Home Assignment 1

In this home assignment, you will implement and evaluate several n-gram language models.

Your tasks are:

1. Choose a text corpus in one language out of English, Hebrew, or Arabic. Wikipedia can be a useful source (e.g., a plain text one: http://mattmahoney.net/dc/textdata.html), but it is too large to use in its entirety. You'll want enough text to implement reasonable language models, but not too much, such that the computations are still efficient. Be thoughtful and make sure to describe your considerations in deciding on the dataset size and composition.
2. Implement basic unigram, bigram, and trigram language models. Report their perplexity values as we've learned. Make sure to apply **proper experimental methodology**.
3. Implement the following modifications on the trigram language model and report their perplexities:
    a. **Add-delta smoothing**. Experiment with delta=1 and delta=2.
    b. **Knesser-Nay smoothing**. See Jurfasky & Martin, Section 3.7, for details on this smoothing technique.
4. Compare and discuss the different language models you have implemented. What are their advantages and disadvantages? How are they reflected in your experimental results?

Deliverables:

1. A Jupyter notebook that contains code for reproducing all your experiments. The notebook should run as is out of the box. It should also download the data from some external repository. You can use libraries we've used in the course, and standard Python libraries, but no other libraries that require installation. You may **not** use existing implementations of the n-gram language models; rather, you should implement them yourself as we've done in the course.
2. A written PDF report in English or Hebrew describing your work, including all the steps: from data collection and preparation, through implementation, running experiments, reporting results, and discussing them. The report should contain mathematical definitions for all the models you have implemented.

Guidelines:

1. You have 48 hours to complete the assignment. Due date: June 27, 2023, 8:00am. We will not accept late submissions.
2. You should complete the assignment by yourself. You may consult the course materials and other external materials. However, make sure to properly reference such materials in your report.
3. You can re-use code from the labs and homework assignments in the course, but make sure to indicate which parts of your code are taken from which course material.
4. Submit the jupyter notebook **and** the PDF report by email to Shaked (shakedbr@campus.technion.ac.il) and Yonatan (belinkov@technion.ac.il).

5. If you have questions about the assignment, you should submit them in writing by email to Shaked (shakedbr@campus.technion.ac.il) and Yonatan (belinkov@technion.ac.il), up to 6 hours from the start time. We will respond by 24 hours from the start time.