Graded Assignment: Machine Learning

Start Assignment

Due Jun 15 by 11:59pm **Points** 10 **Submitting** a text entry box or a file upload
Available May 24 at 12am - Jun 15 at 11:59pm 23 days

Assignment Overview

Welcome Everybody!

I am Ankit, your very own course instructor for Real Time Data Analysis and greeting you to this super interesting assignment exercise that we will solving together (I am saying "we" because I will also be solving it alongside you to improve our previous benchmarks). In fact, this ideology of solving problems together also is a lesson from the military where both the instructor and cadets run together (without any cheating by riding the bike!) to finish the early morning fitness routine.

Now, enough analogies. Let us come back in the context and discuss the assignment. Remember that this assignment is not an examination to determine your abilities because I already know that you have it in you. In fact, this assignment is an attempt to-

- Reveal the ability to do **Hard Work** inside you;
- **Skill** to be beyond the best while doing smart work;
- **Creativity** to rise above the cut by means of thinking outside the box for new solutions;
- **Patience** to fail but never give up;
- **Perseverance** to get up gain and run and;
- **Determination** to stay till the end for your own self and your peers;

Accordingly, all the grading methods are scientifically designed to look for the aforesaid qualities in side you to make you not just a good problem solving engineer but also an efficient team leader. So, do not stress out when you are stuck at any step and just believe in yourself because I know you can do it.

Introduction

As a result of the ongoing Coronavirus Disease (COVID-19) Pandemic, our daily life routines and behavioural patterns changed drastically not only offline but also online. One example of such a change is the change in the reading patterns on Wikipedia and Reddit [1,2]. COVID-19 has also been a hot topic on other social media platforms such as Facebook, Twitter, or YouTube.

To understand the information spreading mechanisms during the COVID-19 pandemic, in this assignment, we focus on the Twitter platform. **Twitter is an online social network** where users can follow each other and share information using short text posts called tweets. The platform **offers a function to retweet a tweet, which means sharing it with your followers without any change**.

Retweeting is a popular function and it has also found its way in other online social networks such as Weibo. **Retweeting can be seen as amplifying the spread of original content** and thus **retweet prediction is a crucial task when studying information spreading processes**. As such, understanding retweet behaviour is useful and has many practical applications, e.g. political audience design [3,4], fake news spreading and tracking [5,6], health promotion [7], mass emergency management [8], etc. **Modelling retweet behaviour has been an active research area and is also especially important during times of crisis, such as the current COVID-19 pandemic**.

Understanding the Problem

In this assignment, you will work towards "Retweet Prediction Problem" based on COVID-19 related tweets. The assignment is based on a dataset containing approximately 8 million COVID-19-related tweets, spanning the period October 2019 to April 2020. For each tweet, the dataset provides metadata and some pre-calculated features such as sentiment scores and entities. ***Given the set of features for a tweet from the dataset, the task is to predict the number of times the tweet will be retweeted*** (this feature is present in the dataset by the name of "***#retweets***").

Understanding the Data

The data to be used for the assignment is provided **here** (https://www.dropbox.com/s/31glrg2mxxlc15l/Assignment-2.zip?dl=0). The data is hosted in the Dropbox Cloud due to the size of the zipped file being ~0.94 GB (You have to click the link provided and download the zip file named "Assignment-2.zip"). Now, let me briefly explain you the contents of the zip file. It consists of the following files that you will discover once you unzip-

- **train.data** - contains training examples.
- **train.solutions** - each line contains an **integer** value representing **the number of retweets** for the corresponding tweet in the **train.data** file.
- **feature.name** - contains tab-separated names of the features in the *.data files

Note: There will be two more files related to testing and validation but for the time being, ignore the testing and validation related files and create your validation and testing files from the train.data only)

Now, the dataset that you have witnessed above consists of multiple rows describing various tweets in terms of multiple features already extracted for you represented by various columns. Features are separated by a tab character ("\t"). The following list indicates the feature indices:

- Tweet_id: Long.
- Username: String. Encrypted for privacy issues.
- Timestamp: Format ("EEE MMM dd HH:mm:ss Z yyyy").
- #Followers: Integer.
- #Friends: Integer.
- #Favorites: Integer.
- Entities: String. For each entity, we aggregated the original text, the annotated entity, and the produced score from the **FEL** (https://github.com/yahoo/FEL) library. Each entity is separated from another entity by char ";". Also, each entity is separated by char ":" in order to store "original_text:annotated_entity:score;". If FEL did not find any entities, we have stored "null;".
- Sentiment: String. **SentiStrength** (http://sentistrength.wlv.ac.uk/) produces a score for positive (1 to 5) and negative (-1 to -5) sentiment. We split these two numbers by whitespace char " ". Positive sentiment was stored first and then negative sentiment (i.e. "2 -1").
- Mentions: String. If the tweet contains mentions, we remove the char "@" and concatenate the mentions with whitespace char " ". If no mentions appear, we have stored "null;".
- Hashtags: String. If the tweet contains hashtags, we remove the char "#" and concatenate the hashtags with whitespace char " ". If no hashtags appear, we have stored "null;".
- URLs: String: If the tweet contains URLs, we concatenate the URLs using ":-: ". If no URLs appear, we have stored "null;".

