

NLP

Natural Language Processing



Céline, Pereg, Ludivine, Ronan

Sommaire

1

Définitions

Data Mining, Text Mining, NLP,
Documents structurés et non structurés

2

Text Mining: Processus et exemples

Apprentissage supervisé, non supervisé,
recherche et extraction de l'information...

3

BOW, réduction de la dimensionnalité

Bag of Words
Stopwords, lemmatization, stemming

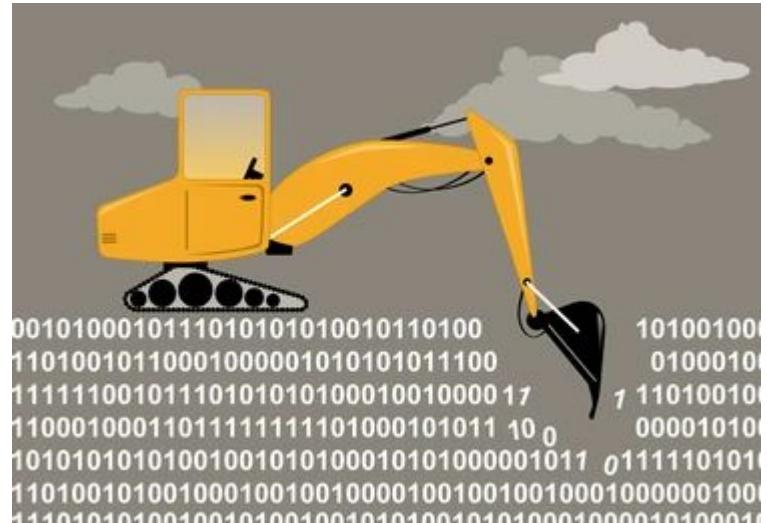
A large yellow geometric shape, resembling a stylized 'L' or a corner, occupies the left side of the slide. It has a diagonal cut across its top-left corner.

1. Définitions

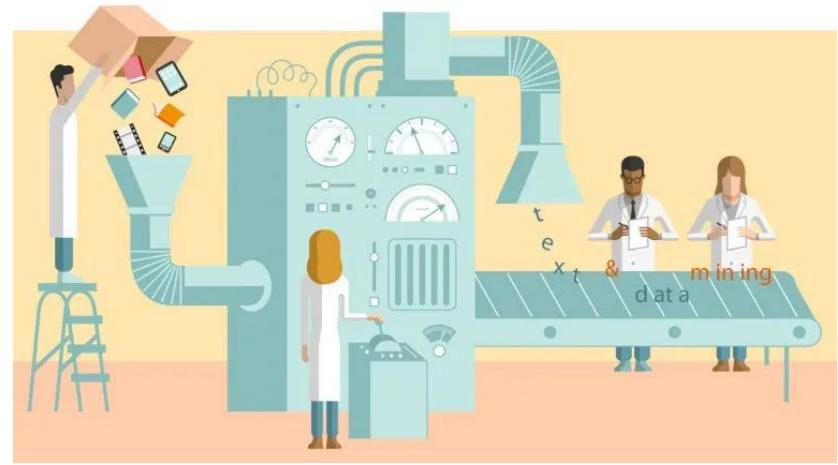
Data Mining, Text Mining, NLP,
Documents structurés et non
structurés

Data Mining

Le Data Mining est le processus de fouille de données. Cela permet d'établir des relations entre les données ou en repérant des patterns.



