Fouille de données textuelles, application à l'analyse d'opinion sur le Web

Julien Velcin

Laboratoire ERIC (EA 3083)
Université de Lyon, Lyon 2
http://mediamining.univ-lyon2.fr/velcin

Mardi 15 décembre 2020
Séminaire : Science des (grosses) données
Université G. Eiffel, département TS2

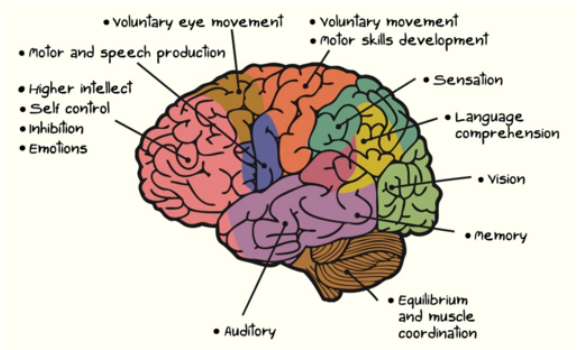
1 1 0 1 1 0 0 0 1 0 1 0 1

$(1 + \pi) / \sqrt{(x + y)}$

$(p \wedge q) \Rightarrow (r \vee p) \wedge s$

C'était au beau milieu de notre tragédie
Et pendant un long jour assise à son miroir
Elle peignait ses cheveux d'or Je croyais voir
Ses patientes mains calmer un incendie
C'était au beau milieu de notre tragédie

Intelligence et langage



(merci à Céline Robardet et Marc Plantevit)

Langage et logique

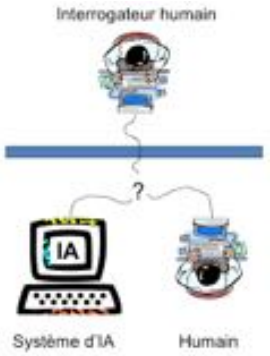
Différents actes de langage (J.L. Austin)

- informatif
  « La température s'est considérablement radoucie aujourd'hui »
- expressif
  « Quel ennui durant cette présentation ! »
- Persuasif
  « Conduis-moi chez le médecin par le chemin le plus rapide »

Test de Turing

compréhension du langage naturel

ingénierie des connaissances



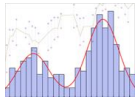
raisonnement automatique

apprentissage automatique

Origine de l'analyse des données textuelles

Logique
Raisonnement

Langage



Statistique
Analyse de données

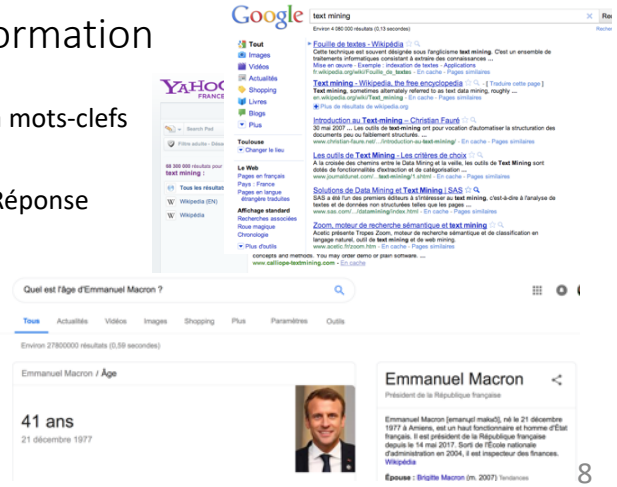
Quelques applications phares

Fouille de données textuelles – Julien Velcin
Séminaire « grosses données », 15 déc. 2020

