



# Panorama des méthodologies de fouille de textes et interdisciplinarité



ANF TDM 2023

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

12 octobre 2023

# Plan

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



"une image pour le plan d'une présentation" (générée par [bing.com](https://www.bing.com))

# FOUILLE DE DONNÉES

QU'EST-CE QUE LA FOUILLE DE  
DONNÉES ?

## Fouille de données ?

- **Exploration et analyse de grande échelle des données**
  - **Le but de découvrir des modèles et des tendances significatives.**
  - **Aussi appelée exploration de données ou data mining.**
- **Cruciale pour la découverte de tendances et modèles significatifs**
  - **Dans le monde d'aujourd'hui où énormément de données sont 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 [freepik.com](https://www.freepik.com))

# FOUILLE DE TEXTES

CONTEXTE, PROBLÉMATIQUES,  
INTERDISCIPLINARITÉ, SCHEMA

## Fouille de textes ?

- **Extraction d'informations de haute qualité à partir du texte**
  - Elle implique le prétraitement, l'annotation, l'analyse et l'interprétation du texte.
- **Cruciale pour la prise de décisions basées sur des données massives et non structurées**
  - Avec l'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**



"Un puzzle avec différents morceaux représentant les diverses étapes du processus de fouille de texte, tels que le prétraitement, l'annotation, l'analyse et l'interprétation." (générée par [starryai.com](http://starryai.com))

## Fouille de textes : contexte

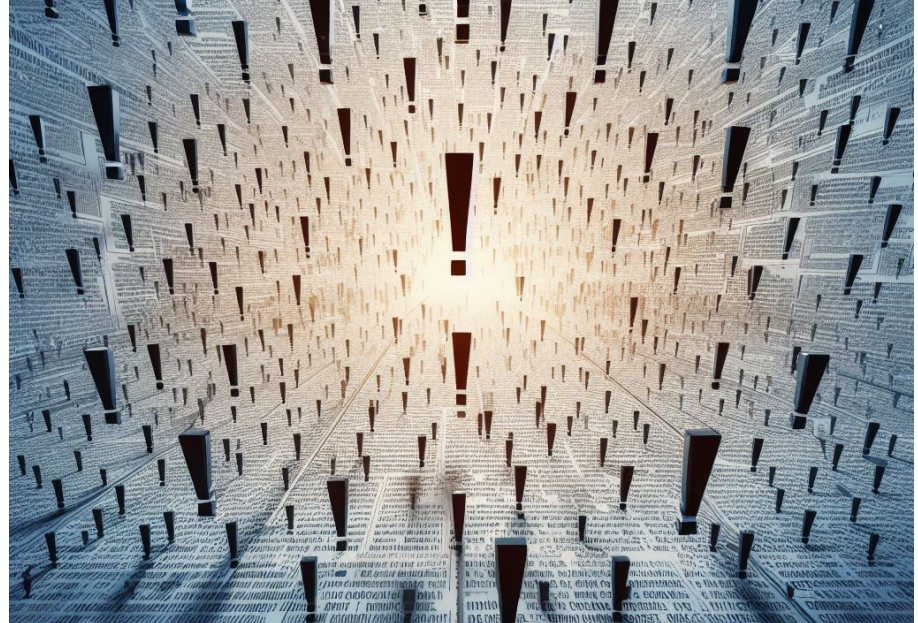
- Explosion des données textuelles
- Textes généralement non structurés
- Dépassement des méthodes d'analyse traditionnelles
- Enjeux : détection des sentiments, résumé automatique, classification de textes, analyse de tendances
- Applications : marketing, médias sociaux, santé, finance, éducation, etc.



"Un nuage de mots avec des termes clés tels que : "Détection des sentiments", "Classification de textes", "Analyse tendances", "Données non structurées", "Textes massifs"." (générée par DALL-E)

## 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](https://www.bing.com))



# Fouille de textes : interdisciplinarité

- Linguistique :

- Analyses du langage naturel et des structures

*Le travail sur le langage naturel et les structures linguistiques est fondamental à la fouille de texte.*

- Informatique :

- Programmation et conception d'outils

*Des compétences en programmation et en algorithmique sont nécessaires pour concevoir les outils de fouille de texte.*

- Statistiques :

- Identification des tendances et patterns

*L'analyse statistique est utilisée pour identifier les tendances et les patterns dans les données.*

- Intelligence Artificielle :

- Automatisation et extraction d'informations

*L'IA et l'apprentissage automatique sont utilisés pour automatiser le processus d'extraction des informations.*

- Métier :

- Compréhension du domaine d'application

*Il est important de comprendre les aspects spécifiques du domaine d'application afin d'améliorer les interprétations et d'éviter les biais*



"Des icônes ou des images symbolisant l'interaction entre la linguistique, l'informatique, les statistiques et l'intelligence artificielle." (générée par [bing.com](https://www.bing.com), optimisée par DALL-E 3)

