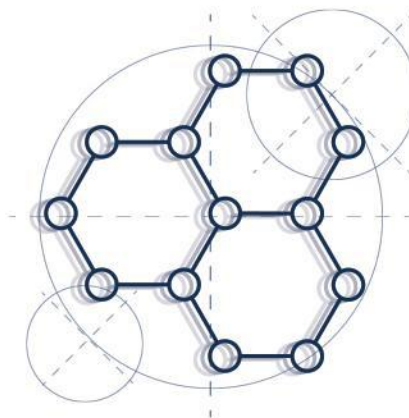


EI RECHERCHE INFORMATION – Introduction au Moteur de Recherche



HeadMind Partners

AI & BLOCKCHAIN





Headmind Partners

Qui sommes-nous ?

ACTEUR MAJEUR DU CONSEIL. 100% INDÉPENDANT FONDÉ EN 2000.

1500

collaborateurs
Monde

130

M€ CA
(+10%/an)

+160

Clients
Grands Comptes

Paris
Bruxelles



*Engagé auprès de ses clients dans la
construction
d'une société numérique sécurisée et
responsable.*



HeadMind Partners
AI & Blockchain

Valoriser les **données** pour
piloter la stratégie Business

HeadMind Partners
Cyber Risk & Security

Maîtriser et sécuriser
sa **transformation** numérique

HeadMind Partners
Digital

Placer le numérique au cœur
des **opérations**

Une organisation 360°

DATA LAB

- Tests de modèles à l'état de l'art
- R&D (NLP, Computer Vision, Time Series, etc.)
- Création de produits

IA LIFE CYCLE

- Conception
- Industrialisation
- Cloud Computing
- MLOPS (CI/CD)
- Monitoring IA

WATCH TOWER

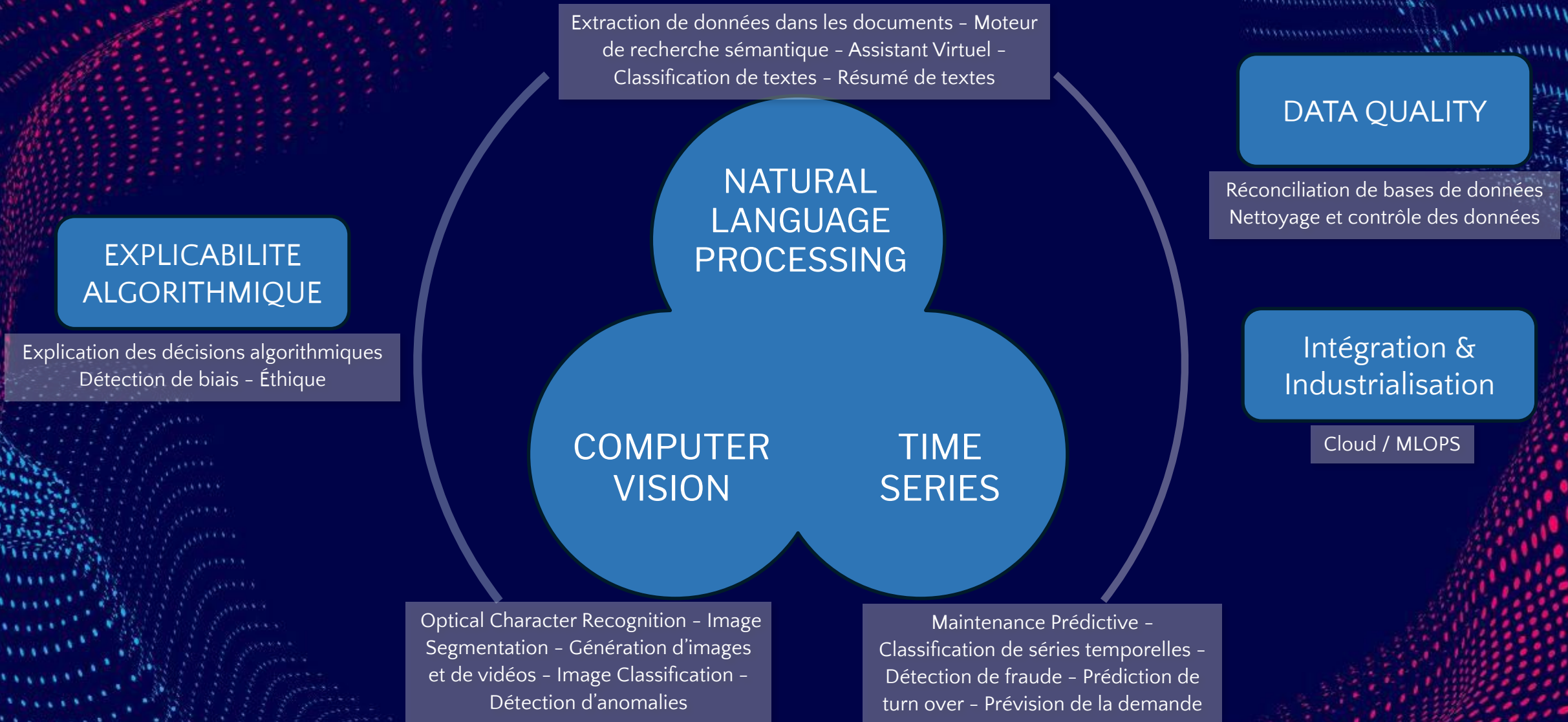
- RetEx IA
- Règlementation IA/ Impacts/ Normes/ Cybersécurité
- Acculturation et cas d'usage IA