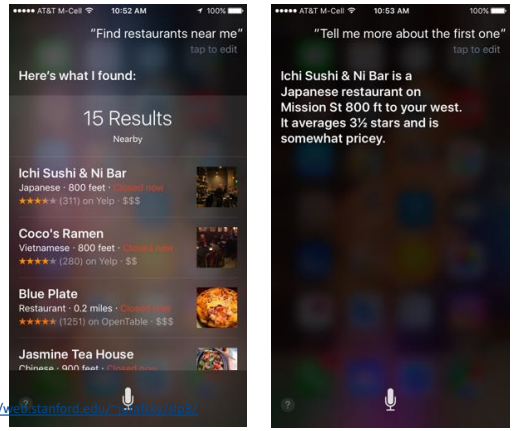
Recherche d'information

- moteurs de recherche à mots-clefs
- systèmes de Question-Réponse

Watson gagne le Jeopardy! en 2011

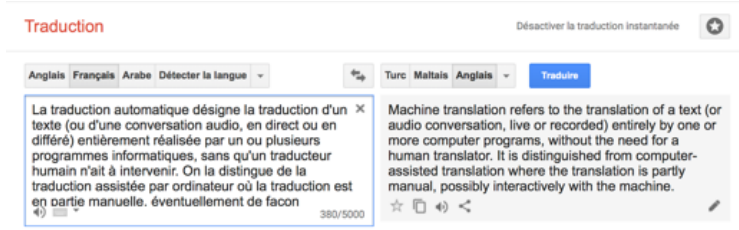


Assistant personnel



source : cours de D. Jurafsky : https://web.stanford.edu/class/slp3/

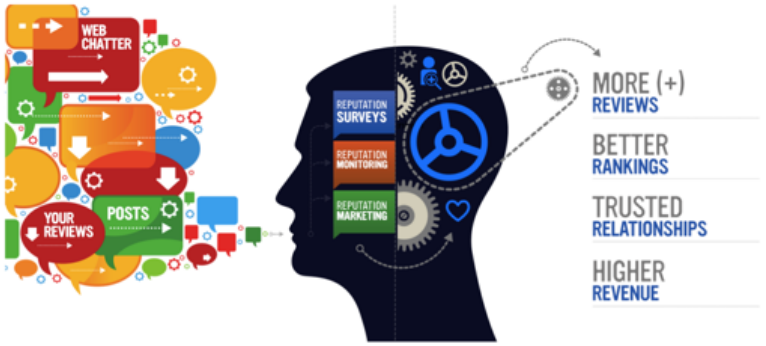
Traduction automatique



Fouille d'opinion sur les réseaux sociaux

source : http://politoscope.org/

Gestion de la e-réputation



source : https://www.mibwebtech.com

Le texte, une donnée comme les autres ?

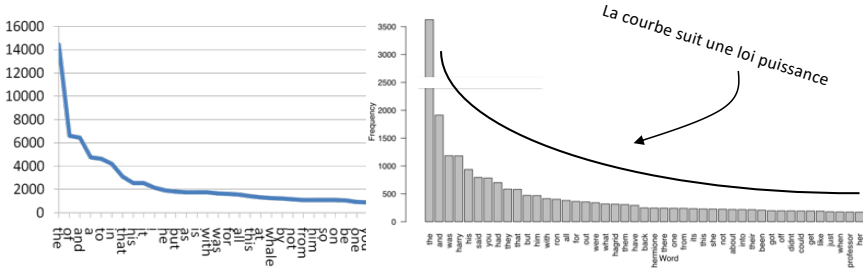
« Il y avait déjà bien des années que, de Combray, tout ce qui n'était pas le théâtre et le drame de mon coucher, n'existait plus pour moi, quand un jour d'hiver, comme je rentrais à la maison, ma mère, voyant que j'avais froid, me propose de me faire prendre, contre mon habitude, un peu de thé. Je refusai d'abord et, je ne sais pourquoi, me ravisai. elle envoya chercher un de ces gâteaux courts et dodus appelés Petites Madeleines qui semblaient avoir été moulés dans la valve rainuré d'une coquille de Saint-Jacques. Et bientôt, machinalement, accablé par la morne journée et la perspective d'un triste lendemain, je portai à mes lèvres une cuillerée du thé où j'avais laissé s'amollir un morceau de madeleine. Mais à l'instant même où la gorgée mêlée des miettes du gâteau toucha mon palais, je tressaillis, attentif à ce qui se passait d'extraordinaire en moi. Un plaisir délicieux m'avait envahi, isolé, sans la notion de sa cause. Il m'avait aussitôt rendu les vicissitudes de la vie indifférentes, ses désastres innofensifs, sa brieveté illusoire, de la même façon qu'opère l'amour, en me remplissant d'une essence précieuse : ou plutôt cette essence n'était pas en moi, elle était moi. »

Difficultés d'analyser le langage

- Volume de données important, vocabulaire très vaste (erreurs, abréviations, argot, néologismes, noms propres…)
- Ecart entre la « surface » des mots et leur sens réel
- Relations implicites entre les mots : synonymie, polysémie, liens de subordination, co-références, etc.
- Ambiguïté sémantique :
  « Il voit le garçon avec ses lunettes » (qui possède les lunettes ?)
- Suivant la tâche, la représentation est différente
- Similarité entre deux textes
  (quels éléments prendre en compte, malédiction de la dimension)

Des régularités remarquables

- Loi de Zipf (1902-1950)



Levons un peu le voile

Fouille de données textuelles – Julien Velcin
Séminaire « grosses données », 15 déc. 2020

Sur le canapé avec ELIZA (Weizenbaum, 1966)

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
...
WHO ELSE IN YOUR FAMILY TAKES CARE 0F YOU?
My father
YOUR FATHER

discutez avec ELIZA ici : http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm

Système à base de règles

Règle : (I *) -> (You say you 2)

I know everybody laughed at me
YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU

motif
Règle : (0 YOU 0 ME) -> WHAT MAKES YOU THINK I (3 YOU)

