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

Summer Semester 2023 from 15.05.23 until 12.06.23
Submission: On `https://moodle.thi.de`, Details below

3 - Word Embeddings and Clustering

Problem Statement:

In a few years, autonomous flying cabs will commute between Ingolstadt and Munich. The flight time will be around 18 minutes. An intelligent personal voice assistant will make the journey as pleasant as possible. Word embeddings will be used as input for various NLP subcomponents of the voice assistant. As an expert in speech understanding, you have been tasked to program and train a simple **word embedding** (word2vec) model. You are also to perform word clustering experiments based on your computed word embeddings and visualize the word clusters appropriately. You will be provided with two datasets of different sizes for your calculations and visual exploration. Since this is a prototype that is in competition with other companies, your program should stand out as much as possible. This is to be achieved by a complete redevelopment of the word2vec calculation using Python/PyTorch. For clustering and visualization you may use open-source software. If you need special Python modules, you may discuss their use with the instructor before submission.

Datasets and Resources:

- Datasets to compute Word Embeddings (different sizes)
- Python, PyTorch, Tools for Visualisierung (t-SNE, ...)

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

- Discussion of the structure and operation of a *word2vec* word embedding and how that can be implemented in PyTorch. Decide whether to use CBOW or Skip-gram.
- Your model should include the following components: Bag-of-Words Encoding, Feed-Forward-Network, Softmax.
- Creating an initial code framework for your model to compute word embeddings using PyTorch.
- Note: You are not allowed to use PyTorch's *Embedding-Layer* – this is to be implemented by yourself in this task.

2. Baseline implementation and experiments - Progress Report

- executable implementation of your neural network. Efficiency of your implementation is a plus
- Calculation of word embeddings with your model on at least one of the two given datasets
- Visualization of Word Embeddings with methods like t-SNE. For this you may use open-source software.

3. Desired Code and Experiment - Final Report

- Find word clusters based on your calculated word embeddings on at least one of the two datasets. You can use existing k-Means implementations here.
- What word clusters result for different choices of k? Presentation of your exploratory clustering experiments in the final report.
- Explain potential problems of your implemented neural network for computing word embeddings and point out suggestions for improvement.
- stretch-goal: independent implementation of a k-means clustering algorithm based on word embeddings
- stretch-goal: Visual comparison of the word embeddings of your network with those you can calculate with other freely available tools

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

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.. [1]

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