Data Science for Digital Humanities 2

Final Project

Prof. Dr. Goran Glavaš
Lennart Keller

July 17, 2023

Overall Goal

- A project in which you need to combine several topics covered during the course
  - Explorative data analysis
  - Graph analysis
  - Machine learning
  - Statistical testing of results

Project: Analysis of Subreddit interactions

- **Task**: Predict positive / negative sentiment in messages from one subreddit referencing another subreddit

- **Data**: data about messages / communication between subreddits
  - Each data point represents one post from one *subreddit* which references another *subreddit*
  - Data given:
    - Source subreddit
    - Target subreddit
    - Binary sentiment label: +1 (positive), -1 (negative)
    - Precomputed features for the actual post (raw text not available)

Project: Analysis of Subreddit interactions

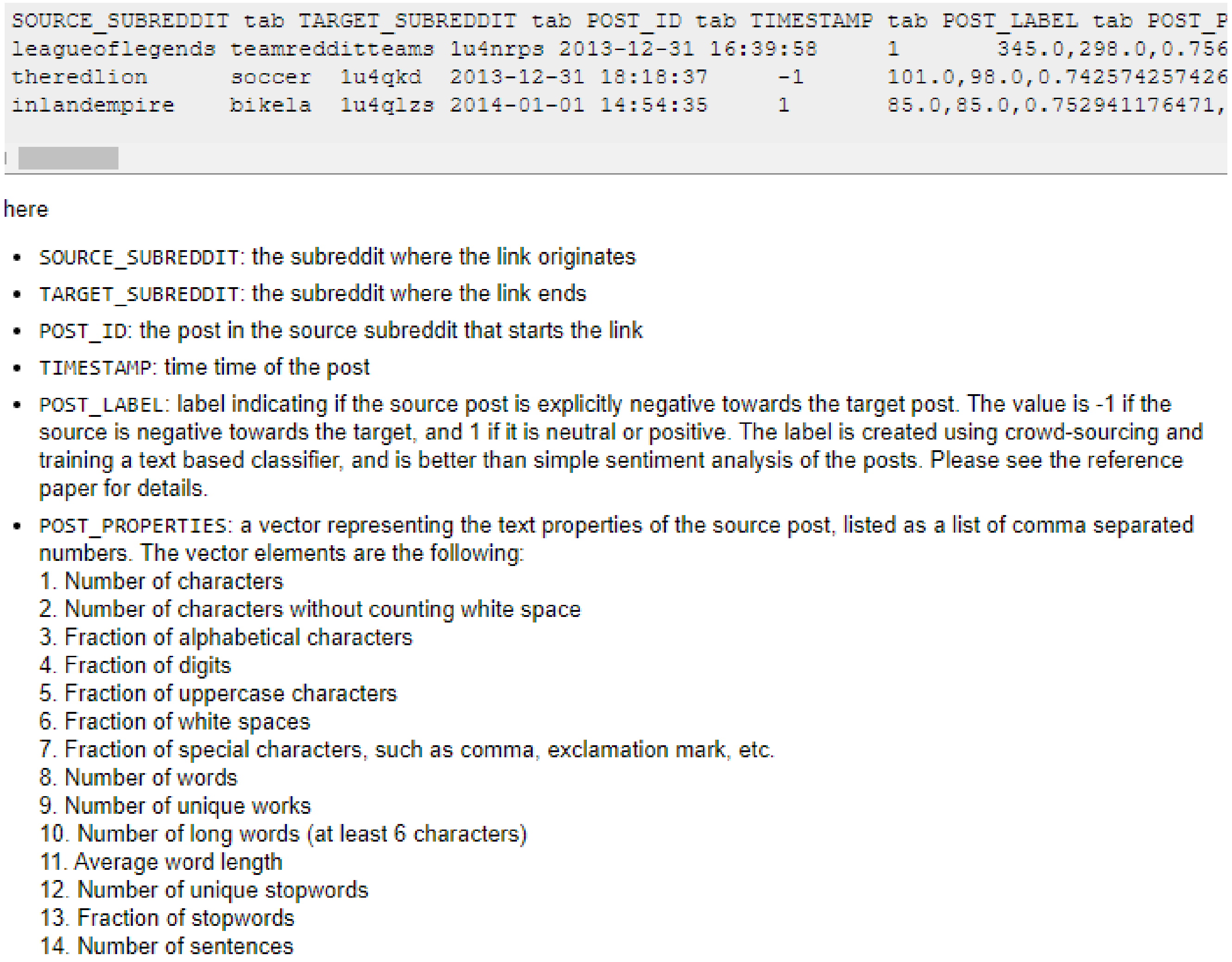
- Preliminary data analysis:
  - Analyze the data
  - E.g., distribution of posts per subreddits (as source and as target)

