**Détection de nouveau document - Rapport vf**

**Introduction :**

Nous avons cherché à détecter si un article scrappé apporte de nouvelles informations (nouveau) ou si ces dernières sont déjà présentes dans un autre document (ancien).

**Méthode 1 (Version 0):**

**Méthodologie**

La démarche choisie afin de résoudre le problème énoncé ci-dessus, a été de prendre différentes représentations vectorielles décrivent les articles. Certains ont été fournis par les M2 non alternants tels que le Term Frequency Inverse Data Frequency (TF-IDF) et le Bag-of-Word (BoW). D’autres ont été produits par le groupe 3 : BoW G3, Word2vec et Glove. Nous avons créé un nouveau jeu de données afin d’étudier la syntaxe en choisissant nous même des features.

Pour prédire les articles atypiques (nouveaux contenus), nous avons déterminé les valeurs aberrantes grâce aux modèles de classification non supervisés puisque nos données ne sont pas labellisées. Un score d'anomalie a été déterminé pour chaque article : cela nous a permis de savoir si le document est nouveau ou ancien. Nous avons ainsi testé nos données sur trois modèles afin d’obtenir les meilleurs résultats possibles : Local Outlier Factor (LOF), Isolation Forest et One Support Vector Machine (One SVM).

Puis, nous avons choisi de représenter nos résultats sous forme de graphiques en 3D pour permettre une meilleure analyse visuelle. De ce fait, nous avons simultanément réalisé une ACP afin de réduire le nombre de dimensions et rendre les graphiques lisibles et interprétables.

Pour confirmer les résultats et vérifier qu'ils sont bons nous avons fait une analyse qualitative en regardant les textes et leurs nuages de mots des articles considérés comme atypiques.

**Résultats**

Pour effectuer nos différents modèles, nous avons, dans un premier temps, utilisé les données fournies par les M2 non alternants : TF-IDF et BoW. Notre groupe G5 a également créé un jeu de données sur des features syntaxiques afin de voir si les résultats pouvaient être intéressants syntaxiquement. Nous pouvons observer quelques résultats ci-dessous grâce à l’ACP.

|  |  |  |
| --- | --- | --- |
| Isolation Forest - Features | LOF - TF-IDF | Isolation Forest - BoW |

Légende : nouveaux articles et doublons (anciens articles)

Les résultats ne sont globalement pas convaincants, l’utilisation du modèle Isolation Forest sur les données BoW est le seul cas qui permette de distinguer correctement les nouveaux documents des doublons.

Nous avons donc choisi de mettre ces jeux données de côté et de se concentrer sur le nouveau jeu de données BoW amélioré et fourni par le G3. De plus, suite aux résultats, nous avons décidé de tester le modèle One SVM en plus du LOF et de l’Isolation Forest.

|  |  |  |
| --- | --- | --- |
| LOF - BoW G3 | Isolation Forest - BoW G3 | One SVM - BoW G3 |

Légende : nouveaux articles et doublons (anciens articles)

Nous constatons que les modèles Isolation Forest et One SVM sont efficaces et donnent de bons résultats pour déterminer les nouveaux documents sur les nouvelles données BoW.

Nous avons également testé nos modèles sur d’autres représentations vectorielles : Glove et Word2vec fournies également par le groupe G3.

|  |  |  |
| --- | --- | --- |
| One SVM - Glove | LOF - Word2vec | One SVM - Word2vec |

Légende : nouveaux articles et doublons (anciens articles)

Nous nous apercevons que les résultats ne sont pas très bons. Les meilleurs restent donc avec le jeu de données BoW du groupe G3.

