

# Master Humanités Numériques

Machine Learning pour les données textuelles  
Introduction générale

Julien Velcin  
Laboratoire ERIC – Université Lyon 2  
<http://eric.univ-lyon2.fr/jvelcin>

## Plan du cours

- De l'analyse des données textuelles
  - exemple de données textuelles
  - définition et principales difficultés
  - quelques applications phares
- Mise en pratique
  - moteur de recherche d'information
  - classification de données textuelles
  - grands modèles de langue

## Plan du cours

- De l'analyse des données textuelles
  - exemple de données textuelles
  - définition et principales difficultés
  - quelques applications phares
- Mise en pratique
  - moteur de recherche d'information
  - classification de données textuelles
  - grands modèles de langue

## De (très) nombreuses sources de données textuelles

- **Sites web :**
  - articles de presse
  - blogs, forums
  - critiques de produits (ebay, amazon, allociné)
  - encyclopédies (wikipedia, freebase)
- **Réseaux et médias sociaux :**
  - Facebook, Twitter, Flickr, LinkedIn ...
- **Données ouvertes (open data) :**
  - data.gov, ParisData...
- **Humanités numériques :**
  - données historiques (patrimoine)
  - nombreux corpus disponibles

**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/](https://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](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](https://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 ...

#journéedelalanguefrançaise

Top Direct Comptes Photos Vidéos Autres options

Suggestions · Actualiser · Tout afficher

Khalil (pilgrim) @sehnaoui Suivre Sponsored

Tom Kenter @TomKenter Suivi par Shiri Dori-Hacohen ... Suivre

Alberto Lumerbras @alberto... Suivi par Bertrand Jouve Suivre

Trouver des amis

Tendances · Modifier

#JournéeDeLaLangueFrançaise  
#BourdinDirect  
#SRFCOL  
#SOSPascal  
#BrunoFunRadio  
Lacazette  
Troyes  
Oliver 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 @Nitneu\_ · 1 min POUAHHAHAH #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? #JournéeDeLaLangueFrançaise

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" #JournéeDeLaLangueFrançaise

sign in subscribe search

UK world sport football opinion culture business lifestyle fashion environment tech travel

home

### headlines

Now 4°C Lyon

12:00	15:00	18:00	21:00
9°C	13°C	9°C	6°C
Cloudy	Cloudy	Cloudy	Cloudy

Climate change / February breaks global temperature records by 'shocking' amount

Warnings of climate emergency after surface temperatures 1.35C warmer than average temperature for the month

Great Barrier Reef Severe coral bleaching worsens

US elections 2016 Clinton and Sanders attack 'pathological liar' Trump

Japan US sailor arrested in Okinawa on suspicion of rape

German elections Anti-refugee AfD party makes dramatic gains

Brazil More than a million protest over 'horror' government

Thailand Eight die in bank after chemical fire extinguisher leak

Ivory Coast Gunmen open fire on tourist resort, killing 16

United Arab Emirates Plane reported missing in Yemen

Egypt Justice minister sacked for saying he would arrest prophet Muhammad

+ More headlines

### highlights

100 Best Nonfiction Books of All Time

No. 7 The Right

Hide

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

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

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 évidemment le reste sera en DLC ce qui fera grimper le jeu à environ 130€ (édition deluxe) donc pas pour moi...

FRONT · ALL · RANDOM · ASKREDDIT · FUNNY · PICs · VIDEOS · TODAYLEARNED · GIFs · NEWS · AWW · WORLDNEWS · MOVIES · GAMING · SHOWERTHOUGHTS · TELEVISION · JOKES · EXPLAINLKEIMFIVE · MILDLYINTERESTING · IAMA · SCI

## THE NEW REDDIT JOURNAL OF SCIENCE



hot new rising controversial top filter by field ▾

Humans have triggered the last 16 record-breaking hot years experienced on Earth (up to 2014), with the new research tracing our impact on the global climate as far back as 1937. The findings suggest that without human-induced climate change, recent hot summers and years would not have occurred. ▶ phys.org

15 hours ago by drevipodee  
3720 comments share

### Top 200 Comments show 500

sorted by: best (suggested) ▾

[+] old-tobe 665 points 13 hours ago

So what can we actually do to combat this? Aside from colonizing space and getting humans off this planet?

permalink

[+] XIIcubed 1957 points 13 hours ago \* @

Switch to nuclear energy.

edit: thanks for the gold nuclear energy fwtw

permalink parent

[+] Mr\_Industrial 939 points 13 hours ago

Good luck convincing several million people that nuclear energy is safer than most other forms of energy. It's not about the facts, it's about perception of the facts.

permalink parent

[+] climbtree 828 points 12 hours ago

You don't have to. The public rarely has input into power plant construction etc. Once they're up and running no-one cares about it anymore.

