Middle East Technical University | Department of Computer Engineering

CENG 463

Introduction to Natural Language Processing

Fall 2021-2022

Assignment 1

Due: 09.12.2021 Thursday, 23:55

In this assignment, you will implement a text classifier for English book descriptions. You are given a data gathered from www.goodreads.com/: Titles and descriptions of books from 8 different genres (philosophy, science-fiction, romance, horror, science, religion, mystery, sports). Based on that data, you will train two different machine learning classifiers: a Naïve Bayes Classifier and a Support Vector Classifier. Then, you will evaluate and compare the performance of your classifiers.

In order to simplify your job, the data is already divided into 3 parts: train(ing), dev(elopement) and test. Moreover, a partially filled code template is provided. You will first apply basic text processing steps that you think are required/effective for the given classification problem. Then you will select features (that classifiers will take into account). Thus, you will be able to train your classifiers. Classifiers will be trained on the training set. In order to fine-tune your classifiers by selecting more effective preprocessing steps and a better set of features etc., you will observe their performances on the development set. Once final versions of your classifiers are ready, you will evaluate them on the test set.

Reports

- Give your implementation details.

- Explain basic text processing operations you have applied. Why do you think they are required/effective/beneficial for the given task?

- Explain your feature selection. Why do you think your features are required/effective/beneficial for the given task?

- Evaluate performance of your classifiers. Give their accuracies, recalls, precisions and F1-scores on each genre and on the overall test set. Comment on those.

- Compare performances of NB and SVC classifiers. Did one of them significantly outperform the other? Why? Is this result expected?

- Give confusion matrix of your Naïve Bayes classifier. Analyse errors through confusion matrices (These kinds of books cannot be classified correctly because ... These genres are difficult to distinguish because…)

Submission

Submit your code and report via odtuclass.
Naming convention: **463_A1_<*your_id*>.py** for code,
463_A1_<*your_id*>_Report.pdf for report.

Notes

- In data files, each book occupies 2 lines. Titles are at odd numbered lines, whereas descriptions are at even numbered lines.

- The template code is expected to significantly reduce your effort. However, you do not have to use it. You are allowed to change it or totally discard it and code everything from scratch.

- Using Python is a must.

- You are encouraged to use Google Colab environment since it prevents OS/hardware driven errors/incompatibilities. (This is not an obligation. Working on your local is also allowed.)

- You can ask anything about the assignment via e-mail or discussion forum.

- You are encouraged to use discussion forum.

- Submit your code and report even if you cannot go far. You can get partial credit from even preliminary stages of implementation or unapplied ideas.

References

NLTK: https://www.nltk.org/book_1ed/ch06.html

https://www.nltk.org/book_1ed/

https://www.nltk.org/

Goodreads: www.goodreads.com/