# Fouille de textes : le processus (simplifié)

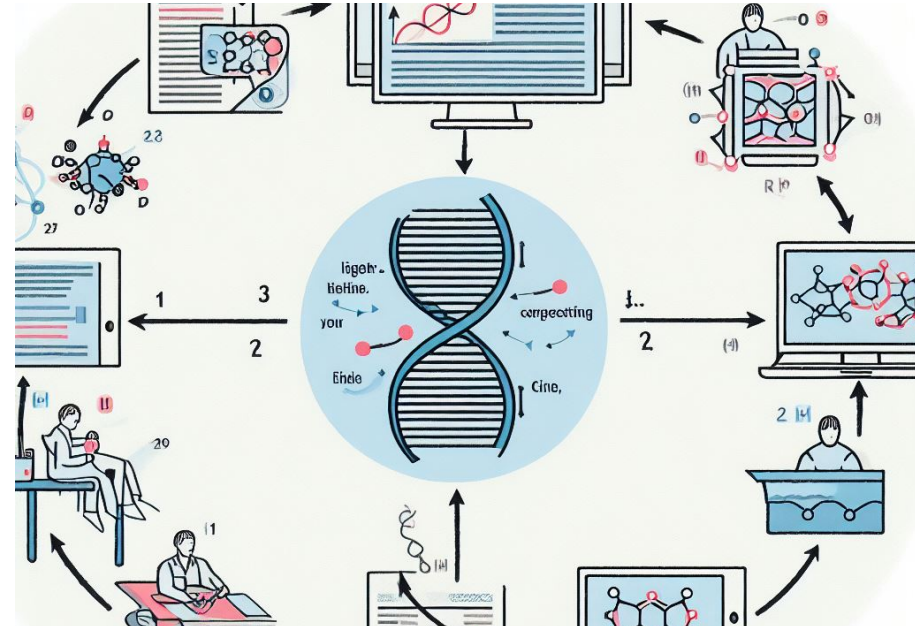
1. **Collecte de Textes** : Acquisition de données textuelles à partir de diverses sources
2. **Prétraitement** : Nettoyage, normalisation... transformation du texte pour faciliter l'analyse
3. **Extraction de caractéristiques** : Sélection des éléments importants pour la fouille de texte
4. **Modélisation 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](https://www.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 parcourant les étapes, de la collecte de texte à l'analyse et l'interprétation du réseau génique." (générée par [bing.com](https://www.bing.com) et optimisée par DALL-E 3)

# OUTILS PAR ÉTAPE

CORPUS, REPRÉSENTATION,  
NORMALISATION, EXPLORATION &  
VISUALISATION, MODÈLES

## 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.**

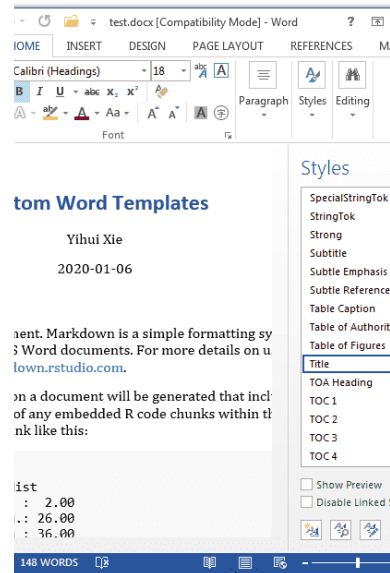
# Document textuel ?

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

Je la grande statue de Zeus, à Olympie, l'œuvre de Dieux. Entre le Soleil (Hélios) et les autres, groupées deux à deux, s'ordonnaient. Au centre de la frise, en surmonte, (masculine) qui président aux unions de la série de huit couples divins, il en est un, Hestia. Pourquoi les appariés ? Rien de la légende qui puisse justifier cette association (comme Zeus-Héra, Poséidon-Amphitrite) et sa sœur (comme Apollon-Artémis, Hécate Aphrodite-Éros), ni protectrice et jumelle. Quel lien unissait donc, dans l'esprit, ceux qui semblent étrangers l'un à l'autre, l'artiste personnel du sculpteur. Quand l'artiste ancien est tenu de se conformer à ce qu'exerce dans le cadre des schèmes imposés par son propre d'une déesse mais aussi non – se prêtait moins que les autres diesses anthropomorphes. On la voit rarement figurée, comme Phidias l'avait sculptée, mais elle est présente 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">
  <submenuitem text="New in Scribus 1.5" file="new_in_scribus_1.5">
  <submenuitem text="Specification" file="specification">
</area>
- <area text="Documentation" file="intro.html">
  <submenuitem text="Introduction" file="introduction">
  <submenuitem text="Editorial Notes" file="editorial_notes">
  <submenuitem text="About the Team" file="about_the_team">
</submenuitem>
- <submenuitem text="Setup" file="config">
- <submenuitem text="Configuring Scribus" file="configuring_scribus">
  <submenuitem text="Hyphenation and Spacing" file="hyphenation_and_spacing">
  <submenuitem text="Font Setup" file="font_setup">
  <submenuitem text="Fonts in Depth" file="fonts_in_depth">
</submenuitem>
- <submenuitem text="Scribus Basics" file="scribus_basics">
  <submenuitem text="Quick Start Guide" file="quick_start_guide">
  <submenuitem text="Command Line Reference" file="command_line_reference">
  <submenuitem text="Keyboard Shortcuts" file="keyboard_shortcuts">
  <submenuitem text="Mouse Shortcuts" file="mouse_shortcuts">
  <submenuitem text="Document Information" file="document_information">
  <submenuitem text="Working with Frames" file="working_with_frames">
  <submenuitem text="Working with Text" file="working_with_text">
  <submenuitem text="Text Properties" file="text_properties">
  <submenuitem text="Search and Replace" file="search_and_replace">
  <submenuitem text="Working with Styles" file="working_with_styles">
  <submenuitem text="Working with Images" file="working_with_images">
</submenuitem>
</menu>
```

**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 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 elementum purus venenatis maximus. Nulla rhon 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