HeadMind Partners
AI & Blockchain

Notre mission : Conseiller les entreprises dans leur transition IA. Travailler sur la conception, l'industrialisation et le monitoring des projets IA.

40 consultants
5 managers

Nos expertises IA



Nos 6 produits phares

Fondé sur l'innovation et l'expertise, notre cabinet de conseil spécialisé en IA & Blockchain fait du potentiel des données des entreprises un levier de croissance et de différenciation.

Virtual Assistant

répond aux questions, détecte des intentions et déclenche des processus

Talent Mind

analyse les données RH pour prédire et expliquer le turnover

Mind Search

explore différentes sources de données de manière sémantique et multilingue

HeadMind Partners
AI & Blockchain

Mind Scan

transforme vos images ou documents faxés en données valorisées

Anti Money Laundering

parcours les transactions et diagnostique les fraudes et anomalies

Maintenance Prédictive

prévient les pannes et incidents et vous informe en temps réel

Nos modes d'intervention et nos prestations pour s'adapter à vos enjeux



FORFAIT

- Prestation sur mesure
- 5 offres packagées



REGIE

- AI Analyste
- ML Engineer
- MLOps
- Data Engineer
- ☐ Support conseil et technologique par
Un manager expert si nécessaire



SOLUTIONS SUR ETAGERES

- Déploiement de nos produits phares dans votre environnement en licence OneShot



CENTRE DE SERVICES

- De 3 à 15 consultants
- Optimisation de vos processus
- Unité d'Œuvre (UO)

Le **CONSEIL IA** pour construire une stratégie IA cohérente avec les ambitions de votre groupe.

Nos **AUDITS** pour renforcer vos solutions data / IA et vous assurer que vos usages IA répondent aux normes.

Nos **FORMATIONS** pour faire monter en compétences vos équipes métier et data sur de nouvelles expertises.

Nos références clients



AIRBUS



CLARINS



BNP PARIBAS



TRACTEBEL





Introduction to search engines

Introduction

Objective:

Make your own search engine on a real web-based data source.

Context:

➤ Web Data : Stack Exchange Forum







Schedule :

- **Day 1:** Exploring and Understanding Data
- **Day 2:** Creation of a search engine in Python
- **Day 3 and 4:** Improvement of the search engine by adding semantic features
- **Day 5:** Preparation of the report and other deliverables, oral presentations

Teams

Group	Member	Mail	Repo
1			
1			
1			
1			
2			

Coding Meteo

	Search Engine	Python	Pandas	NLP	Coding Environment (Colab, IDE, requirements)	Project Management (git, gitlab, readMe...)
Meteo						
Comments				Which Libraries ?		