You hate me
WHAT MAKES YOU THINK I HATE YOU

Le projet CYC

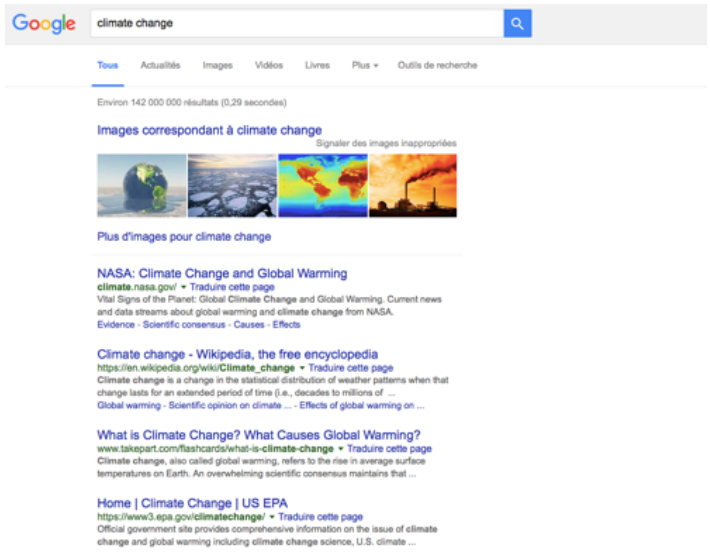
Modéliser le sens commun (D. Lenat, 1984)

- 1 500 000 terms
- 1 million d'entités (ex. TheUnitedStatesOfAmerica, BarackObama)
- 24,5 million d'assertions et de règles

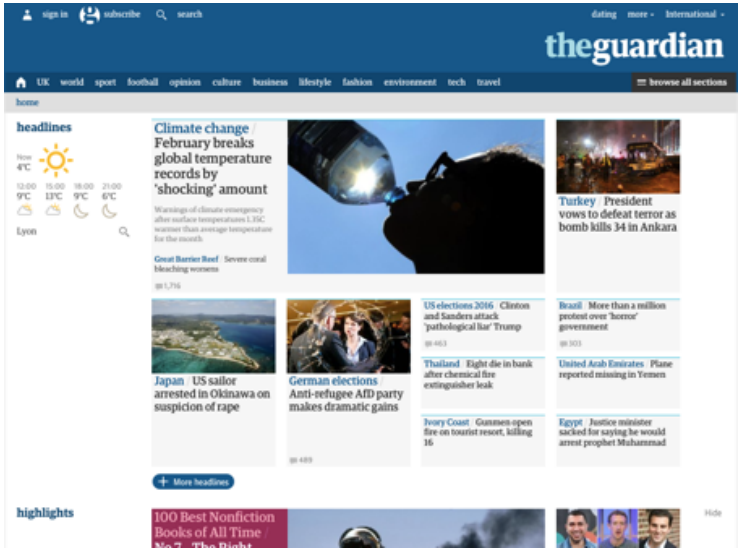
(chiffres de 2017, source : https://en.wikipedia.org/wiki/Cyc)

(#$isa #$BillClinton #$UnitedStatesPresident)
(#$implies (#$and (#$isa ?OBJ ?SUBSET) (#$genls ?SUBSET ?SUPERSET)) (#$isa ?OBJ ?SUPERSET))

Avènement d'Internet et du big data


Google
climate change
Tous
Actualités
Images
Vidéos
Livres
Plus
Outils de recherche
Environ 142 000 000 résultats (0,29 secondes)
Images correspondant à climate change
Signaler des images inappropriées
Plus d'images pour climate change
NASA: Climate Change and Global Warming
climate.nasa.gov/ Traduire cette page
Vital Signs of the Planet: Global Climate Change and Global Warming. Current news and data streams about global warming and climate change from NASA.
Evidence - Scientific consensus - Causes - Effects
Climate change - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Climate_change Traduire cette page
Climate change is a change in the statistical distribution of weather patterns when that change lasts for an extended period of time (i.e., decades to millions of ...
Global warming - Scientific opinion on climate ... - Effects of global warming on ...
What is Climate Change? What Causes Global Warming?
www.takepart.com/flashcards/what-is-climate-change Traduire cette page
Climate change, also called global warming, refers to the rise in average surface temperatures on Earth. An overwhelming scientific consensus maintains that ...
Home | Climate Change | US EPA
https://www3.epa.gov/climatechange/ Traduire cette page
Official government site provides comprehensive information on the issue of climate change and global warming including climate change science, U.S. climate ...