Pour le confirmer, nous observons les nuages de mots des articles contenant des nouvelles informations avec les meilleurs résultats : les modèles (Isolation Forest et One SVM) avec la meilleure représentation : BoW du Groupe 3.

|  |  |
| --- | --- |
| Isolation Forest - BoW G3  One SVM - BoW G3 | Isolation Forest - BoW G3 |

Le premier article, il est le seul à être ressorti dans les 2 modèles comme étant l’un des documents apportant de nouvelles informations. En ce qui concerne les autres articles trouvés, ils sont assez différents. Néanmoins, il y a quelques thèmes qui reviennent en évoquant la même chose comme “La Covid” et “président” en parlant des annonces d'un président ou du premier ministre mais également du confinement, … etc :

|  |  |  |
| --- | --- | --- |
| Isolation Forest - BoW G3 | One SVM - BoW G3 | One SVM - BoW G3 |

Les documents ressortis par l’Isolation Forest sont globalement meilleurs, avec moins de documents évoquant un même sujet contrairement au One SVM.

Néanmoins, nous ne sommes pas satisfait des résultats obtenus et voulons tester une autre méthode.

**Méthode 2 (Version 1) :**

**Méthodologie**

Nous avons choisi de garder le BoW du groupe 3 comme représentation vectorielle car c’est celle qui avait les meilleurs résultats.

Pour cette seconde méthode, nous avons décidé de réaliser une réduction de dimension avec un t-distributed Stochastic Neighbor Embedding (t-SNE) sur notre BoW. Puis d’appliquer l’un des algorithmes de non supervisé utilisé dans la première partie (LOF, Isolation Forest et One SVM).

**Résultats**

Nous avons représenté nos résultats en 3D :

|  |  |  |
| --- | --- | --- |
| LOF - BoW | One SVM - BoW | Isolation Forest - BoW |

Légende : nouveaux articles et doublons (anciens articles)

Les résultats sont peu concluants pour le LOF et le One SVM. Cependant, pour l’Isolation Forest tous les articles atypiques ont été pris en compte.

Ci-dessous, nous regardons toujours l’Isolation forest avec le BoW du groupe 3 mais en sélectionnant moins de nouveau articles, nous obtenons les résultats suivant :

|  |
| --- |
| Isolation Forest - BoW G3 |

Légende : nouveaux articles et doublons (anciens articles)

Les résultats sont toujours aussi bons en détectant beaucoup moins d’articles doublons.

Cela nous confirme comme dans la première méthode que les meilleurs résultats sont obtenus avec l’Isolation forest avec le BoW du groupe 3.

Pour confirmer les résultats, nous regardons des textes et les nuages de mots :

|  |  |
| --- | --- |
| Isolation Forest - BoW G3 | Isolation Forest - BoW G3 |

Les premier résultats obtenus sont des textes très courts. Un article avec très peu d'informations c’est plus difficile de trouver un lien avec d'autres, ce qui semble cohérent.

Si l’on essaye de regarder les articles qui sont ressortis par le modèle de classification non supervisé possédant un peu plus de contenu, nous avons :

|  |  |
| --- | --- |
| “nom clochedate creation secteur activite lutte contre exclusiontaille association antennes locales cloche action association lien social changement regard via trois programmes carillon cloche biscuits clochettes” | “raisons lesquelles recevez aujourd hui cette medaille or cnrs michel lazdunski ressortent assez venons voir entendre vie travaux insisterai pourtant quelques points modestie trop passes sous [...] http www education gouv fr decembre” |
| Isolation Forest - BoW G3 | Isolation Forest - BoW G3 |

Environ pour les 30 premiers articles ressortis par le modèle, les résultats sont différents (ils n'évoquent pas les mêmes informations). De plus, aucun texte ne parle de “La covid” ou mentionne un lien avec comme le confinement, … par exemple.

**Conclusion :**

Pour trouver les documents apportant une nouvelle information, c’est la seconde méthode qui nous en apporte les meilleurs résultats.

Néanmoins, les résultats obtenus sont dépendant des articles scrappés (environ 7 000) mais également des représentations qu’on a choisies et auraient pu être bien meilleur si nous avions eu plus de documents.