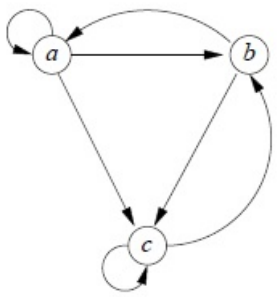
67978: A Needle in a Data Haystack
Introduction to Data Science
Homework #3: Pagerank, Text mining, Viz
Due: **9 Jan, 11:59pm, on Moodle**

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

- There will be two submission links, for theoretical and practical parts.
- 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. **All three 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**.
- Points may be reduced for submissions that fail to comply.
- The **meta** question should be answered through a questionnaire. The link will be posted on the moodle.
- Make sure you follow the News forum and HW forum for any updates.

Problem 1 (PageRank).
Compute the PageRank of each page in the graph below:



(a) Assuming $\beta = 0.6$, regular teleports. (5pt)

(b) Assuming $\beta = 0.6$ and the teleport set is $\{a, c\}$ (5pt)

(c) **[Hard! 5pt]** Construct, for any integer n, a graph G_n with a subset of n nodes S such that depending on β, any of the n nodes can have the highest PageRank among S. G_n can have nodes in it other than S.

(d) Solve (d) for $n = 3$. (5pt)

Problem 2 (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). (5pt)

(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. (5pt)

Problem 3 (NLP Challenges (2)). For each of the following, indicate whether the sentence is structurally ambiguous, lexically ambiguous (a word means different things), or a mixture of both. In each case, paraphrase the possible meanings and draw their corresponding parse trees (To the best of your ability. Don't worry about getting the labels exactly right beyond basic noun/verb/adj/adverb. I care mostly about the basic structure). (10pt)

(a) Hershey Bars Protest

(b) Giant Waves in California

(c) Stolen painting found by local maniac

Problem 4 (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. 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: log frequency, x axis: log rank). Also print a list of the top 20 tokens (separate from the plot). (10pt)

(c) Repeat (b) after removing stopwords. (5pt)

(d) Repeat (b) with stemmed text. (5pt)

(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. (14pt)

(f) Show one example sentence where POS tagging made a mistake (explain). (6pt)

(g) Homographs are words with the same spelling and different POS. Cound number of POS for each word in your text (no stemming or stopword removal). Output the word + list of POS for top 10 (most POS) and bottom 10 (least POS) words. (5pt)

(h) Create a Tag cloud (word cloud) of proper nouns (NNP, NNPS). Does it correspond to what you know about the book? (7pt)

(i) 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 $\backslash 1$ references the first captured match). Run on your text and report any found. (8pt)

Problem 5 (Meta). How long (in hours) did this assignment take? Please answer using the link, not in the pdf.