

Network analysis for information retrieval

Julien Velcin

Master MALIA / MIASHS

2023-2024

Outline

- Motivation
 - ubiquity of information networks
 - applications (in particular to IR)
 - importance of indexing
- Representation of documents
 - sparse representations
 - dense representations
 - topic models
- Network analysis
 - spectral clustering, modularity
 - representation learning for graphs
- Analyzing information networks
 - Graph Neural Networks

(part 1/4)

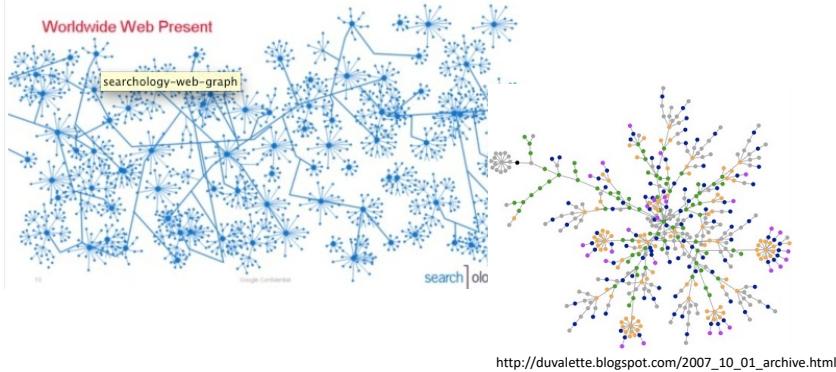
Ubiquity of information networks

- Information networks = networks with information flows
- Information is usually coded as string
- More formally:
 - graph $G = (V, E)$
 - weighted $G = (V, E, W)$
 - Attributes/features associated to E and/or V
- We can model more complex IN (e.g., multiplex networks, heterogeneous networks)

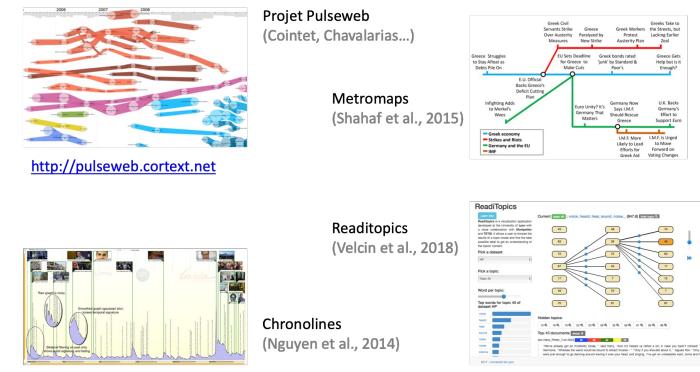
Motivation

Network analysis for information retrieval, M2 MALIA-MIASHS, Julien Velcin

Web graph(s)



Informational landscape



Google climate change

Tous Actualités Images Vidéos Livres Plus Outils de recherche

Environs 142 000 000 résultats (0,29 secondes)

Images correspondant à climate change

NASA: Climate Change and Global Warming

Climate change - Wikipedia, the free encyclopedia

What is Climate Change? What Causes Global Warming?

Home | Climate Change | US EPA

Blogs, forums, chats

Sujet	Auteur	Envoyé le
Victoire idéologique de la gauche ?	The Reporter	02/04/2009 12:27
Re: Et la réponse fut	Flera	05/04/2009 10:04
Re: Et la réponse fut	The Reporter	05/04/2009 10:22
Re: Victoire idéologique de la gauche ?	The Reporter	05/04/2009 09:44
Re: Victoire idéologique de la gauche ?	Fertois	03/04/2009 07:17
Re: Victoire idéologique de la gauche ?	The Reporter	03/04/2009 11:42
Re: Victoire idéologique de la gauche ?	dambrachaine-h	03/04/2009 13:51
Re: Victoire idéologique de la gauche ?	The Reporter	03/04/2009 12:58
Le navire de guerre	dambrachaine-h	03/04/2009 23:46
Re: Victoire idéologique de la gauche ?	Tesla	02/04/2009 14:14
Re: Victoire idéologique de la gauche ?	frecctizen	02/04/2009 17:30
Re: Victoire idéologique de la gauche ?	Tesla	02/04/2009 17:41
Re: Victoire idéologique de la gauche ?	frecctizen	02/04/2009 23:39
Re: Victoire idéologique de la gauche ?	Tesla	03/04/2009 09:13

Visitor has requested translation from language: French

Stephen says:

Hello Marie. My name is Stephen how can I help you?
(Bonjour Marie. Mon nom est Stephen comment ose je vous aide?)