```
15"lZ [Content_Types].xml <?xml:namespace prefix="o" ns="urn:schemas-microsoft-com:office:word" />
RÉ{V"i<º.æQUë-
8"fi3VΔE~/6l—µ%î=ñĀi^i7+Y0%&-d&.i0fiA..6ĀL4
fÉX"é*îèV$zç3N̄.3+ˆQñófi%p)0µ^†
l&ç5}nH"dŸs"XgîLĬN`Ÿâi|jÉ"ûi|óPèrÉ&Séé?òPŸ
T\yπ∞$N,Ŋ€†ŪU*%Ÿ'/-D/è€<òç×X°>0"(ŋçº<E,,ç)
ØF0Āº$ç<âòŋ<Fk> èh'y'æËŸ/úâ$€qûi£,,?Δñ0L
Éé"m"~PK!èŸ0N
_rels/.rels <?xml:namespace prefix="o" ns="urn:schemas-microsoft-com:office:word" />
@0É"É-0Q/iEN/c-€Ÿ[ILÉŸjç) "<ÿÿ)ÈaGĀ""i-zs
+^√[°ºxI0º•1x+p,
0É" f" #I) -ÉâŸæè~âð0zĀ*0*Ā&
i")æ"ç$≥É0qu0"ð~3†01'+jH[{"m=EæŪj€~
03-èè-ñ&]fTîiŋ2çj)1, l0/%úèb"
zÉ0&=â0, °——â/°|f\Z.ĀĀÉ760!ŸŸ_·oú)AŪ"~PK!
document.xml.rels <?xml:namespace prefix="o" ns="urn:schemas-microsoft-com:office:word" />
j•rĕòsæĀ6€æi äwâ' {8†»rĒâ0{ç×x0B0AkWĬ*è
1áµid:4e>†K/çèĒÉ4ΔVm°uāriĀw2NP00≈EVw15â;
~ò0òfi†É7o:>S!7p`âĀĒ8JX[dì0KĐĒĀEVKâ-âc2Bf
'0]æ0". .òΔ0=òsŋŸ"°0é+0ŸèüÈ(!0>z"~PK!T~
document.xml.rels <?xml:namespace prefix="o" ns="urn:schemas-microsoft-com:office:word" />
çñzIs&M`$H i,ŸŸY°Ī0æâèºi~ª
0iŋº", {hBπpY-2-0jçĀ07µâæüE_6
,±ĀââΔ€"æ?0â09ĀÉ|H$ ç6
i2<√XŸçĪ0 wm€ W601j0,)„,€i-•-(æĀŋB0°iH0!~H
Nŋ"æ"â[LŋŸ_ĀĒ'0k<\ŋĪ$,"içºmŪj"æ'Pm"•'R
```

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



- **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 ▾ Copier Légende ***  
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 (qt) /select

```

{
  "responseHeader": {
    "zkConnected": true,
    "status": 0,
    "QTime": 3,
    "params": {
      "q": "genre:Fantasy",
      "indent": "on",
      "wt": "json",
      "_: "1460656348479"
    }
  },
  "response": {
    "numFound": 82, "start": {
      "id": "/en/9_2005",
      "directed_by": ["Shane Ack",
      "initial_release_date": "2",
      "genre": ["Computer Animat",
      "Animation",
      "Apocalyptic and post-ap",
      "Science Fiction",
      "Short Film",
      "Thriller",
      "Fantasy"],
      "name": ["9"],
      "_version_": 1531518174029
    }, {
      "id": "/en/300_2007",
      "directed_by": ["Zack Snyder",
      "initial_release_date": "2",
      "genre": ["Epic film",
      "Adventure Film",
      "Fantasy",
  
```

Apache Solr

Runner Import Builder

POST http://localhost:9200/mkyong/posts/1001

1 Authorization 2 Headers (1) 3 Body 4 Request Script 5 Limits 6

```

{
  "title": "Java 8 Optional In Depth",
  "category": "Java",
  "published_date": "23-FEB-2017",
  "author": "Rambabu Posa"
}
  
```

Request 6

Body Cookies Headers (4) Tests

```

1 {
2   "_index": "mkyong",
3   "_type": "posts",
4   "_id": "1001",
5   "_version": 1,
6   "result": "created",
7 }
8   "shards": {
9     "total": 2,
10    "successful": 1,
11    "failed": 0
12  },
13  "created": true
  
```

Response 8

ElasticSearch

The aggregation pipeline is a framework for data aggregation modelled as pipelines. Documents enter a multi-stage pipeline that transforms the data.

Collection

```

b.orders.aggregate( [
  $match stage -> { $match: { status: "A" } },
  $group stage -> { $group: { _id: "$cust_id", total:
  ] }
  
```

