

# Extraction d'Information (Extraction de relations, etc.)

Master DAC, Sorbonne Université

**Xavier Tannier**

[xavier.tannier@sorbonne-universite.fr](mailto:xavier.tannier@sorbonne-universite.fr)

### ACQUISITION

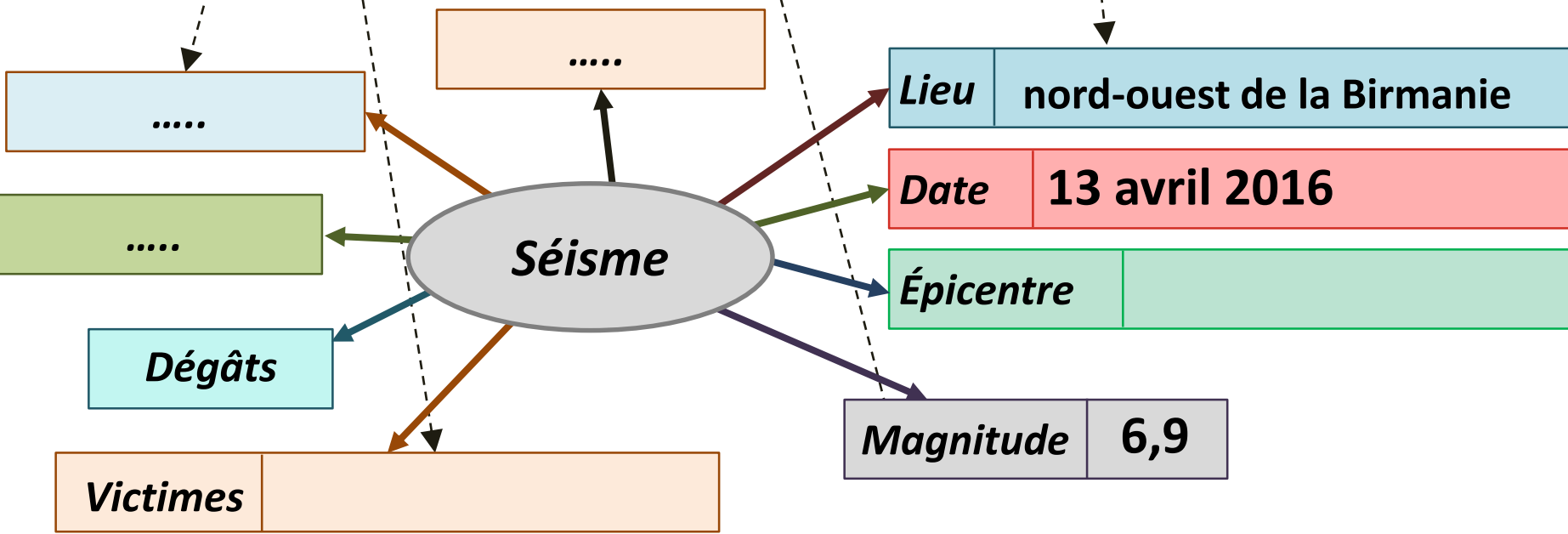
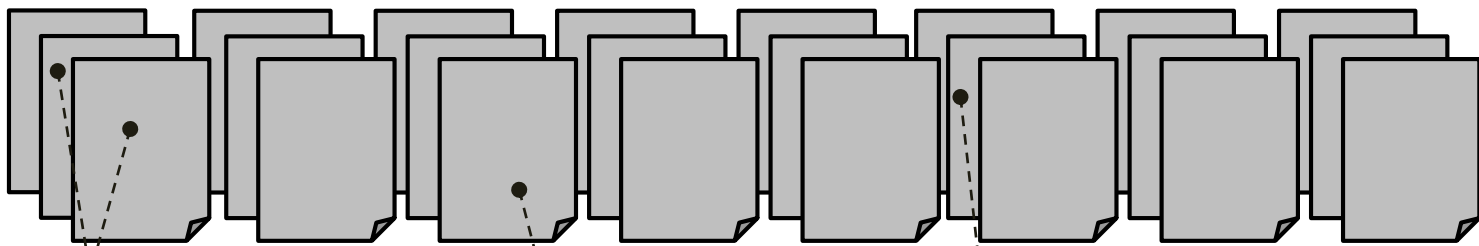
Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

### ACQUISITION

Avant de se faire racheter par Google, YouTube avait déclaré que son modèle économique était basé sur la publicité.

### ACQUISITION

Estée Lauder finalise l'acquisition de Becca



# *Extraction de relations*

- En général, extraction de triplets (*relation, entité1, entité2*)
- Souvent, une relation est entre deux entités nommées
- Permet de créer ou d'enrichir des bases de connaissances
- Les approches :
  - Règles (motifs) écrits « à la main » par des linguistes
  - Apprentissage totalement supervisé
  - Apprentissage semi-supervisé
    - « bootstrapping »
    - supervision distante
    - apprentissage non supervisé sur le web

# *Extraction à base de règles*

ORG1 aux? (annoncer|concrétiser|effectuer|finaliser) (le|l')  
(rachat|acquisition) (d' |de) ORG2  
→ ACQUISITION(ORG1, ORG2)

ACQUISITION



Dassault Systèmes a annoncé le rachat d'Exalead pour environ 135 millions d'euros.

