15.459: Financial Data Science and Computing

Competition: Project D Natural Language Processing and Machine Learning Classification

November 27, 2018

Brain Warm-up

Question:The circumference of the Earth is approximately 40,000 km. If we made a circle of wire around the globe, that is only 10 meters (0.01 km) longer than the circumference of the globe, could a flea, a mouse, or even a man creep under it?
Answer: It is easy to compare old and new perimeter - original perimeter is $2 \times \pi \times R_{old}$, length of wire is $2 \times \pi \times R_{new}$ and find out that the result is about 1.6 m. So a smaller man can go under it and a bigger man ducks.

Competition Guidelines

You will be given a set of new data from the RCV1 corpus and you will need to classify it. Your classification will be based on the four top-level topics (i.e., Corporate/Industrial, Economic, Government/Social, Market). Please be aware that you do not need to change your code from project D, you will only need to run it. No new modeling is permitted. Data format sample are available on the github repository. The data you will be using will be available on the same repository:

https://github.com/ucfbrd/15.459-Playground

Input

You will be provided with a set of data with the following columns. see sample csv file on github

- `id`: Articles IDs.
- `article`: Cell arrays of the articles text that have been conventionally pre-processed: stemming, stop word removal, capitalization, punctuation, etc.

Output

The required output is a csv file that contains the following columns. See sample csv on github

- `id`: Articles IDs.
- `article`: Cell arrays of the articles text that have been conventionally pre-processed: stemming, stop word removal, capitalization, punctuation, etc.
- `CCAT`: Binary indicator
- `ECAT`: Binary indicator
- `GCAT`: Binary indicator
- `MCAT`: Binary indicator

Your file submissions should be in the form of `firstname_lastname.csv`. Only one submission for the team is required.

Evaluation

The evaluation of the performance of your submission will be based on the accuracy of your model to predict if an article belongs to a specific category.

Submission Deadline

The data will be made available Tuesday, November 27 2018 at 9:00AM and you will have until Thursday, November 29 2018 at 09:59AM to submit your predictions to yberrada@mit.edu. The results will be announced on Thursday, November 29 2018 at 11:30AM.

GOOD LUCK !