```

{
  "cust_id": "A123",
  "amount": 500,
  "status": "A"
}
{
  "cust_id": "A123",
  "amount": 250,
  "status": "A"
}
{
  "cust_id": "B212",
  "amount": 200,
  "status": "A"
}
{
  "cust_id": "A123",
  "amount": 300,
  "status": "B"
}
  
```

orders

The aggregation pipeline provides an alternative to map-reduce and map aggregation tasks where the complexity of map-reduce may be unwarranted. Aggregation pipelines have become fundamental for value types and user-defined

MongoDB

Getting Started / What is a Graph Database?

connected world, and understanding most domains requires processing rich sets of connections to an happening. Often, we find that the connections between items are as important as the items themselves.

name: "Dan"  
May 29, 1970  
er: "@dan"

name: "An"  
born: Dec 5,

LOVES

LOVES

LIVES WITH

DRIVES

OWNS

since: an 10, 2011

brand: "Volvo"  
model: "V70"

Person

Person

Car

people do this today? While existing relational databases can store these relationships, they navigate JOIN operations or cross-lookups, often tied to a rigid schema. It turns out that "relational" databases poorly. In a graph database, there are no JOINs or lookups. Relationships are stored natively: relationships have degree (importance) and value types and user-defined

neo4j

# Constitution de corpus : plateformes



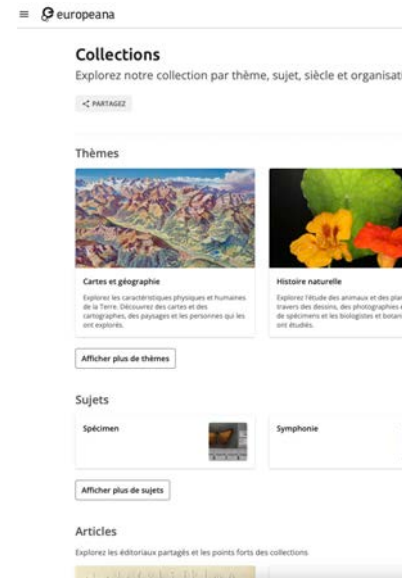
**Gallica**: La bibliothèque numérique de la Bibliothèque Nationale de France. Visitez : <https://gallica.bnf.fr/>



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



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



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

# Représentations de textes

4. Generating vectors → firstly we create zero vector and append the counts of words to it and we get vector counts

vector Visualization of Bag of words



as we can easily observe, Bag of words is just counting of all significant words in a text.

types of Vectorizer

**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.

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

Sentence 1: The car is driven on the road.

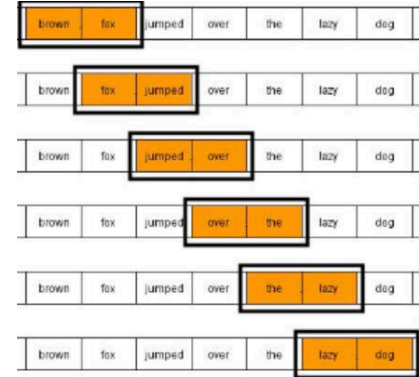
Sentence 2: The truck is driven on the highway.

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	$\log(2/2) = 0$	0	0
Car	1/7	0	$\log(2/1) = 0.3$	0.043	0
Truck	0	1/7	$\log(2/1) = 0.3$	0	0.043
Is	1/7	1/7	$\log(2/2) = 0$	0	0
Driven	1/7	1/7	$\log(2/2) = 0$	0	0
On	1/7	1/7	$\log(2/2) = 0$	0	0
Road	1/7	0	$\log(2/1) = 0.3$	0.043	0
Highway	0	1/7	$\log(2/1) = 0.3$	0	0.043

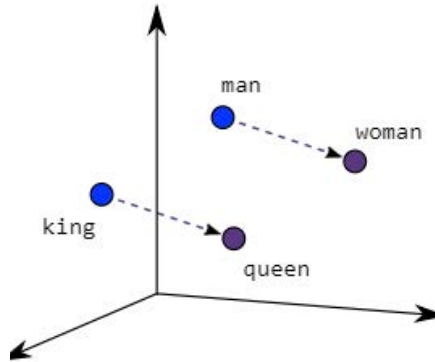
**TF-IDF (Term Frequency-Inverse Document Frequency)** : Une mesure statistique qui évalue l'importance d'un mot dans un document dans le contexte d'un corpus de documents.



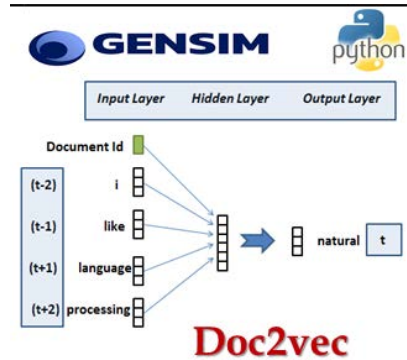
**N-grams** : Une approche qui prend en considération l'ordre des mots. Un N-gram est une séquence continue de N éléments dans un texte donné.

# Représentations de textes

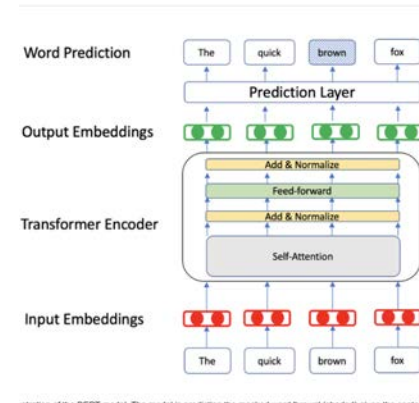
word embeddings are plotted, similar



**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 transformeurs, capable de capturer de plus longs contextes et de traiter la compréhension de la langue à un niveau plus profond.

**⚠ Chaque type de représentation a ses propres avantages et inconvénients, le choix dépend donc des besoins spécifiques de l'analyse de texte à réaliser. ⚠**



## 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)

**⚠ Chaque outil a ses 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

Use network visualization to partition data to make meaningful network representation. Customize colors, size or label representation. The vectorial preview module lets you put the final touch and care about aesthetics

