Practical: Spoken and Natural Language Understanding
Prof. Dr. Georges, Shubham Vijay Kurlekar(M.Sc.), Technische Hochschule Ingolstadt

Summer Semester 2023 from 17.04.23 until 15.05.23
Submission: On `https://moodle.thi.de`, Details below

2 - SMS Spam Detection

Problem Statement:

You are hired as an AI expert in the development department of a telecommunications company. The first thing on your orientation plan is a small project that your boss has assigned you for the following given situation. Your supervisor has given away his private cell phone number on too many websites and is now complaining about daily spam SMS. Therefore, it is your job to write a spam detector in Python *from scratch*. In doing so, you are to implement a Naive Bayes classifier yourself that can handle both bag-of-words (BoW) and tf-idf-*features* as input.
For the evaluation of your spam detector, an SMS collection is available as a dataset – this has yet to be suitably split into train and test data. To keep the costs as low as possible and to avoid problems with copyrights, your boss insists on a new development with Python. If you need special Python modules, you can discuss their use with the instructor before submission.

Datasets and Resources:

- SMS Spam Collection Dataset: `http://archive.ics.uci.edu/ml/machine-learning-databases/00228/`
- Knowledge from lecture: tf-idf, BoW

Milestones for sucessfully passing the Exercise:

Submission of Plan of Work Report

- Introduction: Problem Description, available data, methods and metrics used in evaluation, outline of solution etc.
- Data: Examples/samples of dataset(s), key characteristics, necessary data preparation for further processing etc.

Submission of Progress Report

- Methods,Models: short explanation of how (implemented / proposed) models / methods work etc.
- Experiments: transparent experiment design, choice of (hyper-)parameters etc.

Submission of combined Final Report and Code

- Evaluation and discussion: explanation of evaluation, discussion of results etc.
- Conclusion: what has been done? Drawbacks and potential improvements? etc.
- Code: only Python, no other modules, e.g. NLTK. No jupyter notebooks.

General Tips for the reports

1. Design and data preprocessing - Plan of Work Report

- Define the requirements for your spam detection program using tf-idf and bag-of-words features (hint: *Bayes* classifier)
- Familiarize yourself with the given dataset. Consider how to split the data into train and test data and which text preprocessing is useful.
- Selection of suitable data structures (with regard to performance) and derivation of an initial code framework for the spam detector.

2. Baseline implementation and experiments - Progress Report

- Executable implementation of a baseline model M_1 .[1]
- Training the spam detector on training data and testing on test data.
- stretch-goal: you implement another spam classifier, e.g. K-NN, and compare the classification results from different models and features
- stretch-goal: integration of performance measurements (e.g. runtime, memory, ...)

3. Desired Code and Experiment - Final Report

- Implementation of Naive Bayes spam detector
- Suitable visualization (chart, table, ...) and interpretation of your results with suitable metrics on the given dataset (this must be suitably split into train and test data beforehand)
- Explanation of potential problems of your developed solution and suggestions for improvement.
- stretch-goal: experiment and compare results with other freely available classification datasets of your choice.
- stretch-goal: cross-validation approach in model evaluation

Submission: A 3-Page final report covering all the above points. Final source code should be able to reproduce preprocessing, experiments and obtained results.

[1] The term *Baseline*-Model means it is considered as the reference model, because one would later like to make comparisons with other models.

Requirements for the Reports:

- **PDF-Document**, Font size 11pt
- appropriate visualization of the results (if experiments are required), e.g. in the form of tables or diagram
- The required number of pages must be realized via text (visualizations such as tables, diagrams, ... do not count). In addition, the number of pages must not be exceeded or fall short of the requirement..
- Whenever you use other people's ideas (even if only in spirit), cite them appropriately so that the other sources can be found again. I.e., create a source list or refrences.. [2]

[2] a citation guide: `https://www.slub-dresden.de/en/research/writing-and-publishing/zitieren-gute-wissenschaftliche-praxis`