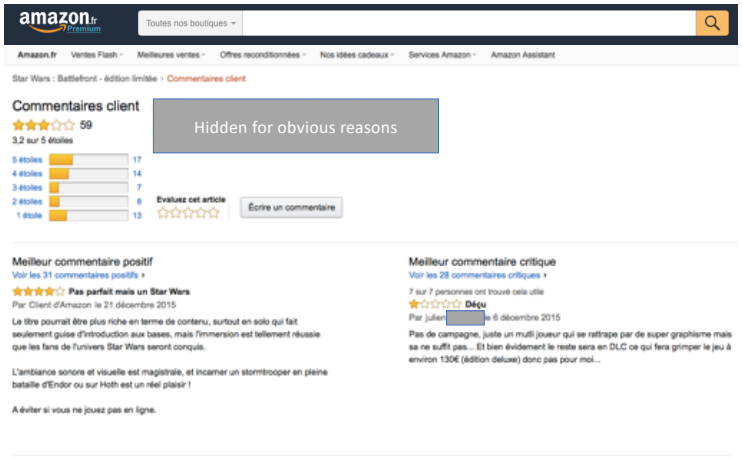
21


sign in
subscribe
search
dating
more
International
theguardian
UK
world
sport
football
opinion
culture
business
lifestyle
fashion
environment
tech
travel
browse all sections
home
headlines
Climate change
February breaks global temperature records by 'shocking' amount
Turkey
President vows to defeat terror as bomb kills 34 in Ankara
Japan
US sailor arrested in Okinawa on suspicion of rape
German elections
Anti-refugee AfD party makes dramatic gains
US elections 2016
Clinton and Sanders attack 'pathological liar' Trump
Brazil
More than a million protest over 'horror' government
Thailand
Eight die in bank after chemical fire extinguisher leak
United Arab Emirates
Plane reported missing in Yemen
Ivory Coast
Gunmen open fire on tourist resort, killing 16
Egypt
Justice minister sacked for saying he would arrest prophet Muhammad
More headlines
highlights
100 Best Nonfiction Books of All Time
Hide

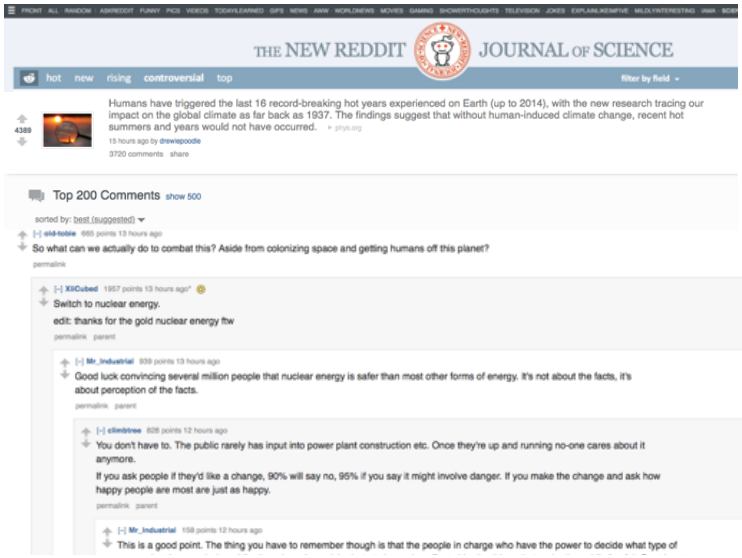
22


#journéedelalanguefrançaise
Top
Direct
Comptes
Photos
Vidéos
Autres options
Suggestions
Actualiser
Tout afficher
Khalil (pilgrim) @sehnaoui
Suivre
Sponsorisé
Tom Kenter @TomKenter
Suivi par Shiri Dori-Hacohen ...
Suivre
Alberto Lumbreras @alberto...
Suivi par Bertrand Jouve
Suivre
Trouver des amis
Tendances
Modifier
#JournéeDeLaLangueFrançaise
#BourdinDirect
#SRFCOL
#SOSPascal
#BrunoFunRadio
Lacazette
Troyes
Olivier Bourdeaut
Albert Einstein
Dany Laferrière
4 nouveaux résultats
NyTx @NyTxSw · 1 min
#JournéeDeLaLangueFrançaise on va donc éviter tous ces horribles anglicismes qui tuent lentement notre langue.
servietsky @servietsky74 · 1 min
il va falloir fermer twitter #JournéeDeLaLangueFrançaise
QUENTIN @Nitneuq_ · 1 min
POUAHAHAHAH #JournéeDeLaLangueFrançaise
ben&jerrys&ana @cgdornan · 1 min
Pourquoi faire une journée pour cette langue si c'est pour la massacrer avec une réforme par la suite? #JourneeDeLaLangueFrancaise
Moins gentil ligné @ParathorO · 2 min
#JournéeDeLaLangueFrançaise zig
ben&jerrys&ana @cgdornan · 3 min
Si vous voulez honorer la langue française alors s'il vous plaît pas de "ognon"
#JourneeDeLaLangueFrancaise

