EEE 486/586

Statistical Foundations of Natural Language Processing

Assignment 1

(Due 12/03/2023, 23:59 PM)

General Instructions

Groups: You are expected to work alone.

Assignment: In this assignment, you will explore some of the fundamental laws of statistical natural language processing. Also, you will gain basic hands-on experience and skills to devise your own language data, form a corpus and perform basic pre-processing on this data. You are also expected to accumulate experience on writing technical documents, develop an efficient organizational structure for your reports where you can present results of your experiments and organize your findings, insights, and conclusions in a technical and precise manner.

Important remarks:

(a) Collaboration and code sharing among students are prohibited.

(b) You are not allowed to use any NLP specific libraries; you can use other common libraries such as "numpy" or "codecs" in the python environment, or any other non-specific libraries in the coding environment of your choice.

(c) Properly label all your figures throughout your report.

(d) Your reports will be evaluated based on the proper completion of tasks, clarity of presentation of results, sufficiency of discussions regarding the results, quality of writing, plots and organization of the report and your possible insights and comments.

(e) In all relevant parts of the assignment, please avoid *poetry*. All your choices should be in *prose* form of literature. Of course, you do not need to check possible sporadic quotations of pieces of poetry within a large text. Once you are sure that the main dominant form is prose, then you are fine. (Prose: a form of language that has no formal metrical structure. It applies a natural flow of speech, and ordinary grammatical structure, rather than rhythmic structure, such as in the case of traditional poetry. Novels, textbooks and newspaper articles are all examples of prose.)

(f) Please see the following for information about academic honesty and plagiarism as well as our Course Syllabus:
http://ascu.bilkent.edu.tr/Academic_Honesty.pdf

Assignment

(a) Go to http://www.gutenberg.org website to find e-books. Then, determine three authors of your choice. Choose three books from each author and download them as UTF-8 text files. Before downloading, shoot a glance at whether there are unreadable characters in the text. Try **not** to choose small-sized books; it will be better if you prefer text files with sizes above 1 MB.

(b) Also, determine three different types of literature (eg. Science-Fiction, Romance, Classic Novels, Horror, Dystopia, etc.) of your choice. Choose three books from each type and do the same procedure in Part (a).

(c) Once you obtained the texts, get rid of Gutenberg Project information which takes place at the top and the bottom of the books. Then, start tokenization and preprocessing the corpus. In tokenization, get rid of every punctuation marks and cast every word to lowercase. It is not so crucial that how you treat words with an apostrophe, it is up to you ("isn't" $\rightarrow$ "is not" / "isn't" $\rightarrow$ "isnt" does not matter so much for our purposes).

(d) Search for and determine a common English Language stop-word list and obtain stop-word-removed versions of your texts above. After doing this, you will have a corpus that contains: three books from three different authors and three books from three different literary types with a total number of 18 books. For each 18 books, you will have versions before and after stop-word removal which brings the total to 36.

(e) While going over your corpus, you should create vocabulary files carrying **the word types** along with their **frequencies**. You will implement this for all books separately, before and after stop-word removal.

(f) For the authors you chose in Part (a), compose corresponding larger *author corpora* by combining all three books of each author. Plot the Zip's Law curves for the three author corpora in linear scale (i.e, one plot with three author corpora on each). Describe what you observed in plots and discuss whether your plots are consistent with the Zipf's Law. Then, for each of the three authors separately, plot only log-log curves for all of the three books for this author in one figure (i.e., one log-log plot for each of the three authors with his/her books are shown separately). Again give your observations briefly.

(g) In this part, you will examine the relationship between the token size in the corpus and the vocabulary size (number of unique words or **types**). You need to keep track of the number of word types with respect to the increasing token size as you traverse along the corpora. You may want to check the number of word types for every 5000-10000 tokens or so. As in the previous part, obtain *author corpora* for all of the three authors by combining their corresponding books. Plot the relationship between vocabulary size and token size (**token size on the horizontal axis**) of three author corpora on two plots for normal and log-log versions (i.e, two plots each carrying three different author corpora). Properly label the figures like previous ones. Give a brief description of the patterns you have seen. Do a little research about this kind of behavior of texts. What is the name of this relation? Explain that relation. What are its parameters?

