CSE 5337/7337 Information Retrieval and Web Search name L.J. Brown
Spring 2018: Homework 4
Due: 3/29/2018
Notation: (5337/7337 points)

1. (20/20 points) Consider these three documents and query:

Doc1: The business of life is a business of everlasting learning
Doc2: The unexamined life is not worth living
Doc 3: Always keep learning
Query: Life Learning

Compute the similarity of the query to each document using ntc.atc weighting, thus determining the order of relevant documents. Show your work.

* table at the end of document

2. (15/15 points) If $|s_i|$ denotes the length of string s_i, show that the levenstein edit distance between s_1 and s_2 is never more than $\max\{|s_1|, |s_2|\}$.

Take either s_1 or s_2, In the worst case there is no overlap between the two strings. So change each of the characters in the string that don't match and the levenstein distance is now $min\{|s_1|, |s_2|\}$. Now either add or remove the reaming characters and the levenstein distance will be, $max\{|s_1|, |s_2|\} - min\{|s_1|, |s_2|\} + min\{|s_1|, |s_2|\} = max\{|s_1|, |s_2|\}$ in the worst case.

3. (15/15 points) Apply Equation 6.6 to the sample training set in Figure 6.5 to estimate the best value of *g* for this sample.

$$(0 + 1) / (0 + 1 + 2 + 1) = 0.25$$

4. (15/5 points) [Do not use examples provided in class]

a) Find two differently spelled nouns that map to same soundex code.

ram and rain

b) Find two phonetically similar nouns that map to different soundex codes.

phew and few

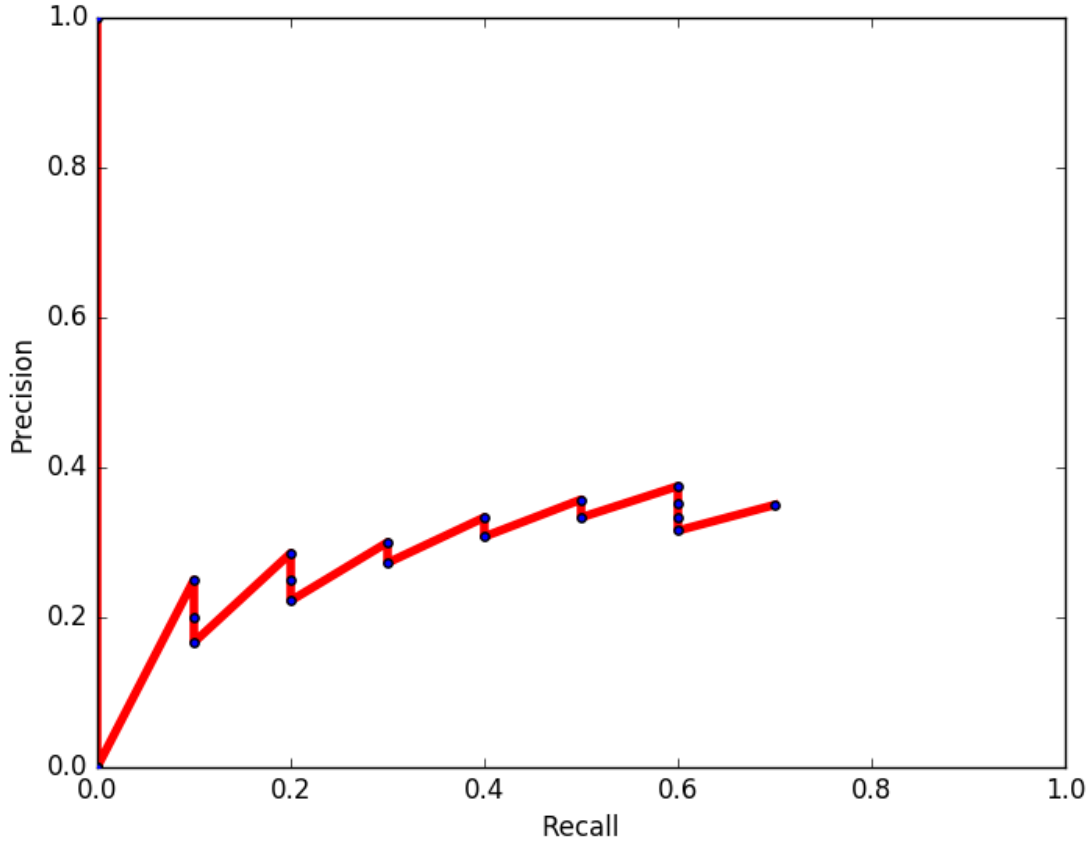
c) Find two words that stem to the same sequence of characters but have different soundex codes.

mules and Mule

5. (15/15 points) Let R_q = { d2, d8, d17, d27, d32, d46, d59, d73, d80, d94 } be the set of relevant documents for query q. Given the following ranking of retrieved documents in the answer set of q:

1. d38	6. d111	11. d30	16. d73
2. d199	7. d94	12. d46	17. d321
3. d7	8. d21	13. d99	18. d58
4. d17	9. d222	14. d80	19. d5
5. d231	10. d27	15. d3	20. d32

Create the Precision versus Recall Curve for q. What is the MAP result?



MAP: 0.285995

6. (15/10 points) Suppose a search engine has just retrieved the top 50 documents from a collection based on scores from a ranking function *R(Q,D)*. The user interface can show only 10 results. Why might you want the system to include a document other than in the top 10 or even top 50 retrieved?

The scores might be equal for the first 11, or 51 results.

7. (5/5 points) Locate a copy of US Patent #7818332, by Microsoft about query speller. What section of the patent considers the query "food explorer"?

section 5

Graduate students:

8. (0/15 points) Exercise 5.2. Estimate the space usage of the Reuter-RCV1 dictionary with blocks of size k=8 and k=16 in blocked dictionary storage.

(END)

Terms	Q	v(Q)	D1	v(D1)	D2	v(D2)
the	0	0.08804563	1	0.17609126	1	0.17609126
business	0	0.08804563	2	0.35218252	0	0
of	0	0.08804563	1	0.47712125	0	0
life	1	0.17609126	1	0.17609126	1	0.17609126
is	0	0.08804563	1	0.17609126	1	0.17609126
a	0	0.08804563	1	0.47712125	0	0
everlasting	0	0.08804563	1	0.47712125	0	0
learning	1	0.17609126	1	0.17609126	0	0
unexamined	0	0.08804563	0	0	1	0.17609126
not	0	0.08804563	0	0	1	0.17609126
worth	0	0.08804563	0	0	1	0.17609126
living	0	0.08804563	0	0	1	0.17609126
always	0	0.08804563	0	0	0	0
keep	0	0.08804563	0	0	0	0
SQRT(SUMS	1.41421356		3.31662479		2.64575131	

D3	v(D3)	v(Q)*v(D1)	v(Q)*v(D2)	v(Q)*v(D3)
0	0	0.01550407	0.01550407	0
0	0	0.03100813	0	0
0	0	0.04200844	0	0
0	0	0.03100813	0.03100813	0
0	0	0.01550407	0.01550407	0
0	0	0.04200844	0	0
0	0	0.04200844	0	0
1	0.17609126	0.03100813	0	0.03100813
0	0	0	0.01550407	0
0	0	0	0.01550407	0
0	0	0	0.01550407	0
0	0	0	0.01550407	0
1	0.17609126	0	0	0.01550407
1	0.17609126	0	0	0.01550407
1.73205081		0.25005785	0.12403253	0.06201626