Text Mining

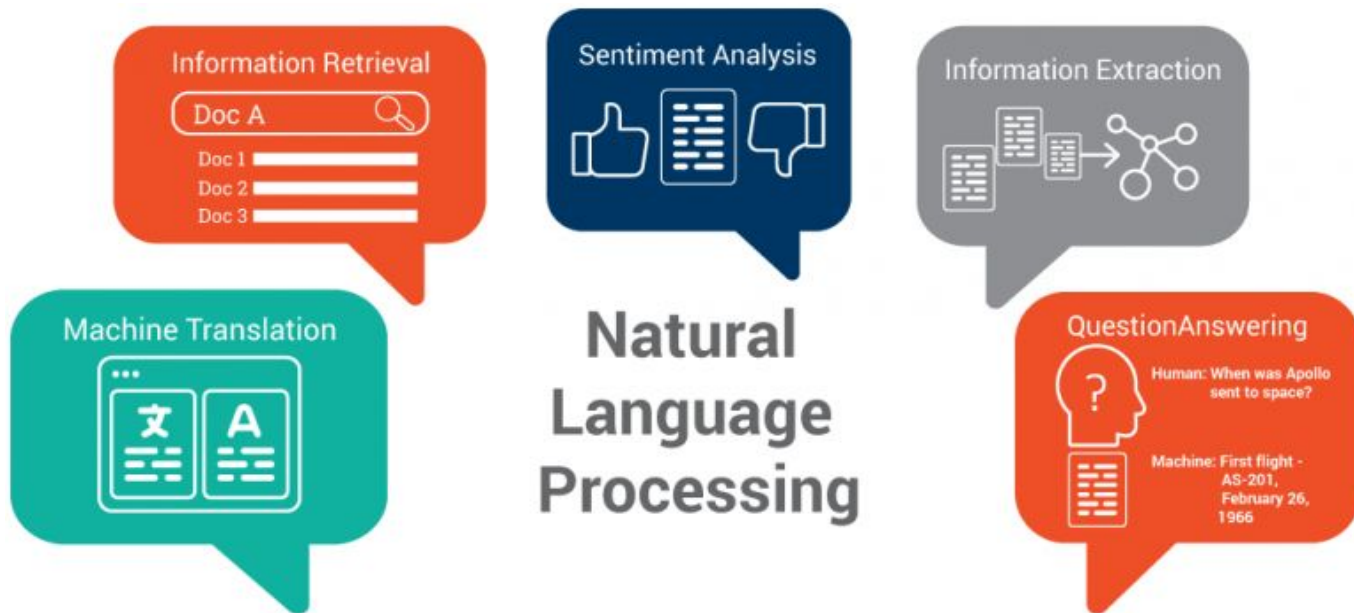
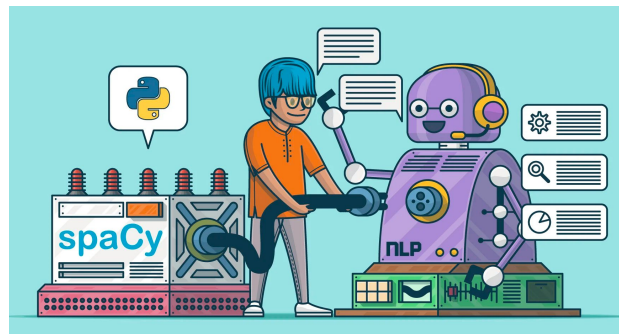


Le Text Mining est un ensemble de méthodes, de techniques et d'outils pour exploiter les documents non structurés que sont les textes écrits, comme les fichiers bureautiques de type word ,les emails, les documents de présentation de type powerpoint ...

le text mining s'appuie sur des techniques d'analyse linguistique. Le text mining est utilisé pour classer des documents, réaliser des résumés de synthèse automatique ou encore pour assister la veille stratégique ou technologique selon des pistes de recherches prédéfinies.

NLP

Natural Language Processing : désigne l'ensemble des tâches permettant à un ordinateur de traiter des données en langage humain.



Documents structurés



Un document structuré est un document électronique dans lequel une méthode de balisage est utilisée pour identifier l'ensemble et les parties du document comme ayant différentes significations au-delà de leur mise en forme.

Exemple : le format XML

Documents non structurés

ESG & NLP

Extraction de données
extra-financières dans des
documents non structurés



WEEFIN

le terme **information non-structurée** décrit les documents binaires (ex. : documents . pdf et . docx) qui sont ajoutés à l'aide d'applications propriétaires telles que Acrobat ou Word.

