Information Retrieval Fall 2016

Homework 2

Assigned: 13th Oct 2016

Due: 21st Oct 2016 (at beginning of class)

Problem 1 (8 points)

A. The following list of R's and N's represents relevant (R) and non-relevant (N) documents in a ranked list of 50 documents. The top of the ranked list is on the left of the list, so that represents the most highly weighted document, the one that the system believes is most likely to be relevant. The list runs across the page to the right. This list shows 10 relevant documents. Assume that there are only 10 relevant documents for this query.

RRNRNRNNNN RNNNRNNNNR RNNNNNNNRN NNNNNNNNNN RNNNNNNNNN

Based on that list, calculate the following measures:

1. Average precision

2. Inerpoloated precision at 50% recall

3. Interpolated precision at 33% recall

4. R-precision

B. Now, Imagine another system retrieves the following ranked list for the same query.

RNNRNNNRNN NNNRNNNNNN NRNNNNNNRN NNRNNNNRNN NNNNRNNNNR

Repeat parts (A.1), (A.2), (A.3), and (A.4) for the above ranked list. Compare the two ranked lists on the basis of these four metrics that you have computed--i.e., if you were given only these four numbers (Mean Average Precision, Precision at 50% recall, Precision at 33% recall, and R-precision) what can you determine about the relative performance of the two systems in general.

C. Plot a recall/precision graph for the above two systems. Generate both an uninterpolated and an interpolated graph (probably as two graphs to make the four plots easier to see). What do the graphs tell you about the system in A and the one in B?

Problem 2 (4 points)

Let 'perfect-retrieval' be defined as a list of ranked documents where

- all the relevant documents are retrieved and

- every relevant document is ranked higher than any non-relevant one.

A. Prove that at a list demonstrates perfect-retrieval if and only if there exists a `cutoff=c` such that the list has `PREC(c)=1` and `RECALL(c)=1`

B. Consider a particular `cutoff c=10`

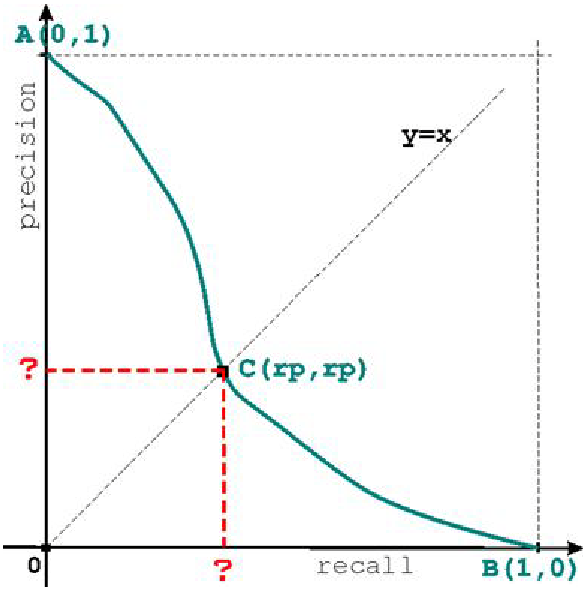
B1. Give example of a list that has `PREC(10)=1` and `RECALL(10)<1`

B2. Give example of a list that has `PREC(10)<1` and `RECALL(10)=1`

C. Prove that a list demonstrates perfect-retrieval if and only if `R_PREC=1`

D. Prove that a list demonstrates perfect-retrieval if and only if `AveragePrecision=1`

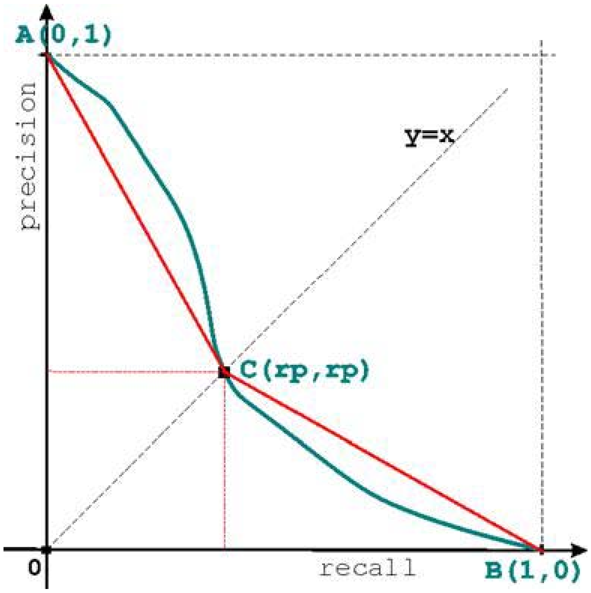
Problem 3 (4 points)



Consider a typical precision-recall curve starting at `A (RECALL=0,PREC=1)` and ending at `B(RECALL=1,PREC=0)`as shown in the plot (1) below.

