3rd Lab Project - Feature Extraction & ML

Web Data Mining - SS 2022

This is a hands-on lab project focused on the feature extraction process that is the stepping stone for effective machine learning models. You will practise how to identify, discover and extract features from Twitter data, that refer either to users or tweets. Each group will :

- **experiment** on tasks that involve feature engineering for the supervised classification of bots in Twitter and for the unsupervised clustering of users based on their activity. An annotated dataset with bots and humans will be available for the supervised machine learning process.
- **prepare a consistent, detailed hands-on tutorial as a Medium[1] article**, focused on the general topic of "Feature extraction and ML for bot and activity classification of Twitter users"; No other/further report will be required. An indicative example of a Medium tutorial article can be found here.

The tutorial will have the following sections:

A. Getting access to Twitter API v1 and v2, using Python and Tweepy (or any other wrapper)
B. Extract a large set of features that will be used for the bot detection and activity clustering processes. The features need to be extensive and language agnostic (they should **NOT** be focused only on the English language). All the features need to be categorised into the following (or more categories; if less please justify the restrictions):
    a. Temporal
    b. Content
    c. Network - Social Neighbourhood
    d. Sentiment
    e. User
    f. +++ (e.g. Graph features?)
C. Evaluation of the time needed to extract these features - Time is crucial for producing near Real-Time results. You can provide a table with the average time needed to extract all the features for X users in a specific machine.
D. Development of a supervised ML model for binary bot-human classification, using features from those identified earlier (a comparison between different algorithms is a plus).
E. Development of an unsupervised ML model for grouping users in correspondence to their activity (e.g. extremely active, very active, active, slightly active, inactive). The ideal number of clusters would be 5, however even less or more are ok.
F. Document the evaluation of models and test them with real accounts (Accuracy, Precision, Recall, F1-Score - charts are a plus)
G. Insightful conclusions with respect to the extracted features and their efficiency (what you've learnt)
H. ***Optional o1****- Plus for anyone interested: Discover possible correlations between types of activity and bot/human accounts.*

[1] https://medium.com

I. ***Optional o2** - Plus for anyone interested: Test your model across all datasets, evaluate its cross performance.*

IMPORTANT REMARKS :

- Document the process and **your conclusions** in your Medium-like article
- **Try to be brief and concise. You don't have to get into extreme details!**
- Share your code with Jupyter Notebook
- Upload/send your solutions by **02/06/2022**

Data

All data is available in Kaggle and in Botometer's Repo. Due to Twitter's terms of services and data sharing policies, the datasets include only the user IDs and the label respectively. Consequently, in order to extract features from each account you will need to download their latest tweets (according to research 200 - 1000 tweets are enough). Please bear in mind that many of the users included in these datasets have already been suspended or deleted by Twitter itself. So a dataset that initially includes 3000 accounts may end up having just 500, which are not enough. Each group will need to collect users-data that include **at least 2000** accounts, maintaining an overall balance between bot and human instances.

Table 1: Team allocations

Team	Datasets
[103 - 88 -101 - 100]	D1 , D2 , D3
[108 -119 - 113 - 106]	Gilani
[111 - 97 - 95 - 117]	Caverlee
[74 - 93 - 98]	Cresci_RTbust, Cresci_2017
[123 - 116 - 91 - 107]	Cresci_Stock

Attention

You don't have to use the whole datasets, you can just use samples of these data as long as you fulfil the 2000 users requirement. Of course, the larger the number of the instances, the better the results! Finally, if the users included in the datasets of each group are less than 2000, you can use any of the datasets of the other groups to complete them. If you , for any reason, face problems with the data collection process do not hesitate to contact us!

Helpful Documents & Repos

Papers: P1, P2, P3 , P4 , P5
Repos: R1 , R2, R3

Real Accounts to test your model on

1. @Aristoteleio: 234343780
2. @ylecun: 48008938
3. @nportokaloglou: 249130209
4. @greeceinfigures: 1348351605654106118
5. @big_ben_clock: 86391789
6. @adonisgeorgiadi: 70340615
7. @kostasvaxevanis: 289527193
8. @vanitysthasmin: 1005766052607938560
9. @akurkov: 62213337
10. @ve10ve_ghost: 1495855082734297095
11. @adarshburman2: 1976335622
12. @bassaces1: 1066275282653536256