Mane On www.whoson.com says:

I need price [J'ai besoin de prix]

Stephen says:

Link Sent: <http://www.whoson.com/installableorder.aspx>

Stephen says:

Ok. For prices please go to <http://www.whoson.com/installableorder.aspx>
(Ok. Pour des prix veuillez voir à <http://www.whoson.com/installableorder.aspx>)

[FRONT](#) [ALL](#) [RANDOM](#) [ASKREDDIT](#) [FUNNY](#) [PICS](#) [VIDEOS](#) [TODAYLEARNED](#) [GIFS](#) [NEWS](#) [AWW](#) [WORLDNEWS](#) [MOVIES](#) [GAMING](#) [SHOWERTHOUGHTS](#) [TELEVISION](#) [JOKES](#) [EXPLAINIKEMPFIE](#) [MILDLYINTERESTING](#) [IAmA](#) [SCIENCE](#)

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](#)

4389 15 hours ago by [drewapode](#) 665 points 13 hours ago 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?

XiiCubed 1957 points 13 hours ago* Switch to nuclear energy. edit: thanks for the gold nuclear energy fwt

Mr_Industrial 659 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.

climbree 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 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.

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

9

sign in subscribe search

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

headlines

Now 4°C Lyon

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

Warnings of climate emergency after surface temperatures 1.3°C warmer than average temperature for the month

Great Barrier Reef: Severe coral bleaching worsens

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

Thailand Eight die in bank after chemical fire extinguisher leak

Ivory Coast Gunmen open fire on tourist resort, killing 16

Brazil More than a million protest over 'horror' government

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

Hide 10

amazon.fr Toutes nos boutiques ▾

Amazon.fr Ventes Flash - Meilleures ventes - Offres recommandées - 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

Évaluez 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.

11

Accueil Notifications Messages #malavita

Top Direct Comptes Photos Vidéos Autres options ▾

Suggestions - Actualiser - Tout afficher 11 nouveaux résultats

K Kaplan International @K... Suivre Sponsorié

Aras BOZKURT @arasbozkurt Suivre

Stéphane Pouyillau @spouyll... Suivre Trouver des amis

Tendances - Modifiez

#EnVolture Sponsorié par Allianz France #ASSEPSG #JacquelineSauvage #Camping #TheVoice #Malavita Milan Bordeaux Ronaldo Florian Thauvin Benoît Violier

Gabi @1gabi_01 - 2 min Et putain... 😂 #Malavita

Mouna Camara @mouna_camara - 2 min Très bon film #Malavita

Stephanie L@Sldlouren - 2 min Après #Malavita place à #LOLUSA

Black Mamba @SmallHawkeye - 2 min Bon ce film était bof. Rien d'exceptionnel, des petits passages marrant. Dommage avec un tel casting... #Malavita

Ree @HirRee - 2 min Film d'action américain en normandie 😂 #Malavita

12

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
* 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 arête au répondeur d'un message à l'autre ; ces arcs sont munis d'une couleur qui indique la relation entre les messages concernés. Cette modélisation permet une meilleure représentation de la dynamique du débat ainsi que l'identification de chaînes de discussion. Nous comparons les deux représentations graphiques représentant les utilisateurs et groupes respectivement par des messages, puis nous analysons la communauté postée 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 discussion 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.

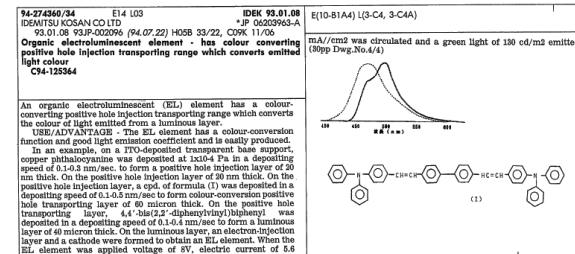
Turge, P. et M. Llinas (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM SIGKDD 2003*, 115-120.
Zhang, J., M. Ackerman, et L. Adams (2007). Expertise networks in online communities: Structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pp. 221-230.

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

Summary

This work discusses are often modelled by 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 node replies to. The edge can be weighted by the keywords that characterize the reply. This allows for a better representation of the dynamics 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.

