MSAI-337: Natural Language Processing
Group Project #1

(5.0 points + 1.25 bonus points)

Winter 2022

Description: For this assignment, your group will construct a corpus from a source file provided to the group. This source file is scraped from Wikipedia based upon results from the Wikidata Query Service (see https://query.wikidata.org/). HTML, images, tables, etc., have been removed. You will implement a sequence of transformations to convert this source text into a corpus suitable for natural language modeling or natural language processing tasks. Certain aspects of this assignment are deliberately under-specified so that you can experience some of the real-world challenges of corpus preparation.

Tasks: Please perform the following tasks to construct your corpus:

1. Tokenize the source text using the NLTK tokenizer (see https://www.nltk.org/) and convert all text to lower case. (0.5 pts)
2. Replace (i) years, (ii) decimals, (iii) date days, (iv) integers and (v) all other numbers with `<year>`, `<decimal>`, `<days>`, `<integer>` and `<other>` tags, respectively. You can write your own rule-based code to perform this task or alternatively formulate "regular expressions" to perform this task. Describe your methodology/expressions in about one paragraph. (1.0 pts)
3. Split the corpus into training, validation and test sets to approximate an 80/10/10 distribution using a methodology of your choice and which makes sense in a natural language processing setting. Please describe your methodology in a few sentences. (0.5)
4. Apply a frequency threshold of three to the corpus and report summary statistics. Summary statistics should include the (i) number of tokens in each split, (ii) the vocabulary size, (iii) the number of `<unk>` tokens, (iv) number of out of vocabulary words, (v) the number of types mapped to <unk>, (vi) the number of stop words in the vocabulary and (vii) two custom metrics of your choice. Summary metrics should be presented in a table. Please provide a few sentences describing and motivating you choice of the two custom metrics. (1.0 pts)
5. Write functions to (i) construct integer representations of an input sequence of text using the corpus vocabulary and (ii) a function to recover the text using the list of integers generated by the first function. Please see `starter.py` for recommended signatures for these functions. (2.0 pts)
6. Bonus: Construct a word-piece tokenization of the source text using either the byte-pair encoding algorithm covered in class or a Huggingface word-piece tokenizer. Either approach should use a target vocabulary of approximately 5,000. There are a large number of tutorials on using Huggingface for this tasks -- a few to get you started include:
    - https://huggingface.co/docs/transformers/tokenizer_summary
    - https://huggingface.co/docs/tokenizers/python/latest/pipeline.html
    - https://towardsdatascience.com/how-to-build-a-wordpiece-tokenizer-for-bert-f505d97dddbb
    - https://www.kaggle.com/funtowiczmo/hugging-face-tutorials-training-tokenizer

    Reproduce the table from #4 above for this word-piece tokenized corpus.
7. Bonus: Specify the query used to produce the source text -- there are four conditions. (0.25 pts)

What to Turn In: Your submission should include (i) a single PDF file for all items #1-#4 and bonus questions, (ii) the train, validation and test corpora text files and (iii) the Python code written for #5 in a file named `group_n.py` where *n* is your group number. All of these files should be submitted as a single zip file.