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

Summer Semester 2023 from 20.03.23 until 17.04.23
Submission: On `https://moodle.thi.de`, Details below

1 - Wikipedia Language Model

Problem Statement:

A customer of yours is dissatisfied with the quality of the speech recognition. After a conversation with the customer, you find out that he dictates books. However, the speech recognition was not built for this purpose. Accordingly, initial investigations on a book excerpt have also shown that the language model used is not suitable. For building a better language model for the application, you have asked your customer to provide a text from the book (see "Datasets and Resources")

In order to save costs and also to avoid problems with copyrights, your company has decided to not use existing solutions for this project. A *new program* for estimating language models is now to be programmed in Python *from scratch*. If you need special Python modules, you can discuss their use with your instructor before submission.

Datasets and Resources:

- WikiText-2 (**raw/unprocessed**), Train, Dev, Test
- Penn Treebank, word-based

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 of your program for estimating a 2-gram language model M_1
- Select appropriate data structures (in terms of performance) and derive an initial code framework. Justification of your design decisions
- Implement data preprocessing for the data sets. Consider what kind of preprocessing is necessary for the calculation of a language model
- stretch-goal: FST for preprocessing of text data
- stretch-goal: Your design takes into account the later implementation of a language model M_2 with back-off

2. Baseline implementation and experiments - Progress Report

- Executable implementation of a baseline language model M_1 (without Backoff) with perplexity calculation.[1]
- Test your implementation of M_1 on the preprocessed train and test datasets.
- Recording of the first perplexity results on the two data sets (incl. interpretation)
- stretch-goal: integration of performance measurements (e.g. runtime, memory, ...)

3. Desired Code and Experiment - Final Report

- Implementation of routines for data preprocessing of the Wikitext2 and PTB dataset.
- Baseline implementation of a 2-gram language model M_1, training and testing with appropriately preprocessed Wikitext-2 data.
- above experiments with at least one further data set of comparable size (e.g. PTB)
- Recording and interpreting your language model evaluation results on the two data sets.
- Explanation of potential problems of your developed solution and suggestions for improvement.
- stretch-goal: Implementation of a back-off variant from lecture to improve the quality of your language model $\Rightarrow$ results in new model M_2. Comparison of the perplexities of your model(s) M_1, M_2 on the test data.

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`