Brevets



Some applications

Network analysis for information retrieval, M2 MALIA-MIASHS, Julien Velcin

Applications (in particular to IR)

- Search engines
- Personal assistants
- Opinion mining / e-reputation
- Community detection
- Multi-document summarization

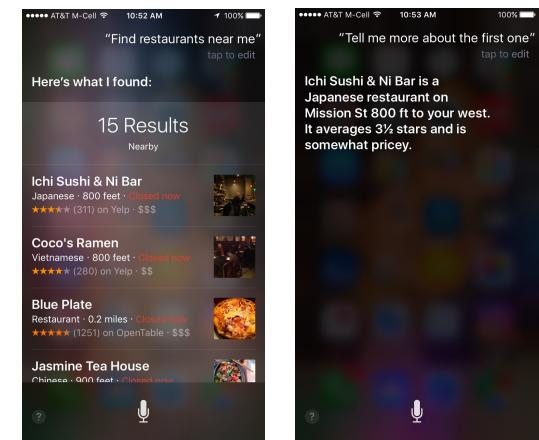
Applications

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



Watson gagne le Jeopardy! en 2011

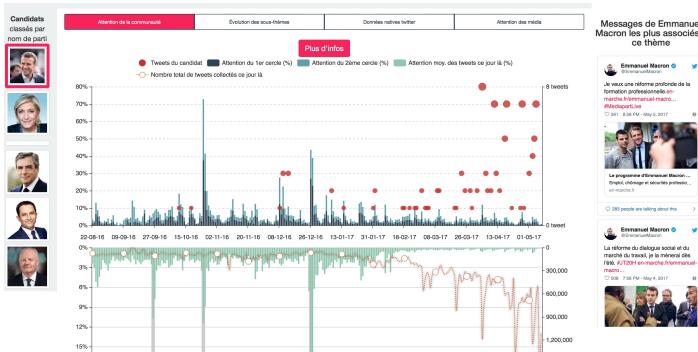
Personal assistants



source : cours de D. Jurafsky : <https://web.stanford.edu/~jurafsky/sip3/>

18

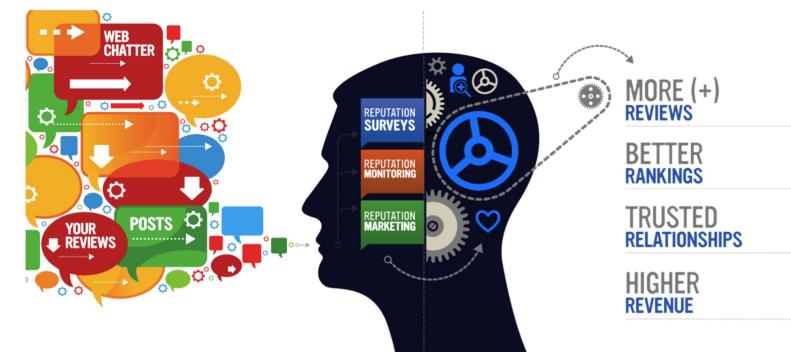
Opinion mining



source : <http://politoscope.org/>

19

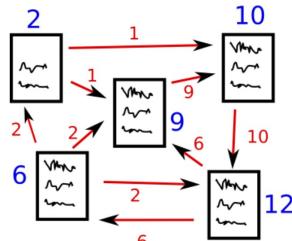
E-reputation management



source : <https://www.mibwebtech.com>

20

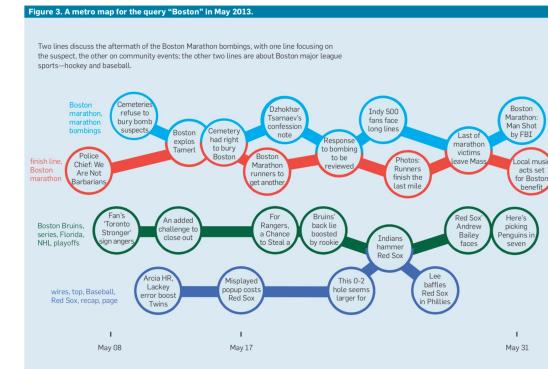
Multi-document summarization



Source: <http://www.scottbot.net/HIAL/>

21

Metromaps (Shahaf et al., 2015)