Others :



Day 1



Vivien Gillo

AI Consultant

NLP Team

Sectors:

- Bank
- Transports



Florent Villenave

AI Consultant

Time Series Team

Sectors:

- Defense
- Telecom

Nice to meet you

HeadMind Partners

Data Exploration

Session objectives:

- ❖ Understand the data
- ❖ Get first insights for creating the Search Engine

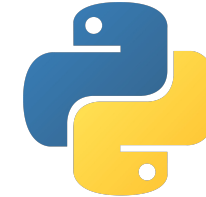
Data:

⇒ Stack Exchange Forum

Topics of the day :

- ❖ SQL Querying
- ❖ Data Extraction
- ❖ Data Exploration
- ❖ Data Visualisation

Technos :





Day 2



Vivien Gillo

AI Consultant

NLP Team

Sectors:

- Bank
- Transports



Florent Villenave

AI Consultant

Time Series Team

Sectors:

- Defense
- Telecom

Nice to meet you

HeadMind Partners

Architecture of Search Engine

1) Processing Data:

- ❖ Extraction
- ❖ Cleaning

2) Indexation

- ❖ By Document
- ❖ Inverted
- ❖ Save and Load

3) Searching

- ❖ By word
- ❖ Boolean
- ❖ Probabilistic

4) Ranking

5) Querying

6) Evaluation

- ❖ NDGC
- ❖ Propositions

01 Indexation

- Indexing is the process by which search engines organize information before a search to enable super-fast responses to queries.
- Index is a new representation of the content, that allows to retrieve and rank information faster than simply just scanning the entire content.
- It exists as many indexing as people on Earth, depending on the information structure and content. Yet, some indexes are the most common :
 - Full-text Index : In this type of indexing, the search engine analyzes and indexes the entire text content of web pages. It extracts and stores keywords, phrases, and other textual elements to facilitate search and retrieval based on text matching algorithms.
 - Inverted Index : It is designed to map terms or keywords to the documents or web pages in which they appear. The name "inverted" stems from the fact that the index inverts the relationship between terms and documents compared to a traditional forward index.
 - Term-Frequency Index : Based on the inverted index, each word is associated to the documents in which they appear, paired with the frequency of that word in the given document. This method give more information than the classic inverted index, which is the frequency of the word in the document, thus the importance of that word in the document.
 - TF-IDF Index : Same thing as the Term-Frequency Index, changing the Term-Frequency value by the TF-IDF.

02 Search

- The search step is a crucial part of any search engine, enabling users to find relevant information from vast amounts of indexed data.
- A well-executed search step helps users find the most relevant information by considering various factors such as keyword matching, document relevance, and user intent. Search engines utilize advanced algorithms to rank and present the most pertinent results at the top, enhancing the user's search experience.
- There are many ways to implement the search step in a search engine, according to the type of index used. The most common are :
 - The search by count : It counts the number of words in common between the query and the documents, and ranks them in descending order. The Term-frequency index makes it easier to count the number of occurrences, while the TF-IDF index can be used to weight the relevance of words.
 - The Boolean Search : Boolean search is a search technique that involves using logical operators to combine or exclude search terms. These operators include AND, OR, and NOT. Boolean search is precise but may require users to have a good understanding of the search domain and construct effective queries.
 - The Probabilistic Search : Probabilistic search, also known as ranked retrieval, is a search technique that ranks documents based on their relevance to a query. It relies on statistical algorithms that calculate the probability of a document's relevance to the search terms. The ranking process considers factors such as term frequency, document length, and the overall significance of the query terms within the document.

Infos

- Assert .sort : renvoie None
- Loi de Zipf : faire après la création de l'index inversé
- Avoir une idée de son moteur de recherche
- Ttable (boolean search) : compatible avec python 3.6, 3.7 et 3.8
- Livrables :
 - Oral : démonstration
 - Rapport : PDF→ Modification du PDF dans le GitLab aujourd'hui
- Evaluation du moteur de recherche : présentation de la méthode
- Notebook Final

03 Evaluation

Notebook : Search_Engine.ipynb

Fichier Excel : evaluation_search_engine_post_queries_ranking_EI_CS.xlsx

PostId	Titre	Query : "..."
123	XXX	2
124	YYY	
125	ZZZ	1

2nd document le + pertinent

Non prise en compte des documents impertinents

Document le + pertinent

Calcul du Score via un NDCG (Normalized Discounted Cumulative Gain)

Limites de l'approche : à vous de proposer des améliorations, des adaptations ou d'autres méthodes.