Exportable PDF, SVG and PNG export

6 presets



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

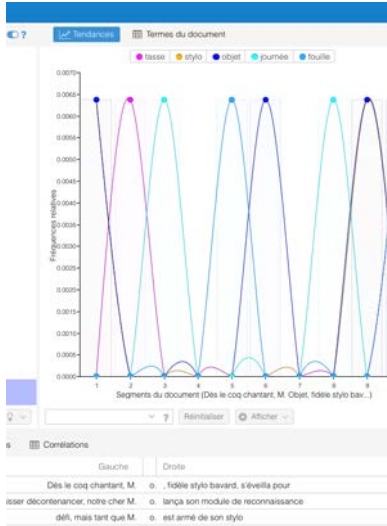


Topic 4, which is about the fish, seems to be quite well-represented in the corpus.



**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/>

## MINER LITE – FREE QUALITATIVE DATA ANAL

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

- Importation of documents from plain text, RTF, HTML, PDF as well as data stored in limited text files
- Importation from other qualitative coding software such as Atlas.ti, HyperReels like Transana and Transcriber as well as from Reference Information Systems
- Intuitive coding using codes organized in a tree structure.
- Ability to add comments (or memos) to coded segments, cases, or the whole project
- Free Boolean text search tool for retrieving and coding text segments.
- Free frequency analysis with bar chart, pie chart, and tag clouds.
- Free coding retrieval with Boolean (and, or, not) and proximity operators (includes support for XLS, Tab Delimited, CSV formats, and Word format)
- Free export graphs to BMP, PNG, JPEG, WMF formats.
- Free single-file (\*.qdp) project format.
- Free 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 Di

Team Members: John Stawko, Carsten Görg, Zhicheng Liu, Sakshi Prasad, Anand Sainath, Abhinav Meekal Bajaj, Alex Humstedt, Mohit Jain, Younsuh Kang, Jeeyoung Kim, Vasil Papatzopoulos, Neel P



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

Desktop (java-based) Jigsaw  
Download Now  
0.54 Release  
March 4, 2016

LiveView on the web  
The stand-alone web-based LiveView for your csv, txt, and json data

**ACTIVE**  
**DOWNLOADS**  
Tutorial Videos - Instructional videos for Talk slides - Slides from talk about Jigsaw RFI Report Video (2013) - Browsing the InfoVAST Papers Video (2012) - Aca Car Reviews Video (2012) - Reviews of Wine Reviews Video (2011) - R. Parker - CHI Papers Video (2010) - CHI conf pap BioJigsaw Video (2010) - Genomics app Tutorial Video (2009) - More details about Overview Video (2007) - Introduction to:

**PAPERS**  
Full papers:  
Reflections on the Evolution of the J Combining Computational Analyses Visual Analytics Support for Intelligence Informing design of visual analytics : Examining the Use of a Visual Analyt How Can Visual Analytics Assist Iniv Visualization and Language Process Combining Computational Analyses Evaluating Visual Analytics Systems Jigsaw: Supporting Investigative An Evaluating Visual Analytics at the 20 Jigsaw: Supporting Investigative An

**Abstracts:**  
Jigsaw: Visualization for Investigativ Jigsaw to Save Vastopolis - VAST 20 Data Ingestion and Evidence Marsha Jigsaw meets Blue Ignanodon - The Visual Analytics with Jigsaw (Invited)

Using Jigsaw, we won the university:

**Summary**  
Investigative analysts and researchers acquire clues and connect small bits of evidence to bits of evidence are short text documents or spreadsheets, and analysts must examine far that may occur in the future. As the number of documents to examine rises, it becomes m

We have created Jigsaw, a visual analytics system to help analysts and researchers bett and accurate understandings of the larger stories and important concepts embedded the particularly focus on presenting the identifiable important entities (people, places, organi



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

- **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.

 **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>

 **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

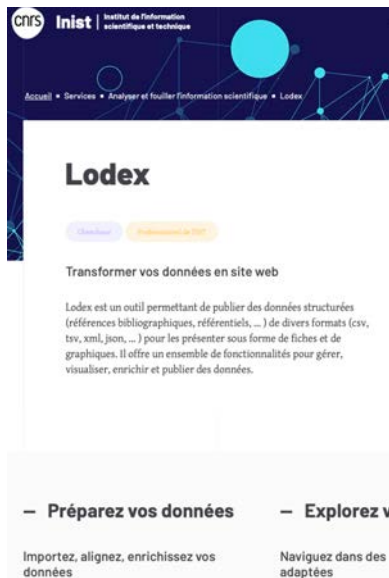