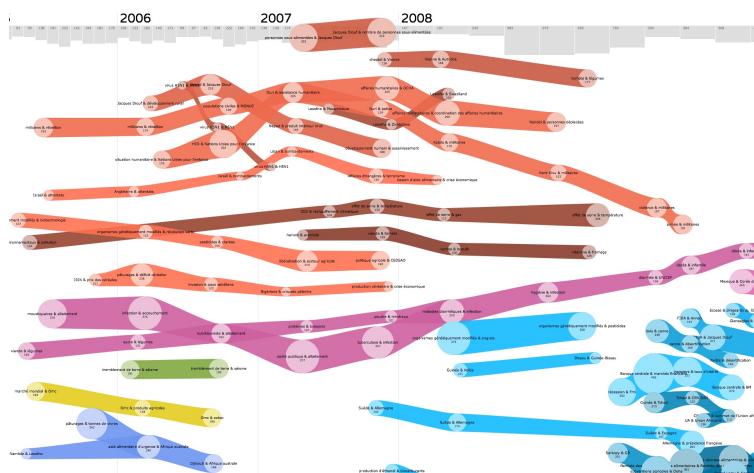


May 08

May 17

May 31

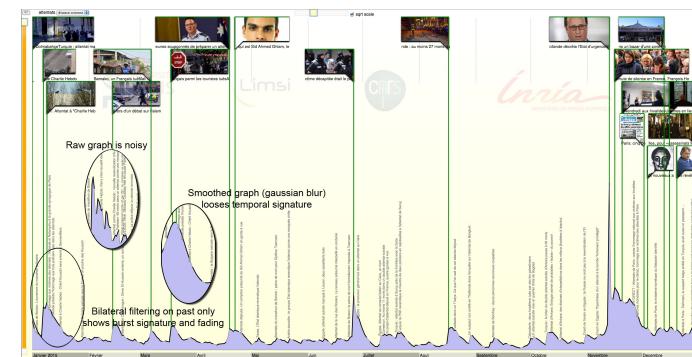
22



Project Pulseweb : <http://pulseweb.context.net>

23

Journalisme de données : exemple du projet Chronolines



Kiem-Hieu Nguyen, Xavier Tannier, Véronique Moriceau. Ranking Multidocument Event Descriptions for Building Thematic Timelines. In *Proceedings of the 30th International Conference on Computational Linguistics (Coling 14)*. Dublin, Ireland, August 2014.

24

5 septembre 2016

LE HUFFINGTON POST
EN ASSOCIATION AVEC LE GROUPE *Le Monde*

Edition: FR ▾ FR

À LA UNE POLITIQUE ÉCONOMIE INTERNATIONAL CULTURE MÉDIAS PEOPLE LE BON LIEN C'EST DEMAIN C'EST LA VIE LE HUFFPLAY

Primaire de la droite • Crise au pouvoir • La preuve en images • Réussir autrement • Alimentation • Sexualité • Déconnecter pour respirer • Ça marche • Vie de bureau

Exemple du projet JADN (2016-2018)

US

EDITION US

THE HUFFINGTON POST
INFORM • INSPIRE • ENTERTAIN • EMPOWER

NEWS POLITICS ENTERTAINMENT WELLNESS WHAT'S WORKING VOICES VIDEO ALL SECTIONS

5 setembro 2016

HUFFPOST BRASIL
ASSOCIADO À Abril

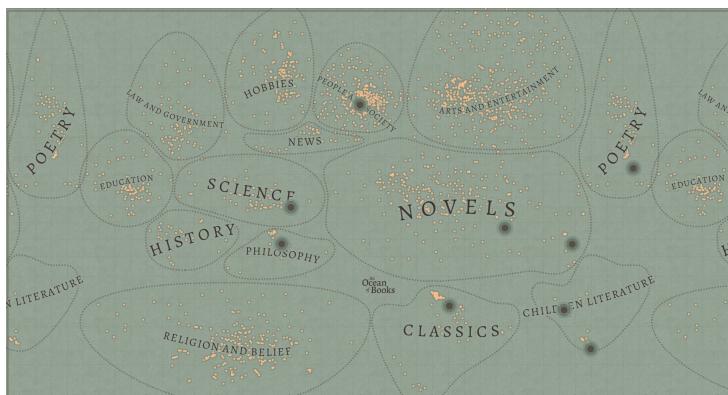
Edição BR ▾ BR

HOME PAÍS MULHERES VOZES DA RUA LGBT RIO 2016 MUNDO EQUILÍBIO LIVROS E HQS VIRAL TEM JEITO!