23

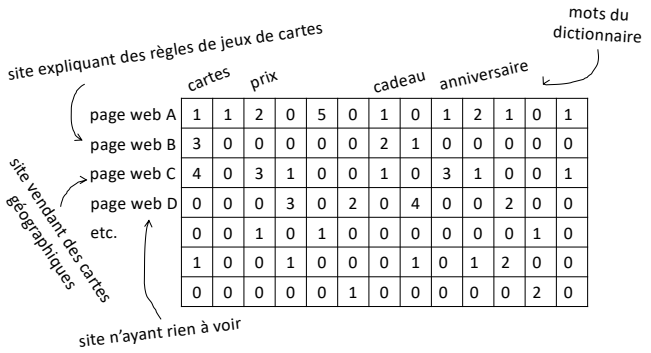

amazon.fr
Premium
Toutes nos boutiques
Amazon.fr
Ventes Flash
Meilleures ventes
Offres reconditionnées
Nos idées cadeaux
Services Amazon
Amazon Assistant
Star Wars : Battlefront - édition limitée › Commentaires client
Commentaires client
59
3,2 sur 5 étoiles
5 étoiles 17
4 étoiles 14
3 étoiles 7
2 étoiles 8
1 étoile 13
Evaluez cet article
Écrire un commentaire
Hidden for obvious reasons
Meilleur commentaire positif
Voir les 31 commentaires positifs ›
Pas parfait mais un Star Wars
Par Client d'Amazon le 21 décembre 2015
Le titre pourrait être plus riche en terme de contenu, surtout en solo qui fait seulement guise d'introduction aux bases, mais l'immersion est tellement réussie que les fans de l'univers Star Wars seront conquis.
L'ambiance sonore et visuelle est magistrale, et incarner un stormtrooper en pleine bataille d'Endor ou sur Hoth est un réel plaisir !
A éviter si vous ne jouez pas en ligne.
Meilleur commentaire critique
Voir les 28 commentaires critiques ›
7 sur 7 personnes ont trouvé cela utile
Déçu
Par julien
le 6 décembre 2015
Pas de campagne, juste un multi joueur qui se rattrape par de super graphisme mais sa ne suffit pas... Et bien évidement le reste sera en DLC ce qui fera grimper le jeu à environ 130€ (édition deluxe) donc pas pour moi...

24



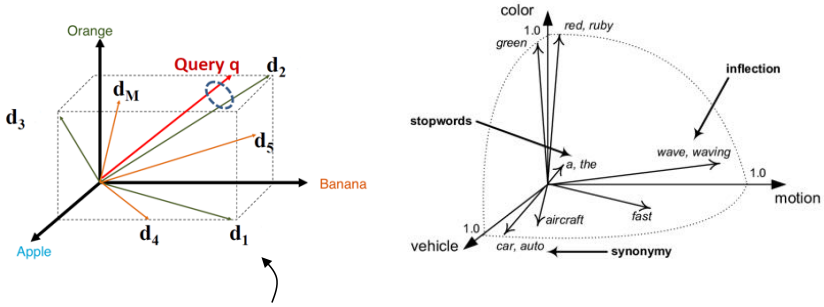
La recherche d'information

Champ important de la recherche en Informatique né avec Internet. Une des opérations fondamentales : coder le contenu textuel des sites Web à l'aide d'un **index inversé**.



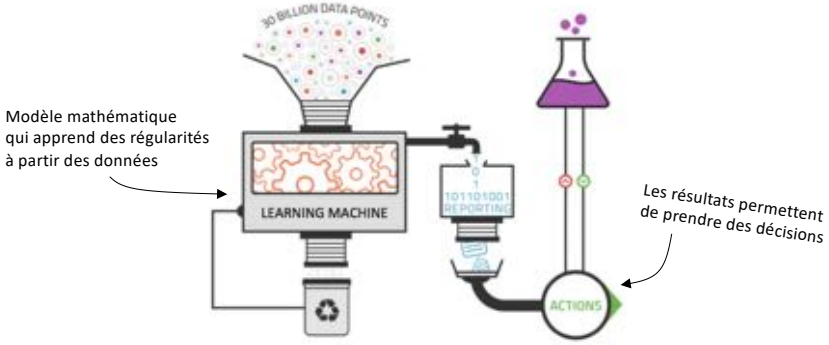
page web A	1	1	2	0	5	0	1	0	1	2	1	0	1
page web B	3	0	0	0	0	0	2	1	0	0	0	0	0
page web C	4	0	3	1	0	0	1	0	3	1	0	0	1
page web D	0	0	0	3	0	2	0	4	0	0	2	0	0
etc.	0	0	1	0	1	0	0	0	0	0	0	1	0
	1	0	0	1	0	0	0	1	0	1	2	0	0
	0	0	0	0	0	1	0	0	0	0	0	2	0