A large yellow geometric shape, resembling a stylized arrow or a corner, occupies the left side of the slide. It has a diagonal edge separating it from the white background.

2. Text Mining: Processus et exemples

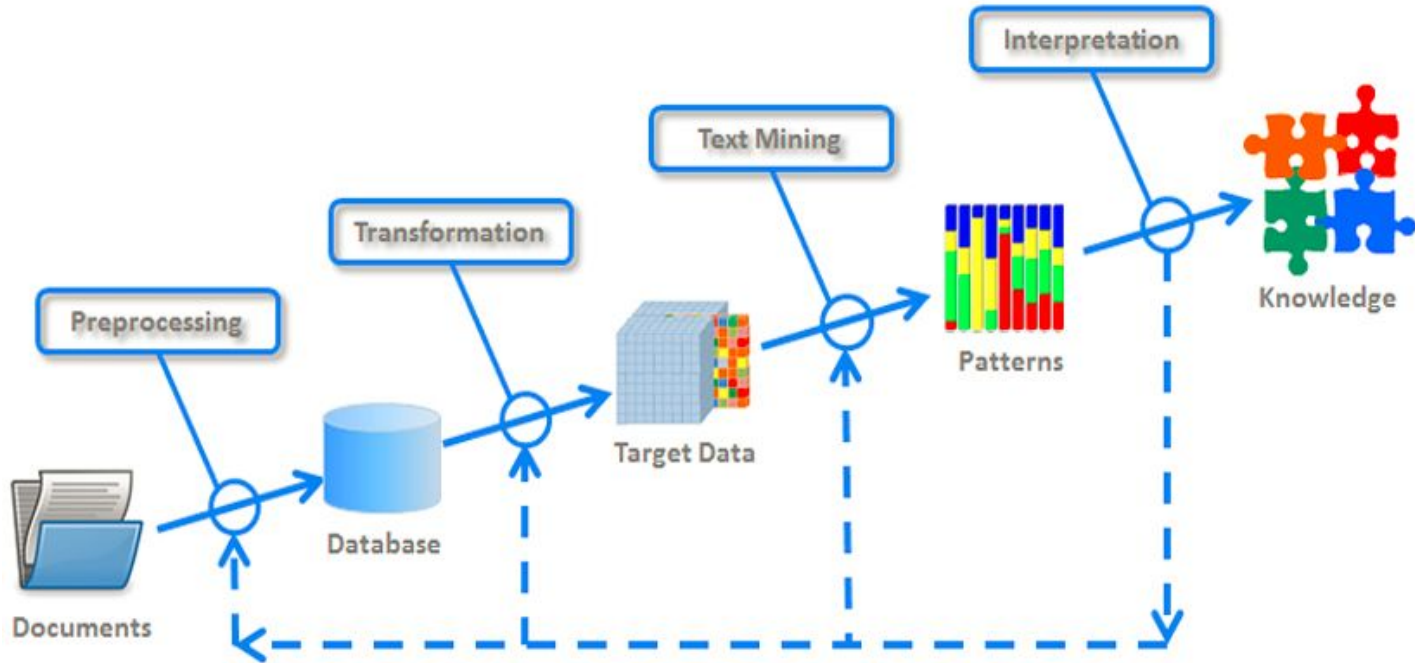
Apprentissage supervisé, non
supervisé,
recherche et extraction de
l'information...