Day 3 & 4



Mehdi Arsaoui

AI Consultant

NLP Team

Secteurs:

- Media
- Finance



Elise Barrère

AI Consultant

NLP Team

Secteurs:

- Bank
- Cybersecurity

Nice to meet you

HeadMind Partners

01 Data Preprocessing

Before using data for analysis or prediction, it is important to preprocess it. To prepare the text data for model development, we perform text preprocessing. This is the **very first step** in NLP projects. Some of the preprocessing steps are: removing punctuation, special characters, stopwords, tokenization... In the context of NLP use cases, the following processing techniques are very useful:

Document parsing

A more difficult task than it seems is the transformation of different types of documents or files (.docx, .pdf, .xml, ...) into template-readable text.

Autocorrection

In a context where textual inputs are sent by users, it can be useful to analyze them and automatically correct spelling errors using semantic methods.

Data cleaning

A common task for any ML model development. You want your data to be easily readable for the model development. In the case of a NLP task, you might want to remove punctuation, special characters, non-relevant words for example.

Tokenization

Once your data is ready to be processed, the real work starts now. Tokenization is the process of splitting your inputs in a list of tokens which will be processed by your model.

This apple was amazing.

□ [This, apple, was, amazing]

Stemming

Stemming is the process of reducing a word to its root form by chopping off the of the word without taking into consideration the context of the word.

[This, apple, was, amazing]

□ [Thi, appl, wa, amaz]

Lemmatization

Lemmatization is the process of reducing a word to its root form based on the dictionary definition.

[This, apple, was, amazing]

□ [This, apple, be, amazing]

02 Vectorization

Word vectorization or word embeddings is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics. After the words are converted to vectors, we need to use some techniques such as Euclidean distance, Cosine Similarity to identify similar words. It is possible to also work at document level.

Bag-of-words

A bag-of-words model is a way of extracting features from text by using a known vocabulary. The representation describes the occurrence of words within the document.

N-Grams Extraction

N-Grams methods are a generalization of bag-of-words. Instead of analyzing one word (or token), one analyzes a group of N words.

TF-IDF

TF-IDF method re-weights the count features of the representation given by a bag of words according to the frequency of a given word in the document and the frequency of that word in all documents.

Word2Vec

Word2Vec is a semantic algorithm trained to reconstruct the context of words. The advantage is that it only trains on the analyzed documents, allowing a better understanding of the context. The two types of word2Vec algorithms are **Common Bag-of-Word** and **Skip-grams**.

Sentence Transformers

Sentence-BERT is an embedding model based on BERT. It uses siamese and triplet networks that is able to derive semantically meaningful **sentence embeddings**. SBERTs are pretrained algorithms.

03 Clustering

Text clustering consists in grouping texts with similarities into clusters. In the context of a search engine, clustering can be useful in order to reduce the field of possibilities during the search step, by searching only in the input cluster. In that case, an **online** model must be used in order to identify the input cluster without recomputing the clustering.

To extend this method, **topic modeling** should be considered. More than clustering, the topic modeling allocates a topic to each document, facilitating the understanding of these documents, and allowing to reduce the list of documents considered during the search step.

KMeans

Most common clustering algorithm, used in any context. Centroid-based, which allows new inputs.

Latent Dirichlet Allocation

Topic modeling algorithm that allows soft clustering. Soft clustering means that the LDA does not allocate an input to a cluster, but gives a probabilistic score for each identified cluster. This decomposition allows to identify topics within the documents.

BERTopic

BERTopic is a topic modeling algorithm based on BERT. It consists of different steps: dimensionality reduction, clustering and cluster tagging. For each step, several methods are possible.

04 Model improvements

Once your data is ready to use, you will need to perform similarity computations in order to make your searches. However, these similarity calculations highly depends on the methods you have used before, each having its advantages and disadvantages. To improve your results, you can use alternative methods such as n-grams, clustering, topic modeling, or a mix of several methods.

For example, you might want to perform a mix of statistical and semantic method as vectorizing. If this is the case, it is important to understand how to merge these methods. (Is an 'average' of both embeddings relevant ?)



