**Rapport v0 - Nouveau document**

**I. Import**

**Import Packages**

Importation de tous les packages nécessaires au bon déroulement du code : pandas, numpy, plotly, sklearn.

**Import Data**

Représentation de word embedding donnés par les M2 non-alternants : Term Frequency Inverse Data Frequency (TF-IDF) et Bag-of-Word (BoW), Features syntaxiques créées par notre groupe (G5) et des données fournies par le groupe (G3) : BoW, Glove et Word2vec.

**II. Functions**

**Function graph\_3d**

Paramètres d’entrée :

* les données que l’on souhaite voir sur le graphique
* les prédictions utilisées
* les valeurs données par la prédiction (anciens articles et nouveaux)
* la transparence des points sur le graphique

Paramètre de sortie : graphique 3D en mouvement des résultats

Cette fonction permet de représenter les articles sur les axes principaux d’une ACP (réduction de dimensions). C’est ce qui nous permet de faire une analyse visuelle.

**Function duplicate\_prediction**

Paramètre d’entrée : le chemin vers les données (.csv ou .json).

Paramètre de sortie : les prédictions (liste de -1 et 1 qui détermine les nouveaux articles et les doublons).

Cette fonction permet de donner les prédictions des données en entrée.

**III. Modèles**

**Local Outlier Factor (LOF)**

Modèle non supervisé (utilisé sur des données non labellisées) servant à la détection de valeurs aberrantes. Le score d'anomalie de chaque échantillon nous permet de détecter quels articles sont nouveaux et lesquels sont considérés comme des doublons. Son calcul se fait grâce à la mesure de l'écart local de densité d'un individu donné par rapport à ses voisins.

Le paramètre n\_neighbours correspond au nombre de voisins à utiliser.

Le paramètre de contamination correspond au degré de contamination de l'ensemble de données, c'est-à-dire la proportion de valeurs aberrantes dans l'ensemble de données.

**Isolation Forest**

Modèle non supervisé servant à détecter des anomalies dans un jeu de données. Il isole les données atypiques, autrement dit celles qui sont trop différentes des autres données. Le score d’anomalie calculé nous permet de déterminer si un article est considéré comme nouveau ou si c’est un doublon.

Le paramètre de contamination correspond au degré de contamination de l'ensemble de données, c'est-à-dire la proportion de valeurs aberrantes dans l'ensemble de données.

**One SVM**

Modèle non supervisé servant à la détection de valeurs aberrantes. Le score obtenu nous permet de savoir si l’article est une nouveauté ou non.

Le paramètre nu correspond au degré de contamination de l'ensemble de données, c'est-à-dire la proportion de valeurs aberrantes dans l'ensemble de données.

**IV. Résultats**

1. **Analyses TF-IDF, BoW et Features syntaxiques**

Pour nos représentations nous avons choisi le graphique en 3D sur nos données BOW en entrée et nos prédictions BOW, TF-IDF ou Features entraînées avec un des deux modèles (LOF ou Isolation Forest). Cela nous a donné différents résultats :

| **Modèle LOF** | **Modèle Isolation Forest** |
| --- | --- |
|  |  |
| Prédictions : BOW | |
|  |  |
| Prédictions : TF-IDF | |
|  |  |
| Prédictions : Features | |

Les représentations graphiques ci-dessus ont été réalisées sur l’ACP de BOW et les prédictions indiquées. Il s’est avéré que la combinaison avec les données et les prédictions BOW pour l’Isolation Forest donnait les meilleurs résultats : les nouveaux articles et les doublons sont correctement distincts.

**2) Analyses faites sur le BoW du groupe 3**

Pour nos représentations nous avons choisi le graphique en 3D sur les données du nouveau BoW entraînées avec 3 modèles : LOF, Isolation Forest et One SVM.

| **Données du nouveau BOW** |
| --- |
|  |
| Modèle LOF : contamination = 0.1 et n\_neighbors = 3 |
|  |
| Modèle Isolation Forest : contamination = 0.015 |
|  |
| Modèle One SVM : nu = 0.03 |

Les modèles Isolation Forest et One SVM, contrairement au LOF, nous montrent que les articles nouveaux et les doublons sont correctement pris en compte dans le nouveau jeu de données BOW.

**3) Analyses faites sur les données GLOVE**

Nous avons représenté les résultats des données Glove en 3D pour mieux les visualiser. Nous les avons entraînées avec 3 modèles : LOF, Isolation Forest et One SVM.

| **Données GLOVE** |
| --- |
|  |
| Modèle LOF : contamination = 0.015 |
|  |
| Modèle Isolation Forest : contamination = 0.015 |
|  |
| Modèle One SVM : nu = 0.05 |

Pour les trois modèles non supervisés ci-dessus, ils confirment que les nouveaux articles ont bien été détectés (rouge) et que les doublons sont aperçus comme des anciens articles (bleu).

**4) Analyses faites sur les données word2vec**

Nous représentons les données en 3D grâce à une réduction de dimension (comme ACP). Nous avons entraîné les données word2vec avec 3 modèles de non-supervisés : LOF, Isolation Forest et One SVM. Cela nous a donné différents résultats :

| **Données GLOVE** |
| --- |
|  |
| Modèle LOF : contamination = 0.03 |
|  |
| Modèle Isolation Forest : contamination = 0.015 |
|  |
| Modèle One SVM : nu = 0.03 |

Pour les trois modèles non supervisés ci-dessus, ils confirment que les nouveaux articles sont bien traités comme tel (rouge) et que les doublons sont aperçus comme des anciens articles (bleu).