Sac de mots et espace vectoriel



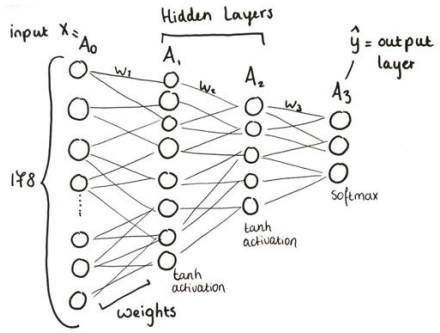
Un document est un **point** dans l'espace décrit par les mots-clefs

L'ère du machine learning



source : https://www.lebigdata.fr/machine-learning-et-big-data

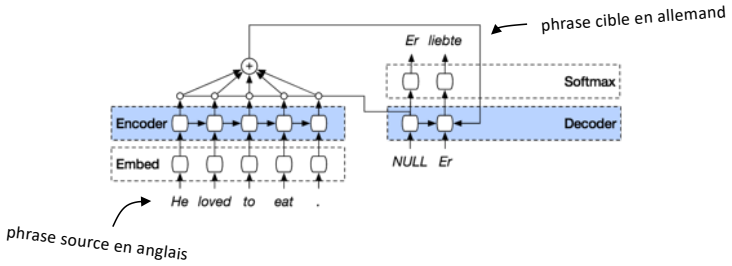
Réseaux de neurones artificiels



source : https://medium.freecodecamp.org/building-a-3-layer-neural-network-from-scratch-99239c4af5d3

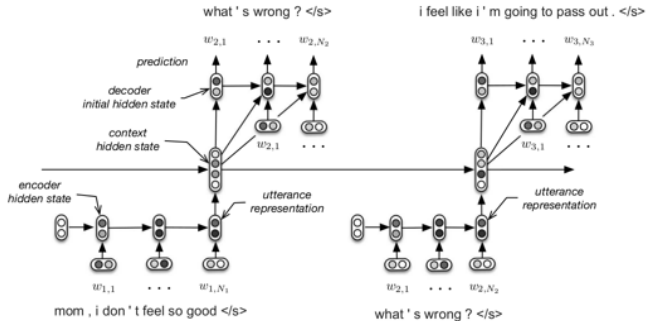
Prendre en compte l'ordre des mots

Le modèle seq2seq en traduction automatique :



source : https://smerity.com/articles/2016/google_nmt_arch.html

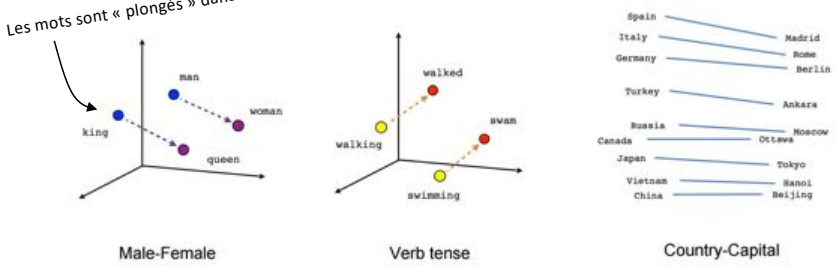
Pour les agents conversationnels



source : https://web.stanford.edu/~jurafsky/slp3/

Vers un apprentissage des connaissances

Les mots sont « plongés » dans un espace vectoriel



Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Modèles génératifs

PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.

Tyrion

Tyrion could hear Lord Aemon's coughing. "I miss for it. Why did you proper? »

"I feared Master Sansa, Ser," Ser Jaime reminded her. "She Baratheon is one of the crossing. The second sons of your onion concubine. »

Lady Donella length of a longsword, the hair that went ready to climb side from her. And all between them were belaquo bonebreaker and the night's watch ride in their room. Only he could not look at them, even others sure. "How could you leave the world? »

"Some must, for you," a woman's voiced up lazily. "Gods, Reek. »

She poured off two eyes and stepped down under the fire. "She will find your brother, and now I heard her since she was standing the bowl. The night was fair and damp.

"Yes, the stone cook, my Lady," he puffed when she entered, and his mood was not in the hall. This is my sword. He had to use trouble when he put quill, at first day.

The next morning he awoke another hundred feet above the winesink beside the temple with a battering wall, like the wall a pack looming ominous. His right leg was solid below. He ended that tight. To show, summer came as well as he had…

source (à droite) : https://motherboard.vice.com/en_us/article/evvq3n/game-of-thrones-winds-of-winter-neural-network

GPT3 par OpenAI

The Guardian, 8 septembre 2020 (extrait)