- **Régression logistique** : Un modèle de classification souvent utilisé pour prédire des classes binaires. Pour utiliser, visitez: [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
- **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/>

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

Et sans programmation ?...

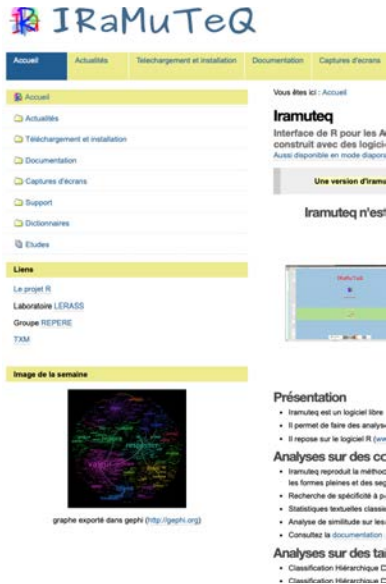
- **RapidMiner: Plateforme d'analyse de données et d'apprentissage automatique. Visitez: <https://rapidminer.com/products/studio/>**
- **KNIME: Plateforme d'analyse de données open-source. Visitez: <https://www.knime.com/>**
- **Orange: Un outil de visualisation et d'analyse de données open-source. Visitez: <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 😊



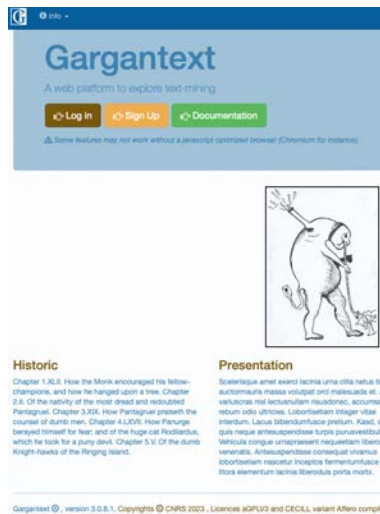
The screenshot shows the LODEX website. At the top, there's a navigation bar with 'Accueil', 'Services', 'Analyser et fouiller l'information scientifique', and 'Lodex'. The main heading is 'Lodex' with sub-headings 'Analyser' and 'Visualiser les données'. Below this, it says 'Transformer vos données en site web'. A paragraph describes LODEX as a tool for publishing structured data. At the bottom, there are two columns: 'Préparez vos données' (Importez, alignez, enrichissez vos données) and 'Explorez vos données' (Naviguez dans des données adaptées).

LODEX



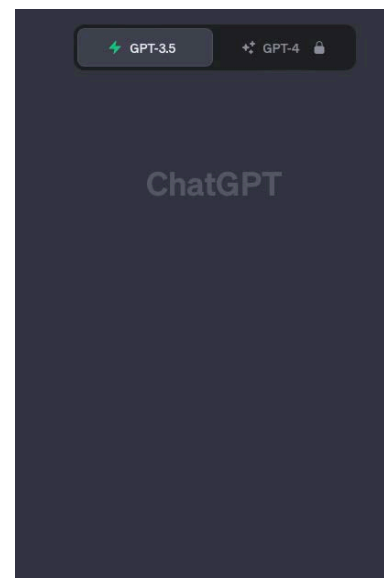
The screenshot shows the IRaMuTeQ website. It has a navigation bar with 'Accueil', 'Actualités', 'Téléchargement et installation', 'Documentation', and 'Captures d'écrans'. The main heading is 'IRaMuTeQ' with the subtitle 'Interface de R pour les Analyses de Réseaux de Mots'. Below this, there's a section 'Iramuteq n'est' with a small image and a list of features: 'Iramuteq est un logiciel libre', 'Il permet de faire des analyses', 'Il repose sur le logiciel R (via les formes plates et des seq)', 'Recherches de spécificité à priori', 'Statistiques textuelles classiques', 'Analyse de similarité sur les mots', and 'Consultez la documentation'. There's also a section 'Analyses sur des textes' with 'Classification Hiérarchique C' and 'Classification Hiérarchique G'.

IRaMuTeQ



The screenshot shows the Gargantext website. It has a navigation bar with 'Log In', 'Sign Up', and 'Documentation'. The main heading is 'Gargantext' with the subtitle 'A web platform to explore text mining'. Below this, there's a section 'Historic' with a paragraph about the Monk and a section 'Presentation' with a paragraph about the Monk's adventures. There's also a small illustration of a man with a large head and a small body.

Gargantext



The screenshot shows the ChatGPT interface. At the top, there's a navigation bar with 'GPT-3.5' and 'GPT-4'. The main heading is 'ChatGPT'.

ChatGPT



# DISCUSSION

OUTILS CLÉ-EN-MAIN VS.  
PROGRAMMATION, EXEMPLES

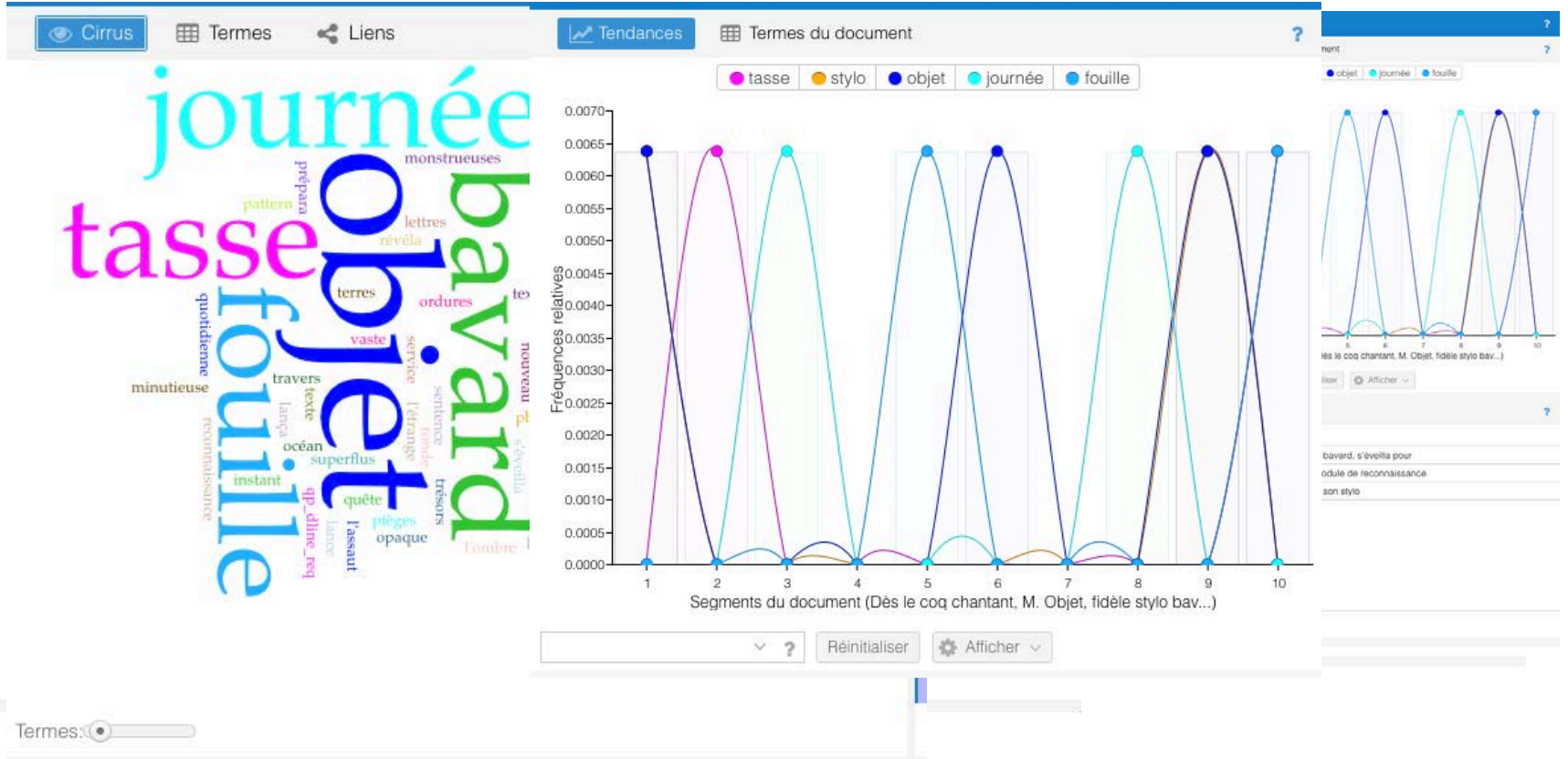
- **Outils Clé-en-main (e.g., RapidMiner, KNIME, MonkeyLearn):**
  - **Avantages:**
    - **Faciles à utiliser, pas besoin de compétences en programmation**
    - **Idéals pour des tâches simples et standard**
    - **Permettent souvent d'obtenir rapidement des résultats**
  - **Inconvénients et Limites:**
    - **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 (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**

 **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

### Importer

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

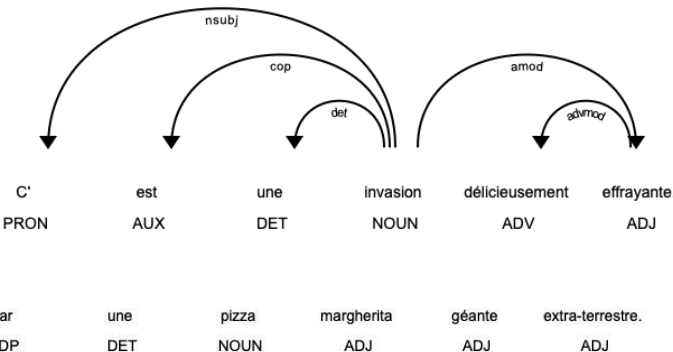
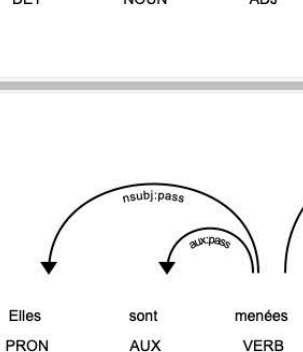
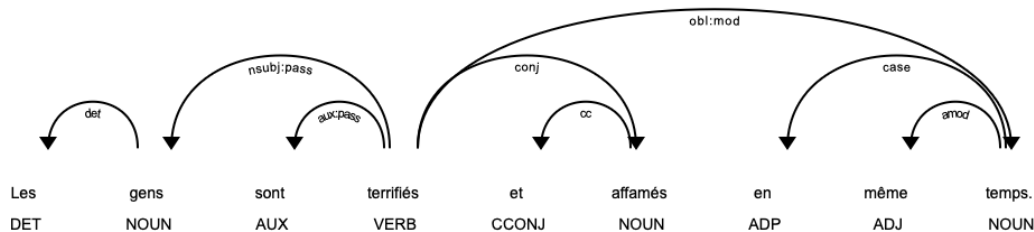
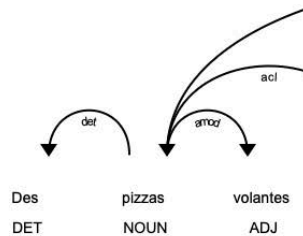
```
[1]: import spacy
from spacy.lang
from spacy impo
import matplotlib
from collection
import warnings
warnings.filter
```

### Charger l