If you ask people if they'd like a change, 90% will say no, 95% if you say it might involve danger. If you make the change and ask how happy people are most are just as happy.

permalink parent

[+] Mr\_Industrial 158 points 12 hours ago

This is a good point. The thing you have to remember though is that the people in charge who have the power to decide what type of

```

174 <token rang="8">on</token>
175 <token rang="9">sait</token>
176 <token generique="que" rang="10">que::</token>
177 </productionVerbale>
178 <productionVerbale pseudo="M" rang="5">
179 <espace longueur="10" rang="1"/>
180 <chevauchement type="debut" position="Externe" rang="2">[</chevauchement>
181 <token rang="1">oui</token>
182 <espace longueur="3" rang="4"/>
183 <chevauchement type="fin" position="Externe" rang="5">]</chevauchement>
184 <espace longueur="2" rang="6"/>
185 </productionVerbale>
186 <productionVerbale pseudo="C" rang="6">
187 <token rang="1">on</token>
188 <token rang="2">a</token>
189 <token rang="3">positionné</token>
190 <token rang="4">tous</token>
191 <token rang="5">les</token>
192 <token rang="6">parkings</token>
193 <token rang="7">d</token>
194 <token generique="attente" rang="8">ATTENTE</token>
195 <token rang="10">on</token>
196 <token rang="11">de</token>
197 <token generique="deserte" rang="12">deSSERte</token>
198 <token rang="13">l</token>
199 <token rang="14">accès</token>
200 <token generique="principal" rang="15">princiPAL</token>
201 <token rang="16">l</token>
202 <token rang="17">accès</token>
203 <token generique="instant" rang="18">instant</token>
204 <token rang="19">service</token>
205 <token elision="1" rang="19">d</token>
206 <token rang="19">service</token>
207 <token rang="19">service</token>
208 <pause type="courte" duree="" rang="20">>(. )</pause>
209 <token generique="h" rang="21">.h::</token>
210 <token rang="22">la</token>
211 <token rang="23">seule</token>
212 <token rang="24">chose</token>
213 <pause type="courte" duree="0.2" rang="25">>(0.2)</pause>
214 <token rang="26">qui</token>
215 <token rang="27">s'et</token>
216 <token rang="28">ce</token>
217 <token rang="29">modifier</token>
218 <token rang="30">pour</token>
219 <token rang="31">l</token>
220 <token generique="instant" rang="32">instant</token>
221 <token rang="33">.h</token>
222 <token rang="34">c</token>
223 <token rang="35">est</token>
224 <token rang="36">on</token>
225 <token rang="37">sait</token>

```



# Articles scientifiques

## Un cadre pour la représentation et l'analyse de débats sur le Web

Anna Stavrianou\*, Julien Velcin\*\*, Jean-Hugues Chauchat\*\*  
ERIC Laboratoire - Université Lumière Lyon 2, Université de Lyon,  
5 avenue Pierre Mendès-France 69676 Bron Cedex, France  
\* anna.stavrianou@univ-lyon2.fr  
\*\* julien.velcin@univ-lyon2.fr  
\*\*\* jean-hugues.chauchat@univ-lyon2.fr

**Résumé.** Les débats en ligne sont souvent modélisés par des réseaux sociaux d'utilisateurs représentés sous forme de graphes, chaque noeud correspondant à l'un des intervenants du débat. Ici, nous proposons un nouveau modèle basé sur un graphe de messages : chaque noeud du graphe correspond à l'un des messages échangés et chaque arc relie un message à celui auquel il répond ; ces arcs peuvent être caractérisés par les mots-clés résumant la relation entre les messages connectés. Cette modélisation permet une meilleure représentation de la dynamique du débat ainsi que l'identification de chaînes de discussion. Nous comparons les deux représentations, graphes représentant les utilisateurs et graphes représentant les messages, puis nous analysons la connaissance qui peut être extraite à partir de chacun d'eux. Nos expériences sur des débats réels valident le modèle proposé et montrent les informations complémentaires qui sont apportées par le graphe des messages.

### 1 Introduction

Le développement du Web2.0 a provoqué la création d'un grand nombre de blogs, de forums et de discussions en ligne. L'analyse de ce type de discussions est très intéressante, tant pour des organismes publics que pour l'industrie ou le secteur commercial. Les discussions en ligne contiennent les intérêts et les avis des internautes, des critiques de produits, la présentation d'idées politiques, etc. En conséquence analyser ces données devient une problématique stratégique.

