



# Panorama des méthodologies de fouille de textes et interdisciplinarité



ANF TDM 2024

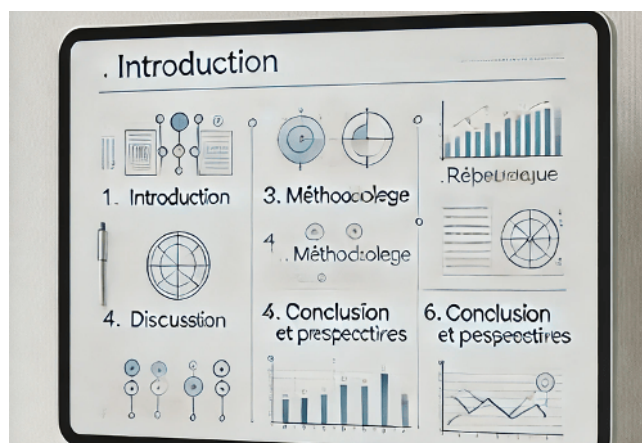
Adrian-Gabriel CHIFU ([adrian.chifu@univ-amu.fr](mailto:adrian.chifu@univ-amu.fr)) - LIS/AMU

03 octobre 2024

1

## Plan

- Fouille de données
- Fouille de textes
- Outils
- Discussion
- LLMs (IA générative)
- Conclusions



"une image pour le plan d'une présentation" (générée par ChatGPT 4o)



# FOUILLE DE DONNÉES

QU'EST-CE QUE LA FOUILLE DE DONNÉES ?



Aix-Marseille université  
Sociétalement engagée



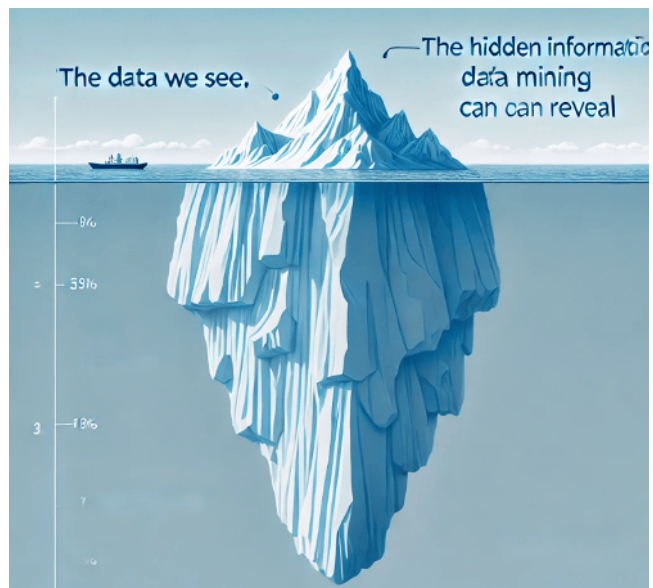
PANORAMA DES MÉTHODOLOGIES DE FOUILLE DE TEXTES ET INTERDISCIPLINARITÉ

3

3

## Fouille de données ?

- Exploration et analyse de grande échelle des données
  - Découvrir des modèles et des tendances significatives.
  - Exploration de données ou Data mining.
- Cruciale pour la découverte de tendances et modèles significatifs
  - Énormément de données produites chaque jour ⇒ la fouille de données est essentielle pour extraire des informations précieuses.
- Utilisée dans le marketing, la détection de fraude, la recherche médicale, etc.



"Un iceberg avec la partie visible représentant "les données que nous voyons" et la partie immergée représentant "les informations cachées que la fouille de données peut révéler"" (générée par ChatGPT 4o)



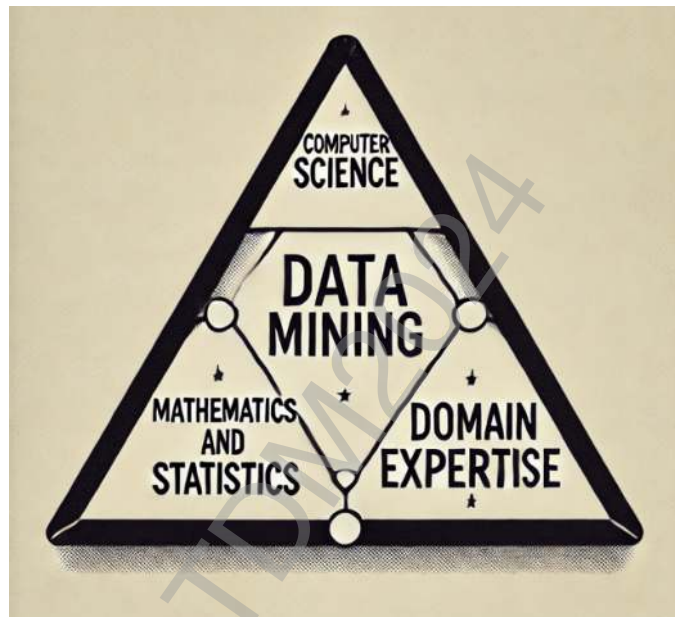
PANORAMA DES MÉTHODOLOGIES DE FOUILLE DE TEXTES ET INTERDISCIPLINARITÉ

4

4

## Fouille de données ▲

- À l'intersection :
  - Informatique : Algorithmes, structures de données
  - Mathématiques : Probabilité, statistiques
  - Application métier : bibliométrie, biologie, chimie, etc.



PANORAMA DES MÉTHODOLOGIES DE FOUILLE DE TEXTES ET INTERDISCIPLINARITÉ

5



5

## FOUILLE DE TEXTES

CONTEXTE, PROBLÉMATIQUES,  
INTERDISCIPLINARITÉ, SCHÉMA



Aix-Marseille  
université  
Sociétalement engagée



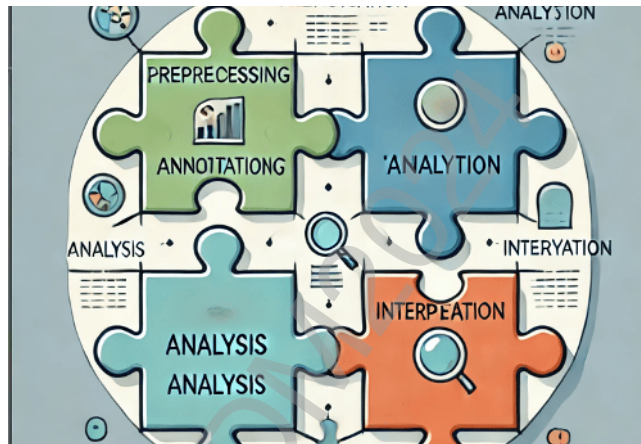
PANORAMA DES MÉTHODOLOGIES DE FOUILLE DE TEXTES ET INTERDISCIPLINARITÉ

6

6

## Fouille de textes ?

- Extraction d'informations de haute qualité à partir du texte
  - Prétraitement, l'annotation, l'analyse et l'interprétation du texte.
- Cruciale pour la prise de décisions basées sur des données massives et non structurées
  - Explosion des données textuelles non structurées, la fouille de texte est devenue une technique clé pour extraire des connaissances précieuses et pour aider à la prise de décisions.
- Applications : Tendances des médias sociaux, lutte contre la fraude, amélioration du service client, recherche médicale, etc.



