Tools for NLP Project

February 5th to March 31st, 2024

Introduction

This final project will combine the majority of the assignment tasks and lessons that have been done within the class. The deadline is on **31.03.2024 at 23:59** with zero option for extensions.

For the final project, you are required to find a partner **among the course participants**. If you haven't managed to find a partner and have already tried the course Teams chat without success, please contact us as soon as possible.

Make a single GitHub repository shared between you and your partner.

Please refer to the following GitHub documentation: and add **`IuliiaZaitova`**, **`Jihaiku`**, and **`pyRis`** as collaborators to your repository.

How to Submit the document

- ☐ Make a folder named: `btnlp_project`
- ☐ Place the `Report.pdf` file into the folder
- ☐ Place the `python_codebase.py` or `.ipynb` into the folder
- ☐ Place both yours and your partner's Assignment 5 Application package code into the folder
- ☐ Place the `python_codebase.pdf` into the folder
- ☐ Zip the folder and submit it to Teams
- ☐ Upload these to your shared private repository
- ☐ Your GitHub repository must present the summary of the report as its front page

Programming Requirements

Find a partner among the course students to collaborate with. Create one single GitHub repository where you both upload your version of the preprocessing package (`application05.py`) as of Assignment 5 completion. You will then create a bash file `combine_versions.sh` where both versions of your code are called and ran.

Ideal Specific Criteria

These tasks are assumed to have been completed for your **individual** codebases
(meaning that these should have been completed via the Assignments).

1. Preprocess your Twitter Dataset as in Assignment 3 and 4
2. Create a functioning class that packages all the items and does everything you implemented as in Assignment 3
3. Properly use PEP8 guidelines with Autoformaters and Pylinters to make your code readable as in Assignment 4
4. Use both NLTK/VADER and SpaCy/Textblob Sentiment Analysis on it as in Assignment 5
5. Run and show the results via visualizations (*e.g.* Confusion matrix / bar plots) comparing the results of each of your preprocessing phases as in Assignment 5

These are assumed to be for the **combined** codebase of you and your partner
(doing these is required by the project).

1. Combine your code with a partner and call within the `combine_versions.sh` both of your packages
    - Note: You will need to submit your original code within the folder with your new codebase
2. Preprocess the dataset at least with one version
3. Use the NLTK and SpaCy Sentiment Analysis libraries on the preprocessed dataset

Report Requirements

The file `report.pdf` has to describe your findings on the differences between sentiment analysis tools that you have used for the Twitter Dataset and how various aspects of preprocessing were either helpful or not nearly as useful as you expected.

Specific Criteria

- You must use the EACL 2024 template from Assignment 5 (you can in theory expand upon that very report and resubmit it)
- This must be 3 - 4 pages without a Title or Citation page.
- There should be sections roughly equating to:
    1. Introduction
    2. Dataset
    3. Methodology for Preprocessing
    4. NLTK/VADER Sentiment Analysis Scoring and Results
    5. SpaCy/Textblob Sentiment Analysis Scoring and Results
    6. Comparison of the two techniques for Sentiment Analysis
    7. Discussion + Conclusion
- It must include a .bib file for all references.

Grading Criteria

There are three sections to your grading, the only required section is the one noted as *Required.* You must complete this section to be considered for assessment for the final project. If you complete everything within the Required Section then you will pass with a 4.0.

The Optional section is items that you have in theory worked on throughout the five assignments which is labelled as *Optional.* And the section noted as *Bonus* is items that you were not tasked with doing while within the class lectures or assignments.

Each point obtained by doing the Optional/Bonus assignments will improve your grade by one scoring (*e.g.* 4.0 to 3.7, 3.7 to 3.3, 3.0 to 2.7 etc) and they are pass/fail with no partial points. In order to get 1.0, you must obtain 8 total points from either the *Optional* or *Bonus* sections to score a 1.0 for your final grade.

Required for the passing grade: (10 points possible)
(You must complete this section, not completing this will result in automatic failure)

- ☐ Submit a paper using the EACL Template (1 point)
- ☐ Submit a paper that is 3 - 4 pages long (1 point)
- ☐ Complete four of five assignments (4 points)
- ☐ Work with a partner via GitHub (1 point)
- ☐ Preprocess the dataset (1 point)
- ☐ Use both NLTK and SpaCy Sentiment Analysis tasks on the dataset (2 points)

Optional: (8 possible points)
(For each point completed, you will obtain an increase in score)

- ☐ Use an Autoformater on your new codebase
  (1 point)
- ☐ Score above a 7/10 on the Pylinter (ignore odd/wrong issues via the ignore option) on your new codebase
  (1 point)
- ☐ Each person has their own version of the preprocessing present
  (1 point)
- ☐ Each person has their own version of the Sentiment Analysis tasks and 2 visualizations (each including both models) present
  (1 point)
- ☐ You have a screenshot proving you used Coli Cluster
  (1 point)
- ☐ You combined/improved the preprocessing for the new codebase
  (1 point)
- ☐ You specify the differences between the Sentiment Analysis tasks scoring
  (1 point)
- ☐ Include a Discussion/Conclusion that includes meaningful points from previous Assignments
  (1 point)

Bonus: (8 possible points)
(For each point completed, you will obtain an increase in score)

- ☐ Complete Bash Script Optional Assignment
  (4 points)
- ☐ Additionally to NLTK/VADER and SpaCy/Textblob, implement a HuggingFace Sentiment Analysis
  (4 points)