Familia • Homem Moderno • Animais • Tech • Meio Ambiente • Ciência • Saúde • Educação • Arte • Diversão • Comportamento • Esportes • LGBTfobia • 100% Carioca

Projet LIFRANUM

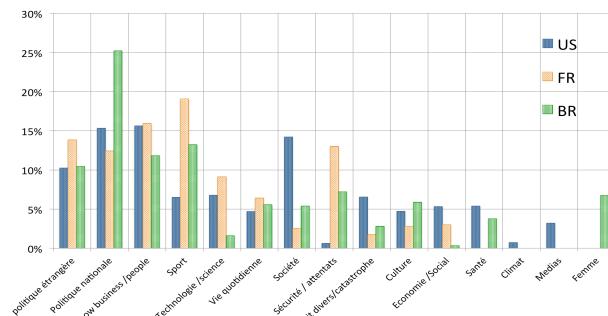
- Projet ANR LIFRANUM MARGE, ERIC, BnF



<https://artsexperiments.withgoogle.com/ocean-of-books>

27

Exemple du projet JADN (suite)



Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post, par J. Velcin, J.C. Soulages, S. Kurpiel, D.L. Otavio, M. Del Vecchio et F. Aubrun. Atelier Journalisme computationnel, adossé à la conférence EGC, Grenoble, 2017.

26

Table 1.1 Categorized NLP applications

Search	Web	Documents	Autocomplete
Editing	Spelling	Grammar	Style
Dialog	Chatbot	Assistant	Scheduling
Writing	Index	Concordance	Table of contents
Email	Spam filter	Classification	Prioritization
Text mining	Summarization	Knowledge extraction	Medical diagnoses
Law	Legal inference	Precedent search	Subpoena classification
News	Event detection	Fact checking	Headline composition
Attribution	Plagiarism detection	Literary forensics	Style coaching
Sentiment analysis	Community morale monitoring	Product review triage	Customer care
Behavior prediction	Finance	Election forecasting	Marketing
Creative writing	Movie scripts	Poetry	Song lyrics

tiré de (Lane et al., 2019)

28

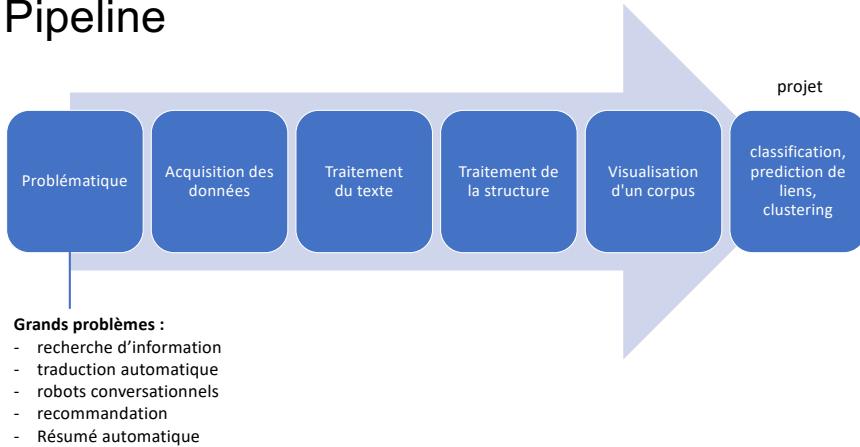
Importance of indexing

- What we're looking for?
 - words (and n-grammes)
 - authors
 - sources
 - categories (predefined? discovered? both?)
- An heterogenous perspective : building/using (latent) spaces that link various objects (eg., documents and authors, topics and words, etc.)