A. Stavrianou et al.

Turney, P. et M. Litman (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM TOIS* 2(4), 315–346.  
Zhang, J., M. Ackerman, et L. Adams (2007). Expertise networks in online communities: Structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pp. 221–230.

Zhou, D., I. Couicoll, H. Zha, et C.-L. Giles (2007). Discovering temporal communities from social network documents. In *ICDM: International Conference on Data Mining*, pp. 745–750. IEEE Computer Society.

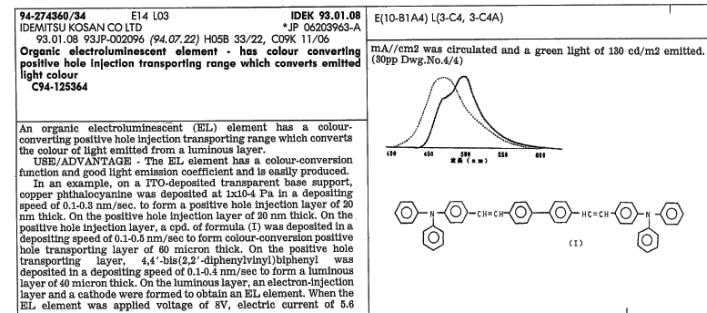
### Summary

The online discussions are often modeled as a social network of users and they are represented by a graph where each node denotes a participant of the discussion. In this paper, we propose a new framework for discussion analysis. It is based on a graph of messages: each node corresponds to a message of the discussion and each edge (directed) points out which message the specific edge refers to. The edges can be weighted according to their characteristics and messages, users and links can be analyzed to infer the context of the discussion and facilitates the identification of discussion chains. We compare the two representations: the user-based and the message-based graph and we analyze the different information that can be extracted from them. Our experiments with real data validate the proposed framework and show the additional information that can be extracted from a message-based graph.

# Plan du cours

- De l'analyse des données textuelles
  - exemple de données textuelles
  - **définition et principales difficultés**
  - quelques applications phares
- Mise en pratique
  - moteur de recherche d'information
  - classification de données textuelles
  - grands modèles de langue

# Brevets



# Big data, le Web et tout ça...

- Big data :
  - V de Volume
  - V de Vélocité
  - V de Variété (**texte**, image, vidéo, son, tags...)
- Etc.
- Le WWW est une source phénoménale de données, en particulier textuelle
- Mais il existe beaucoup d'autres sources : mémoire d'entreprise, données du patrimoine...

## De quel volume parle-ton ?

- Techniquement infini, on parle de « big data »
- 16 à 18 milliards de pages indexées par Bing et entre 45 et 50 milliards par Google au 14/03/16  
(source : <http://www.worldwidewebsize.com>)
- 1 million de serveurs à travers le monde traitent ~1 milliards de requêtes par jour  
(source : <http://atkinsbookshelf.wordpress.com/tag/how-many-servers-does-google-have/> au 3/01/14)
- 175 millions de tweets envoyés chaque jour en 2012

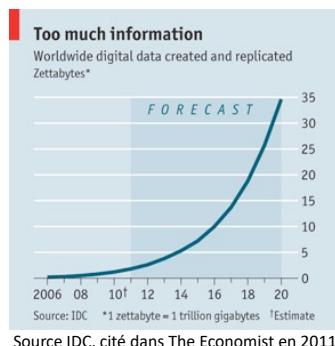
## Surcharge d'information



Image credit: Go-Globe.com

## Et ce n'est pas fini...

- De plus en plus de données numériques :



1 zettabyte = 1 000 000 000 000 000 000 byte

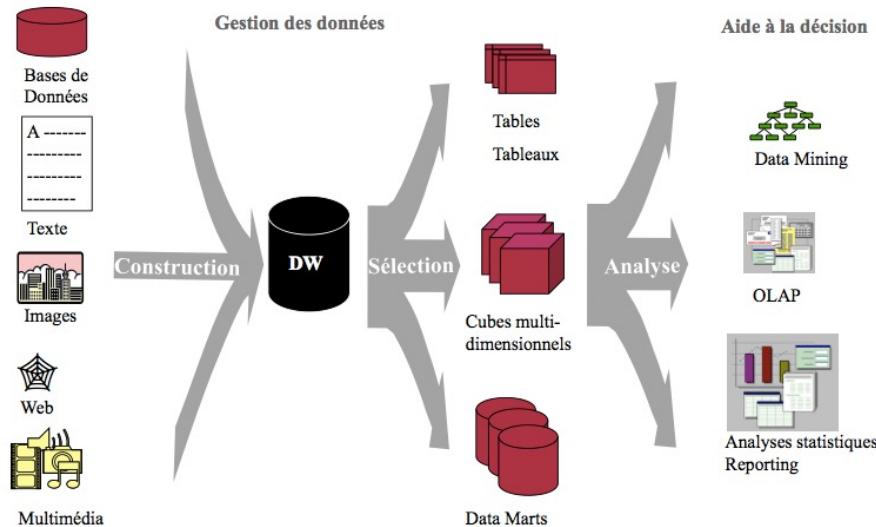
## Une solution : la science des données (*data science*) !