(h) Now, we will proceed with only log-log plots. For each of the three authors, plot word type-token relations in log-log format where you will show the entire nine curves (each corresponding to a book) with different coloring for the authors. Repeat with three author corpora (i.e, one more plot with three curves corresponding to each author). Again briefly explain your observations.

(i) For all of the curves in Part (h), find the slopes of best-fitting lines. (This has become a hint.) Present the slopes in proper tables and also denote some of them on the plots of Part (h) in order to sufficiently convey your findings. Comment on your observations.

(j) Instead of using authorship, repeat Parts (h) and (i) by using the literary types corpora you composed.

(k) By using your findings in Parts (g), (h), (i), and (j), do you think you can derive simple clustering methodologies in which you automatically group books so that the members of each group belong to the same author or literary type? (Three groups of three members each for the author-based 9 books; and three groups of three members for the literary type based 9 books.) Discuss and elaborate on with your previous experimental results and findings. What about doing this with five different authors?

(l) Does removing stop-words significantly change your findings and observations in Parts (g), (h), (i), (j), and (k)? Perform some experiments and report your conclusions with appropriate numbers of plots and tables, and explanations that you deem sufficient to convey your results to the reader.

(m) Briefly read the paper "W. Li, "Random texts exhibit Zipf's-law-like word frequency distribution," in IEEE Transactions on Information Theory, vol. 38, no. 6, pp. 1842-1845, Nov. 1992." You may also want to search for and have a look at more modern papers on this issue to get the general idea. Then, construct a randomly generated corpus for yourself with a size of approximately equivalent to the combined size of two novels that you used in previous parts. Does it show a Zipfian behavior? Does it show the behavior that you observed in Part (g). Comment on.

Report Preparation

1. **The deadline for the Final Report is 12 March 2022, 23:59PM.** Late submissions will be penalized according to the policy in the Course Syllabus.
2. Each report should be typeset, **no handwriting is allowed**.
3. I recommend using LaTeX, though it is not mandatory. If you did not use it before, take this as a chance for getting used to that good practice.
4. Each report must be uploaded to Moodle.
5. **A single PDF file** should be submitted titled `'name_lastname_studentid.pdf'`.
6. Print out of all of your code should be appended to the report as an Appendix.
7. Each report should contain the following sections clearly separated with headings:

Abstract: A one-paragraph summary of all major aspects of your report from Introduction to Discussion. What is this report about? No references should be given, and the abstract should be self-contained.

Introduction: Briefly state the general topic and essence of the assignment. If possible establish the topics under study by briefly over-viewing the existing literature. State the purpose of your work. Give a general description of what the rest of the report will be about.

Corpus Construction and Implementation: Explain the corpus that you will be composing and using. Explain the pre-processing tools and/or methods that you used to process your corpus with proper references. Describe all qualitative/quantitative properties of your corpora such as the size, contents and other information about the corpus. Briefly comment about your choices in constructing your corpus. Give a summary on your implementation details such as which programming infrastructure you used in performing your experiments and analysis, etc.

Results: You need to mention all parts of the assignment with separate subsections where you see necessary. Present your key results, illustrate your outputs visually with the help of figures and tables. Detailed numbers/plots should be provided in illustrations, and each figure or table should contain a paragraph-long, self-contained caption that explain the contents. Figures/tables should be referenced in appropriate sections and the main trends/results should be stated in the text.

Discussions & Conclusions: Interpret your results in light of experimental findings. Which parts of your analyses worked, and which failed? Does the methodology have drawbacks, flaws and rooms for improvement? By considering the assignment as a whole, what did you learn? Try to encapsulate the central point of the entire assignment and summarize the findings in a concise way.

References: A list of all referenced material formatted according to standard conventions used in journals. Crude, unformatted lists are not acceptable.