Summation of project(French politic newspaper bias group)

Members: Boyuan Han, Xiaozhong He

Dataset

* We have totally 2456 political articles from 5 websites.
* Left (<https://www.francetvinfo.fr/>)
* Middle(<https://www.afp.com/fr/plateforme-de-linfo>,<https://fr.euronews.com/>)
* Right(<https://fr.aleteia.org/>,<https://www.cjnews.com/category/en-francais)>
* We collect articles by using python to download articles from website RSS feed.Using library(BeautifulSoup and newspaper)to reduce useless ingredients to get pure articles.In our code,computer can automatically download articles without duplicated items every half hour(realized by timer)
* Then we put these articles in one table, and save it as csv format(UTF-8)

Data preprocessing

* Tool : nltk
* Remove URL symbols and number
* Remove stopwords
* Lowercase letters
* Extract stems

Features

* Use TF-IDF further process
* Then use chi-test to get feature (also used for building dictionary with chi scores higher or equal to 7.81)
* After feature extraction, we transfer data into vectors prepared for machine learning algorithm.

Learning Model

* 1. For the general machine learning(SVM , Naïve Bayes( Multinomial) and Logistic Regression),we use k-fold cross validation to training the model(cv=10).To get better machine learning models, we grid search to find the optimal parameters(such as c in svm …. L1 and L2 regularization)At last, we use voting strategy to find best algorithm by comparing the results of different models.
* 2. Use deep learning method (CNN, LSTM).Both of them needs to preprocess data as sequential data, then use keras to build neuro networks structure and get the learning model.

Results



Accuracy of CNN is 0.827 and accuracy of LSTM is 0.69.