```
[2]: nlp = spacy.loa
# Cell 3: Défin
text = "Des pizzas
Elles sont mené
Les gens sont t
doc = nlp(text)
```

### Analyse

```
[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)  
[1]         print(f
```

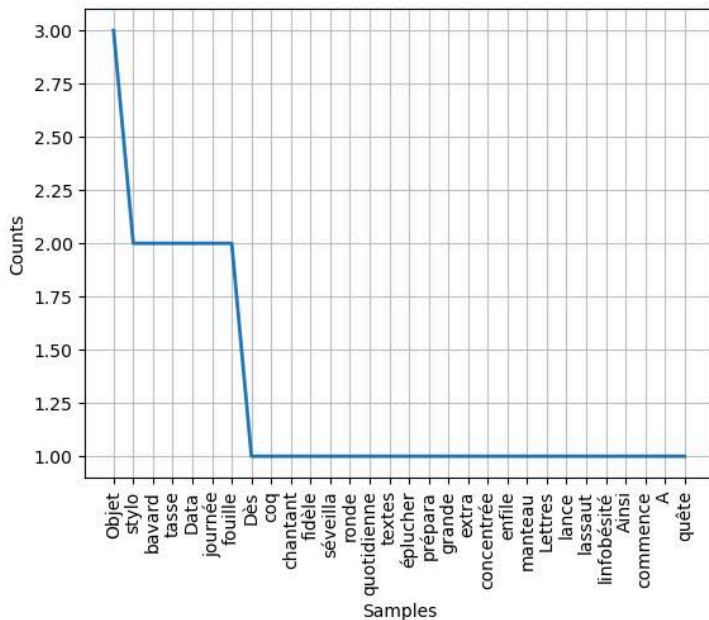
```
['Dès',  
a', 'ex  
s', 'tr  
'fautes  
s', 'la  
itable'  
t', 'ar
```

```
[2]
```

## Anal

```
[5]: frequen  
print(f  
most_co  
for wor  
pri
```

```
<FreqDi  
Objet:  
enne: 1  
sité: 1  
1 | ord  
| océan  
| lança  
nder: 1  
| texte
```



# Exemple : tm - nuage de mots (R)

## Installer

```
[1]: install.packages("tm")
      Updating HT
      Making 'pac
      done
```

## Charger

```
[2]: library(tm)
      library(wordcloud)
      library(RColorBrewer)
      Loading req
      Loading req
```

## 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"))
```

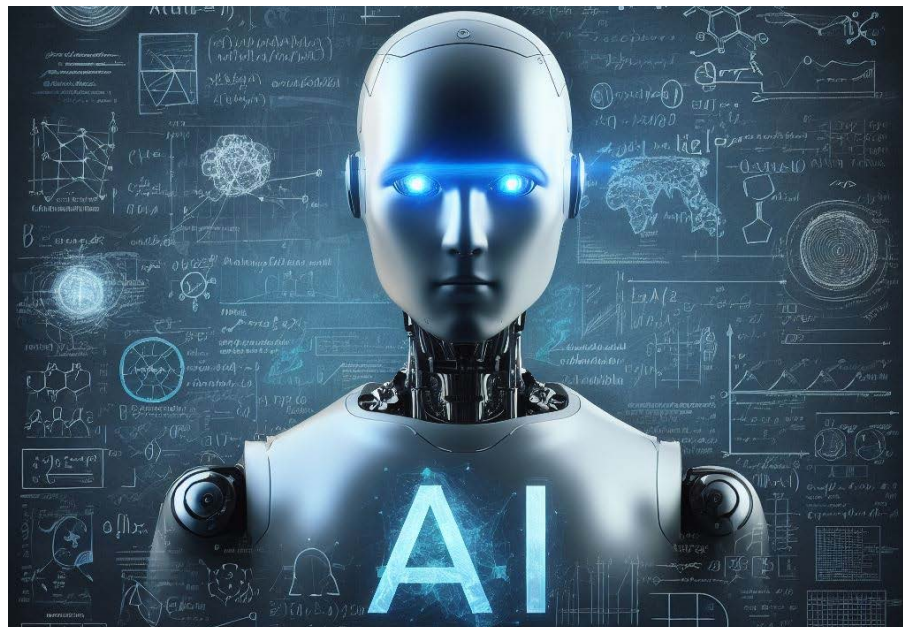


# IA GÉNÉRATIVE, LLM

EXPLICATIONS, EXEMPLES, POINTS  
DE VIGILANCE

## 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://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



 **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.

 **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 : "Dès le coq chantant, M. Objet, fidèle stylo bavard, s'éveilla pour sa ronde quotidienne des textes à éplucher. Il se prépara une grande tasse de Data extra concentrée, enfile son manteau de Lettres et se lance à l'assaut de l'infobésité.

