Instructors: Ebru Sezer, Hayriye Çelikbilek
Course: AIN442 Fall 2021
Assignment: LSA and Word2Vec
Given: 22.12.2021
Due: 30.12.2021, 23:00 or 07.01.2022

Report Format

The following headings/sections are required in the project report, which will be relatively long. Each section should have reasoning and explanations, especially in the development section.

1) Cover: Your name, student number, course name, delivery date.
2) Introduction: Define the project in your own words. What is the problem description? Project goals?
3) Data: What are your materials? Define the data, its columns, units, statistics.
4) Method: What are your methods to solve the problem? Which classifiers have you selected? Why?
5) Development
    1) Plan: What is the project plan? Your requirements. (For example; your hardware requirements might affect the number of selected hyperparameters)
    2) Analysis: Analyze the features of data. Pairwise and individual graphics related to them. Which features have you selected? Why? What kind of analysis have you done? Why?
    3) Design: What is your solution overflow? Why? Which methods have you chosen? Why? What are their functions/properties? Why are they beneficial to our problem? Which other methods may be chosen? Why are they eliminated? How do you tune your solution?
    4) Implementation: Flow of the implementation. Explain your code parts such as functions, selected libraries/methods, Jupyter blocks.
    5) Programmer Catalog: Write down the time you spend for analysis, design, implementation, testing, and reporting. Can other developers reuse your code/solution? When? How? Your suggestions to other programmers. Add the source code here in the report.
    6) User Catalog: How can one use your code for other purposes / for other data? Prepare the user manual of the program (might use screenshots). State the restrictions, if any—your suggestions to users.
6) Results, Discussion and Conclusion: Your results. Which metrics have you used? Why? Visual statements are preferable. Performance evaluation and comments might be listed. Discuss your results and conclude with several statements about several aspects of your solution. You may use whatever is necessary to support/explain your statements/hypothesis. For example, a word cloud may show the word frequencies; later, you need a statement using frequency information. You also need to compare your results with other data and methods. Why did it work? Can we be sure that it worked? How?
7) References: Resources you have used to prepare the project and the report. Use numbers that are indicated in the text.

Assignment Instructions

Read and handle "turkish_dataset.csv" file properly, remember your prior knowledge from AIN214, AIN440, AIN442 courses, and your projects (wrong speling, pre-procesing, data-splits and their numbers, selecting classifier and its hyperparameters, etc.). Notice the number of topics given in the dataset. By using the Zemberek[1,2] library, apply necessary functions (spell checking, normalization, tokenization, stemming, lemmatizing) to the data.

Project's main problem is to represent the given texts by using LSA, Word2Vec and both combined [3, notice that it is a summation] and measure these three representations using your selected classifier (NB, RF, LR etc). Apply hyperparameter tuning, data splitting properly and compare your prediction results over a properly selected metric. Enrich your project with visualization of vector representations and your results.

Make sure to explain the given problem carefully and your project report is sufficient/enough to understand your problem, effort, and solution. Give a sufficient amount of evidence and support your conclusions. Lastly, your problem is a multivalue classification problem according to category labels given by the data. If you are using some artificial learning/modeling, make sure that you are using the sklearn library. Other obligatory libraries are pandas, NumPy, matplotlib, nltk, gensim, zemberek. Please remain within the course's library scope. You may take advice and ask questions during my office hours (lab course time) or any time that is not the last minute.

You will submit your code as "LSA2Vec.ipynb", data as given in the same directory, and your report as "Report.pdf" onto the codepost website (3 files). Please do not forget to write your name and student number at the beginning of your code. Grading will be based on 30% Report, 70% Code. Oral presentations of the code will be held in the next lab session one by one. The impact of this project on your overall grade is **twofold**.

Good Luck.

References

[1] https://github.com/ozturkberkay/Zemberek-Python-Examples
[2] https://github.com/ahmetaa/zemberek-nlp
[3] You may inspire from : https://github.com/cemoody/lda2vec