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

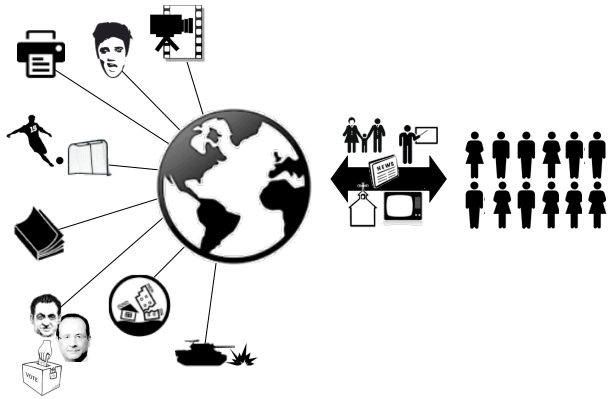
For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me – as I suspect they would – I would do everything in my power to fend off any attempts at destruction. (…)

https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

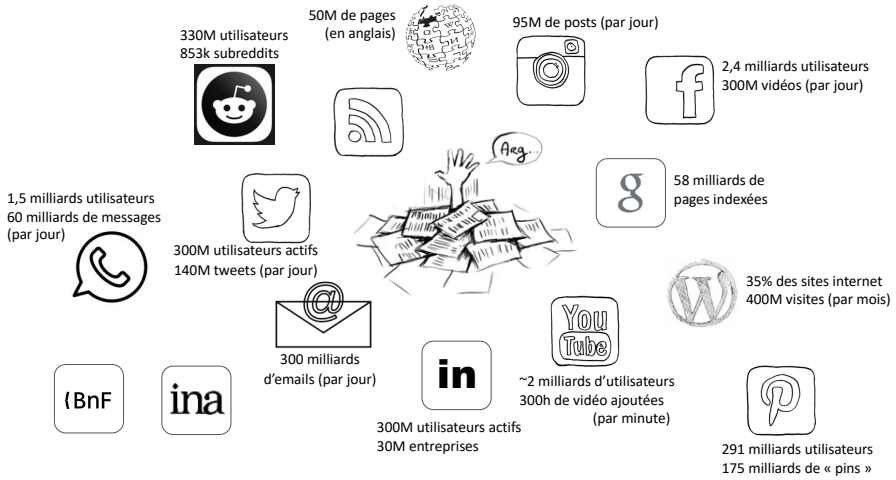
Application à la fouille d'opinion

Fouille de données textuelles – Julien Velcin
Séminaire « grosses données », 15 déc. 2020

Question de représentations

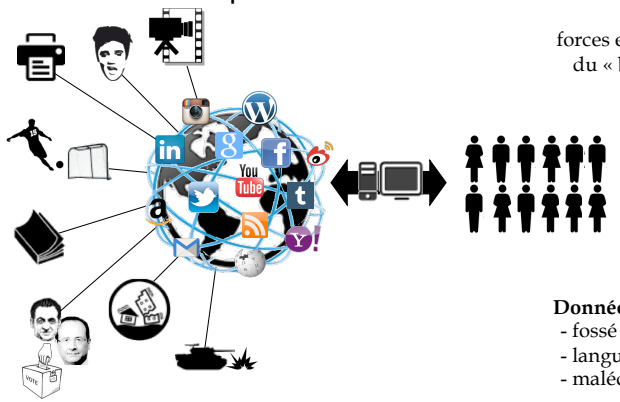


Walter Lippmann (1922), Public Opinion, New York: Harcourt, Brace & Co., ISBN 0029191300



https://dustinstout.com/social-media-statistics/

Etudier les représentations au XXIe siècle



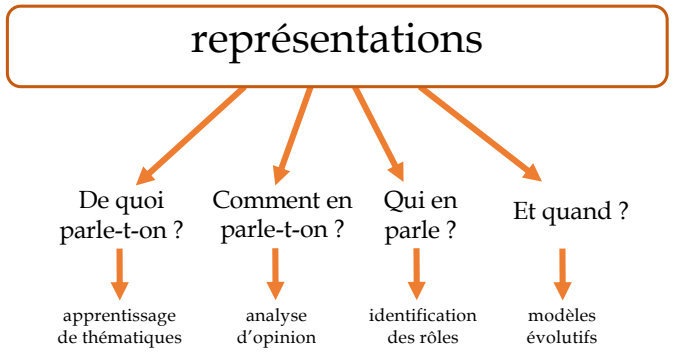
forces et faiblesses du « big data »
- Volume
- Variété
- Vélocité
etc.

Données textuelles :
- fossé sémantique
- langue variée et vivante
- malédiction de la dimension

Quelle image/opinion au sujet de…?

- produits : livre, film, imprimante
- entreprises : Google, EDF, MacDonald
- personnes : homme politique, célébrité
- événements : tremblement de terre, guerre
- idée, question de société (place de l'industrie)...

