Projet Data Mining

6 juillet 2022

Ce projet permet aux étudiants de mettre en pratique et de valider les connaissances acquises durant les cours de Data Mining. Les points qui suivent détaillent les modalités du projet.

1 SeneWeb

Seneweb est un site d'actualité. Les internautes peuvent y lire et ajouter des articles. Pour faciliter le repérage de l'information, la catégorisation des articles peut être utilisée. Cette catégorisation consiste à classer les articles similaires dans une même catégorie.

2 Kaggle

Kaggle est une plateforme web organisant des compétitions en science des données. Elle offre aussi un ensemble d'outils pour la science des données.

3 Datasets

Le jeu de données **Articles** contient un ensemble d'articles du site d'actualité **Seneweb**. Il peut être téléchargé ici.

Il s'agit de 7838 documents correspondant à des articles dans six domaines d'actualité. Ces domaines sont les suivants :

- Politique (3423)
- Sante (1519)
- Sport (1355)
- Justice (650)
- Economie (505)
- Education (386)

la structure du jeu de données est la suivante :

	identifiantArticle	contenuArticle	thematiqueArticle
0	https___actu24.net_13-personnes-convoquees-a-k...	Après l'accalmie de ces derniers jours avec la...	Politique
1	http___sentv.info_actualite-nationale_politiqu...	SENTV.info : L'attribution des marchés des den...	Politique
2	http___www.seneweb.com_news_justice_saccage-de...	Saccage de ses magasins : Auchan traîne en jus...	Justice
3	http___www.seneweb.com_news_politique_issa-sal...	Issa Sall nommé (depuis des mois) ministre con...	Politique
4	http___www.seneweb.com_news_sante_les-vaccins-...	Les Vaccins chinois surs et efficaces Les vacc...	Sante
...	...	...	...
7833	https___www.seneweb.com_news_politique_saccage...	Bachirou Ba, collectif Aar sunu momel Le coord...	Politique
7834	https___www.seneweb.com_news_politique_report-...	Report des Elections locales Le report des éle...	Politique
7835	https___www.seneweb.com_news_justice_plainte-c...	Plainte contre Prési Cissé et Cie : Dakaractu ...	Justice
7836	https___www.seneweb.com_news_sante_covid-19-5-...	Covid-19 : 5 décès, 52 nouveaux tests positifs...	Sante
7837	https___www.seneweb.com_news_sante_etude-ameri...	Étude américaine : 'La qualité du sperme dégri...	Sante

7838 rows × 3 columns

FIGURE 1 – Structure du jeu articles

3.1 Travail attendu

- Analyse exploratoire des données
- Pré-traitement des données
- Représentation des textes :
  - Word Count Vectors
  - TF–IDF Vectors
- Modèles
  - Random Forest
  - Support Vector Machine
  - K Nearest Neighbors
  - Multinomial Naïve Bayes
  - Multinomial Logistic Regression
- Mesures d'évaluation
  - Précision
  - Rappel
  - F-score

Comparer les différents modèles en utilisant les mesures d'évaluations et les deux structures de représentation de textes ci-dessus. Le travail doit se faire sur la plateforme kaggle.

3.2 Liens

- Analyse exploratoire des données

3.3 Date limite de rendu des projets

-