Processus



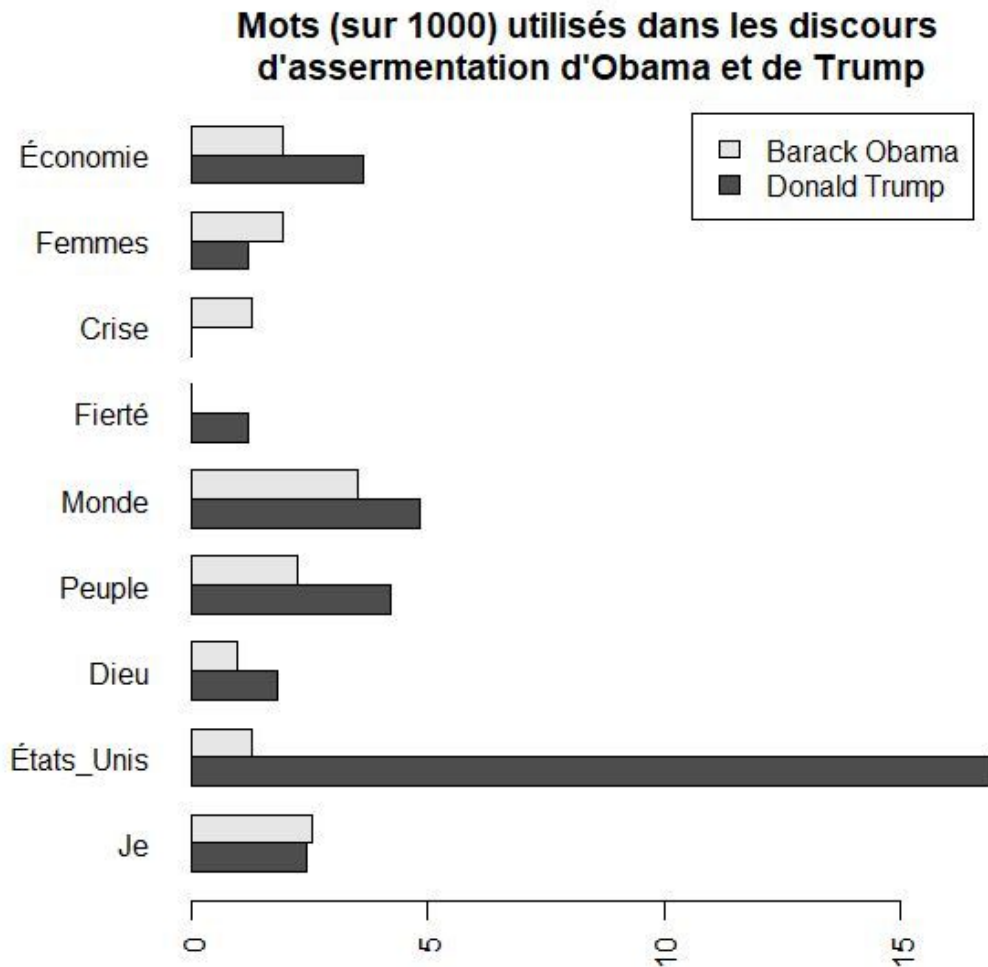
Le **processus** de **Text Mining** consiste à analyser des ensembles de documents textuels afin de capturer les concepts et thèmes-clés, et de découvrir les relations et les tendances cachées. Il ne nécessite pas que vous connaissiez les mots ou les termes précis utilisés par les auteurs pour exprimer ces concepts.

Processus



Processus

Analyse de la fréquences d'une série de mots d'un dictionnaire dans deux textes. Ici, le discours inaugural et de Trump avec celui d'Obama.



Exemple

File View Mode

Document PMID 10861484

Show PubMed Entry

Cyclophosphamide enhances anti - tumor effect of wild - type p53 - specific CTL .

