

[illegible]

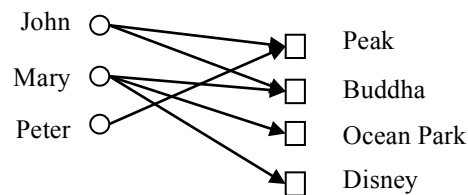
2. [25] Fill in the precision, recall and fallout values in the following table. There are a total of 100 documents, and all of the relevant documents are shown in the table; they are marked with a \checkmark in the first column. Draw the precision/recall graph and fallout/recall graph as in slide 11 of the lecture notes (i.e., no need to interpolate or smooth the graph).

	Rank	doc ID	Recall	Precision	Fallout
	1	1001			
	2	2873			
	3	3916			
	4	0983			
	5	8310			
\checkmark	6	7892			
	7	4562			
\checkmark	8	4921			
\checkmark	9	7934			
	10	9248			
...					
	98	1688			
\checkmark	99	0926			
	100	3861			

Given that the top 10 documents are retrieved. Compute:

- the average precision $Ap(q)$ and explain what is “q” in the question, i.e., “q is the query which ...”
- $NDCG_{10}$ assuming that relevant documents get scores of 1 and non-relevant documents get scores of 0.

3. [25] The bipartite graph represents the places visited by a traveler.



A link indicates that the person has visited a place. Define the **interestingness** of a place as the summation of the **experience** of travelers who visited the place, and the experience of a traveler as the summation of the interestingness of places he/she visited.

- (a) Compute the interestingness of each place and experience of each traveler using the hub/authority metaphor. Assuming at the beginning all hub and authority values are one. Perform 2 iterations.
- (b) Given that Peak and Buddha both have two travelers explain why they get the interestingness values obtained in (a).
- (c) In plain English, how would you describe an interesting place and an experienced traveler, each in one sentence? I.e., “A place is interesting if it ... A traveler is experienced if he/she ...”

4. [20] The following sentences are extracted from Wikipedia and considered as two documents.

Document 1: Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources.

Document 2: Automated information retrieval systems are used to reduce what has been called "information overload".

After removing all punctuation marks and words with three characters or less:

- (i) Show the bigrams of these two documents. Suppose an index is built on the bigrams (no need to show the index).

For the query **information retrieval systems** (note: no double quote applied):

- (ii) Explain how you would transform the query before the index is searched.
(iii) Assuming that TF is used as term weight (i.e., IDF is not used) what is the similarity of the query to the two documents using inner product similarity?
(iv) Using this example, explain the advantage of bigram compared to unigram (i.e., single term) indexing in terms of search quality (this part is an open discussion; a few sentences are enough).