



# Introduction au TAL (NLP)

# En quoi consiste le traitement automatisé du langage (en anglais Natural Language Processing) ?

# Répondre à une question formulée en langage naturel

- Le programme d'IA Watson d'IBM a gagné le jeu Jeopardy aux USA en février 2011!

**“Voyage dans la Valachie et la Moldavie” de William Wilkinson a inspiré à cet auteur son plus célèbre roman**



Bram Stoker

# Extraire des informations

Sujet: **réunion programme d'études**

Date: 29/07/2020

À: Paul Bismuth

Réunion : Pr  
d'études

Date : 30

Début : 10

Fin : 11

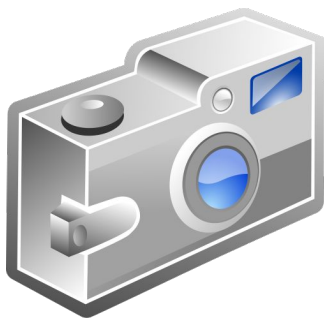
Lieu : Sal

Salut Paul, la réunion a été programmée demain. Ce sera en salle 159 de 10h à 11h.

Pauline

Créer la réunion dans l'agenda

# Extraire des informations et analyser les avis



Caractéristiques :

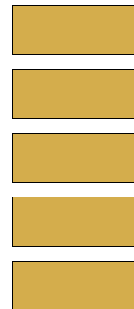
zoom

prix

taille et poids

flash

simplicité



## Taille et poids

- ✓ compact et agréable à porter !
- ✓ puisque l'appareil est petit et léger, je n'ai pas eu de problème à transporter ces appareils professionnels
- ✗ l'appareil photo semble fragile, il est en plastique et très léger



# Traduction automatique

- Entirement automatique
- Assistance aux traducteurs

Texte en entrée :

这不过是一个时间的问题。

Traduction proposée par Stanford's *Phrasal* :

This is only a matter of time.

Enter Source Text:

تعرض الرئيس اللبناني اميل لحود لـ حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت الي " محاكمة " لـ رئيس الجمهورية علي موقف +ه من المحكمة الدولية و " الملاحظات " التي ادلي بـ +ها حول هذا الموضوع .

Translate Clear

Enter Translation:

lebanese

- president
- suffered
- exposed
- president emile
- before
- presented
- offer

Done!

# Etat de l'art

## Mature

### Identification de spam

Let's go to Agra!



Buy V1AGRA ...



### Étiquetage grammatical

Pronom Verbe Article Nom Adjectif

Je suis un génie méconnu.

### Reconnaissance des noms propres

PERSONNE

ORGANISME

Einstein a rencontré les officiels de l'ONU à  
New York

VILLE

## En cours de progrès

### Analyse d'avis

Les meilleures quenelles de Lyon !



Le serveur nous a ignorés pendant 20 min



### Résolution de coréférence

Macron a dit à Trump qu'il devrait démissionner

### Identification des homonymes

Il faut que je change la pile de ma *souris*.



### Analyse syntaxique

Je peux voir la Tour Eiffel de la fenêtre !

### Traduction automatique

第13届上海国际电影节开幕...



Le 13e festival international du film de Shanghai...

### Extraction d'informations

Vous êtes invité à déjeuner le  
27 mai à 12h30



créer

Déjeuner  
27 mai

## Toujours difficile

### Réponse à une question en langage naturel

Q. Qui peut sauter plus haut que la Tour Eiffel ?

### Paraphrase

XYZ a racheté ABC hier

ABC a été vendu à XYZ

### Résumé

Pas de pluie prévue

Ciel dégagé

Températures douces



Le temps est agréable

### Dialogue

Où passe le dernier Disney ce soir ?



Cinéville à 20h. Voulez-vous réserver une place ?



# Ambiguity makes NLP hard: “Crash blossoms”



Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

# Ambiguity is pervasive

*New York Times* headline (17 May 2000)

Fed raises interest rates

Fed raises interest rates

Fed raises interest rates 0.5%




# In-video quizzes!

- Most lectures will include a little quiz
  - Just to check basic understanding
  - Simple, multiple-choice.
  - You can retake them if you get them wrong

# Pour quelles autres raisons la compréhension du langage naturel est-elle difficile ?

## langage SMS et non standard

Troooooop contente du bak  
@copinesdeterm enf1 en vac' 

## segmentation

the New York-New Haven Railroad  
the New York-New Haven Railroad

## expressions idiomatiques

casser les pieds  
perdre la face  
monter sur ses grands chevaux

## néologismes

retweeter  
franglais  
docufiction

## contexte

Marie et Jeanne sont soeurs.  
Marie et Jeanne sont mères.

## noms propres trompeurs

*Les Misérables* ont été écrits  
par Victor Hugo.

Mais c'est ce qui la rend passionnante !

# Pour avancer sur ce problème...

- La tâche est ardue ! De quels outils avons-nous besoin ?
  - Connaissances sur le langage
  - Connaissances sur le contexte
  - Un moyen de combiner les sources de connaissances
- Manière habituelle de procéder :
  - des modèles probabilistes construits à partir de données linguistiques
    - $P(\text{"maison"} \rightarrow \text{"house"})$  élevée
    - $P(\text{"L'avocat général"} \rightarrow \text{"the general avocado"})$  basse
  - Par chance, les fonctions de texte brut font souvent la moitié du travail

# This class

- Teaches key theory and methods for statistical NLP:
  - Viterbi
  - Naïve Bayes, Maxent classifiers
  - N-gram language modeling
  - Statistical Parsing
  - Inverted index, tf-idf, vector models of meaning
- For practical, robust real-world applications
  - Information extraction
  - Spelling correction
  - Information retrieval
  - Sentiment analysis

# Skills you'll need

- Simple linear algebra (vectors, matrices)
- Basic probability theory
- Java or Python programming
  - Weekly programming assignments

# **Introduction to NLP**

What is Natural  
Language Processing?