COS 435, Spring 2016 - Problem Set 1, Part 2

Due 11:59 pm Wednesday February 17 by DropBox submission
Due at 5:00 pm, Wednesday, Feb. 17, 2016 if submitting handwritten work on paper.

Collaboration and Reference Policy

You may discuss the general methods of solving the problems with other students in the class. However, each student must work out the details and write up his or her own solution to each problem independently. For each problem, list the students with whom you discussed general methods of solving the problem.

Some problems have been used in previous offerings of COS 435. You are NOT allowed to use any solutions posted for previous offerings of COS 435 or any solutions produced by anyone else for the assigned problems. You may use other reference materials; you must give citations to all reference materials that you use.

Lateness Policy

A late penalty will be applied, unless there are extraordinary circumstances and/or prior arrangements:

- Penalized 10% of the earned score if submitted by 10am Thursday (2/18/15).
- Penalized 25% of the earned score if submitted by 4:30 pm Friday (2/19/15).
- Penalized 50% if submitted later than 4:30 pm Friday (2/19/15).

Submission

Submit your solutions as a PDF file using the Computer Science Department DropBox submission system for COS435 at https://dropbox.cs.princeton.edu/COS435_S2016/HW1Part2 Name your file HW1Part2.pdf. If you have not used this facility before, consult the instructions at
https://csguide.cs.princeton.edu/academic/csdropbox - student
Note that you are automatically enrolled in CS DropBox using the registrar's COS435 enrollment list.

You may hand write your solutions as long as they are legible. In this case, you may either scan your writing to produce a PDF file for submission through DropBox or turn in your document by 5:00 PM Wed. Feb. 17 in the bin outside Prof. LaPaugh's office.

Problem 1 of Part 2 (2012 exam problem):

PageRank is usually applied to the graph of a collection of documents without regard to the content of the documents. In this problem we will change that. For each pair of documents, there is a pre-computed real-valued similarity measure ranging between 0 and 1 (e.g. cosine similarity in the vector model). Let sim(i,j) denote this similarity between the ith and jth documents. We use this value to modify the PageRank calculation, which we call **pr-s** for "PageRank with similarity". For the graph of a collection of **n** documents, the new PageRank equation is:

$$\textbf{pr-s}_{\text{new}}(\mathbf{k}) = \alpha/\mathbf{n} + (1-\alpha)\sum_{\text{i with edge from i to k}} ((1+\text{sim}(\mathbf{i},\mathbf{k})) * \textbf{pr-s}(\mathbf{i}))$$

Part a: Even if the underlying graph has no sinks, PageRank with similarity is not guaranteed to converge. Why not? Be specific.

Part b: Modify the definition of PageRank with similarity so that the idea of using similarity in this way is retained but the resulting calculation does converge assuming no sinks. Give the new equation and an argument (not necessarily a proof) that it converges.

Problem 2 of Part 2:

Below is the input for the discounted cumulative gain (DCG) example from lecture, modified to give a score instead of a gain for the document at each rank. Scores range from 0 to 2. Calculate the expected reciprocal rank (ERR) with n=10 for this example. You are welcome to write a small program to do this, but it is not required. You are not allowed to use any pre-existing software tool that computes ERR.

rank	**score**
1	**2**
2	**0**
3	**0**
4	**1**
5	**2**
6	**0**
7	**0**
8	**0**
9	**1**
10	**1**