Day 5



Mehdi Arsaoui

AI Consultant

NLP Team

Secteurs:

- Médias
- Finance



Valentin Gorce

AI Senior Consultant

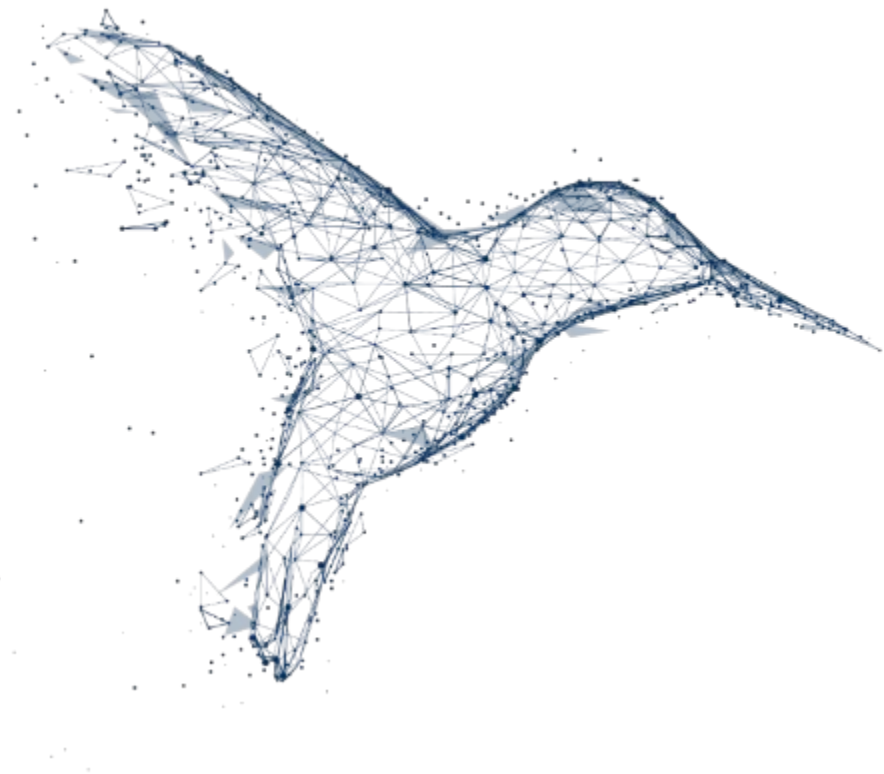
Time Series Team

Secteurs:

- Industrie

Nice to meet you

HeadMind Partners

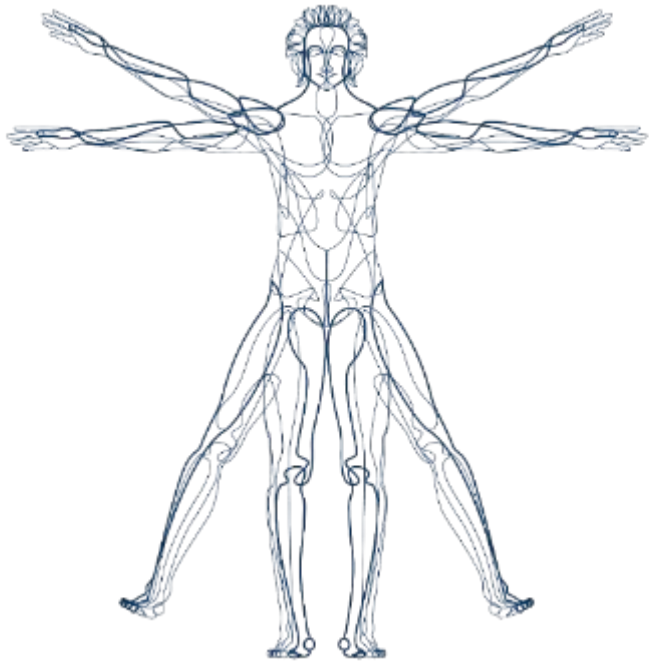


Speakers

Attendance

	Vendredi 26/05	Mardi 30/05	Mercredi 31/05	Jeudi 01/06	Vendredi 02/06	Contact
Valentin Gorce						vgorce320@headmind.com
Vivien Gillo						vgillo155@headmind.com
Elise Barrère						ebarrere220@headmind.com
Mehdi Arsaoui						marsaoui463@headmind.com
Florent Villenave						fvillenave318@headmind.com

