COMP 150: Natural Language Processing

Spring 2020

Due 11:59pm Feb 6, 2020

Question 1. [Document similarity] Suppose our pets have produced two documents:

D1 = [woof woof meow] D2 = [woof woof squeak]

(a) What is the cosine similarity of D1 and D2, not using idf weighting? [15 points]

(b) What is the cosine similarity if idf weighting is used? [15 points]

(c) How would the answer to (b) change if we added a third document: D3 = [meow squeak] to the collection? [10 points]

Question 2. [Evaluation metrics] Assume you are given an inquiry application and you wish to evaluate its performance. As we saw in class, accuracy is not always a good measure (think about why). For this reason, you have decided to calculate *precision*, *recall* and *f-measure* in order to score the following system against the answer key. Assume any item reported by the system and found in the answer key is correct. [20 points]

Hint: Think about which variables are required for these calculations (write them out!). Then, frame the inquiry problem in such a way that you can count the metrics for each of the required variables.

1. *Jay Leno attacked Conan O'brien.*
2. *attacks by the U.S.-backed rebels*
3. *the latest in a series of attacks in the 10-year-old civil war.*
4. *Mr. Baldwin is also attacking the greater problem: lack of ringers.*
5. *the criminals were convicted for bombings.*
6. *The broadway musical "Bridges of Madison County" bombed.*
7. *Groupon fires CEO Andrew Mason.*

The answer key includes the following strings of words describing attack events:

1. *the martians bombarded the Earth with death rays*
2. *attacks by the U.S.-backed rebels*
3. *the latest in a series of attacks in the 10-year-old civil war.*
4. *the criminals were convicted for bombings.*

5. *the allies launched a missile at the enemy stronghold.*

Question 3. [Naive Bayes and smoothing] Do exercises 4.1 and 4.2 in third (on-line) edition of the textbook (https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf) (page 81). Show the intermediate steps in your calculation. Compute using probabilities, not logs of probabilities (so instead of adding logs of probabilities, you multiply probabilities). [20 points for 4.1] and [20 points for 4.2]