Abstract The tumor suppressor protein p53 is overexpressed in up to 50 % of all human malignancies , both in solid tumors as well as hematological malignancies , and is therefore an attractive target for immunotherapy . We have recently shown that cytotoxic T lymphocytes (CTL) , raised in p53 gene deficient (p53 - / -) mice and recognizing a murine wild - type (wt) p53 peptide , were able to eradicate a mutant p53 - induced and overexpressing tumor in p53 + / + nude mice . These CTL also prevented the outgrowth of a more aggressive p53 - overexpressing tumor in immunocompetent C57BL / 6 mice . Importantly , this occurred in the absence of demonstrable damage to normal tissue . Possibly due to the aggressive nature of the latter tumor , adoptive transfer of wtp53 - specific CTL did not result in the eradication of established tumors , either in nude or immunocompetent mice . Therefore , we explored whether the cytotoxic drug cyclophosphamide (CY) could potentiate the therapeutic activity of wtp53 - specific CTL . We show here that CY acts synergistically with adoptively transferred wtp53 - specific CTL in controlling the growth of an aggressive mutant p53 - induced and overexpressing tumor . Previously described mechanisms underlying the synergism between CY and immune T cells were evaluated , but were not found to be operational in this model .

Adjuvants ; Immunologic ; Antineoplastic Agents ; Antineoplastic Agents , Alkylating ; Interferon - alpha ; Tumor Suppressor Protein p53 ; perfoslamide ; Cyclophosphamide ; Interferon - beta ; ras Proteins ; Adjuvants , Immunologic ; pharmacology ; Adoptive Transfer ; Animals ; Antineoplastic Agents ; pharmacology ; Antineoplastic Agents , Alkylating ; metabolism ; pharmacology ; Cell Line ; Cyclophosphamide ; analogs & derivatives ; metabolism pharmacology ; Dose - Response Relationship , Drug ; Immunohistochemistry ; Immunotherapy , Adoptive ; Interferon - alpha ; pharmacology ; Interferon - beta ; pharmacology ; Mice ; Mice , Inbred C57BL ; Mice , Nude ; Neoplasm Transplantation ; Neoplasms , Experimental ; genetics ; therapy ; T - Lymphocytes , Cytotoxic ; immunology ; Time Factors ; Transfection ; Tumor Cells , Cultured ; Tumor Suppressor Protein p53 ; immunology ; ras Proteins ; metabolism ;

Annotation

Concepts Interactions

Reload Export Help

Conf	Type 1	Name 1	Type 2	Name 2					
0.08	chem	Cyclophosphamide	disease	Neoplasms					
0.08	chem	Cyclophosphamide	gene	TRP53					
0.06	disease	Neoplasms	gene	TRP53					
0.05	chem	Cyclophosphamide	gene	TP53					
0.04	chem	Cyclophosphamide	gene	IFNB1					
0.04	disease	Neoplasms	gene	TP53					
0.04	disease	Neoplasms	gene	IFNB1					
0.03	chem	Cyclophosphamide	gene	P53					

Documentation as PDF :: Contact: odini@ontogene.org :: For project information visit www.ontogene.org :: Logged in as: FR

A large yellow geometric shape, resembling a stylized 'L' or a corner, occupies the left side of the slide. It has a diagonal cut across its top-right corner.

3. Bag of Words

Bag of Words

Bag of Words

Document	the	cat	sat	in	hat	with
<i>the cat sat</i>	1	1	1	0	0	0
<i>the cat sat in the hat</i>	2	1	1	1	1	0
<i>the cat with the hat</i>	2	1	0	0	1	1

Utilisé en NLP pour des tâches comme la classification de texte

Bag of Words : en python (avec Keras)

```
from keras.preprocessing.text import Tokenizer
```

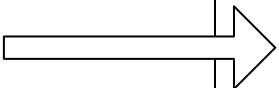
```
docs = [  
    'the cat sat',  
    'the cat sat in the hat',  
    'the cat with the hat',  
]
```

```
## Step 1: Determine the Vocabulary
```

```
tokenizer = Tokenizer()  
tokenizer.fit_on_texts(docs)  
print(f'Vocabulary: {list(tokenizer.word_index.keys())}')
```

```
## Step 2: Count
```

```
vectors = tokenizer.texts_to_matrix(docs, mode='count')  
print(vectors)
```



