



Objectif et Motivation

Dans le cadre de ce projet, il est prévu de classer les films en catégories sur la base de scénarios d’écrans open source. Les techniques de Traitement du Langage Naturel ont été implémenté pour effectuer cette recherche.

Description des Données

L’ensemble de données utilisé dans le projet consiste en des scénarios de film. Ces scénarios; extrait du site (www.imsdb.com) sur internet. Ces scénarios comprennent environ 2 500 films et sont des documents à source ouverte.

Sujet	Nombre de films	Nombre de mots	Nombre de lignes
Action	290	6.751.146	519.318
Adventure	166	3.801.832	247.294
Animation	35	613.873	50.468
Biography *	3	76.112	5.803
Comedy	347	7.378.082	597.497
Crime	201	4.771.146	298.323
Drama	174	4.053.475	273.170
Family	39	709.031	53.718
Fantasy	113	2.423.609	220.328
Film_Noir *	4	114.948	10.172
History *	3	66.285	5.651
Horror	149	3.254.284	250.329
Music*	5	117.169	11.835
Musical	22	425.017	44.973
Mister	107	2.603.253	203.712
Romance	192	4.279.482	185.802
Sci_Fi	155	3.524.646	232.547
Short*	3	13.761	1.390
Sport*	2	52.916	5421
Thriller	373	8.660.527	721.710
War	26	689.262	53.020
Western	13	351.757	25.247

* Catégories supprimées car le nombre de films est insuffisant

Technologies Utilisées et Méthodes

Les technologies et bibliothèques utilisées sont les suivantes;

Langue du logiciel: Python

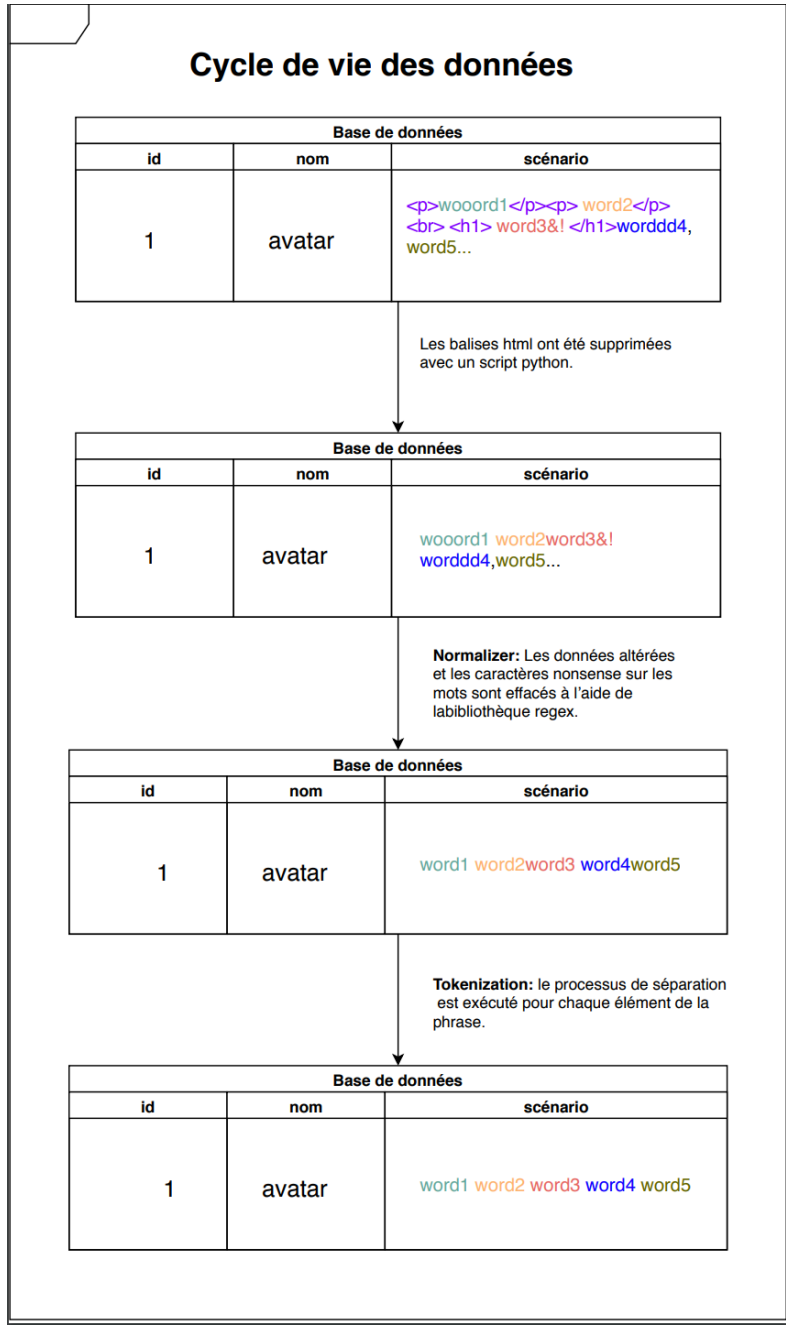
Bibliothèques utilisées: Regex, Pandas, Nltk, Scikit, Pickle

Algorithmes préférés: Vectorisation (Countvectorizer, TFIDFVectorizer,) Algorithmme Naive Bayes



Les Etapes du Projet

Si le processus global du projet est pris en compte; Il se compose de deux étapes de base. Ce sont: la prétraitement et le créer un modèle d'apprentissage machine

