

## HW 4 - PageRank

Total points: 100

Issued: 10/25/2018 Due: 11/08/2018

The goal of this programming assignment is to use PageRank to derive a ranking of words in a document based on their PageRank scores. The PageRank score of a word serves as an indicator of the importance of the word in the text.

For this assignment, you will use a collection consisting of titles and abstracts of research articles published in the WWW conference. The dataset is available for download from blackboard. Each document is POS tagged.

Tasks:

1. Write a program that loads each document into a word graph (either directed or undirected). In doing so, tokenize on whitespace, remove stopwords, and keep only the nouns and adjectives corresponding to {NN, NNS, NNP, NNPS, JJ}. Apply a stemmer on every word. For each candidate word, create a node in the graph. Add an edge in the graph between two words if they are adjacent in the original text.
2. Run PageRank on each word graph corresponding to each document in the collection as follows:
  - Initialization:  $\mathbf{s} = [s(v_1), \dots, s(v_n)] = [\frac{1}{n}, \dots, \frac{1}{n}]$ , where  $n = |V|$ .
  - Score nodes in a graph using their PageRank obtained by iteratively computing the equation below. Assume the weight  $w_{ij}$  of an edge  $(v_i, v_j)$  is calculated as the number of times the corresponding words  $w_i$  and  $w_j$  are adjacent in text.

$$s(v_i) = \alpha \sum_{v_j \in \text{Adj}(v_i)} \frac{w_{ji}}{\sum_{v_k \in \text{Adj}(v_j)} w_{jk}} s(v_j) + (1 - \alpha)p_i, \quad (1)$$

where  $\alpha$  is a damping factor ( $\alpha = 0.85$ ) and  $p_i = \frac{1}{n}$ .

3. After the PageRank convergence or a fixed number of iterations is reached (e.g., 10 iterations), form n-grams of length up to 3 (unigrams, bigrams and trigrams) from words adjacent in text and score n-grams or phrases using the sum of scores of individual words that comprise the phrase.
4. Calculate the MRR for the entire collection for top- $k$  ranked n-grams or phrases, where  $k$  ranges from 1 to 10, using the gold-standard (author annotated data) provided in blackboard:

- Mean reciprocal rank, MRR

$$MRR = \frac{1}{|D|} \sum_{d=1}^{|D|} \frac{1}{r_d}$$

$r_d$  is the rank at which the first correct prediction was found for  $d \in D$ .

5. [\[Extra-credit - 30 points\]](#) Compare the MRR of the above PageRank algorithm with the MRR of a ranking of words based on their TF-IDF ranking scheme. Calculate the TF component from each document and the IDF component from the entire collection.

Submission instructions:

1. Write a README file including:
  - A detailed note about the functionality of each of the above programs,
  - Complete instructions on how to run them
  - MRR values for each  $k$  ranging from 1 to 10.
2. Make sure you include your name in each program and in the README file.
3. **Make sure all your programs run correctly on the CS machines. You will lose 40 points if your code is not running on these machines. The path to the data should be an input parameter, and not hardcoded.**
4. Submit your assignment through blackboard.