"Un puzzle avec différents morceaux représentant les divers aspects du processus de fouille de texte, tels que le prétraitement, l'annotation, l'analyse et l'interprétation." (générée par StarV&T Color)



## Fouille de textes : problématiques

- Prétraitement : Nettoyage & Normalisation
- Gestion de la diversité linguistique : Sémantique & Syntaxe
- Interprétation du contenu non structuré
- Analyse des sentiments et subjectivité
- Extraire des informations pertinentes dans une masse de données
- L'accès et le stockage de grandes quantités de données



"Des points d'exclamation représentant chacune des problématiques sur un arrière-plan de texte indistinct pour illustrer le défi de trouver des informations pertinentes dans une grande masse de données textuelles." (générée par Bing.com/40)



## Fouille de textes : le processus (simplifié)

1. **Collecte de Textes** : Acquisition de données textuelles à partir de diverses sources
2. **Prétraitement** : Nettoyage, normalisation... transformation du texte pour faciliter l'analyse
3. **Extraction de caractéristiques** : Sélection des éléments importants pour la fouille de texte
4. **Modélisation et Analyse** : Utilisation de techniques (symboliques, statistiques, etc.) pour identifier patterns
5. **Interprétation & Evaluation** : Résultats obtenus sont interprétés, évalués et utilisés pour prendre des décisions

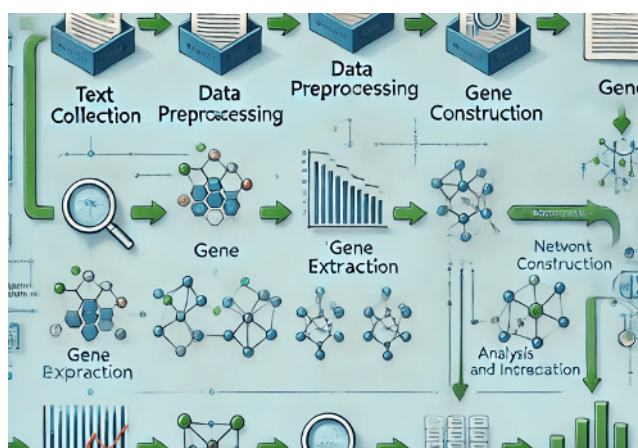


"Un diagramme ou un schéma linéaire qui illustre les étapes du processus de fouille de textes, de la collecte de textes à l'interprétation et à l'évaluation." (générée par [bing.com](#) et optimisée par DALL-E 3)



## Exemple (domaine Bio-médical)

1. **Collecte de Textes** : Extraction d'articles de recherche sur le cancer du sein à partir de la base de données PubMed.
2. **Prétraitement** : Nettoyage des articles pour supprimer les informations inutiles, normalisation du texte en utilisant le Medical Subject Headings (MeSH).
3. **Extraction de caractéristiques** : Identification des noms de gènes, de protéines, de médicaments avec le logiciel d'annotation bio-médicale, PubTator.
4. **Modélisation et Analyse** : Construction de réseaux d'interactions entre les gènes identifiés avec le logiciel Cytoscape.
5. **Interprétation & Evaluation** : Les réseaux sont analysés pour découvrir de nouvelles relations entre les gènes du cancer du sein pour la recherche sur de nouvelles thérapies.



"Schéma illustrant un article de recherche bio-médicale par les étapes de la collecte de textes à l'analyse de réseaux." (générée par [bing.com](#) et optimisée par DALL-E 3)



## OUTILS PAR ÉTAPE

CORPUS, REPRÉSENTATION,  
NORMALISATION, EXPLORATION &  
VISUALISATION, MODÈLES



Aix-Marseille  
université  
Sociétalement engagée



PANORAMA DES MÉTHODOLOGIES DE FOUILLE DE TEXTES ET INTERDISCIPLINARITÉ

11

11

### Corpus ?

- Un ensemble organisé de textes utilisé pour extraire des informations dans le contexte de la fouille de textes.
- Peut être spécialisé (centré sur un sujet ou une source spécifique) ou généralisé (couvrant une variété de sujets ou de sources).
- Précieux pour identifier des motifs, des tendances ou des anomalies dans les données textuelles.
- En fouille de textes, permet l'analyse de gros volumes de données textuelles de manière automatisée pour extraire des informations pertinentes.
- Aide à découvrir ou à vérifier des hypothèses sur la manière dont les informations sont exprimées dans le texte, à comprendre le contexte et le sentiment de l'information, et à construire des modèles d'information pour la prise de décision.



PANORAMA DES MÉTHODOLOGIES DE FOUILLE DE TEXTES ET INTERDISCIPLINARITÉ

12

12

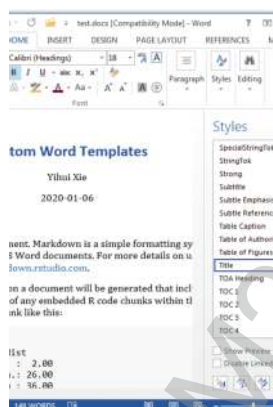
## Document textuel ?

m ipsum dolor sit amet, consectetur adipiscing elit. Sed vehicula, ex a eleifend sodales, massa lorem inter quam, sit amet tincidunt nisi neque vitae nunc. Aenean venenatis scelerisque eros, in molestie arcu euismod in. Donec lacinia imperdiet risus sed aliquam. Cras ac malesuada odio. Etiam dui dui, gravida in vestibulum id, cursus vitae lacus. Maecenas eget ligula semper ligula pretium imperdiet non sed erat. Cras interdum tellus in massa accumsan posuere. Fusce laoreet enim element purus venenatis maximus. Nulla rhon accumsan ligula vitae facilisis. Se sodales, dui venenatis accumsan varius, justo ipsum egestas leo, a placerat orci lorem eu sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Donec posuere faucibus metus ac tempor.

**Texte Brut** : Suite de caractères sans formatage, souvent en .txt

de la grande statue de Zeus, à Olympie, entre le Soleil (Hélios) et les autres Dieux. Entre le Soleil (Hélios) et les autres Dieux, groupés deux à deux, s'ordonnaient. Au centre de la frise, en surmontant, se trouve une figure masculine (masculine) qui préside aux unions de huit couples divins, il en est un, Hestia. Pourquoi les appeler ? Rien de la légende qui puisse justifier cette association (comme Zeus-Héra, Poséidon-Amphitrite et sa sœur (comme Apollon-Artémis, Héra et Aphrodite-Eros), ni protectrice et sœur). Quel lien unissait donc, dans l'esprit de l'artiste, les deux personnages qui semblent étrangers l'un à l'autre ? L'artiste ancien est tenu de se conformer à ce qu'exerce dans le cadre des schèmes imposés par une déesse mais aussi non se prêtait moins que les autres divinités anthropomorphes. On la voit rarement figurer, comme Phidias l'avait sculptée, dans l'art plastique, l'association

**Image** : Texte scanné ou capture d'écran



**Formaté** : Avec balises de structure et formatage (ex. Word, PDF, rendu HTML)

```
<menu>
<area text="Welcome" file="index.html">
<submenuelem text="New in Scribus 1.5" file="
</submenuelem text="Specification" file="
</area>
<area text="Documentation" file="intro.html
<submenuelem text="Introduction" file="d
<submenuelem text="Editorial Notes" file="
<submenuelem text="About the Team" file="
</submenuelem>
<submenuelem text="Setup" file="config.f
<submenuelem text="Configuring Scribus
<submenuelem text="Hyphenation and S
<submenuelem text="Font Setup" file="f
<submenuelem text="Fonts in Depth" file="
</submenuelem>
<submenuelem text="Scribus Basics" file="
<submenuelem text="Document Informa
<submenuelem text="Working with Fram
<submenuelem text="Working with Text"
<submenuelem text="Text Properties" file="
<submenuelem text="Search and Replac
<submenuelem text="Working with Style
</submenuelem text="Working with Imag
```

**Structuré** : Organisé en champs spécifiques (ex. bases de données, XML)



## Brut ? Binaire ?

🤔 Quelle est la différence entre un fichier en format texte brut et un fichier en format binaire ? 🤔



## Texte brut vs. fichier binaire

m ipsum dolor sit amet, consectetur adipiscing elit. Sed vehicula, ex a eleifend sodales, massa lorem interdum, sit amet tincidunt nisi neque vitae nunc. Aenean venenatis scelerisque eros, in molestie arcu euismod in. Donec lacinia imperdiet risus sed aliquam. Cras ac malesuada odio. Etiam dui dui, gravida in vestibulum id, cursus vitae lacus. Maecenas eget ligula semper ligula pretium imperdiet non sed erat. Cras interdum tellus in massa accumsan posuere. Fusce laoreet enim elementum purus venenatis maximus. Nulla rhoncus accumsan ligula vitae facilisis. Sed sodales, dui venenatis accumsan varius, justo ipsum egestas leo, a placerat orci lorem eu sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Donec posuere faucibus metus ac tempor.

