CSE 472: Social Media Mining

Project II

Prof. Huan Liu
Due on Dec 01, 2019 at 11:59 pm

This is a ***2-member*** project assignment. Each group is supposed to work on the steps together, including the writeup of the report. Please submit one set of your report and related files per group on Canvas.

Please form a different team from Project I so that you can make more friends :)

1 The Default Project

All undergraduate and MCS students should work on this project. As for Ph.D. and Masters students with thesis, there are two other options available:

1. Proposing your own research projects.

2. Working with DMML members for promising research projects.

You still need to form a ***2-member*** group for the above options. We will release the details soon.

1.1 Task: Fake News Classification

Social media has become one of the major resources for people to obtain news and information. For example, it is found that social media now outperforms television as the major news source. However, because it is cheap to provide news online and much faster and easier to disseminate through social media, large volumes of fake news or misinformation are produced online for a variety of purposes, such as financial and political gain. The extensive spread of fake news/misinformation can have a serious negative impact on individuals and society: (i) breaking the authenticity balance of the news ecosystem; (ii) intentionally persuading consumers to accept biased or false beliefs; and (iii) changing the way people interpret and respond to real news and information. Therefore, it is important to detect fake news and misinformation in social media.

We formally define the task as follow. Given the title of a fake news article A and the title of a coming news article B, participants are asked to classify B into one of the three categories:

- **agreed:** B talks about the same fake news as A.

- **disagreed:** B refutes the fake news in A.

- **unrelated:** B is unrelated to A.

1.2 File Descriptions

In the attached folder, you are provided with 4 CSV files:

- **train.csv:** Training data
- **test.csv:** Test data
- **validation.csv:** Validation Data
- **sample_submission.csv:** Expected submission format

The training data and the validation data include the "label" of each news pair, while the test data doesn't. Students should use the training data to train a classifier and evaluate their model's performance with the validation file. Finally, by using the trained model, you are required to predict the results for the test data. The format of your output file should be the same as "sample_submission.csv". The columns in train, validation and test data is as follows:

- **id**: the id of each news pair.
- **tid1**: the id of fake news title 1.
- **tid2**: the id of news title 2.
- **title1_zh**: the fake news title 1 in Chinese.
- **title2_zh**: the news title 2 in Chinese.
- **title1_en**: the fake news title 1 in English.
- **title2_en**: the news title 2 in English.
- **label**: indicates the relation between the news pair: agreed/disagreed/unrelated.

The English titles are machine translated from the related Chinese titles. Students can use either Chinese Version or English version or both to finish the task.

1.3 Submission

Students are supposed to submit the result file (named ***"submission.csv"***), source code and report in one *.zip* file named LASTNAME1_LASTNAME2_PJ2 (Instead of LASTNAME1 and LASTNAME2 type the lastname of each member).
The submitted results should be reproducible with the submitted code/data. Moreover, do not change the name of the files as your submitted .csv file will pass an automatic program.
The report should not be less than 2 pages and should include description of the data pre-processing, model, and validation results.
Use a "Reference" section and cite all the papers, tutorials, packages, software and libraries you used for your program.