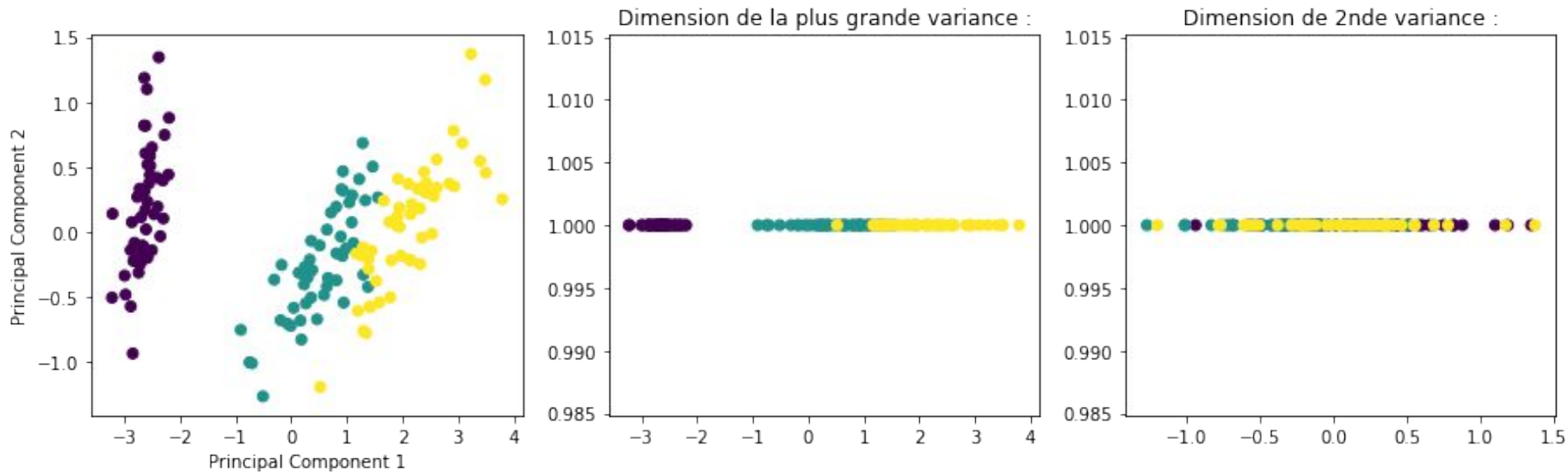
Vocabulary: ['the', 'cat', 'sat', 'hat', 'in', 'with']
[[0. 1. 1. 1. 0. 0. 0.]
 [0. 2. 1. 1. 1. 1. 0.]
 [0. 2. 1. 0. 1. 0. 1.]]

A large yellow geometric shape, resembling a stylized 'L' or a corner, occupies the left side of the slide. It has a diagonal edge separating it from the white background.

4. Réduction de la dimensionnalité

Retrait des stopwords,
lemmatization, stemming

Réduction de la dimensionnalité



Retrait des stop words

le traitement automatique du langage naturel vise à créer
des outils de traitement de la langue naturelle pour
diverses applications .

traitement automatique langage naturel vise créer outils traitement
langue naturelle diverses applications

La Lemmatization



La **lemmatisation** désigne un traitement lexical apporté à un texte en vue de son analyse. Ce traitement consiste à appliquer aux occurrences des lexèmes sujets à flexion (en français, verbes, substantifs, adjectifs) un codage renvoyant à leur entrée lexicale commune (« forme canonique » enregistrée dans les dictionnaires de la langue, le plus couramment), que l'on désigne sous le terme de *lemme*.

Les lexèmes (lemmes) d'une langue connaissent éventuellement plusieurs formes en fonction de leur genre (masculin ou féminin), leur nombre (un ou plusieurs), leur personne (moi, toi, eux...), leur mode (indicatif, impératif...). On rencontre ainsi plusieurs formes pour un même lemme. On désigne ces formes comme des flexions, ou formes fléchies.

FIN