Texte Brut

```
5"LZ.[content_Types].xml <?xml encoding="UTF-8" >
<document xmlns="http://www.openoffice.org/document/1.0" >
  <meta:author " " />
  <meta:creator " " />
  <meta:documentname " " />
  <meta:generator " " />
  <meta:language " " />
  <meta:revision " " />
  <meta:version " " />
  <text:p>m ipsum dolor sit amet, consectetur adipiscing elit. Sed vehicula, ex a eleifend sodales, massa lorem interdum, sit amet tincidunt nisi neque vitae nunc. Aenean venenatis scelerisque eros, in molestie arcu euismod in. Donec lacinia imperdiet risus sed aliquam. Cras ac malesuada odio. Etiam dui dui, gravida in vestibulum id, cursus vitae lacus. Maecenas eget ligula semper ligula pretium imperdiet non sed erat. Cras interdum tellus in massa accumsan posuere. Fusce laoreet enim elementum purus venenatis maximus. Nulla rhoncus accumsan ligula vitae facilisis. Sed sodales, dui venenatis accumsan varius, justo ipsum egestas leo, a placerat orci lorem eu sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Donec posuere faucibus metus ac tempor.</text:p>
</document>
```

Fichier binaire (fichier .docx ouvert avec un éditeur de texte basique)



## Document : structure

- **Structuré** : Bases de données relationnelles
  - Ces documents suivent une organisation spécifique où l'information est divisée en champs clairement définis.
- **Semi-structuré** : Emails, pages HTML, documents JSON, documents XML
  - Ces documents ont à la fois des parties structurées (ex. : balises) et des parties non structurées (ex. : corps d'un message email)
- **Non-structuré** : Articles de blog, publications sur les réseaux sociaux, livres numériques, documents Word, PDFs sans champs de remplissage spécifiques
  - Ces documents ne suivent pas une structure clairement définie et leur contenu consiste principalement en du texte libre





## Documents : formats

- **CSV** : Fichier structuré avec des champs séparés par des virgules

```
Plain Text
Nom, Prenom, Age
Dupont, Jean, 45
Martin, Elise, 30
```

- **JSON** : Fichier semi-structuré avec des paires clé-valeur

```
{
  "employés": [
    {"prénom": "Jean", "nom": "Dupont"},
    {"prénom": "Elise", "nom": "Martin"},
    {"prénom": "Pierre", "nom": "Lemaire"}
  ]
}
```

- **XML** : Fichier semi-structuré avec des balises définissant les données

```
<employés>
  <employé>
    <prénom>Jean</prénom>
    <nom>Dupont</nom>
  </employé>
  <employé>
    <prénom>Elise</prénom>
    <nom>Martin</nom>
  </employé>
</employés>
```



## Documents : stockage

- **Bases de Données Relationnelles** : MySQL, PostgreSQL, etc. Elles sont idéales pour les données structurées.
- **Systèmes de Gestion de Bases de Données Non Relationnelles (NoSQL)** : MongoDB, Couchbase, etc. Elles peuvent stocker des données non structurées et semi-structurées de manière efficace. 4 types de NoSQL : paires clé-valeur, famille de colonnes, document, graphes
- **Bases de données d'objets sous forme de documents** : Elasticsearch, Solr. Ces systèmes sont optimisés pour les opérations de recherche de texte.
- **Stockage sur le Cloud** : Amazon S3, Google Cloud Storage, etc. Ces services offrent une grande échelle et une grande flexibilité pour stocker des documents textuels de toutes sortes.
- **Stockage Hadoop (HDFS)** : Solution idéale pour le stockage et le traitement de grands volumes de données textuelles.
- **Logiciels de Gestion de Contenu (CMS)** : WordPress, Joomla, etc. Ils sont parfaits pour stocker et gérer les contenus textuels web.



## Quelques exemples

Request-Handler (qz) /select

```

common:
  <responseHeader>{
    <status>:0,
    <@time>:"1",
    <params>:{
      <?>:"Fantasy",
      <indent>:"on",
      <wt>:"json",
      <?>:"154669348279"}},
    <response>:{<numFound>:82,<start
  <?>:"2005",
  <directed_by>:[<?>:"Shane Acker",
  <initial_release_date>:"2",
  <genre>:[<?>:"Computer Animation",
  <?>:"Animation",
  <?>:"Analytical and post-4",
  <?>:"Science Fiction",
  <?>:"Short Film",
  <?>:"Thriller",
  <?>:"Fantasy",
  <name>:[<?>:""],
  <?>:"1541518174800"},
  <?>:"1986-2007",
  <directed_by>:[<?>:"Zack Snyder",
  <initial_release_date>:"2",
  <genre>:[<?>:"epic Film",
  <?>:"Adventure Film",
  <?>:"Fantasy",
  
```

Apache Solr

POST http://localhost:9200/\_mapping/indexes/1

```

{
  "index": "elasticsearch",
  "type": "document",
  "properties": {
    "name": "Elasticsearch",
    "version": "7.10.2"
  }
}

```

Response

```

{
  "index": "elasticsearch",
  "type": "document",
  "id": "1",
  "score": 1,
  "source": {
    "name": "Elasticsearch",
    "version": "7.10.2"
  }
}

```

ElasticSearch

```

b.orders.aggregate(
  { $match: { status: "A" } },
  { $group: { _id: "$cut_id", total:
    { $sum: "$total" } } }
)

```

Collection

```

{ "cut_id": "I012", "amount": 500, "status": "A" },
{ "cut_id": "I012", "amount": 200, "status": "A" },
{ "cut_id": "I012", "amount": 200, "status": "A" },
{ "cut_id": "I012", "amount": 200, "status": "A" },
{ "cut_id": "I012", "amount": 200, "status": "A" }

```

MongoDB

```

MATCH (p:Person)-[:LOVES]->(c:Car)
MATCH (p)-[:LIVES WITH]->(c)
MATCH (p)-[:DRIVES]->(c)
MATCH (p)-[:OWNS]->(c)

```

Graph visualization showing relationships between a person (Dan) and a car (Volvo V70).

neo4j

## Constitution de corpus : plateformes (proposées en 2023)

LES ACTUALITÉS

200 ans de ballon ovale

Pastiches de presse

Gallica: La bibliothèque numérique de la Bibliothèque Nationale de France. Visitez : <https://gallica.bnf.fr/>

Le plus vaste réservoir d'archives scientifiques et universitaires de toutes les disciplines.

28,2 M de documents

9 492 auteurs

438 909 articles

IStEX: Une plateforme numérique offrant un accès à des millions de publications scientifiques et universitaires de toutes les disciplines. Visitez : <https://www.istex.fr/>

Welcome to Project Gutenberg

Project Gutenberg is a library of over 70,000 free eBooks

Choose among free audio and Kindle eBooks, download them or read them online. You'll instantly appreciate the eBooks, for you to enjoy.