### Valoriser les grandes masses de données :

- Rechercher l'information du haut vers le bas (*top down*)
- Extraire des connaissances utiles (pépites) à partir des données (*bottom up*)
- De plus en plus des **approches hybrides** basées sur le **machine learning**

## Liens avec le data mining et l'informatique décisionnelle



## Problématiques

- Récupération, extraction
- Stockage des données
- Représentation, indexation
- Analyse des données
- Visualisation, exploration
- Evaluation
- Prise de décision

## Pour les données textuelles

- Extraction, stockage des données :
  - ➔ Comment gérer l'hétérogénéité des formats ?
  - ➔ Quelle structure de stockage ?
- Représentation, indexation :
  - ➔ Quelle est la meilleure représentation ?
  - ➔ Comment indexer les données de manière efficace ?
- Analyse des données :
  - ➔ Comment comparer des données textuelles ?
  - ➔ Quels algorithmes choisir ?

## « Quelques » difficultés

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

## Un cas d'étude : le « HuffPos »

- En lien avec les réseaux sociaux
  - Organisé en thématiques
  - Articles commentés
  - Communauté de bloggers
  - Le journaliste peut jouer à la fois le rôle de curateur et de *community manager*



2

## Fouille de textes : origines

- Intelligence artificielle (IA)
    - Traitement Automatique des Langues (TAL)
  - Statistiques
    - Statistiques textuelles
  - Linguistique
    - Linguistique computationnelle
  - Puis :
    - Bases de données, fouille de données...

April 19, 2012

# THE HUFFINGTON POST

THE INTERNET NEWSPAPER: NEWS BLOGS VIDEO COMMUNITY

Login with Facebook to see what your friends are reading

**Rep. Gwen Moore**  
 U.S. Representative for Wisconsin's  
 4th Congressional District

GET UPDATES FROM REP. GWEN MOORE

FAN

RSS

EMAIL

Follow

Recency | Popularity

Page: 1 2 3 Next > Last > (3 total)

**bluespagan**  
*Love is the Law, Love under Will*  
 140 Fans

4 hours ago (7:22 AM)

Rep. Moore, I am a victim of sexual abuse (from my biological father so just as you no dark back alley stranger). I want to say thank you.

And to those out there who are questioning her story, I never turned my abuser in. Why? Because I was ashamed, scared and felt alone. He left me feeling as though I (and I went to undermine and bid this to bring attention to the "I portion here) had done something wrong, that I had asked for it and that I deserved to be punished for it as well. I was 8 years old at the time and I went to him when I really was able to sit down and talk to him. That in and of itself is sad. Sady, it didn't end there, it simply ended with that person. I went on to abuse myself and allowed others to abuse me. Thankfully, about 18 years worth of abuse, I found a man that I opened up to about what happened to me. He listened, he understood and he held me as I cried. He then helped me to get away from her. He has been my rock ever since. Now I can look at my son and my husband and we have a beautiful little girl who I will ensure will know that she can come to me with anything, something I didn't have. So thank you Rep. Moore for standing up for women like me.

↳ Reply

Permalink | Share it

**HUFFPOST SUPER USER**  
**Shawn Hunt**  
*Liberty and Justice for ALL!!!!*  
 1059 Fans

1 hour ago (8:58 AM)

I see you. I am happy that you no longer live in fear and that you can trust someone. good life. Your story of survival truly made my day.

↳ Reply

Permalink | Share it

**bluespagan**  
*Love is the Law, Love under Will*  
 140 Fans

2 hours ago (8:34 AM)

Thank you. It is just easier to type out when you are anonymous. I still have flashbacks of it all over and in person. Hopefully one of these days I will no longer feel the shame I still feel at times and will be able to freely share my story.

