Wikipedia Article Classification

End-to-End Project (regression)

Question: Perform a binary classification on feature and non-feature Wikipedia articles.

Details: Provided a repository of Wikipedia articles, your task is to create a binary classification model to classify the articles as featured and non featured.

Data Extraction: You have to use Python's Wikipedia API to extract the article data in raw format. Moreover, you need to create the labeling using the following dataset. Columns are separated by "@$@" delimiter. For each article's name, there exists a labeling denoting whether it is a featured article or not (please look into the Class column). You need to use this dataset to create your own labeling and perform the classification.

Feature Engineering: For each article, try to extract the features which might correlate with the target label using the Wikipedia API. The first logical step will be to create sampled articles from the following dataset. This will help you reduce the sample size and get only those articles using API for which you can perform the classification. You can create your own features while doing the EDA. Please refer to the API and see what kind of features can be used to perform the classification.

Post-Evaluation questions:

1. What's the accuracy you are getting?
2. What if you choose SVM instead of logistic regression?
3. What features do you think are important? Can you get the same accuracy using only a single feature?