Project Gutenberg: Cette bibliothèque numérique contient de nombreux livres électroniques gratuits. Visitez : <https://www.gutenberg.org/>

Collections

Explorez notre collection par thème, sujet, siècle et organisation

Thèmes

Cartes et géographie

Histoire naturelle

Europeana: Une plateforme qui offre un accès à des numérisations de patrimoine culturel européen. Visitez : <https://www.europeana.eu/>

## Constitution de corpus : plateformes (proposées aujourd'hui)

**Hathitrust Digital Library** : Une immense bibliothèque numérique, surtout orientée vers la littérature scientifique, les ouvrages académiques, et les textes anciens. Visitez : <https://www.hathitrust.org>

**CORE** : Un agrégateur de contenu scientifique en libre accès. Il offre des millions d'articles de revues et de thèses. Visitez : <https://core.ac.uk>

**arXiv** : Pour constituer des corpus dans des domaines scientifiques et techniques, arXiv propose des articles prépubliés en libre accès. Visitez : <https://arxiv.org/>

**Perseus Digital Library** : Idéale pour les textes classiques en latin, grec ancien, et d'autres langues historiques, notamment pour l'analyse de corpus linguistiques. Visitez : <http://www.perseus.tufts.edu/hopper/>



## Représentations de textes

**Raw Text**: A bag is hard enough to drive than a dog.

**Bag of words vector**

Dog	0
need	2
Cat	1
than	0
it	1
heat	2
needs	0

Let's take an example to get a clearer understanding.

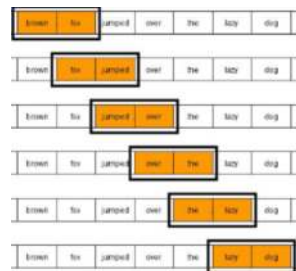
Sentence 1: The car is driven on the road.

Sentence 2: The truck is driven on the highway.

In this example, each sentence is a separate document.

We will now calculate the TF-IDF for the above two documents, which represent our corpus.

Word	TF		IDF	TF-IDF	
	A	B		A	B
The	1/7	1/7	$\log_2(2/1) = 0$	0	0
Car	1/7	0	$\log_2(2/1) = 0.3$	0.043	0
Truck	0	1/7	$\log_2(2/1) = 0.3$	0	0.043
Is	1/7	1/7	$\log_2(2/2) = 0$	0	0
Driven	1/7	1/7	$\log_2(2/2) = 0$	0	0
On	1/7	1/7	$\log_2(2/2) = 0$	0	0
The	1/7	1/7	$\log_2(2/2) = 0$	0	0
Road	1/7	0	$\log_2(2/1) = 0.3$	0.043	0
Highway	0	1/7	$\log_2(2/1) = 0.3$	0	0.043



**Sac de mots (Bag of Words)** : Les textes sont représentés comme un sac de ses mots, ignorant la grammaire et l'ordre des mots mais gardant la multiplicité.

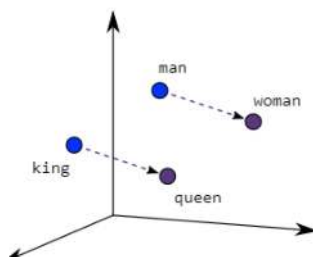
**TF-IDF (Term Frequency-Inverse Document Frequency)** : Une mesure statistique qui évalue l'importance d'un mot dans un document dans le contexte d'un corpus de documents.

**N-grams** : Une approche qui prend en considération l'ordre des mots. Un N-gram est une séquence continue de N éléments dans un texte donné.

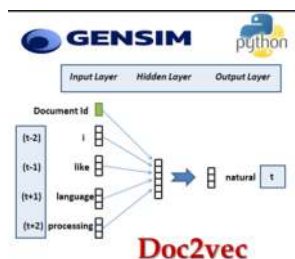


## Représentations de textes

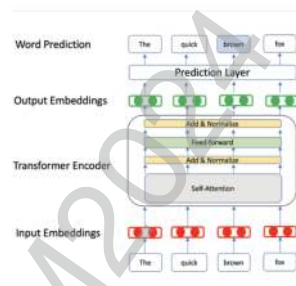
word embeddings are plotted, similar



Représentations distribuées (Word2Vec, GloVe). Ce sont des modèles de représentation de mots qui capturent une multitude de caractéristiques syntaxiques et sémantiques.



Encodage de phrases et de documents (Doc2Vec). Une extension du word2vec qui encode non seulement des mots, mais aussi des phrases, des paragraphes et des documents entiers.



BERT (Transformers): Un modèle de représentation de langage utilisant la technologie des transformateurs, capable de capturer de plus longs contextes et de traiter la compréhension de la langue à un niveau plus profond.



## Outils de représentation de textes

- **Scikit-learn** :
  - Offre des fonctionnalités pour la représentation de sac de mots et de TF-IDF. Visitez : <https://scikit-learn.org/stable/>
- **Gensim** :
  - Une bibliothèque Python pour la modélisation de thèmes, l'indexation de documents et la similitude de documents. Visitez : <https://radimrehurek.com/gensim/>
- **Spacy** :
  - Un outil puissant de traitement du langage naturel qui offre des fonctionnalités pour la représentation de Word2Vec. Visitez : <https://spacy.io/>
- **CamemBERT** :
  - Une variante de RoBERTa pré-entraînée pour le français, offerte par Facebook's Fair. Elle est disponible sur HuggingFace's Transformers. Visitez : [https://huggingface.co/transformers/model\\_doc/camembert.html](https://huggingface.co/transformers/model_doc/camembert.html)
- **BertForSequenceClassification** de Hugging Face :
  - Modèle de classification de texte basé sur BERT. Visitez : [https://huggingface.co/transformers/model\\_doc/bert.html#bertforsequenceclassification](https://huggingface.co/transformers/model_doc/bert.html#bertforsequenceclassification)



## Outils de représentation de textes

⚠ Chaque type de représentation et chaque outil ont leurs propres avantages, le choix dépend des besoins spécifiques de l'analyse de texte à réaliser. ⚠



## Visualisation de données textuelles

Using a mask you can generate wordclouds in arbitrary shapes.

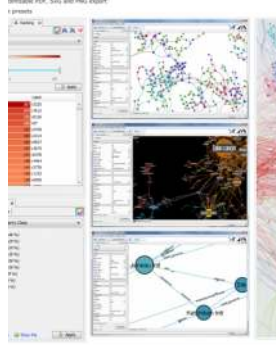


**WordCloud:** Cet outil génère des « nuages de mots » à partir d'un texte, où la taille de chaque mot indique sa fréquence ou son importance. Visitez : [https://amueller.github.io/word\\_cloud/](https://amueller.github.io/word_cloud/)

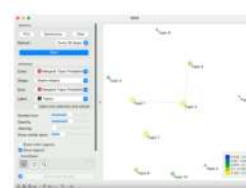
cartography

Up to partition maps to make meaningful the network representation. Customise colors, size or label representation. The central preview module lets you get the final layout and save about distribution

importable PDF, SVG and PNG export



**Gephi:** Utilisé pour la visualisation de réseaux de textes tels que les graphes de cooccurrence de mots. Visitez : <https://gephi.org/>



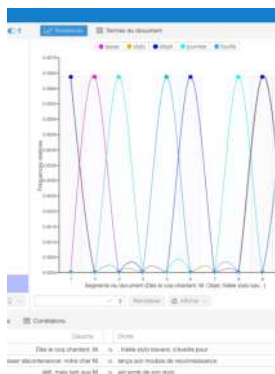
Topic & LDAvis is ideal for this, seems to be quite well represented in the corpus.