↳ Reply

Permalink | Share it

**llip3558**

39 Fans

6 hours ago (8:08 AM)

Great article, I agree wholeheartedly and I am so sorry for everything you went through. God Bless you and thanks for being a voice for all women.

Julien Velion - présentation ARC6 18 Octobre 2012

## Women Are Waiting...

Posted: 04/18/2012 4:21 pm

React > Important Funny Typical Scary Outrageous Amazing Innovative

Follow > Congress Sexual Abuse Values Violence Violence Against Women Act, V

loveliberalchick

6 Fans

10 hours ago (1:45 AM)

fanned

↳ Reply

Permalink | Share it

HUFFPOST PUNDIT

Braintrust

2325 Fans

8 hours ago (7:53 PM)

Thank you, Gwen Moore. I'm confident that as long as there are people like you who care, there's a

possibility to improve things. Yes we can!!

↳ Reply

Permalink | Share it

Submit this story

oldwolf40

Religion is a tool of the evil.

730 Fans

8 hours ago (3:55 AM)

The actual bill, lots of yadda yadda yadda:

These <http://www.govtrack.us/congress/bills/112/h4271>

these <

↳ Reply

Permalink | Share it

Permalink | Share it

24

Julien Velcin - présentation ARC6 18 Octobre 2012

26

# Traitement Automatique des Langues

## *Natural Language Processing*

- Le TAL (NLP) est un champ de recherche et un ensemble de technologies qui permet à une machine de comprendre, interpréter, classer, générer des textes
  - Le TAL est un défi dès les premiers travaux en IA
  - Le TAL permet de :
    - décomposer un texte en ses **constitutants**
    - identifier (découvrir) le **sens** des mots et des expressions
    - découvrir des **motifs** (*patterns*) pour classer les textes ou générer du texte
  - Liens avec la **Recherche d'Information** (*Information Retrieval*)

## Exemples d'applications

- **chercher** de l'information dans les BD et le Web (moteurs de recherche)
- **traduire** automatiquement des textes
- **classer** des textes en fonction de sa thématique, de l'opinion véhiculée...
- **résumer** un document
- **dialoguer** pour répondre à des questions...

## Plan du cours

- De l'analyse des données textuelles
  - exemple de données textuelles
  - définition et principales difficultés
  - **quelques applications phares**
- Mise en pratique
  - moteur de recherche d'information
  - classification de données textuelles
  - grands modèles de langue

### Une première application phare : la Recherche d'Information (1)

- Les moteurs de recherche modernes utilisent les dernières innovations en RI
- Données textuelles *et* structure
- Ces moteurs combinent :
  - indexation des données du Web
  - enrichissement de la requête formulée
  - estimation de la fiabilité des pages

### Recherche d'information (2)

- Des robots (*crawler, spider*) indexent :
  - mots-clés
  - concepts
- Créditabilité d'un site Web (PageRank, HITS)
- Différents critères :
  - correspondance entre la requête et la page
  - structure, richesse, diversité
  - mise à jour régulière, nouveautés etc.

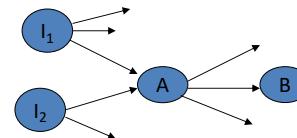
## Recherche d'information (3)

- Modèle de correspondance :
  - Document D = un ensemble de mots clefs pondérés
  - Requête Q = un ensemble de mots clefs non pondérés
  - $R(D, Q) = \sum_i w(t_i, D)$ , où  $t_i$  est dans Q
- De nombreux modèles possibles :
  - booléen (0 ou 1),
  - vectoriel,
  - probabiliste...

33

## Recherche d'information (4)

- PageRank de Google :



$$PR(A) = (1-d) + d \sum_i \frac{PR(I_i)}{C(I_i)}$$

- Assigne une valeur numérique à chaque page, en fonction des liens entre pages
- $d$ : damping factor (0.85)
- D'autre critères possibles, par ex. La proximité entre les mots clefs (« ...information retrieval ... » mieux que « ... information ... retrieval ... »)

34

## Recherche d'information (5)



35

## Recherche d'information (6)

- Utiliser les *snippets* retournés par les moteurs de recherche :
- 1 [Text mining - Wikipedia, the free encyclopedia](#)
- Text mining, sometimes alternately referred to as **text data mining**, roughly equivalent to **text analytics**, refers to the process of deriving high-quality ...  
http://en.wikipedia.org/wiki/Text\_mining [Ask, Entrieweb, Google, Wikipedia, Yahoo]
- Text/Document clustering pour organiser les snippets avec une méthode de clustering
  - Etiquette les catégories avec des expressions fréquentes, mais d'autres solutions existent (ex. : entités nommées)

36

# Résumé automatique

Cracks Appear in U.N. Trade Embargo Against Iraq.

