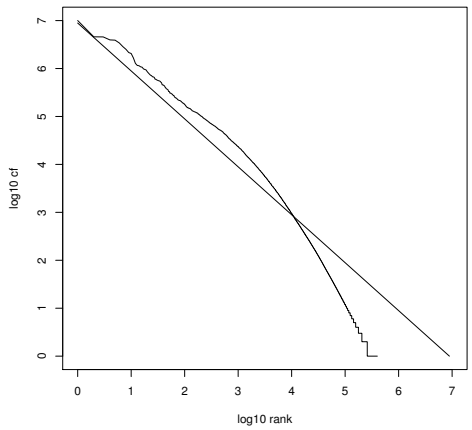
03-60-538: Final Exam 2017

Last Name: _____ First Name: _____ Student ID: _____

This is an open book exam. You can use either pen or pencil, and write on the back of the paper if more space is needed. You must submit your exam paper to me in office before Dec 15 noon.

1. (5 points) Suppose that the precision is 0.5, recall is 0.4. Write the corresponding F1 score.

2. (5 points) Given the following plot for the Zipf's law. What are the approximate range of the frequencies of the top 10 popular terms?



3. (5 points) Heaps' law predicts the number of distinct terms, or the vocabulary size (**M**), based on the total number of tokens in a data collection (**T**). Please write the formula for Heaps' law.

4. (10 points) Documents are indexed as posting lists. One example of two posting lists is as below.

Brutus $\longrightarrow$ $\boxed{1} \rightarrow \boxed{2} \rightarrow \boxed{4} \rightarrow \boxed{11} \rightarrow \boxed{31} \rightarrow \boxed{45} \rightarrow \boxed{173} \rightarrow \boxed{174}$

Calpurnia $\longrightarrow$ $\boxed{2} \rightarrow \boxed{31} \rightarrow \boxed{54} \rightarrow \boxed{101}$

(a) What is the result of running a boolean query "BRUTUS AND CALPURNIA"?

(b) The following algorithm is used to obtain the intersection of two lists. For the above two lists,

i. write the sequence of the elements accessed;

ii. what is the complexity of the algorithm?

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

5. (20 points) Given the following term-document matrix C. Let the rows represent terms, columns represent documents. $C_{ij} = k$ if term i occurs in document j for k times.

	d1	d2	d3	d4	d5	d6
t1	2	0	1	0	2	5
t2	0	1	10	0	0	0
t3	0	1	0	0	2	2
t4	2	0	1	1	1	2
t5	0	0	1	1	0	1

(a) Write the raw score of document frequency for each term;

df(t1)=

df(t2)=

df(t3)=

df(t4)=

df(t5)=

(b) Suppose that the TF-IDF weighting is defined by the following formula:

$$w_{t,d} = (1 + \log_{10} tf_{t,d}) \log_{10} \frac{N}{df_t}. \tag{1}$$

Fill in the following table for the weight of each cell. You can leave the fraction and log function in the table, if no further simplification can be made. Note that log 1 is zero, and you should simplify this at the least.

	d1	d2	d3
t1			
t2			
t3			

(c) Write the Cosine similarity between

- d_1 and d_2 (columns 1 and 2);

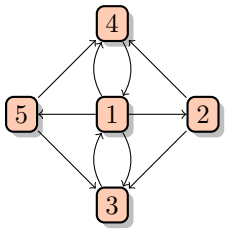
- d_2 and d_3.

(d) Give the Jaccard similarities S(x,y) for the following document pairs. Suppose that the shingle length is one, i.e., we compare the similarity based on terms, not a sequence of terms. Also, we assume set-of-words model according to the definition of Jaccard similarity. You only need to list the fractions, hence calculator is not required.

- d_1 and d_2 (columns 1 and 2);

- d_2 and d_3.

6. (20 points) Given the following graph. Please answer the following questions:



(a) Fill in the following table so that it is an adjacency matrix for the above graph .

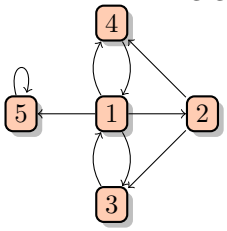
	node 1	node 2	node 3	node 4	node 5
node 1					
node 2					
node 3					
node 4					
node 5					

(b) Fill in the following table so that it is a column stochastic matrix for the above graph.

	node 1	node 2	node 3	node 4	node 5
node 1					
node 2					
node 3					
node 4					
node 5					

(c) Using the power method, calculate the page rank values in the first two iterations, including the initial values.

(d) For the following graph, write the page rank values for all the nodes if random teleporting is **not** used.



7. (10 points) Recall the "export"/POULTRY contingency table in IIR book for calculating MI (mutual information):

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{export} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{export} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

(a) Compute a similar contingency table for "Kyoto"/JAPAN based on the data given below.

	docID	words in document	in c = Japan?
training set	1	Kyoto Osaka Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Shanghai	no
	5	London	no

	$e_c = e_{japan} = 1$	$e_c = e_{japan} = 0$
$e_t = e_{kyoto} = 1$	$N_{11} =$	$N_{10} =$
$e_t = e_{kyoto} = 0$	$N_{01} =$	$N_{00} =$

(b) Make up a contingency table for which MI is 0 – that is, term and class are independent of each other.

8. (15 points) The following questions are related to the Naive Bayes classifier.

   (a) Write the Bayes' rule for the conditional probability $P(A|B)$

   (b) When we derive the Naive Bayes classifier for documents, the probability of a document is simplified as the production of the probabilities for each terms as below:

$$P(d|c) = P(\langle t_1, \ldots, t_{n_d} \rangle | c) = \prod_{1 \leq k \leq n_d} P(X_k = t_k | c) \tag{2}$$

      i. Why do we need to have such simplification?

      ii. Whether such simplification is valid in theory?

   (c) The probability of each term is estimated using the following formula. Note that $\hat{P}$ means the estimated value of P.

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B}$$

   Explain what is B and why there is "+1".