**LDavis :** Il s'agit d'un outil d'exploration interactive de modèles de sujets, conçu pour gérer la visualisation de collections volumineuses de documents textuels. Visitez : <https://cran.r-project.org/web/packages/LDAvis/vignettes/details.pdf>



## Visualisation de données textuelles



**Voyant Tools**: Un outil web pour l'analyse de texte qui offre un large éventail de visualisations, de la fréquence des mots aux graphiques de dispersion. Visitez : <https://voyant-tools.org/>

### QDA Miner Lite - FREE QUALITATIVE DATA ANALYSIS

QDA Miner Lite is a free and easy-to-use version of our popular computer-aided software for the analysis of textual data such as interview and news transcripts, or video or still images. It offers basic CAQDAS features such as:
 

- Importation of documents from plain text, RTF, HTML, PDF as well as data in delimited text files
- Importation from other qualitative coding software such as Atlas.ti, HyperLynx or like Transana and Transcriber as well as from Reference Information System
- Full-text coding using codes organized in a tree structure
- Ability to add comments for memos to coded segments, cases, or the whole project
- Boolean text search tool for retrieving and coding text segments
- Code frequency analysis with bar chart, pie chart, and tag clouds
- Advanced retrieval with Boolean (and, or, not) and proximity operators (includes point tables to XLS, Tab, Delimited, CSV formats, and Word format)
- Export graphs to BMP, PNG, JPEG, WMF formats
- Single-file (zipped) project format
- Interface and help file in English, French, and Spanish.



**QDA Miner Lite** : Un logiciel gratuit d'analyse de données qualitatives. Il offre des fonctionnalités pour le codage manuel des documents et l'exécution de requêtes textuelles. Visitez : <https://provalisresearch.com/products/qualitative-data-analysis-software/freeware/>

### Jigsaw: Visual Analytics for Exploring and Understanding Documents

**Jigsaw**: Visual Analytics for Exploring and Understanding Documents

**Open Members**: John Shook, Corbin King, Qinghui Liu, Sukhinder Parra, Anand Rastogi, Anand Venkatesh, Anu Hiranmayi, Matt Jen, Yoon-Oh Kang, Jayashree Kulkarni, Yael Partonopoulos, Neil P.

**REACTIVE**  
**REACTIVE** is a free and easy-to-use version of our popular computer-aided software for the analysis of textual data such as interview and news transcripts, or video or still images. It offers basic CAQDAS features such as:
 

- Importation of documents from plain text, RTF, HTML, PDF as well as data in delimited text files
- Importation from other qualitative coding software such as Atlas.ti, HyperLynx or like Transana and Transcriber as well as from Reference Information System
- Full-text coding using codes organized in a tree structure
- Ability to add comments for memos to coded segments, cases, or the whole project
- Boolean text search tool for retrieving and coding text segments
- Code frequency analysis with bar chart, pie chart, and tag clouds
- Advanced retrieval with Boolean (and, or, not) and proximity operators (includes point tables to XLS, Tab, Delimited, CSV formats, and Word format)
- Export graphs to BMP, PNG, JPEG, WMF formats
- Single-file (zipped) project format
- Interface and help file in English, French, and Spanish.

**NEW**  
 Web Jigsaw is here! Try out the web-based version of Jigsaw that runs in a web browser (Chrome) April 24, 2023

**Download Now**  
 6.04 Release March 4, 2016

**Abstract**  
 Jigsaw: Visualization for Investigative Jigsaw for Same Instances - MIT 2012 Data Integration and Evidence Mining Jigsaw meets Blue Spines - The Visual Analytics with Jigsaw (preprint) Using Jigsaw, we won the university

**Summary**  
 Investigative analysis and visualization allow users to connect small bits of evidence to help find evidence on short and documents or spreadsheets, and analysts must examine the full range of data. All the number of documents to examine them. A document may have many different Jigsaw. A visual analytics system to help analysts and researchers find and accurate understanding of the major events and important concepts embedded in the particularly focused on presenting the identifiable important entities (people, places, organizations)

**Jigsaw**: Un système de visualisation pour comprendre et analyser les documents et les collections de textes. Visitez : <http://www.cc.gatech.edu/gvu/ijigsaw/>



## Enrichissement et annotations

- **BRAT (Brat Rapid Annotation Tool)**: Un outil d'annotation de texte basé sur le Web pour l'annotation d'entités nommées, les relations, etc. Visitez : <http://brat.nlplab.org/>
- **GATE (General Architecture for Text Engineering)**: Un cadre pour l'extraction d'informations, l'annotation sémantique et d'autres tâches de traitement du langage naturel. Visitez: <https://gate.ac.uk/>
- **Le module NER de Spacy** : Spacy est une bibliothèque Python pour le traitement du langage naturel et possède un module d'extraction d'entités nommées (NER) qui peut être formé pour le français. Visitez : <https://spacy.io/api/annotation>
- **UDPipe** : Il offre des outils pour l'annotation morphologique, syntaxique et lexicale des textes, y compris pour le français. Visitez : <http://ufal.mff.cuni.cz/udpipe>
- **WebAnno** : Un outil d'annotation web pour le crowdsourcing. Supporte le français en permettant une annotation manuelle. Visitez : <https://webanno.github.io/webanno/>



💡 Ces outils peuvent aider à enrichir les documents textuels et à extraire plus d'informations pertinentes pour la fouille de textes. 💡

## Types de modèles de fouille de textes

- **Modèles Statistiques** : Ces modèles sont basés sur des statistiques et permettent d'extraire des informations à partir de données textuelles. Par exemple, la fréquence des termes, les modèles de co-occurrence, etc.
- **Modèles d'Apprentissage Automatique** : Ces modèles sont formés sur des données textuelles pré-annotées pour apprendre des tâches spécifiques comme la classification de textes, l'analyse de sentiments, l'identification d'entités nommées, etc.
- **Modèles de Traitement du Langage Naturel Profond** : Ces modèles utilisent les techniques de l'apprentissage profond pour comprendre les sémantiques et la syntaxe du texte. Exemples : les Réseaux Neuronaux Récurrents (RNN), les Réseaux de Neurones Convolutionnels (CNN) et les Transformers comme BERT.
- **Modèles de Topic Modeling** : Ces modèles identifient les sujets dominants dans un corpus de textes. Un exemple courant est Latent Dirichlet Allocation (LDA).
- **Modèles de Graphes** : Ces modèles représentent les textes sous forme de graphes pour identifier les relations entre différentes entités ou concepts dans le texte.