A. Every point on the precision-recall curve corresponds to a particular rank (or `cutoff`). Intersect the curve with the main diagonal given by the equation **`y=x`** at point `C (rp,rp)`. If the query in question has R relevant documents, at what rank does this intersection point corresponds to?

B. As a consequence of the above, the quantity `rp` represents a standard retrieval metric. Explain which one and why.



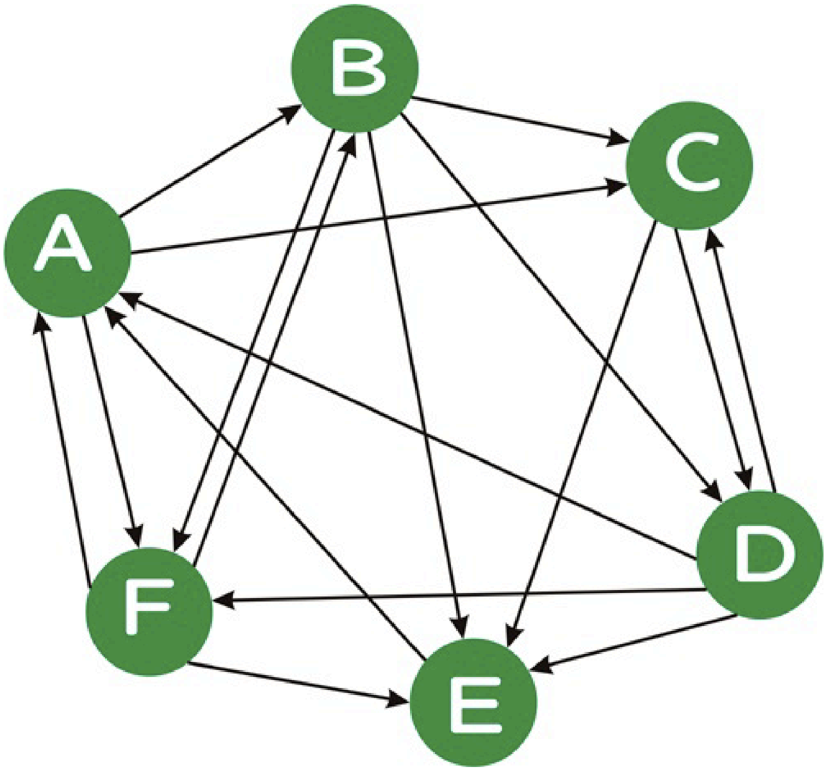
C. As shown in plot (2), we now add the segments [AC] and [CB]. We are interested in the area under the **blue** precision-recall curve and we can approximate it by the area under the **red** lines ACB. Compute this approximation area as a function of `rp`**.**

D. Explain how the `average precision` and `R-precision` measures are thus related.

Problem 4 (2+3+4+4 = 13 Points)

PageRank and Markov Chains

Consider the following directed graph:



a) Treat the above graph as a Markov chain, assuming a uniform distribution on the edges outgoing from each vertex. (In this problem part, you should *not* use any "teleportation.") Give the state transition matrix P of this Markov chain.

b) Compute the stationary distribution of this Markov Chain. This is a distribution π over the vertices such that

   a. $\pi = \pi P$.

   Note: In order to solve for π, you will need to solve six equations in six unknowns. Feel free to use a tool such as MatLab, if you like; otherwise, solve the equations by hand, eliminating one variable at a time. Also recall from class that the six equations given from $\pi = \pi P$ are not linearly independent; you will need to use five of these equations, together with the equation which specifies that the sum of the π probabilities must be 1.

c) Starting with the uniform distribution

   a. $\pi^{(0)} = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$

as an initial "guess", multiply $\pi^{(0)}$ by P to obtain a new "guess" $\pi^{(1)}$. Repeat this process, obtaining $\pi^{(n)}$ from $\pi^{(n\text{-}1)}$ via

$$\pi^{(n)} = \pi^{(n\text{-}1)}\ \mathrm{P}$$

until each of the π values are accurate within two decimal places (i.e, ± 0.01) of the values you solved for above. How many iterations are required?

Note: You should probably write a short program to do this, and this program will be useful for the problem part below as well.

d) Consider the PageRank formula as described in class and at the Wikipedia PageRank page. In particular, consider the following PageRank formula described on that page

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

(formula image courtesy Wikipedia). Let d = 0.85 be the damping factor.

Demonstrate that for the graph above, this formula is equivalent to computing the stationary distribution of a Markov chain described by transition matrix P', where each entry p'_{ij} in P' is obtained from the corresponding entry p_{ij} in P as follows:

$$p'_{ij} = (1\text{-}d)/N + d\ p_{ij}.$$

Using the matrix P' and your code from the problem part above, solve for the PageRank values of each vertex. (Start with a uniform distribution for $\pi^{(0)}$ and repeatedly multiply by P' until the π values "converge", e.g., they no longer change in the second decimal place). How do the PageRank values compare to the original stationary distribution values you computed above (and why)?