Cracks appeared Tuesday in the U.N. trade embargo against Iraq as Saddam Hussein sought to circumvent the economic noose around his country. In response, the United Nations said it would increase its aid to countries hardest hit by the sanctions. Hoping to defuse criticism that it is not doing its share to oppose Saddam, the United Nations' most powerful audience — the Security Council — will be composed of foreign ministers from developing nations.

U.S. troops have been sent to the Saudi Arabian desert indefinitely. "I cannot predict just how long it will take to convince Iraq to withdraw from Kuwait," Bush said. More than 150,000 U.S. troops have remained in the Persian Gulf region to deter a possible Iraqi invasion of Saudi Arabia. Bush's aides said the president would follow his address to Congress with a televised message for the Iraqi people, declaring the world is united against their government's invasion of Kuwait. Saddam had offered Bush time on Iraqi TV, *"to express our thanks and our appreciation for your support."* Namibia, the first of the developing nations to respond to an offer Monday by Saddam of free oil, *"in exchange for sending their own tankers to get it."* said in a statement. Saddam has offered to maintain their mill 200,000 barrels a day to the two countries annoucement, said the exporter itself, it can reported that following James Baker's demand to withdraw from Kuwa Shevardnadze told him he said his heart went world will not be bla developed in the U.S. All their husbands behir Evacues spoke of fo Thuraya, 19, who wo U.S. or other countries known how many me Syria. He said there v billion-a-month estin 1. Cheney promised responding to shov from abroad is gettin the World Bank and tour seeking \$10.5 b food, fuel and supplies from the Middle East. Japanese territory, except for ceremonial occasions. On Monday, Saddam offered developing nations free oil if they would send their tankers to pick it up. The first two countries to respond Tuesday, the Philippines and Namibia, said no. Manila said it had already fulfilled its oil requirements, and Namibia said it would not "sell its sovereignty" for Iraqi oil.

Venezuelan President Carlos Andres Perez dismissed Saddam's offer of free oil as a "propaganda ploy." Venezuela, an OPEC member, has led a drive among producing nations to boost prices by 10 percent over the last year. Up to 20 percent of oil produced by OPEC members, mostly Saudi Arabia, has higher reserves. But according to the State Department, Cuba, which faces an oil deficit because of reduced Soviet deliveries, has received a shipment of Iraqi petroleum since U.N. sanctions were imposed five weeks ago. And Romania, it said, expects to receive oil indirectly from Iraq. Romania's ambassador to the United States, Virgil Constantinescu, denied that claim Tuesday, calling it "absolutely false and without foundation."

Human written summary

7800 chars, 1300 words

37

## Exemple : the babelfy project



39

# Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.  
people    organizations    places
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.

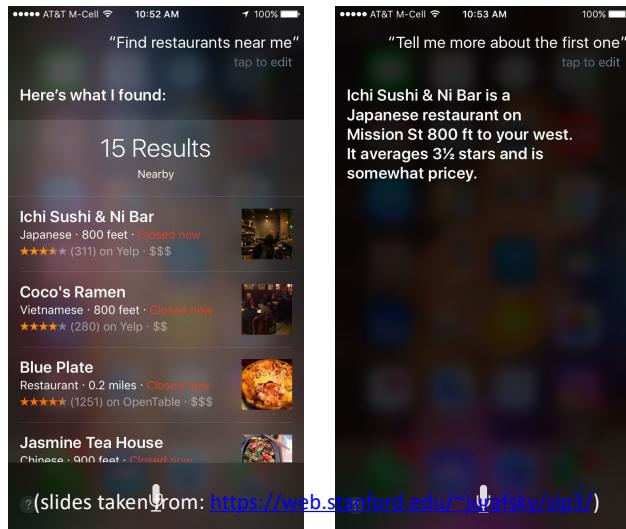
38

## Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the Web).
  - When was Barack Obama born?
    - August 4, 1961
  - Who was president when Barack Obama was born?
    - John F. Kennedy
  - How many presidents have there been since Barack Obama was born?
    - 9

40

# Dialog Systems and Chatbots



41

## 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 OF YOU?**

My father

**YOUR FATHER**

(slides taken from: <https://web.stanford.edu/~jurafsky/slp3/>)

You are like my father in some ways.

**WHAT RESEMBLANCE DO YOU SEE**

You are not very aggressive but I think you don't want me to notice that.

**WHAT MAKES YOU THINK I AM NOT AGGRESSIVE**

You don't argue with me.

