67978: A Needle in a Data Haystack
Introduction to Data Science
Homework #3: Viz,Text Mining
Due: **18 Jan, 11:50pm, on Moodle**

THE HEBREW UNIVERSITY OF JERUSALEM

Instructions: Unless stated otherwise, submission is done **individually**. We rely on trust. You may discuss assignments verbally, but do not share solutions with other students. Cheating is a serious offense and will result in severe penalties.

Files should be zipped to **ex_3_First_Last.zip** (with your first and last name). The theoretical part should be a single file in pdf format only (no docx or jpg). If you're submitting handwritten answers, make sure they are crystal clear (both the handwriting and the scan – please avoid shaky cell pictures taken in a dark room :)). Submissions which fail to comply will not be graded.

Problem 1 (Stemming).

The following is part of the beginning step of the stemming algorithm developed by Porter (1980). Porter's algorithm consists of a set of rules, in the form of S1 → S2, which means that if a word ends with the suffix S1, then S1 is replaced by S2. If S2 is empty, then S1 is deleted:

$$\begin{gathered} \text{IES} \rightarrow \text{I} \\ \text{SS} \rightarrow \text{SS} \\ \text{S} \rightarrow \end{gathered}$$

In a grouped set of rules written beneath each other (as above), only one is applied, and this will be the one with the longest matching S1 for the given word.

(a) What is the purpose of including an identity rule such as SS → SS?

(b) Given the above set of rules, what do the following words map to?

PONIES, TIES, CIRCUS.

(c) What rules should you add to correctly stem CARESSES?

(d) True or false (and why): Stemming increases the size of the vocabulary.

(e) True of false (and why): In a search engine, if you stem the documents and the query, it never lowers precision.

(f) True of false (and why): In a search engine, if you stem the documents and the query, it never lowers recall.

Problem 2 (The Seven Basic Types of Data Stories).

Read about The Seven Basic Types of Data Stories.
(`http://dataremixed.com/2015/03/tapestry-2015-seven-data-story-types/`, `http://mediashift.org/2015/06/exploring-the-7-different-types-of-data-stories/`). Basically, they divide data story visualizations into

(a) Narrate change over time (how has x changed over the last decade?)

(b) Start big and drill down (how much x is there in the world? How much in my zipcode?)

(c) Start small and zoom out (there is this much x in your zipcode. See how much there is in the world)

(d) Highlight contrasts (see how far apart the highs and lows of x are)

(e) Explore the intersection of trends (what does it mean when x grows to be greater than y?)

(f) Dissect the factors (see the how much of xâs growth is caused by y and z)

(g) Profile the outliers (see how x is not at all like z)

Download the Olympics data from class: `tinyurl.com/67978olympic`. Pick **three** types of data stories. For each type, submit

- The name of the data story type.
- A visualization telling such a story (can be excel, no need to go fancy).
- One line explaining the story the visualization is telling.

Problem 3 (Text Mining (Coding Question)).

Download a book from Project Gutenberg `https://www.gutenberg.org`. You should submit two things: The code and the answers. You can use existing packages (recommended: NLTK for python, Stanford CoreNLP for java).

(a) Which book? :)

(b) Tokenize the text. Count occurrences for each token. Plot the results (y axis: log frequency, x axis: log rank). Also print the top 20 tokens.

(c) Repeat (b) after removing stopwords.

(d) Repeat (b) with stemmed text.

(e) Run POS-tagging on the original text. Count occurrences for each POS. Plot the 10 most frequent POS and their frequency (the names of the POS should appear in the figure).