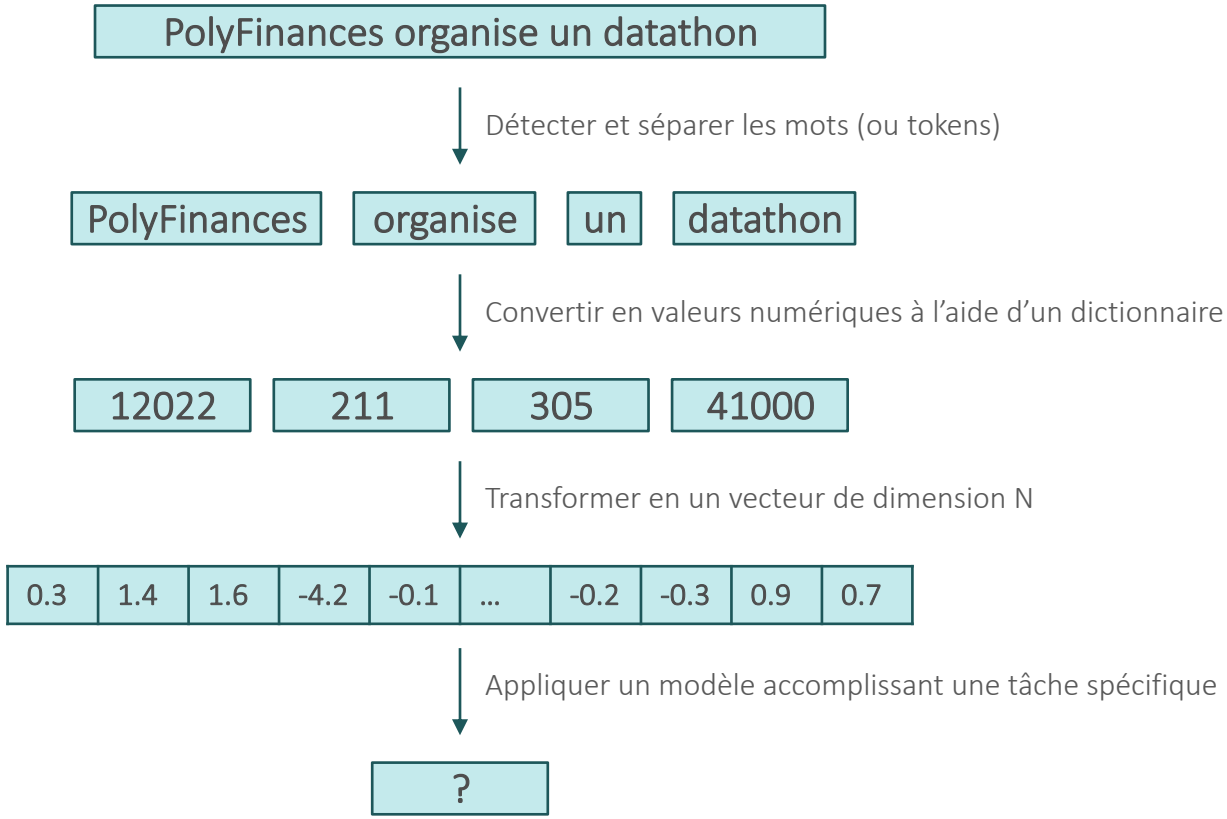
SR.ai

# Datathon PolyFinances
# NLP Crash Course

# NLP, comment ça marche?

*L'objectif est de transformer du texte (chaîne de caractères) en information pertinente pour notre cas d'usage. Il s'agit de le convertir en valeurs numériques. Pour cela, il faut passer par une ou plusieurs transformations.*




- *D'autres étapes peuvent être ajouter au processus pour obtenir un vecteur contenant une meilleure information.*
- *Le vecteur de dimension N est appelé « embeddings ». Il appartient à un espace représentationnel abstrait, où les mots/phrases ayant une signification similaire sont proches.*
- *Cette transformation peut être effectuée par de nombreux modèles d'approche, de taille et d'efficacité différentes.*
- *Il existe aussi un grand nombre de tâches différentes: classification de texte, analyse de sentiment, détection d'entités, etc.*

# Quelques suggestions

*Le but n'étant d'apprendre tout sur le NLP, mais bien de développer une solution au problème, voici une liste de packages et modèles qui pourraient vous être utiles!*

| Package/modèle | Description | Pours | Contres |
|---|---|---|---|
| TF-IDF (sklearn) | Transforme un texte en un vecteur de fréquences de mots | • Rapide et efficace | • Ne tient pas compte des relations entre les mots<br>• N'est pas pré-entraîné<br>• Requiert du nettoyage de texte |
| USE (tfhub) | Transforme le texte en un vecteur représentationnel de dimension 512. Utilise un deep averaging network. | • Relativement rapide<br>• Pré-entraîné<br>• Tient compte des relations inter-mots<br>• Pas de limites de taille de texte | • Moins bonne performance que des modèles Transformer |
| BERT (huggingface) / Transformer (huggingface) | Transforme le texte en un vecteur représentationnel. Utilise une architecture "Transformer", utilisant le self-attention. Huggingface possède une librairie de 80k+ modèles pré-entrainés! | • State-of-the-art<br>• Pré-entraîné<br>• Tient compte des relations inter-mots<br>• Transfer learning + peut être utilisé pour de nombreuses tâches | • Lent<br>• Requiert des GPUs pour le fine-tuning<br>• Limite de taille du texte |
| spaCy (spacy) | Effectue des tâches précises de NLP comme la détection d'entités (NER) avec des modèles pré-entraînés. | • Rapide et efficace<br>• Pré-entraîné | • Moins bonne performance que des modèles Transformer |
| nltk (nltk) | Package très puissant pour le nettoyage de texte et pour d'autres fonctions simples. | | |
| re (python) | Rien de plus basique que le regex. Peut être très efficace, surtout en combinaison avec d'autres méthodes. | | |

# Bonne chance!

paul@sr-ai.co

https://srinvesting.ai