## Types de modèles de fouille de textes

💡 Chaque type de modèle est adapté à différents types de tâches et à différents niveaux de complexité de l'information à extraire. 💡



## Supervision ? Ou pas...

- **Clustering** : Les algorithmes de clustering, tels que K-means ou DBSCAN, regroupent les documents textuels en clusters basés sur leur similarité. Outils:
  - KMeans de Scikit-Learn, visitez : <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
  - DBSCAN de Scikit-Learn, visitez : <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>
- **Latent Dirichlet Allocation (LDA)** : Un modèle de topic modeling qui identifie les sujets dominants dans un ensemble de documents. Outil : Gensim LDA, visitez : <https://radimrehurek.com/gensim/models/ldamodel.html>.
- **Réduction de dimension** : Les techniques comme l'Analyse en Composantes Principales (PCA) ou l'Embedding Stochastic Neighbor (t-SNE) sont utilisées pour visualiser les données textuelles de haute dimension dans un espace bidimensionnel ou tridimensionnel. Outils:
  - PCA de Scikit-Learn, visitez : <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
  - t-SNE de Scikit-Learn, visitez : <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>
- **Modèles de représentation de texte** : Ces modèles, tels que Word2Vec ou Doc2Vec, apprennent des représentations vectorielles des mots ou des documents à partir de grands corpus de texte non étiquetés. Outils:
  - Word2Vec de Gensim, visitez : <https://radimrehurek.com/gensim/models/word2vec.html>
  - Doc2Vec de Gensim, visitez : <https://radimrehurek.com/gensim/models/doc2vec.html>





## Modèles non-supervisés

💡 Utilisez ces modèles lorsque vous avez de grandes quantités de données textuelles non étiquetées et que vous voulez extraire des informations utiles sans supervision. 💡



## Modèles supervisés

- **Régression logistique** : Un modèle de classification souvent utilisé pour prédire des classes binaires. Pour utiliser, visitez : [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
- **SVM (Support Vector Machines)** : Un algorithme de classification qui peut être utilisé pour la classification de textes. Pour utiliser, visitez : <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>
- **Arbres de décision et forêts aléatoires** : Des modèles d'apprentissage automatique qui peuvent être utilisés pour la classification et la régression. Pour plus d'informations, visitez :
  - Pour les arbres de décision : <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
  - Pour les forêts aléatoires : <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- **Réseaux neuronaux** : Des modèles d'apprentissage en profondeur utilisés pour une variété de tâches de traitement du langage naturel. Pour plus d'informations, visitez :
  - Pour Keras : <https://keras.io/>
  - Pour TensorFlow : <https://www.tensorflow.org/>
- **BERT et autres Transformers** : Ces modèles de traitement du langage naturel en profondeur sont pré-entraînés sur de grands corpus de texte et peuvent ensuite être affinés pour des tâches spécifiques avec des données annotées. Pour utiliser, visitez : <https://huggingface.co/transformers/>



## Modèles supervisés

💡 Ces modèles supervisés sont utiles lorsque vous avez de larges quantités de données textuelles annotées pour l'entraînement. 💡



## Biais !

### Explain the Prediction



## Biais !



## Et sans programmation ?...

- **RapidMiner**: Plateforme d'analyse de données et d'apprentissage automatique. Visitez: <https://rapidminer.com/products/studio/>
- **KNIME**: Plateforme d'analyse de données open-source. Visitez: <https://www.knime.com/>
- **Orange**: Un outil de visualisation et d'analyse de données open-source. Visitez: <https://orange.biolab.si/>
- **MonkeyLearn**: Plateforme basée sur le cloud pour le traitement du langage naturel. Visitez: <https://monkeylearn.com/>
- **Google Cloud Natural Language**: Service basé sur le cloud pour l'analyse de sentiment et d'entité. Visitez: <https://cloud.google.com/natural-language>
- **IBM Watson Natural Language Understanding**: Service pour l'analyse de sentiment et de texte. Visitez: <https://www.ibm.com/cloud/watson-natural-language-understanding>
- **Amazon Comprehend**: Service basé sur le cloud qui utilise l'apprentissage automatique pour découvrir des insights à partir de texte. Visitez: <https://aws.amazon.com/comprehend/>
- **Weka**: Suite logicielle pour l'apprentissage automatique et la fouille de données. Elle propose une gamme de techniques prétraitement de données et de classification, notamment pour les données textuelles. Visitez: <https://www.cs.waikato.ac.nz/ml/weka/>

## Pas (trop) de spoil 😊



**NooJ**  
A Corpus Processor - A Linguistic DE  
for developing Natural Language Pre

Help | WikiBook | Screenshots | Downloads | Linguistic Help

[Présentation en français](#)

### NooJ: a linguistic development environ

- Develop linguistic resources to formalize various linguistic phenomena (morphological, lexical, morphological, syntactic and semantic) in any natural language.
- Create your own corpus of texts, apply these linguistic resources and then do formal corpus statistical analyses in [Corpus Linguistics](#).
- Use NooJ's engine and linguistic resources to construct [Natural Processing](#) applications.

NooJ's linguistic engine offers the four types of grammars of the Chomsky hierarchy: Regular grammars, Context-Free Grammars, Context-Sensitive Grammars and Unrestricted Grammars. NooJ's parsers are formally identified by unambiguous finite state

posit PRODUCTS SOLUTIONS LEARN & SUPPORT

DEMO GAG

## RStudio Desktop

Loved by millions of people worldwide, the RStudio integrated development environment (IDE) is a set of tools built to be more productive with R and Python.

Don't want to download or install anything? Get started RStudio on [Posit Cloud](#) for free. If you're a professional scientist, looking to download RStudio and also need its enterprise features, don't hesitate to [book a call with](#).

want to learn about core or advanced workflows in R? Explore the [RStudio User Guide](#) or the [Getting Started](#)

### 1: Install R

**Gargantext**  
A web interface for linguistic processing

Historic: Gargantext is a web interface for linguistic processing...  
Presentation: Gargantext is a web interface for linguistic processing...

- GPT-4o  
Idéal pour la plupart des tâches
- o1-preview  
Utilisation du raisonnement avancé
- o1-mini  
Raisonnement plus rapide
- GPT-4o mini  
Plus rapide pour les tâches au quotidien
- GPT-4  
Ancien modèle

Explorer les GPT

Chat éphémère

NooJ

RStudio

Gargantext

ChatGPT



## DISCUSSION

OUTILS CLÉ-EN-MAIN VS.  
PROGRAMMATION, EXEMPLES



Aix-Marseille  
université  
Sociétalement engagée



## Outils clé-en-main

- Outils Clé-en-main (e.g., RapidMiner, KNIME, MonkeyLearn):
  - Avantages:
    - Faciles à utiliser, pas besoin de compétences en programmation
    - Idéals pour des tâches simples et standard
    - Permettent souvent d'obtenir rapidement des résultats
  - Inconvénients et Limites:
    - Peu de flexibilité pour personnaliser les algorithmes ou le processus d'analyse
    - Les modèles plus avancés ou l'analyse de grands ensembles de données peuvent ne pas être pris en charge



## Outils de programmation

- Outils de Programmation (e.g., Python avec pandas, NLTK, scikit-learn, TensorFlow):
  - Avantages:
    - Flexibilité presque illimitée pour manipuler les données, choisir et adapter les algorithmes
    - Prend en charge des modèles d'apprentissage automatique/deep learning avancés
    - Capacité à gérer de grands ensembles de données
  - Inconvénients et Limites:
    - Nécessite des compétences en programmation
    - Plus de temps pour développer et optimiser le code



## Clé-en-main vs. programmation

💡 En somme, le choix entre les outils clé-en-main et les outils de programmation dépendra des compétences de l'équipe, de la complexité du problème, de la taille des données et du degré de personnalisation nécessaire. 💡



## Exemple : Voyant Tools



## Exemple : Spacy - dépendances, entités nommées (python)

L'arbre des dépendances pour chaque phrase

```

[7]: sentences = list(doc.sents)
for sent in sentences:
    displacy.render(sent, style='dep',

```

```

[1]: import spacy
from spacy.lang
from spacy.lem
import matplotlib
from collections
import warnings
warnings.filter