**WHY DO YOU THINK I DON'T ARGUE WITH YOU**

You are afraid of me.

**DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU**

My father is afraid of everybody.

**WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER**

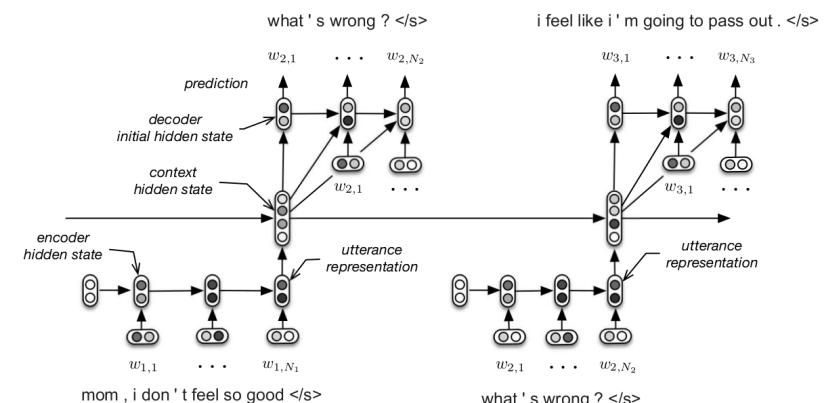
Bullies.

**DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE**

(slides taken from: <https://web.stanford.edu/~jurafsky/slp3/>)

## Architectures Seq2seq

Serban, Iulian V., Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models."



(slides taken from: <https://web.stanford.edu/~jurafsky/slp3/>)

# Opinion mining

This film is an instant classic.  
The scene between Christine Bale and Casey Affleck in their kitchen is a masterpiece of modern film.  
Casey Affleck gives a visceral performance and an important one.  
The rest of the film is an emotional American tragedy, and one that should not be missed.  
An also incredibly intense scene between Affleck and Harrelson was also compelling, with both Harrelson and Affleck doing some of their most important work in this film to date.  
Defoe, Saldana, Whitaker - all giving incredibly strong performances which so absolutely immerses the audience - its just tremendous.  
This film has heart, and that heart is trampled, beaten and broken - an American tragedy.

Luckily I had seen four previous great movies, so I figured 4/5 intelligent, engrossing movies is not too bad in the last few months.  
This movie was gross, boring, raunchy, disgusting & really sickening. I am by no means a prude or turned off by slapstick, but this film had no redeeming qualities.  
Feces, farts, penises, beer, & constant shallow dialogue was not one bit funny.  
When I saw "Borat" with Sacha Baron Cohen a number of years ago, I laughed so much that tears were running down my face.  
There was a message in that movie. In fact, grandpa, there is no message.  
It is empty, bathroom humour constantly.  
The movie tries so hard to copy "Borat" & to be funny, but in my opinion, it fails miserably.  
I should have walked out after the first half hour & asked for a refund.  
What a complete waste of an afternoon!



(excerpt from the talk of R. Lebret at the ERIC lab, 2016)

45

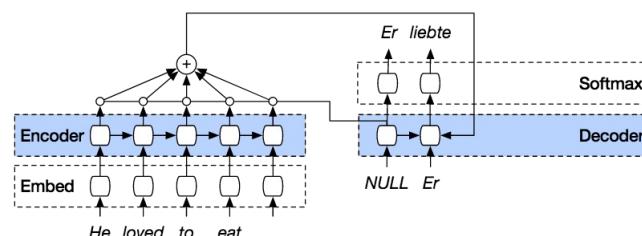
# Traduction automatique (1)

The screenshot shows the Google Translate interface. The source text is in French: "La traduction automatique désigne la traduction d'un texte (ou d'une conversation audio, en direct ou en différé) entièrement réalisée par un ou plusieurs programmes informatiques, sans qu'un traducteur humain n'ait à intervenir. On la distingue de la traduction assistée par ordinateur où la traduction est en partie manuelle, éventuellement de façon interactive avec la machine." The target language is English: "Machine translation refers to the translation of a text (or audio conversation, live or recorded) entirely by one or more computer programs, without the need for a human translator. It is distinguished from computer-assisted translation where the translation is partly manual, possibly interactively with the machine." The interface includes language selection dropdowns (Anglais, Français, Arabe, Déterminer la langue), a toolbar with icons for copy, paste, and search, and a status bar indicating 380/5000 characters.

