HW 2 - PageRank

Knowledge Discovery in Social and Information Networks - 2018
Summer 2018
Total points: 100
Issued: 06/18/2018 Due: 06/21/2018

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 the WWW collection consisting of titles and abstracts of research articles published in the WWW conference. The dataset is available for download from sicuaplus. 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 words, create a node in the graph. Add an edge in the graph between two words if they are adjacent in the original text. 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.

2. Run PageRank on each word graph corresponding to each document in the collection as follows:

   - Initialization: $\mathbf{s} = [s(v_1), \cdots, s(v_n)] = [\frac{1}{n}, \cdots, \frac{1}{n}]$, where $n = |V|$.
   - Score nodes in a graph using their PageRank obtained by recursively computing the equation:

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

   where α is a damping factor ($\alpha = 0.85$) and $\mathbf{p} = [\frac{1}{n}, \cdots, \frac{1}{n}]$.

3. After the PageRank convergence or a fixed number of iterations is reached, 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, as follows, using the gold-standard (author annotated data) provided in sicua:

   - 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 - 50 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 virtual machines.
4. submit your assignment through sicuaplus.