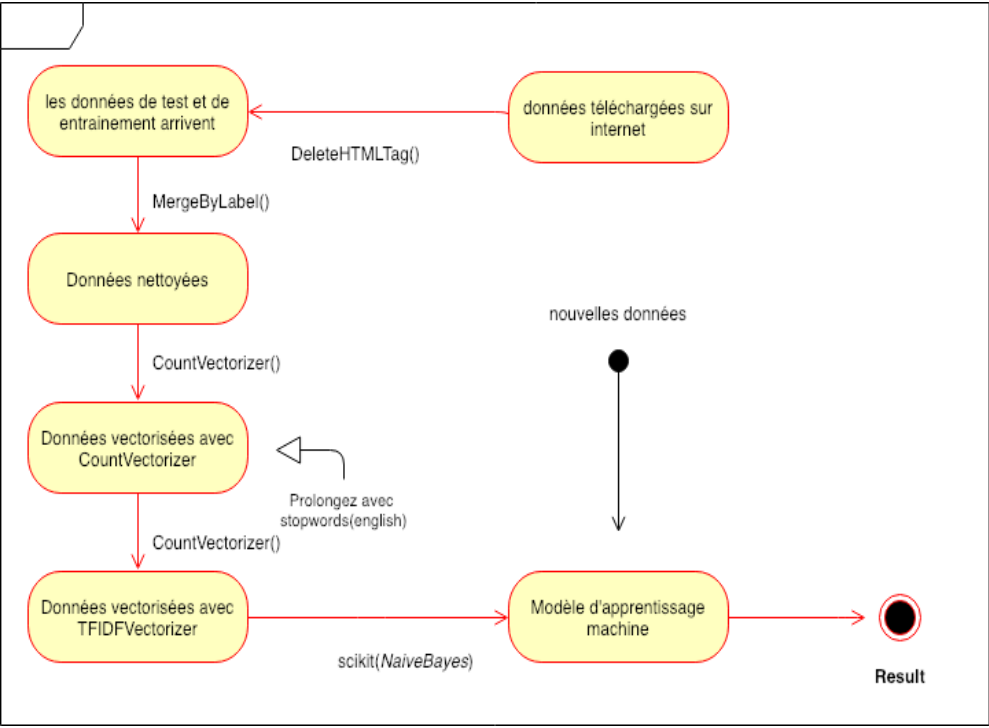


1.Prétraitement

L’ensemble de données, qui est également téléchargé via Internet à l’aide de la méthode DeleteHtmlTag() et qui contient également des balises HTML, est composé de mots. Cet ensemble de données est ensuite transmis aux opérations de normalisation et de tokenization.

2.Créer un modèle d'apprentissag e machine

Dans cette étape du projet, les processus suivants ont été appliqués pour créer un modèle.



Conclusion

Une fois le film sélectionné inséré dans le modèle, les résultats sont extraits pour chaque étiquette. À la suite de ces résultats, des préférences de catégorie sont définies.

```
2018-12-26 09:47:54.234342 : Score for 0.label : 0.0250390363596
2018-12-26 09:47:54.240593 : Score for 1.label : 0.00094802587553
2018-12-26 09:47:54.246045 : Score for 2.label : 0.0
2018-12-26 09:47:54.251159 : Score for 3.label : 0.0
2018-12-26 09:47:54.255559 : Score for 4.label : 0.0555989292884
2018-12-26 09:47:54.260170 : Score for 5.label : 0.0
2018-12-26 09:47:54.264674 : Score for 6.label : 0.0
2018-12-26 09:47:54.269135 : Score for 7.label : 0.0
2018-12-26 09:47:54.273504 : Score for 8.label : 0.0
2018-12-26 09:47:54.278670 : Score for 9.label : 0.0
2018-12-26 09:47:54.283094 : Score for 10.label : 0.0
2018-12-26 09:47:54.287448 : Score for 11.label : 0.0
2018-12-26 09:47:54.291867 : Score for 12.label : 0.0
2018-12-26 09:47:54.296549 : Score for 13.label : 0.0
2018-12-26 09:47:54.300958 : Score for 14.label : 0.0
2018-12-26 09:47:54.305348 : Score for 15.label : 5.57662279723e-05
2018-12-26 09:47:54.310075 : Score for 16.label : 0.000167298683917
2018-12-26 09:47:54.315226 : Score for 17.label : 0.0
2018-12-26 09:47:54.324247 : Score for 18.label : 0.0
2018-12-26 09:47:54.329588 : Score for 19.label : 0.918190943565
2018-12-26 09:47:54.334026 : Score for 20.label : 0.0
2018-12-26 09:47:54.338398 : Score for 21.label : 0.0
Label name is : Adventure, Drama
```

%80 train - %20 test

	precision	recall	f1-score	support
0	0.42	0.87	0.57	34993
1	0.56	0.23	0.32	31334
2	0.62	0.16	0.26	25303
micro avg	0.45	0.45	0.45	91630
macro avg	0.54	0.42	0.38	91630
weighted avg	0.52	0.45	0.40	91630

0.45484011786532796
2018-12-26 10:43:43.874518 : ***** code stop ***

%85 train - %15 test

	precision	recall	f1-score	support
0	0.42	0.87	0.57	26194
1	0.56	0.22	0.32	23587
2	0.61	0.16	0.26	18942
micro avg	0.45	0.45	0.45	68723
macro avg	0.53	0.42	0.38	68723
weighted avg	0.52	0.45	0.40	68723

0.4537491087408873
2018-12-26 10:42:21.511109 : ***** code stop **

%90 train - %10 test

	precision	recall	f1-score	support
0	0.42	0.87	0.57	17476
1	0.57	0.23	0.32	15649
2	0.62	0.17	0.26	12690
micro avg	0.46	0.46	0.46	45815
macro avg	0.54	0.42	0.38	45815
weighted avg	0.53	0.46	0.40	45815

0.4563570882898614
2018-12-26 10:39:19.802984 : ***** code stop **