Attendance Schedule : 9h – 17h

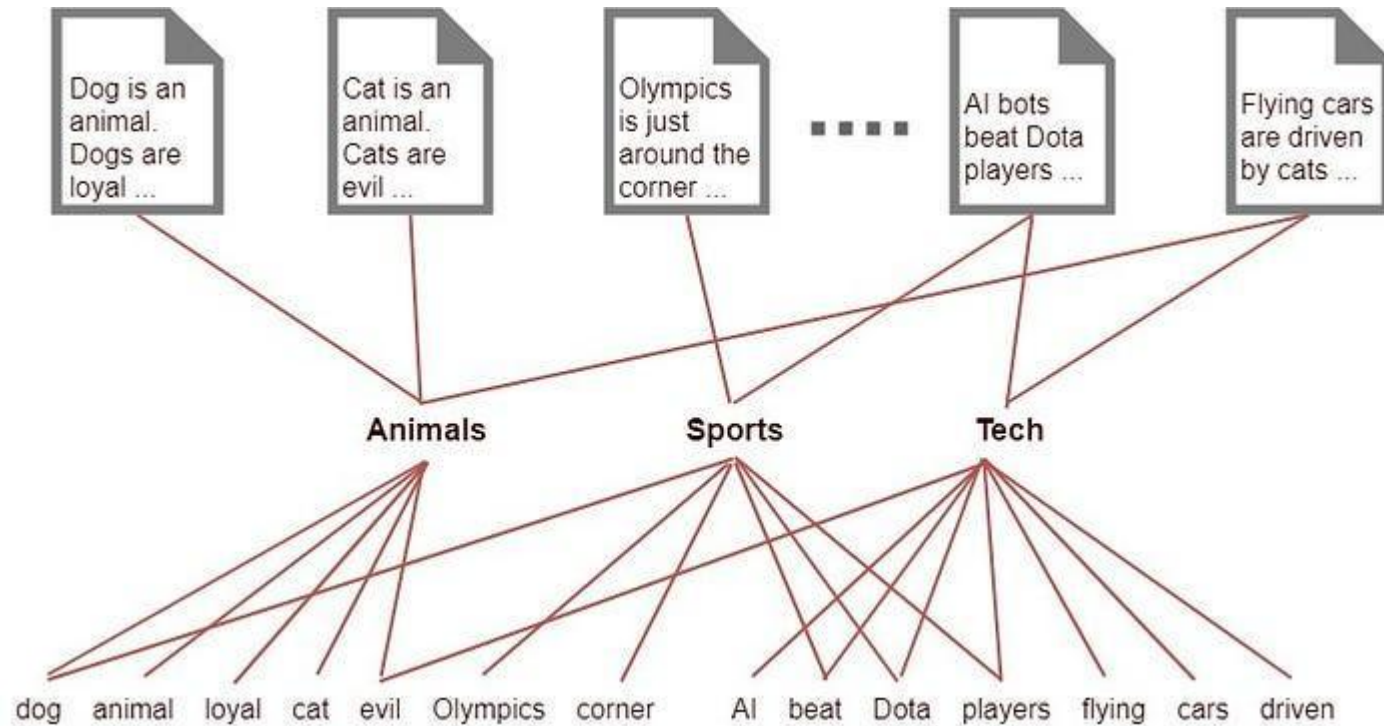


Appendix

Project Meteo Icon



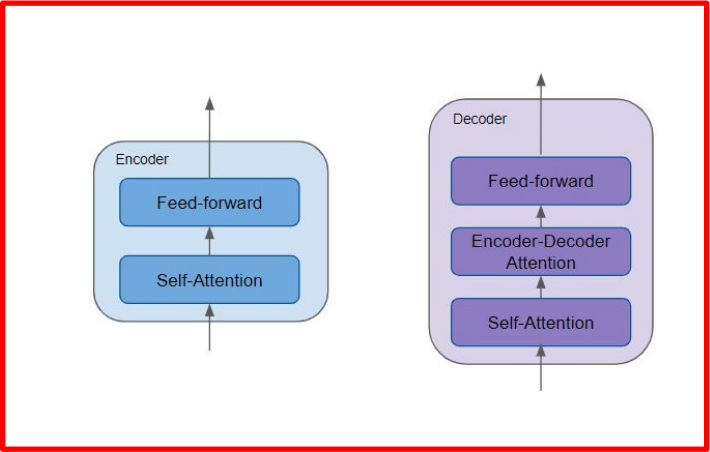
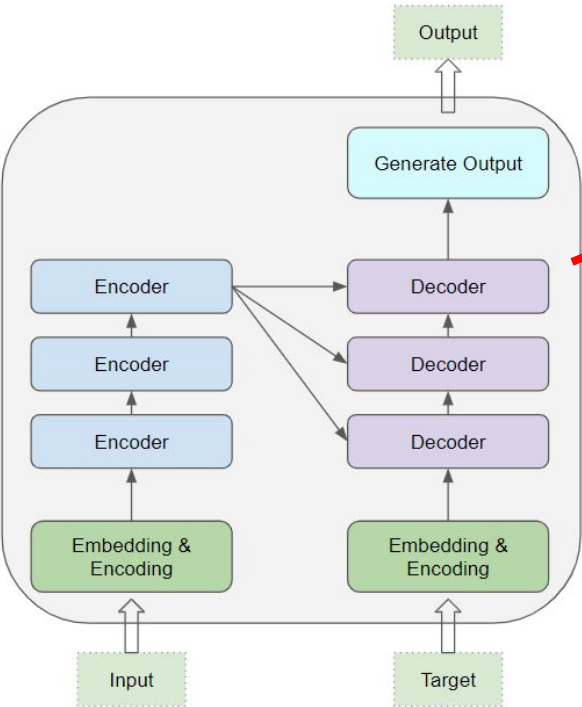
Latent Dirichlet Allocation



Each document can be described by a distribution of topics and each topic can be described by a distribution of words. The vocabulary comes from the documents.

Transformers, BERT models

Transformers architecture



The key to the Transformer's performance : **Attention**
While processing a word, **Attention** enables the model to focus on other words in the input that are closely related to that word:

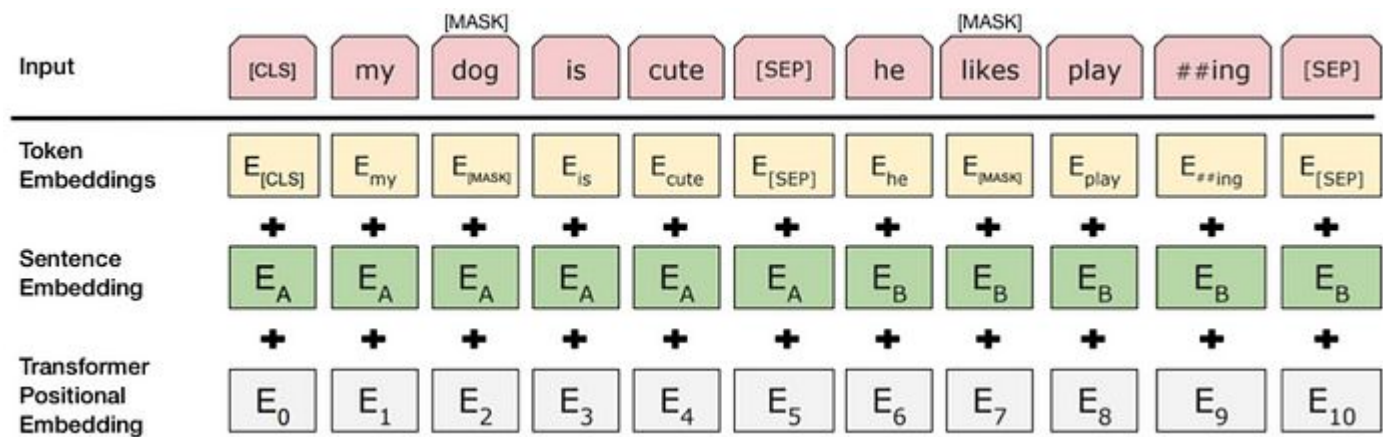
The boy is holding a blue ball

The diagram shows the sentence "The boy is holding a blue ball" with red curved arrows indicating attention. A solid red arrow points from "blue" to "ball", and another solid red arrow points from "holding" to "ball". A dashed red arrow points from "The" to "ball".

Transformers, BERT models

BERT (Bidirectional Encoder Representations from Transformers architecture)

As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).



Topic Modeling, BERTopic