46

## Traduction automatique (2)

Main trend relies on encoder-decoder ANN with an attention mechanism



(source: [https://smerity.com/articles/2016/google\\_nmt\\_arch.html](https://smerity.com/articles/2016/google_nmt_arch.html))

47

## Elmo, BERT et les autres



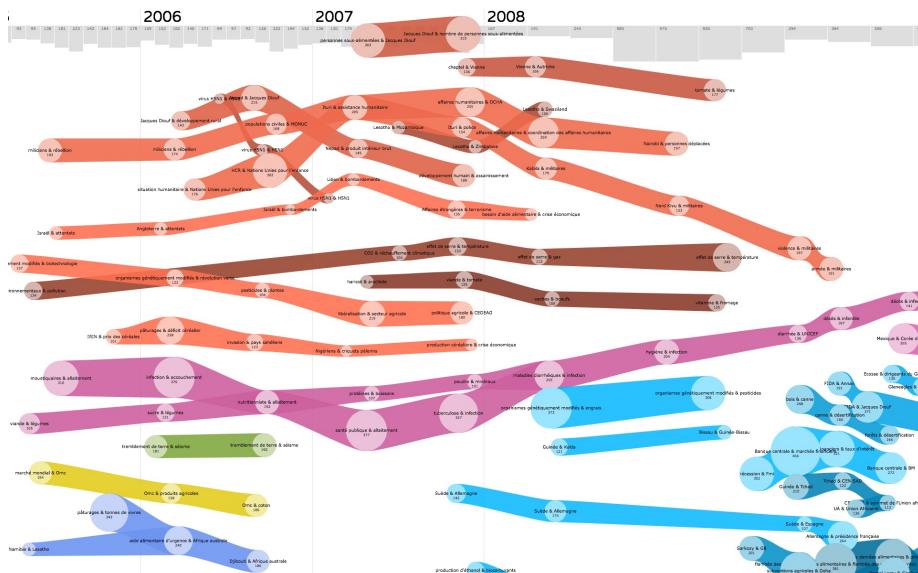
Nous reviendrons dessus en détails...

Mais également :  
analyse des discussions en ligne

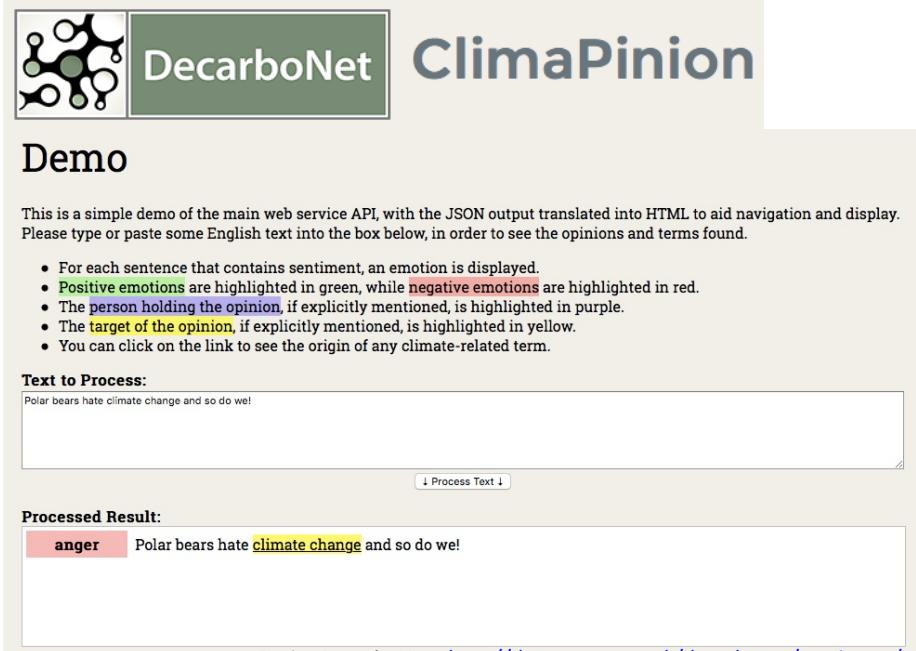


- Motivation :
    - beaucoup de données disponibles, souvent sous-exploitées
    - crucial pour capter l'opinion des internautes
  - Contributions :
    - recommander des messages clefs (Stavrianou et al.,09,10)
    - extraire le réseau social latent (Forestier et al.,11)
    - détecter des célébrités dans les forums (Forestier et al.,12)
    - identifier les rôles dans les discussions (Anukhin et al.,12)

49



Projet Pulseweb : <http://pulseweb.cortex.net>



Projet DecarboNet : <http://demos.gate.ac.uk/decarbonet/sentiment/>