ACQUISITION



Avant de se faire racheter par Google, YouTube avait déclaré que son modèle économique était basé sur la publicité.

ACQUISITION



Estée Lauder finalise l'acquisition de Becca

# *Extraction à base de règles*

- Généralement très précis mais peu couvrant
- Un gros travail d'expert
- Peuvent être conçus pour des domaines très spécifiques ayant peu de données

# *Extraction supervisée*

- Les étapes couramment employées
  1. Trouver les entités nommées
  2. Entre chaque paire d'entités nommées (souvent dans une seule phrase), décider si il existe une relation ou pas (classifieur « NIL vs. RELATION »)
  3. Pour les paires qui ont une relation, trouver laquelle (classifieur RELATION)
- Pourquoi 2 et 3 sont-ils séparés ?
  - Déséquilibre des classes
  - Traits/features différents

# *Extraction de relation : les traits*

Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

E1: Dassault Systèmes

E2: Exalead

H1: Dassault

H2: Exalead

H1-H2: Dassault-Exalead

Les mots et « têtes » des entités, seules et combinées,  
unigrammes et bigrammes



# *Extraction de relation : les traits*

Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

E1: Dassault Systèmes

Before.E1.1: -

E2: Exalead

Before.E1.2: -

H1: Dassault

After.E1.1: annonce

H2: Exalead

After.E1.2: le

H1-H2: Dassault-Exalead

Before.E2.1: d'

Before.E2.2: rachat

After.E2.1: pour

After.E2.2: environ

Les mots avant et après les entités

# *Extraction de relation : les traits*

Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

E1: Dassault Systèmes

E2: Exalead

H1: Dassault

H2: Exalead

H1-H2: Dassault-Exalead

Before.E1.1: -

Before.E1.2: -

After.E1.1: annonce

After.E1.2: le

Before.E2.1: d'

Before.E2.2: rachat

After.E2.1: pour

After.E2.2: environ

Bow: {annonce, le, rachat, d'}

Tous les mots entre les deux entités (sac de mots)

# *Extraction de relation : les traits*

Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

E1: Dassault Systèmes

E1Type:ORG

E2: Exalead

E2Type:ORG

H1: Dassault

E1E2Type:ORG-ORG

H2: Exalead

H1-H2: Dassault-Exalead

Before.E1.1: -

Before.E1.2: -

After.E1.1: annonce

After.E1.2: le

Before.E2.1: d'

Before.E2.2: rachat

After.E2.1: pour

After.E2.2: environ

Bow: {annonce, le, rachat, d'}

Les types des entités

# *Extraction de relation : les traits*

Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

E1: Dassault Systèmes      E1Type:ORG  
E2: Exalead      E2Type:ORG  
H1: Dassault      E1E2Type:ORG-ORG  
H2: Exalead  
H1-H2: Dassault-Exalead  
Before.E1.1: -  
Before.E1.2: -  
After.E1.1: annonce  
After.E1.2: le  
Before.E2.1: d'  
Before.E2.2: rachat  
After.E2.1: pour  
After.E2.2: environ  
Bow: {annonce, le, rachat, d'}

**CHUNKS: VP NP**

Les informations syntaxiques :

- Les « chunks » (constituants) entre les deux entités

# Extraction de relation : les traits

Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

E1: Dassault Systèmes

E1Type:ORG

E2: Exalead

E2Type:ORG

H1: Dassault

E1E2Type:ORG-ORG

H2: Exalead

CHUNKS: VP NP

H1-H2: Dassault-Exalead

Before.E1.1: -

Before.E1.2: -

After.E1.1: annonce

After.E1.2: le

Before.E2.1: d'

Before.E2.2: rachat

After.E2.1: pour

After.E2.2: environ

Bow: {annonce, le, rachat, d'}

DEP: Exalead  $\xleftarrow{\text{prep\_de}}$  rachat  $\xleftarrow{\text{obj}}$  annonce  $\xrightarrow{\text{subj}}$  Dassault

Les informations syntaxiques :

- Les « chunks » (constituants) entre les deux entités
- Le chemin en dépendances d'une entité à l'autre

# *Extraction de relation : les traits*

Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

E1: Dassault Systèmes	E1Type:ORG
E2: Exalead	E2Type:ORG
H1: Dassault	E1E2Type:ORG-ORG
H2: Exalead	CHUNKS: VP NP
H1-H2: Dassault-Exalead	DEP: rachat annonce
Before.E1.1: -	
Before.E1.2: -	
After.E1.1: annonce	
After.E1.2: le	
Before.E2.1: d'	
Before.E2.2: rachat	
After.E2.1: pour	
After.E2.2: environ	
Bow: {annonce, le, rachat, d'}	

TriggerAchat: {rachat, achat, acquisition, ...}

Des informations issues de lexiques

# *Extraction de relation : les traits*

Dassault Systèmes annonce le rachat d'Exalead pour environ 135 millions d'euros.

