# EVALUATION NLP GROUPE A

Etudiant : NONO KAKEU Andy

## 1. Définis ce qu'est le NLP en une phrase

Ensemble des étapes de traitement du langage naturelle dans le but de faire comprendre ce langage à l'ordinateur.

## 2. Donne deux exemples d'applications du NLP

**Exemple 1 :** Le sujet concerne les tweets postés par des utilisateur sur un réseau social. L'objectif est d'utiliser le NLP pour traiter ces postes afin de détecter le sentiment de l'utilisateur

**Exemple 2 :** Le sujet concerne la classification des documents. On peut se servir du NLP pour traiter des textes inscrits sur des documents dans le but de catégoriser des document (article de presse, poème, rapport etc…)

## 3. Cite les quatre étapes principales d'un pipeline NLP

- Récupération des données
- Prétraitement des données
- Choix et entrainement du modèle
- Evaluation du modèle

## 4. Que veut dire prétraiter un texte ?

C'est l'ensemble des étapes suivantes

- **Tokenisation :** Découper la phrase en mots ou le mot en sous mot
- **Nettoyage :** Ces traitement consistes à éliminer les caractères non utiles ou à uniformiser les éléments de mon texte (enlever les accents, enlever la ponctuation etc…)
- **Lemmatisation :** Consiste à associer chaque mot de mon texte à une racine plus significative et plus générale compréhensible par les ordinateurs
- **Représentation :** Consiste à transformer les mots ou sous mot de mon texte en des représentation numériques (vecteurs), afin de les rendre traitable par les ordinateurs.

## 5. Donne un exemple de stopword et explique pourquoi on le retire.

Un article comme ‘les’ en français est un stopword

## 6. À quoi sert la tokenization ?

La tokenisation sert à sectionner le texte en vecteur de mot ou de sous mots.  Ça permet de faciliter la compréhension de la phrase au modèle en lui facilitant la mise en relation entre les mots.

## 7. Quelle est la différence entre Bag-of-Words et TF-IDF ?

Le Bag-of-Words forme le vecteur sur la base de la fréquence d’apparition du mot dans la phrase tandis que le TF-IDF attribut des poids au mot en fonction de son important dans la phrase.

## 8. Pourquoi utilise-t-on des embeddings ?

On utilise un embeddings car il prend en compte le sens des mots et de la phrase en attribuant  des vecteurs similaires à des mots ayant le même sens

## 9. Donne un exemple de modèle de Machine Learning utilisé pour le texte.

SVM

## 10. Que mesure la précision (precision) d’un modèle ?

Mesure la proportion de prédiction position correcte (vrai positif)

## 11. Que mesure le rappel (recall) ?

Mesure la proportion de positif trouvé par le modèle

## 12. Pourquoi le F1-score est-il utile ?

Le F1-score fait une sorte de compromis entre le recall et la precision, il est utilisé pour les classes déséquilibrés

## 13. Qu’est-ce qu’un dataset d’entraînement et un dataset de test ?

Le dataset d’entrainement est l’ensemble des données utilisées pour entrainer notre modèle tandis que le dataset de test est l’ensemble des données utilisées pour tester la performance du modèle en sa capacité à prédire la classe correcte.

## 14. Que veut dire overfitting?

C'est lorsque le modèle apprend tellement bien le dataset d'entrainement qu'il a retenu les prédictions par cœur. Il n'a pas pu former les classes en fonction des caractéristiques.

## 15. Donne une solution pour éviter l'overfitting.

On peut faire la cross-validation, au lieu de diviser le dataset en train et test tout simplement, on peut sélectionner des groupes de data sauf 1 sur lequel on va entrainer le modèle.

## 16. Qu'est-ce qu'un Transformer?

C'est une architecture spéciale d'un modèle basé sur un réseau de neurone, à l'origine des LLMs. Sa particularité est qu'il traite très efficacement les données de façon séquentielle.

## 17. Quelle est la différence principale entre BERT et GPT ?

BERT est le modèle crée par google en 2018, il est bidirectionnel et excelle dans sa compréhension du sens des mots dans les deux sens c'est-à-dire capables de prédire le mot d'avant et après.

GPT est un modèle qui excelle dans sa capacité à prédire séquentiellement le prochain mot, il est plus utilisé pour les problèmes de génération.

## 18. Que veut dire Prompt Engineering ?

C'est la partie de l'ingénierie IA qui étudie les meilleures méthodes de requêtage des LLM afin d'avoir la meilleure réponse possible.

## 19. Qu'est-ce qu'un LLM ? Donne un exemple de modèle connu

Ce sont les modèles d'IA générative basés sur une architecture de transformer. On peut citer ollama ou Gémini

## 20. Cite un avantage et une limite des LLMs.

**Avantage :** le requêtage se fait en langage naturelle

**Inconvénient :** pas très efficace pour les problèmes classiques de classification contrairement au NLP classique