```

Charger l

```

[2]: nlp = spacy.lob
# Cell 3: Défini
text = "Des pizzas volantes"
Elles sont menées
Les gens sont t
doc = nlp(text)

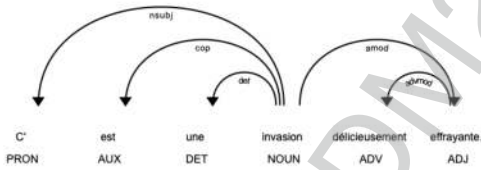
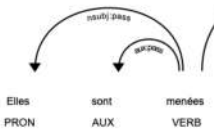
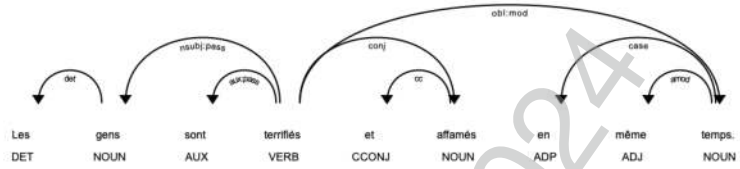
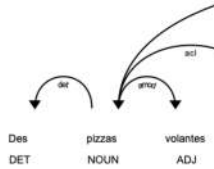
```

Analyse c

```

[3]: words = [token.
word_freq = Cou
print(word_freq
Counter({'pizza
1, 'extra-terre

```



## Exemple : NLTK - analyse fréquentielle (python)

Sup Visualisation des mots les plus courants

```

[4]: stop_wo [6]: frequency_dist.plot(30, cumulative=False)
plt.show()

```

```

[1]: filtere
print(f
['Dès',
'a', 'ex
s', 'tr
'fautes
s', 'la
itable'
t', 'ar

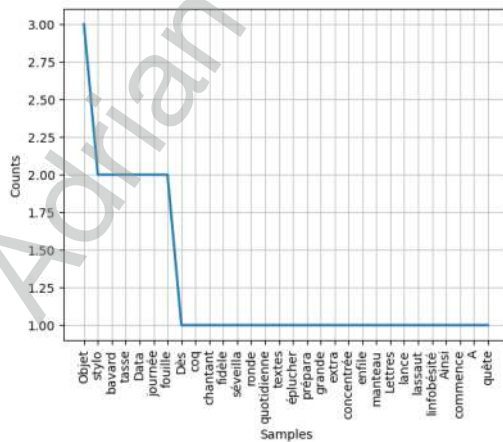
```

Anal

```

[5]: frequen
print(f
most_co
for wor
pr)
<FreqDist
Objet:
enne: 1
sité: 1
l | ord
l océan
l lança
nder: 1
l texte

```



## Exemple : tm - nuage de mots (R)

### Installer

```
[1]: install.packages("tm")
```

### Charger

```
[2]: library(tm)
library(wordcloud)
library(RColorScales)
```

### Définir

```
[3]: text <- "L'
```

### Définir les "mots vides"

```
[6]: stopwords <- c("un", "à", "de", "les", "la", "des", "et", "dans", "se", "pour")
```

### Créer le nuage de mots

```
[7]: wordcloud(words = corpus,
  scale=c(5,0.5),
  min.freq = 1,
  max.words=200,
  random.order=FALSE,
  rot.per=0.35,
  colors=brewer.pal(8, "Dark2"))
```



## IA GÉNÉRATIVE, LLM

EXPLICATIONS, EXEMPLES, POINTS DE VIGILANCE



Aix-Marseille université  
Sociétalement engagée





## IA ?

- L'intelligence artificielle est un domaine de l'informatique qui crée des systèmes capables d'apprendre, de prendre des décisions et d'interagir.
- À l'intérieur de l'IA :
  - Approche symbolique : basée sur des règles et de la logique.
  - Approche basée sur l'apprentissage : comprend le Machine Learning et l'IA générative (notre sujet).



"generate an image to introduce AI" (générée par [bing.com](https://www.bing.com), optimisée par DALL-E 3)



## IA générative ?

- L'IA générative, également appelée apprentissage génératif, fait référence aux systèmes d'IA qui peuvent générer du contenu, des informations ou des résultats en s'appuyant sur les données qu'il a apprises.
- Exemples d'utilisation de l'IA générative :
  - Génération d'images
  - Rédaction de textes
  - Fouille de textes
  - Création de musique
  - Modélisation 3D
  - Synthèse vocale



"génère une image qui illustre l'IA générative, dans le style de Picasso" (générée par [bing.com](https://www.bing.com), optimisée par DALL-E 3)



## Mécanismes de l'IA générative

- Les modèles génératifs apprennent les patterns spécifiques à partir des données d'entrée.
- Ils utilisent ensuite ces patterns pour générer du contenu qui correspond à ce qu'ils ont appris.
- Exemple célèbre : les Réseaux Antagonistes Génératifs (GANs).



## Potentiel et défis de l'IA générative

- Potentiel de l'IA générative :
  - Créations artistiques
  - Solutions de problèmes complexes
  - Personnalisation
- Défis de l'IA générative :
  - Difficulté de contrôle sur le contenu généré
  - Problèmes de droits et de vie privée
  - Problèmes d'éthique



## IA générative

💡 L'IA générative joue un rôle clé dans la nouvelle génération de technologies d'IA, avec un impact potentiel significatif sur les industries créatives, le commerce électronique, et bien d'autres secteurs. Malgré ses défis, elle ouvre des possibilités infinies pour l'avenir de l'IA. 💡



## Les Modèles de Langage à Grande Échelle (LLM)

- Les modèles de langage à grande échelle (LLM) sont un type particulier d'IA générative. Ils ont été entraînés sur d'énormes quantités de texte pour apprendre des statistiques linguistiques.
- Ils peuvent générer du texte qui peut paraître remarquablement humain.
- Des exemples notables incluent ChatGPT de OpenAI, BERT de Google, et T5 de Google.



## LLMs : avantages et limites

- **Avantages des Modèles de Langage à Grande Échelle :**
  - **Polyvalence :** Ils peuvent être utilisés pour de nombreux cas d'utilisation, comme répondre à des questions, résumer du texte, traduire des langues, et bien d'autres.
  - **Efficacité :** Le modèle n'a besoin d'être formé qu'une seule fois, puis peut être réglé pour diverses tâches.
  - **Innovation :** Ils sont à la pointe de la recherche en IA et ont ouvert de nouvelles opportunités.
- **Limites des Modèles de Langage à Grande Échelle :**
  - **Manque de compréhension :** Ils ne 'comprennent' pas vraiment le texte qu'ils génèrent.
  - **Coût élevé :** Le coût d'entraînement de ces modèles est très élevé.
  - **Risques éthiques :** Ils peuvent générer du contenu offensant, trompeur ou biaisé.
  - **Manipulabilité :** Ils pourraient être utilisés à des fins malveillantes, comme la désinformation.



## LLMs

💡 **Les Modèles de Langage à Grande Échelle sont des outils puissants avec un potentiel considérable. Cependant, comme toute technologie, ils viennent avec leurs propres défis. Il dépendra de nous de les utiliser de manière responsable.** 💡



## IA générative et fouille de textes (GPT-4)

Pour la phrase suivante, identifie les aspects et indique quelle est la polarité du sentiment pour chaque aspect (positive, négative, ou neutre). Affiche uniquement les aspects et leur polarités correspondantes. La phrase est : « Le plat a été délicieux, mais le service laisse à désirer. ».

Aspect: Le plat  
Polarité du sentiment: Positive

Aspect: Le service  
Polarité du sentiment: Négative

Play Settings

Le même exercice pour la phrase : « Je n'achèterai pas cet ordinateur, même si le clavier est plutôt pas mal. »

Aspect: L'ordinateur  
Polarité du sentiment: Négative

Aspect: Le clavier  
Polarité du sentiment: Positive

Analyse de sentiment

Identifie les entités nommées de la phrase suivante, ainsi que leur type : "Henri a acheté 300 actions de la société AMD en 2006."