Suggested pipeline

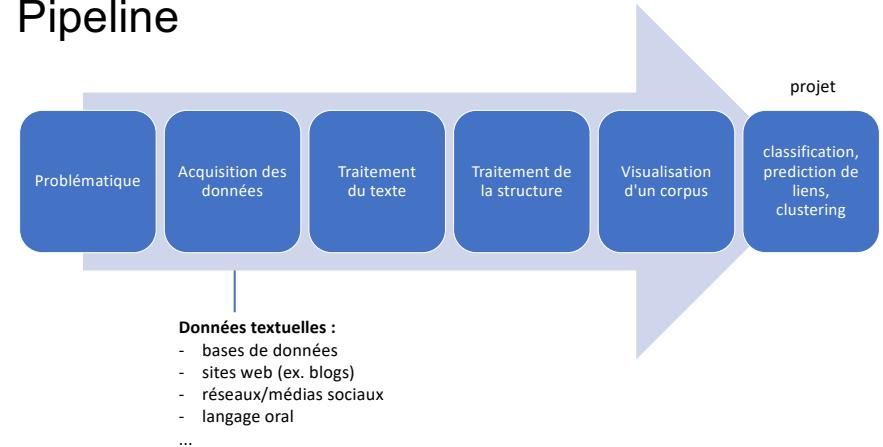
Network analysis for information retrieval, M2 MALIA-MIASHS, Julien Velcin

Pipeline



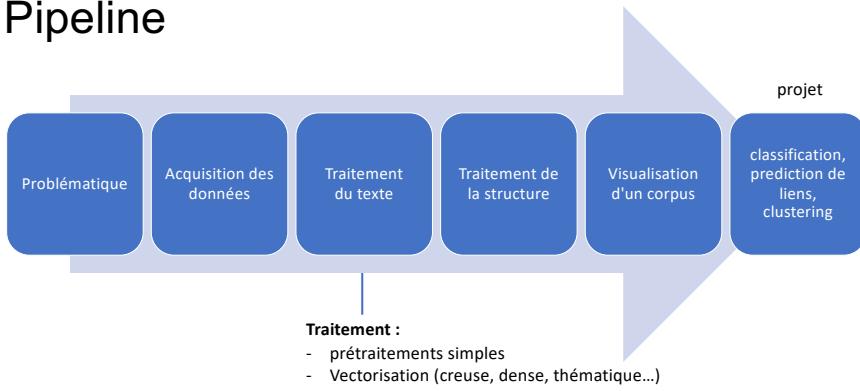
31

Pipeline

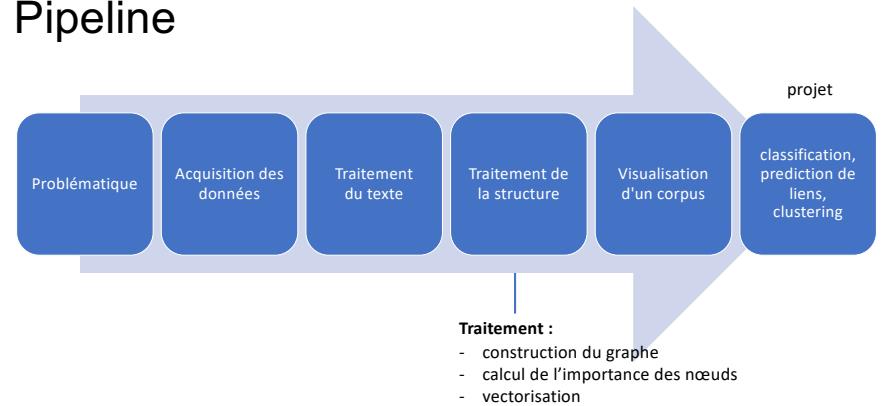


32

Pipeline



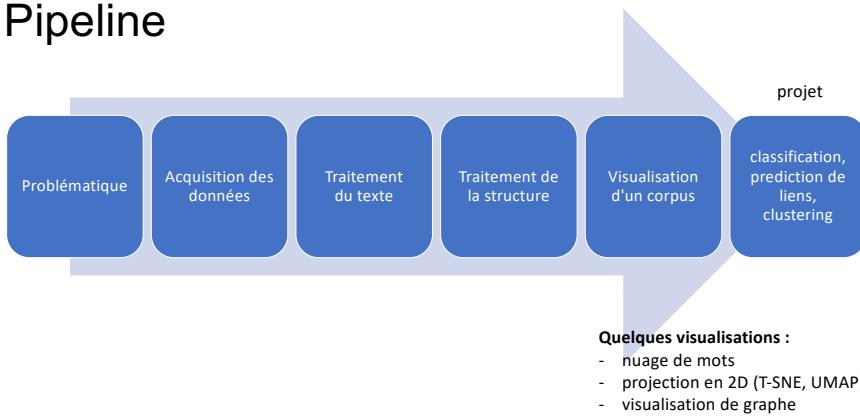
Pipeline



33

34

Pipeline



35