Guidelines

The guidelines mentioned here are providing a rough framework of the criteria which will be used to score your submissions in line with the idea of pushing you to reveal the qualities we had described while introducing the assignment. However, such guidelines are not rigid and can be skipped if there are compelling reasons which needs to recorded and

communicated in writing in the detailed project report. The common scoring framework is described as follows-

1. You will submit a detailed project report (.pdf) not exceeding 10 pages including references along with your codes in the form of a jupyter-notebook (both .ipynb and .pdf versions will be needed).
2. Ideally, your code may exhibit your skills that you have learnt so far in this masters especially in the context of Real Time Data Analysis. Therefore, I would like to highly encourage you to use Apache-Spark as much possible and benchmark against scikit-learn.
3. The detailed project report must consist of the sections-
    1. **Introduction**- start by describing the problem that is available to solve and your own ideas about the relevance of this problem (Quick Tip: Approach this section in the end when you are fully done with everything); (**1/10 Point**)
    2. **Related work**- start by describing any previous approaches taken by other authors to solve similar problems (you need not to go in super depth but just perform a google search and see if others also tried to solve this problem) and end this section by writing your own understanding of the approaches taken previously to solve similar problems; (**1/10 Point**)
    3. **Methods**- this section will focus about the specific methods in detail that you are going to use (if possible, some mathematical explanation but not mandatory). It can be methods that you will use to perform feature extraction from strings to any transformation of other features or even some ideas about models that you re going to use; (**1/10 Point**)
    4. **Experimentation**- this section will describe a set of experiments you have designed to determine the quality of your models. For example, if you are using the TF-IDF for feature extraction then describe how your you will experiment with the parameters of TF-IDF such as max_features or vocabulary etc. In fact you will get help with the timing of this assignment and the chapter Natural Language Processing that you will study in Advanced Data Analysis class; (**1/10 Point**)
    5. **Results-** start comparing how the various experiments designed by you in previous section leads to change in performances of various models proposed by you to solve the problem. For example, you may think of comparing a Random Forest Model performances based on different values of parameters you have chosen for TF-IDF algorithm in previous section; (**1/10 Point**)
    6. **Discussions**- finish your report by making attempts to discuss what more you could do to improve your solutions by means of further experimentation of choosing an entirely different set of methods. Feel free to go as in-depth you like; (**1/10 Point**)
    7. **Conclusions**- conclude your report based on the learnings you have acquired while solving this problem. If possible, attempt to describe how your work is important and can be applied in some other areas to solve problems of similar nature or magnitude; (**1/10 Point**)
4. The code that needs to be submitted by you must correspond to the detailed project report submitted by you. Accordingly, every claim that you will make in the report must be accompanied by a section of the code in the jupyter-notebook. You will be scored about the readability of the code; (**3/10 Points**)

References

[1] Gozzi, N., Tizzani, M., Starnini, M., Ciulla, F., Paolotti, D., Panisson, A. and Perra, N., 2020. Collective response to the media coverage of COVID-19 Pandemic on Reddit and Wikipedia. *arXiv preprint arXiv:2006.06446*.
(https://t.co/7fdCERVjkQ?amp=1)

[2] Ribeiro, M.H., Gligorić, K., Peyrard, M., Lemmerich, F., Strohmaier, M. and West, R., 2020. Sudden Attention Shifts on Wikipedia Following COVID-19 Mobility Restrictions. *arXiv preprint arXiv:2005.08505*.

[3] Stieglitz, S. and Dang-Xuan, L., 2012, January. Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Sciences* (pp. 3500-3509). IEEE.

[4] Kim, E., Sung, Y. and Kang, H., 2014. Brand followers' retweeting behavior on Twitter: How brand relationships influence brand electronic word-of-mouth. *Computers in Human Behavior*, *37*, pp.18-25.

[5] Lumezanu, C., Feamster, N. and Klein, H., 2012, May. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[6] Vosoughi, S., Roy, D. and Aral, S., 2018. The spread of true and false news online. *Science*, *359*(6380), pp.1146-1151.

[7] Chung, J.E., 2017. Retweeting in health promotion: Analysis of tweets about Breast Cancer Awareness Month. *Computers in Human Behavior*, *74*, pp.112-119.

[8] Kogan, M., Palen, L. and Anderson, K.M., 2015, February. Think local, retweet global: Retweeting by the geographically-vulnerable during Hurricane Sandy. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 981-993).

Technical Support

In case you face any problems while solving the problem, feel free to drop a mail to **ankit.tewari@bts.tech (mailto:ankit.tewari@bts.tech)** and I am sure we will be able to figure out how to steer out of the situation clear.