Entité nommée: Henri  
Type: Personne

Entité nommée: 300  
Type: Quantité

Entité nommée: AMD  
Type: Organisation

Entité nommée: 2006  
Type: Date

Reconnaissance d'entités nommées

Résume le texte suivant : "C'est le cog charismatique, M. Otsch, l'ancien styliste bernard, s'élève pour sa tenue quotidienne des notes à répétition. Il se présente une grande base de données extra-consciente, entre son niveau de lecture et sa base à l'assaut de l'orthographe."

Ainsi commence sa journée. A la suite des travaux réalisés à travers des termes décodés d'amais supérieurs, combinant des ordres textuels non structurés et définis, les juges sélecteurs des notes d'orthographe, il trouve dans la base son d'informations.

Un nombre émerge une certaine qualité, "R1143QP\_Dine\_Reg.12.A". Sans se laisser déconcentrer, mais chez M. Otsch, l'ancien styliste, son module de reconnaissance de pattern. En un instant, il lit la phrase originale et la vérifie à l'identique. "Demande son extension de ligne pour le service client le 12 août".

Chez M. Otsch, il y a un nouveau défi, mais tant que M. Otsch est prêt de son style bernard et de sa base de données, aucun texte ne peut échapper à sa fouille minutieuse."

M. Otsch, qui est personnellement comme un analyse de données complexes, commence sa journée en se battant contre l'orthographe. Il cherche des informations précieuses cachées dans des tonnes d'a-mais supérieurs, évite les pièges des erreurs d'orthographe et découvre des phrases cryptées comme "R1143QP\_Dine\_Reg.12.A". Malgré ces défis quotidiens, armé de sa capacité d'analyse et de sa persévérance, il parvient toujours à accomplir son travail.

Résumé automatique



## Hallucinations et limites

Quelle est la population de la Roumanie en octobre 2024 ?

En octobre 2024, la population de la Roumanie est estimée à environ **19 millions** d'habitants. Ce chiffre reflète une légère diminution par rapport aux années précédentes, avec une baisse due à des facteurs comme un faible taux de natalité et une émigration importante.

Play Settings

Quelle est la population de la Roumanie au mois d'octobre 2023 ?

Je m'excuse, mais en tant qu'intelligence artificielle, je ne peux pas fournir les informations en temps réel ni prévoir les données futures. Je vous suggère de consulter les sources de données officielles les plus récentes pour obtenir cette information.

Récence

Explique pourquoi 1187 n'est pas un nombre premier.

Le nombre 1187 n'est pas un nombre premier parce qu'il est divisible par d'autres nombres que 1 et lui-même. En particulier, 1187 est divisible par 13. Voici le raisonnement :

- Diviser 1187 par 13 :  $1187 \div 13 = 91,3$

Cela donne un quotient entier de 91, donc 1187 est divisible par 13.

- Comme 1187 est divisible par 13, il possède au moins un autre diviseur que 1 et lui-même (13), ce qui signifie qu'il n'est pas premier.

Ainsi, 1187 est un **nombre composé**, car il peut être factorisé en  $13 \times 91$ .

Certaines explications (surtout mathématiques...)

Explique pourquoi le Brésil est le premier producteur mondial de mangues. Sois bref.

Le Brésil est le premier producteur mondial de mangues en raison de son climat tropical favorable, de vastes terres agricoles et de sa capacité à cultiver des variétés de mangues adaptées à différentes régions du pays. De plus, des investissements dans les technologies agricoles et les infrastructures de transport facilitent l'exportation massive des mangues vers les marchés internationaux.

Brésil a fait des investissements significatifs dans la technologie agricole et dans les infrastructures qui permettent une production efficace et une distribution rapide.

Ces facteurs combinés expliquent pourquoi le Brésil est un producteur majeur de mangues.

Hallucinations



## CONCLUSIONS

FOUILLE DE TEXTES, OUTILS,  
SUGGESTIONS



Aix-Marseille  
université  
Sociétalement engagée



### Conclusions

- La fouille de textes est une pratique essentielle permettant de découvrir des informations précieuses à partir de vastes quantités de données textuelles. C'est une tâche qui est de plus en plus importante avec le développement de sources de données volumineuses.
- L'utilité de la fouille de textes :
  - La fouille de textes représente un outil précieux pour déceler des tendances, découvrir des connaissances et générer des insights à partir de sources de données non structurées.
  - Elle peut être utilisée pour générer des prédictions, apportant ainsi une grande valeur ajoutée dans des domaines tels que la recherche scientifique, le marketing, les médias sociaux, etc.



## Conclusions : programmer ?

- La programmation est une compétence fondamentale pour réussir dans la fouille de texte. Elle permet de manipuler, nettoyer et analyser efficacement des données volumineuses.
- La connaissance de certains langages comme Python ou R, qui ont d'importantes bibliothèques de traitement de texte, peut s'avérer extrêmement utile.



## Astuces pour débiter

- Commencez petit: initiez-vous avec des projets simples et des jeux de données de petite taille pour vous familiariser avec les techniques de base.
- Entraînez-vous à nettoyer et préparer les données: ces compétences sont aussi importantes que l'analyse elle-même.
- Utilisez les ressources en ligne: il existe de nombreux cours gratuits, tutoriels et forums où vous pouvez apprendre et poser des questions.
- Soyez patient et persévérant: l'apprentissage de la fouille de texte et de la programmation est un voyage, pas une destination.



## Conclusion

💡 Dans un monde où les données regrettent à extraire et à exploiter des informations à partir de textes, vous êtes les super héros armés de scripts Python et de machines à café ! L'investissement dans l'apprentissage de ces compétences ouvrira d'innombrables opportunités et avantages. Alors, allumez vos ordinateurs, ouvrez vos esprits, et plongez-vous dans le merveilleux monde de la fouille de texte ! 💡

- GPT-4o, oct. 2024



## Show must go on...

Aa Nom	🔗 Lien	☰ Accès	☰ Commentaires
Prodigy	<a href="https://prodi.gy/">https://prodi.gy/</a>	payant	annotation
Brat	<a href="http://brat.nlplab.org/index.html">http://brat.nlplab.org/index.html</a>	open-source	annotation
Label Studio	<a href="https://labelstud.io/">https://labelstud.io/</a>	open-source	annotation
PACTE	<a href="http://pacte.crim.ca/">http://pacte.crim.ca/</a>	accès sur demande	annotation collaboration
HTRC Analytics	<a href="https://analytics.hathitrust.org/datasets#top">https://analytics.hathitrust.org/datasets#top</a>	accès sur demande accès institutionnel	analyse visualisation
CORLI (huma-num)	<a href="http://explorationdecorpus.corpusecrits.huma-num.fr/">http://explorationdecorpus.corpusecrits.huma-num.fr/</a>	accès libre	exploration de corpus recensement d'outils
Nooj	<a href="https://nooj.univ-fcomte.fr/index.html">https://nooj.univ-fcomte.fr/index.html</a>	open-source	ressource linguistique analyse annotation
IRaMuTeQ	<a href="http://www.iramuteq.org/">http://www.iramuteq.org/</a>	open-source	analyse visualisation
CORLI "Corpus écrits"	<a href="http://explorationdecorpus.corpusecrits.huma-num.fr/outils-logiciels-corpus-ecrits/">http://explorationdecorpus.corpusecrits.huma-num.fr/outils-logiciels-corpus-ecrits/</a>	inventaire de ressources	logiciels





**MERCI !**

- Questions ?
- [adrian.chifu@univ-amu.fr](mailto:adrian.chifu@univ-amu.fr)
- <https://adrianchifu.com>
- <https://linktr.ee/adrianchifu>



"create a thank you image for the end of a presentation in a funny and cartoonish style" (générée par Bing Chat, optimisée par DALL-E 3)