- Subtask #1: Build the (directed) social graph of subreddits
  - Subreddits are nodes, aggregate information about messages between two subreddits are properties of edges
  - Predict sentiment for the test set posts (5K) purely based on the (heuristics on the) graph, **do not** use numeric descriptors of post properties

Project: Analysis of Subreddit interactions

- Subtask #2: Machine learning to predict message sentiment
    - Use all available information as features and try to learn to predict the sentiment label with a supervised machine learning model
    - Do use given features, but feel free to compute additional features from the graph structure
    - Play with different ML models (in sklearn)
- Statistical comparison of models/results
    - Determine if one model is statistically dsignificantly better than another
    - Pick a suitable statistical test and apply it!

Project Data



```
SOURCE_SUBREDDIT tab TARGET_SUBREDDIT tab POST_ID tab TIMESTAMP tab POST_LABEL tab POST_P
leagueoflegends teamredditteams 1u4nrps 2013-12-31 16:39:58     1       345.0,298.0,0.756
theredlion      soccer  1u4qkd  2013-12-31 18:18:37      -1      101.0,98.0,0.742574257426
inlandempire    bikela  1u4qlzs 2014-01-01 14:54:35      1       85.0,85.0,0.752941176471,
```

here

- `SOURCE_SUBREDDIT`: the subreddit where the link originates
- `TARGET_SUBREDDIT`: the subreddit where the link ends
- `POST_ID`: the post in the source subreddit that starts the link
- `TIMESTAMP`: time time of the post
- `POST_LABEL`: label indicating if the source post is explicitly negative towards the target post. The value is -1 if the source is negative towards the target, and 1 if it is neutral or positive. The label is created using crowd-sourcing and training a text based classifier, and is better than simple sentiment analysis of the posts. Please see the reference paper for details.
- `POST_PROPERTIES`: a vector representing the text properties of the source post, listed as a list of comma separated numbers. The vector elements are the following:
  1. Number of characters
  2. Number of characters without counting white space
  3. Fraction of alphabetical characters
  4. Fraction of digits
  5. Fraction of uppercase characters
  6. Fraction of white spaces
  7. Fraction of special characters, such as comma, exclamation mark, etc.
  8. Number of words
  9. Number of unique works
  10. Number of long words (at least 6 characters)
  11. Average word length
  12. Number of unique stopwords
  13. Fraction of stopwords
  14. Number of sentences

Project Resources

- Dataset:
  - Training data: soc-redditHyperlinks-body.tsv: 313,600,538 posts!
  - Test data: soc-redditHyperlinks-body-test.tsv: 5000 posts
    - Evaluate predicted sentiment on test messages against gold labels

- Website of the dataset/task:
  - http://snap.stanford.edu/data/soc-RedditHyperlinks.html

Final Project Report

- You need to submit:

  (1) All of your code

  (2) The final report

  - At most 8 pages

- In the report, you need to
  - Describe the problem
  - Describe the data (based on your analysis)
  - Describe (in detail) methods you applied (graph analysis, ML)
  - Discuss the results and findings

- Submission date: as late as possible, probably beginning of Sep

All It Takes is Dedication and Effort!

Questions?

Awọn ibeere?

ਸਵਾਲ?

Küsimusi?

Turite klausimų?

Pitanja?

Sorusu olan?

Dúvidas?

質問は？

有问题吗？

Fragen?

¿Preguntas?

Frågor?

Questions?

Pytania?

Domande?

Питання?

Vragen?

Ερωτήσεις;

Porandukuéra?

คำถาม?

أسئلة؟