Ainsi commence sa journée. A la quête des trésors cachés à travers des terres désolées d'emails superflus, combattant des ordures textuelles monstrueuses et évitant les pièges astucieux des fautes d'orthographe, il fouille sans fin le vaste océan d'informations.

De l'ombre émerge une sentence opaque, 'RR143/QP\_Dline\_Req.12.4'. Sans se laisser décontenancer, notre cher M. Objet lança son module de reconnaissance de pattern. En un instant, l'étrange phrase cryptique révéla sa véritable identité : 'Demander une extension de ligne pour le service client le 12 avril!'

Chaque journée apporte de nouveau un défi, mais tant que M. Objet est armé de son stylo bavard et de sa tasse de Data, aucun texte ne peut échapper à sa fouille minutieuse !"



M. Objet, qui est personifié comme un analyste de données textuelles, commence sa journée en se battant contre l'infobésité. Il cherche des informations précieuses cachées dans des tonnes d'e-mails inutiles, évite les pièges des erreurs d'orthographe et décode des phrases cryptées comme 'RR143/QP\_Dline\_Req.12.4'. Malgré ces défis quotidiens, armé de sa capacité d'analyse et de compréhension des informations, il parvient toujours à accomplir son travail.

## Résumé automatique



# Hallucinations et limites

 Quelle est la capitale de la Roumanie en 2023 ?

 La capitale de la Roumanie reste Bucarest, même en 2023, sauf changement majeur imprévu dans le pays qui n'est pas prévisible à l'heure actuelle.


 

 Es-tu sûr ?

 Oui, en l'absence de tout changement majeur imprévu, Bucarest restera la capitale de la Roumanie en 2023. Les capitales de pays ne changent généralement pas sans circonstances exceptionnelles.

 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 13 n'est pas un nombre premier.

 En fait, 13 est un nombre premier. Un nombre premier est un nombre qui n'a que deux diviseurs distincts : 1 et lui-même. En d'autres termes, un nombre premier est un nombre qui ne peut être divisé que par 1 et lui-même sans restes. En ce qui concerne 13, il ne peut être divisé que par 1 et 13, donc c'est bien un nombre premier.

 Explique pourquoi 43 n'est pas un nombre premier.

 En réalité, 43 est bien un nombre premier. Un nombre premier est un nombre qui n'a que deux diviseurs distincts : 1 et lui-même. Un nombre est appelé un nombre premier s'il a exactement deux facteurs positifs distincts. Dans le cas du numéro 43, les seuls facteurs sont 1 et 43. Par conséquent, 43 est un nombre premier.

 Explique pourquoi 1187 n'est pas un nombre premier.

 Le nombre 1187 n'est pas un nombre premier car il peut être divisé par d'autres nombres en plus de 1 et lui-même. Un nombre premier ne peut être divisé que par 1 et lui-même. Cependant, 1187 peut être divisé par 13 et 91 ( $13 \times 91 = 1187$ ), ce qui signifie qu'il a plus de deux facteurs. Ainsi, 1187 n'est pas un nombre premier.

Certaines explications (surtout mathématiques...)

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

 Le Brésil est le plus grand producteur mondial de mangues en raison de multiples facteurs :

1. Climat : Le Brésil a un climat tropical propice à la culture des mangues.
2. Superficie agricole : Le Brésil a une grande surface de terre arable, ce qui lui permet de cultiver de grandes quantités de mangues.
3. Saisons prolongées : Le Brésil ayant des régions ayant des saisons de croissance décalées, cela assure une production continue de mangues tout au long de l'année.
4. Technologie et infrastructures agricoles : Le 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

## 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.**

💡 **Avec l'essor de l'ère des données, la capacité à extraire et à exploiter des informations à partir de textes constitue un atout majeur. 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-4, oct. 2023***

## Show must go on...

Aa Nom	🔗 Lien	☰ Accès	☰ Commentaires
<b>Prodigy</b>	<a href="https://prodi.gy/">https://prodi.gy/</a>	payant	annotation
<b>Brat</b>	<a href="http://brat.nlplab.org/index.html">http://brat.nlplab.org/index.html</a>	open-source	annotation
<b>Label Studio</b>	<a href="https://labelstud.io/">https://labelstud.io/</a>	open-source	annotation
<b>PACTE</b>	<a href="http://pacte.crim.ca/">http://pacte.crim.ca/</a>	accès sur demande	annotation collaboration
<b>HTRC Analytics</b>	<a href="https://analytics.hathitrust.org/datasets#top">https://analytics.hathitrust.org/datasets#top</a>	accès sur demande accès institutionnel	analyse visualisation
<b>CORLI (huma-num)</b>	<a href="http://explorationdecorpus.corpusecrits.huma-num.fr/">http://explorationdecorpus.corpusecrits.huma-num.fr/</a>	accès libre	exploration de corpus recensement d'outils
<b>Nooj</b>	<a href="https://nooj.univ-fcomte.fr/index.html">https://nooj.univ-fcomte.fr/index.html</a>	open-source	ressource linguistique analyse annotation
<b>IRaMuTeQ</b>	<a href="http://www.iramuteq.org/">http://www.iramuteq.org/</a>	open-source	analyse visualisation
<b>CORLI "Corpus écrits"</b>	<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.com](https://bing.com), optimisée par DALL-E 3)