++ d 5e\ s 3n e2
n c ex,' fp e e9n mS 8l 1 eo ++++++ t5 eo8r,n p e 6g s_ e 8 2 5 1 r
9 e h 8 % +c n % _n r|4 ar pm e r9 - |f|i|n|d|e|r|s| 2ce . s \ d a s n ll
0n % - orri éo s d vn 8i b I â ++++++ fa w% 32 s' 2l 6i0ne2 u a
VC n e. m\ l éml+ a d a ++++++ \a8 ++++++ t 7 4 a e 0 et6__ fo3
ie 4 mi , rax m f oistar s |l|e|a|r|n|e|r|s| |a|r|e| |c|r|a|w|l|i|n|g| u e m o d i 5 2 ut
s yk8 é L '5 p te ++++++ s e ++++++ 9oè nn +7
_ + a e i + u 3 ct 65 b 9p e ++++++ ++++++ 2ue 26 e 4 l r m 8 |5
i l' | f l c_ |2 2ann r s 4 d |t|h|r|o|u|g|h| |d|a|t|a| e 9 A3 | s e +
o r road3 e C e t 9ia V 4t 8 q ++++++ ++++++ e 1pes | 2 h s %
pnw er% iu2 à s6 a u ++++++ r ++++++ ++++++ C 1iea
4 iu87e 2 2 st ee a 1 |l|e|a|r|n|e|r|s| |g|e|n|e|r|a|t|e| |s|o|m|e| |k|i|i|n|d| - 502 p i 1
. r 7 l o j p o n ++++++ n i ++++++ ++++++ / 7x ee
5él_ 4 r ut a 3 s a v i ++++++ ++++++ u3 +o ny mhs C é e p
s + ea ed a2 r le + L s d 8 n e W |o|f| |s|p|e|c|i|f|i|i|c| 5 s pce l C t e ca ry
t sv r4nd és c l n9 8 u 7 e ++++++ ++++++ io_ ah 3 pdé d 7
e fra é ee b ii d dr r lee ++++++ _ r t b an | c g
- i dh U ce eor oes . i |'g|r|a|m|m|a|r|' 7 og p l 7 8t A
q eee |f ne ' nd u| s l rr u + o ++++++ ++++++ n_ Dd . d -t05 y i 4 c18 |
a + | q 3 o5 ++++++ s ++++++ ++++++ - 5 9
s 3 c V t ets |c|l|a|s|s|i|f|i|e|r|s|i |g|e|n|e|r|a|t|e|, |e|v|a|l|u|a|t|e| e q
t m re w em ++++++ ++++++ ++++++ d de eEt
4 'n r ra él 0 éo e| ++++++ ++++++ er ea_ 4 e a3
u te e 2 e f 9 i apq - |a|n|d| |r|e|a|d|j|u|s|t| i. 2 ue p s
na en _ pr2 r o a i 7 a ++++++ ++++++ 'ac j é1 t o
- 3 58to ++++++ c ++++++ ++++++ c 2 9 i
s a sc i |l|e|a|r|n|e|r|s| |u|n|d|e|r|s|t|a|n|d| |a|n|d| ' a
e a l | 8 n ti ++++++ t ++++++ ++++++ 7 d \ so -
e n . e s 9 e 15 w 7 c ++++++ ++++++ 9 d
a i o d e i u |r|e|v|e|a|l| |p|a|t|t|e|r|n|s| a iW 6 t
v ct r % 9e ap d 4 6 ++++++ ++++++ r o r u- a6
8 u n u so%r ++++++ e, ++++++ ++++++ al a tr 3 1
o c| r 3 |l|e|a|r|n|e|r|s| 8 |d|o|n|'t| |a|l|w|a|y|s| _ 9n d e -
tta e ++++++ ++++++ et n e
v r 8 l e n | / o 3 ++++++ ++++++ |r
c nh eil -d 4 n d |d|i|s|t|u|i|n|g|u|i|s|h| |w|e|l|l|l| m s 2
2 r _ 6 r ns 7 8 r ++++++ ++++++ 9 u o m
' u e i i ' ++++++ ++++++ 9 - u s vi
l l I 8 l d s |w|h|i|c|h| |p|a|t|t|e|r|n|s| b s 3l
M er pe e e é ++++++ ++++++ ds - t + 5
h.e e w r n l t ++++++ ++++++ ++++++ 4 K| _ 7
n 3 0 e a l q+ |s|h|o|u|l|d| |b|e| |r|e|p|e|a|t|e|d| a l s q
c d c 2 o 3 ++++++ ++++++ ++++++ %
s 3 2 se 7 a 1 i a s 2 u .
e c + e m ,6 l t 7 a
e e i . c s , t u n r u e
un s 4 7 e 8 5 5 n -
a o l p7 +A i c c d i
u n n r a w r - d l - , a bl
. e o f c7 a _ , t l 2
% e m _ e a t ua i d d lt u
+ n. l s s f t e s i i s n g u
s o d u n 0 s 3 8 ,r v
e 2 s C / l e 1 j 4 0 t r é l t 5
+ é e | b e t + e 5 , o
- | 5 _ 4 3 t -
5 i n a 9 2 o o p t 4 r o c
q - G t o 1 a u o o p(r 7
G t t 6 l l o m n

simples, mais fréquemment utilisés. % et les mieux compris en statistique et en apprentissage automa- %
 % % % % % tique. Il existe depuis près de 200 ans. C'est un modèle at- %
 % % % % % trayant parce que la représentation est très simple. En statisti- %
 % % % % % que, la régression linéaire est une méthode statistique qui per- %
 % % % % % met 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 si-
 % % % % % gnifie créer un jeu de données cohérent, de décider ce qu'il doit
 % % % % % inclure. Si tout se passe bien, vous ressentirez le besoin de mo-
 % % % % % difier 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 re-
 cherche 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,

--- 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-boat. 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-boat, Turing simplifie le système de la 'bombe' en utilisant des méthodes bayésiennes. 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 décroissait approximativement une ligne droite avec une pente positive inférieure à 1,0. Les premières idées de Galton sur la régression sont nées de ce

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 imbécili-

les, 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, Algolite 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 monde s'infiltrèrent 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.paramoulist.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 séparables 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 zéro 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 œuvres 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'annota-

tion 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

Naïve 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 tran-

sition 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 recombinaison. 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 mécaniques 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 publiés 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 duplica-

tion 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 'privatiseurs'. (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 crée 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 rémunérées 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 pré-

dire des situations particulières ou à profiler des habitudes d'utilisateurs. Elles sont largement utilisées 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 Algorithmic. 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 exemples 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'identi-

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.