E1: Dassault Systèmes	E1Type:ORG
E2: Exalead	E2Type:ORG
H1: Dassault	E1E2Type:ORG-ORG
H2: Exalead	CHUNKS: VP NP
H1-H2: Dassault-Exalead	DEP: rachat annonce
Before.E1.1: -	TriggerAchat: {rachat, achat, acquisition, ...}
Before.E1.2: -	
After.E1.1: annonce	
After.E1.2: le	
Before.E2.1: d'	
Before.E2.2: rachat	
After.E2.1: pour	
After.E2.2: environ	
Bow: {annonce, le, rachat, d'}	

Et tout autre trait qui vous plaira !

# *Et le classifieur ?*

- Tous les classifieurs peuvent faire l'affaire, avec leurs avantages et leurs inconvénients :
  - SVM
  - MaxEnt
  - Naive Bayes
  - Arbres de décision
  - Réseaux de neurones
  - ...



# **Classification de texte**

# Classification automatique

Quel thème ?  
(international,  
politique, sports,  
people, sciences,  
économie, ...)

## Fillon-Juppé: quel est leur programme en éducation ?

Par **Le Figaro Etudiant** • Publié le 21/11/2016 à 16:43 • Mis à jour le 21/11/2016 à 16:44



LE FIGARO PREMIUM  
> 1 mois d'essai offert



François Fillon affrontera Alain Juppé au deuxième tour de la primaire de la droite et du centre. Hausse de la rémunération des enseignants, réforme du bac, sélection à l'université : comparaison de leurs programmes sur l'éducation.

Pour Alain Juppé, la réforme de l'éducation est la « mère de toutes les réformes ». Son rival au second tour de [la primaire de la droite et du centre](#), François Fillon, veut lui « faire à nouveau de l'école la première marche de l'unité républicaine ». Alors que les deux prétendants au rôle de candidat de la

# *Classification automatique*

Spam ou pas spam ?

WITH DUE RESPECT




Spam x



**Zumba Kabale** <zumbakabale0003@gmail.com>

15 sept.



À cci : moi 

Dear Friend,

I know that this mail will come to you as a surprise as we have never met before, but need not to worry as I am contacting you independently of my investigation and no one is informed of this communication. I need your urgent assistance in transferring the sum of \$11.3million immediately to your private account. The money has been here in our Bank lying dormant for years now without anybody coming for the claim of it.

I want you to corporate with me for the release of this money into your private bank account as the relative to our deceased customer (the account owner) who died a long ago and with her supposed NEXT OF KIN since 31 January 2000. The Banking law here does not allow such money to stay more than 15 years, because the money will be recalled to the Bank treasury account as unclaimed fund.

By indicating your interest I will send you the full details on how the business will be executed.

Best Regqrds  
Mr.Zumba Kabale

# Classification automatique

*“Une arnaque!!!”*

Content  
ou pas content ?

J'avais réservé une chambre double en demandant des lits séparés car partageais la chambre avec la collègue. Notre demande avait été confirmée par la réception de l'hôtel quelques jours avant notre arrivée. Le jour J: ascenseur dont la tapisserie se décolle, chambre avec lit king seize et baignoire trônant au milieu de la pièce...notre demande n'avait finalement pas été prise en compte. Nous avons demandé à être changées de chambre et avons tout simplement fini... À la cave! Chambre accessible par l'extérieur, odeur de moisi, bouteille de jus de fruit périmée dans le frigidaire, finitions douteuse: moisissure sur la poignée de porte des toilettes, saleté dans la baignoire et la cerise sur le gâteau: impossible de fermer la porte de douche car cette dernière cogne dans le pommeau!!la porte de douche a visiblement été montée avant la colonne de douche! Et le pommeau qui se trouve à hauteur de les chevilles, le cordon de douche n'est pas suffisamment long pour le permettre d'atteindre mes cheveux! Deuxième demande auprès de la réception afin d'être changées de chambre, à plus de 200€ la nuit, nous attendons un minimum de propreté et de fonctionnalité! Nous arrivons finalement dans une troisième chambre, correspondant davantage à nos attentes, mais toujours: baignoire sale (morceaux de peau et de cheveux, beurk), table de nuit abîmée. En clair: un nom, une déco surchargée qui masque des finitions inexistantes et hôtel qui ne bénéficie pas du tout du standing qu'il veut se donner.

# *Classification de texte*

- Classification de textes :
  - Attribution de thème, de genre
  - Détection de spam
  - Analyse d'opinion, de sentiment
  - Identification des auteurs
  - Identification de la langue d'un texte
  - ...