# Semaines 04/11 & 11/11

## Tâches de la quinzaine :

**Adjimon Jérôme :**

- Approfondissement des connaissances sur les embeddings (vecteurs). Une compréhension approfondie a été développée sur l'importance des multiples répétitions dans les données, permettant de varier les conclusions. Par exemple, des variations telles que "je suis content" et "content je suis" sont reconnues par le modèle, illustrant ainsi sa capacité à comprendre différentes structures syntaxiques (comme celles évoquées dans Star Wars).
- Structuration du code et conversion des notebooks vers des scripts Python pour une meilleure gestion et réutilisabilité.

**Anas et Aya :**

- Travail sur la distribution multi-sujets avec la réalisation de visualisations, notamment la création de graphes d’aires. Cette étape permet de se familiariser avec les outils et méthodologies. Une fois ces bases maîtrisées, des applications plus avancées seront envisagées.

**Noah et Amine :**

- Exploration de BERTopic, un modèle préconçu pour le traitement des textes. Les tâches incluent :
    - Exécuter des exemples sur des jeux de données textuels pour valider le fonctionnement du modèle.
    - Extraire les données pertinentes pour l’analyse.
    - Tester l’intégration de nouveaux textes dans le modèle afin d'évaluer ses capacités d’adaptation et d'apprentissage.

# NOTES REUNION :

- Sélection du modèle : réalisé après plusieurs tests, au jugé.
- Gestion des langues : classification des langues présentes dans les données, en identifiant celles qui apparaissent le plus souvent. Décision à prendre sur l'uniformisation : " conserver les données en français ou tout traduire en anglais".
- Comparaison des nuages de mots : analyser le nuage de mots en français et le comparer avec un nuage en anglais.

  Extraction des mots-clés pour évaluer la pertinence des résultats.

- Validation de BERTopic : évaluation de l'apport réel de BERTopic.
- Vérification avec Sabrina : comparaison des nuages de mots à l'aide de méthodes classiques (LLP classique) et BERTopic, en anglais et sans traduction. Analyse des résultats pour détecter d'éventuelles anomalies.
- Tâches intermédiaires :
  - Classification des langues.
  - Identifier la langue quand il n'y a pas de texte, en se basant sur la géolocalisation ou des indices contextuels.
- Gestion de projet : suivi des tâches sur GitHub et répartition des responsabilités.

**Code d'extraction de données avec BERTopic (selon M. Collin) :**

- Évaluation de la méthodologie actuelle, trop restreinte, se limitant à l'extraction simple de topics. Étudier les versions avec et sans traduction pour enrichir les mots-clés extraits (les mots-clés ne sont pas des topics).
- Proposition d'élargir les résultats pour inclure une génération de mots-clés plus diversifiée.

**Partie générative :**

- Rattachement du texte à des éléments plus génératifs et exploration des différentiels possibles dans les résultats.

**Analyses comparatives :**

- Création d’un tableau classique avec les résultats de BERTopic (3 approches distinctes) et les réponses associées au questionnaire.
- Comparaison des modèles et des jeux de données dans un tableau d’environ 10 colonnes. Calcul de scores pour évaluer les performances.

**Documentation des résultats :**

- Structuration des données dans un texte balisé. Ajout, pour chaque ligne, d’informations telles que la langue majoritaire détectée ou la langue identifiée par BERTopic. Le tout au format markdown, conformément aux recommandations du tuteur.

**Tests et classements :**

- Test des classements selon les 3 méthodes (rank).
- Étude des performances selon deux approches : classification par espace ou traitement global des données. Évaluer laquelle offre les meilleurs résultats.