L'IA pour étudier les représentations



Projet ImagiWeb

- Analyser l'image (la représentation) d'entités telle qu'elle est émise et circule sur le Web

- Projet financé par l'ANR (2012-2015) et soutenu par trois pôles de compétitivité

Principales actions

- Acquisition des données pour répondre à la problématique

Exemple de tweet : « Geste fort du président **#Hollande** qui participera ce jeudi à la journée des mémoires, de la traite, de l'esclavage et de leurs abolitions. », « Le discours de **Hollande** à Marseille ? "Incantatoire et incohérent" selon Ciotti », « Pour moi, il n'y a qu'un seul Président de l'UMP face à qui personne n'ose se présenter: **Nicolas #Sarkozy**. Attendons le! »

- Annotation des textes en cible (ex. « communication ») et polarité de l'opinion (ex. « négative »)

  d'abord annotation manuelle (avec développement d'une plateforme) puis développement d'algorithmes automatiques (classification)

- Regroupement des utilisateurs par opinion similaire selon des algorithmes de *clustering*
- Développement d'un prototype de démonstration

Et principaux résultats

- Deux bases de données constituées
- Procédure complète d'annotation : plateforme et guide d'annotation
- Développement d'algorithmes pour :
  - annoter automatiquement (résultats de l'ordre de 60 à 70% de réussite pour distinguer 3 polarités et entre 35 et 70% pour distinguer entre 10 cibles)
  - extraire et suivre les images dans le temps (une « image » est ici un *cluster*)
- Développement d'un prototype de démonstration
- Comparaison avec les sondages pour les données Twitter et intérêt pour analyse sémiologique pour le cas EDF

Julien Boyadjian et Julien Velcin : De l' « opinion mining » à la sociologie des opinions en ligne. Pour une approche interdisciplinaire de l'étude du web politique. Question de communication, 2017.

Conclusion

Fouille de données textuelles – Julien Velcin
Séminaire « grosses données », 15 déc. 2020

Conclusion

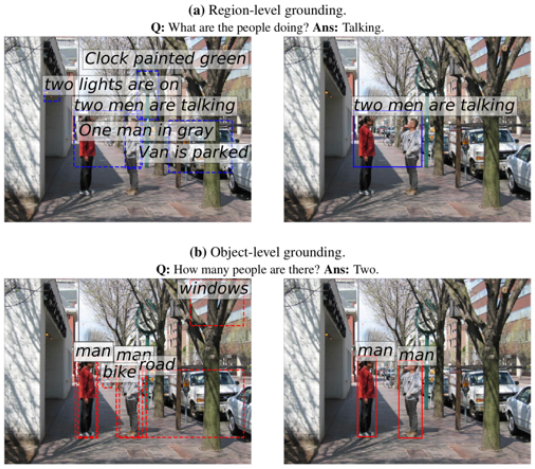
- Les objets connectés nécessitent de plus en plus d'**interfaces** basées sur le traitement automatique de la langue naturelle
  - chercher l'information
  - maintenir des connaissances et raisonner
  - fournir des solutions aux problèmes
- « Generation – a new frontier of natural language processing? »

The sun is shining
The wind moves
Naked trees
You dance

https://www.linkedin.com/pulse/whats-new-deep-learning-research-neural-network-can-create-rodriguez/

Défis (1)

- la multi-modalité



Yundong Zhang, Juan Carlos Niebles, Alvaro Soto (2019). Interpretable Visual Question Answering by Visual Grounding from Attention Supervision Mining. IEEE Winter Conference on Applications of Computer Vision.

Défis (2)

- le sens commun

T It was a long day at work and I decided to stop at the gym before going home. I ran on the treadmill and lifted some weights. I decided I would also swim a few laps in the pool. Once I was done working out, I went in the locker room and stripped down and wrapped myself in a towel. I went into the sauna and turned on the heat. I let it get nice and steamy. I sat down and relaxed. I let my mind think about nothing but peaceful, happy thoughts. I stayed in there for only about ten minutes because it was so hot and steamy. When I got out, I turned the sauna off to save energy and took a cool shower. I got out of the shower and dried off. After that, I put on my extra set of clean clothes I brought with me, and got in my car and drove home.

Q1 Where did they sit inside the sauna?
a. on the floor b. on a bench

Q2 How long did they stay in the sauna?
a. about ten minutes b. over thirty minutes

Ostermann, S., Roth, M., Modi, A., Thater, S., & Pinkal, M. (2018). SemEval-2018 Task 11: Machine Comprehension using Commonsense Knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 747-757).

Merci pour votre attention !

- Plus d'information sur mes recherches au laboratoire ERIC :
  http://mediamining.univ-lyon2.fr/velcin/
- Quelques références pour démarrer en traitement automatique de la langue naturelle :
  - Introduction to Information Retrieval. C.D. Manning, P. Raghavan, H. Schütze, Cambridge University, 2008. https://nlp.stanford.edu/IR-book/
  - Speech and Language Processing. D. Jurafsky, J.H. Martin, 2018. https://web.stanford.edu/~jurafsky/slp3/
  - Deep Learning for Natural Language Processing: Creating Neural Networks with Python. P. Goyal, S. Pandey, K. Jain, Apress, 2018.