47717: Data Mining
Homework #3: Text mining
Due: **20 June, 11:59pm, on Moodle**

THE HEBREW UNIVERSITY OF JERUSALEM

Read the instructions **carefully** (that's a good idea in general).

- Each person submits their own theoretical part. The theoretical part should be a single file in pdf format only (no docx or jpg) named **ex3t_ID.pdf** (ID is your ID).
- If you are submitting handwritten answers, make sure they are crystal clear.
- Only **one** person from each group should submit the practical part code and answers. In the practical part submission link, the person who submits should enter the partner's name (using the add partner button). In addition, *both* people should write the name and id of their partners in the theoretical part pdf.
- For the practical part, the person who submits should submit a single ZIP file named **ex3p_ID.zip** (where ID is the ID of the person submitting). The zip should contain a folder named **code** and a folder called **output**.
- The meta question should be answered through a questionnaire. The link will be posted on the moodle.
- Points may be reduced for submissions that fail to comply.
- Make sure you follow the News forum for any updates.

Problem 1 (NLP Challenges (1)). For this question you can use online demos.

(a) Find a set of three English words that stem to the same form, but really shouldn't – as in, they all have very different real roots (for example: University, universe, universal). Write what they stem to. (7pt)

(b) Find three pairs of words that should have the same root but that the Porter stemmer stems to different roots (e.g., matrix and matrices). Write what they stem to. (7pt)

Problem 2 (NLP Challenges (2)). For each of the following ambiguous sentences: (1) Indicate whether the sentence is structurally ambiguous, lexically ambiguous (a word means different things), or both, (2) paraphrase the possible meanings. (21pt)

(a) Hershey Bars Protest

(b) Giant Waves in California

(c) New Vaccine May Contain Rabies

Problem 3 (Text Mining (Coding Question)).

Download a book from Project Gutenberg `https://www.gutenberg.org`, preferably one that you know. You should submit two things: The code and the answers. You can use existing NLP and viz packages. I recommend using NLTK or spaCy. For the tag clouds, you can use online tools.

(a) Which book? :)

(b) Tokenize the text. Count occurrences for each token. Plot the results (y axis: frequency, logarithmic scale. x axis: rank, logarithmic scale; in other words, the tokens are sorted along the x-axis so that the frequencies are decreasing, axes are in logscale. See `https://www.intmath.com/blog/mathematics/zipf-distributions-log-log-graphs-and-site-statistics-702`). Also print a list of the top 20 tokens (separate from the plot). (10pt)

(c) Repeat (b) after removing stopwords. (8pt)

(d) Repeat (b) after removing stopwords and stemming the text. (8pt)

(e) Run POS-tagging on the original text. Extract all the *adj+noun phrases*. For this exercise, we will define it as one or more adjectives, followed by one or more nouns, including proper nouns (the longest such sequence). For example – delicious peanut butter cookies, oval office. Repeat (b) using adj+noun phrases as tokens (that is, count occurrences of each adj+noun phrase, print the top 20 phrases and plot the log-log graph of counts). (14pt)

(f) Show one example sentence where POS tagging made a mistake (explain). (9pt)

(g) Create a Tag cloud (word cloud) of proper nouns (NNP, NNPS). I like `http://www.wordle.net/`, but you can use anything. Does it correspond to what you know about the book? (8pt)

(h) Write a regular expression to find the set of all strings with two consecutive repeated words, with potential punctuation between them (e.g., "well well", "so so", and "no, no"). (Hint: $\backslash b$ is a word boundary, and \1 references the first captured match). Run on your text and report any found. (8pt)

Problem 4 (Meta). How long (in hours) did this assignment take? Please answer using the link, not in the pdf.