Homework 3 (CS 1671)

Assigned: November 27, 2018

Due: December 6, 2018

1. Naive Bayes Classification

Suppose that we are using a Naïve Bayes Classifier to tell if the topic of a document is Pittsburgh. Assume all the training set is included in the table below. All the words used as features are also included in the table.

- Can you tell if the test set document is about Pittsburgh? Show how you get the result. Use add-one smoothing if needed.

	Keywords in Document	If topic is Pittsburgh
Training set	Pittsburgh Pirates	True
	Pittsburgh Pittsburgh Pitt	True
	Pittsburgh Strip	True
	Pittsburgh Cleveland Cincinatti	False
Test set	Pittsburgh Pittsburgh Pittsburgh Pitt	?

2. Vector Semantics

Consider the table below of term frequencies for 2 documents denoted Doc1 and Doc2:

	Doc1	Doc2
car	27	4
auto	3	33
insurance	0	33

- Compute the tf-idf weights for the terms car, auto, insurance for each document, using the document frequencies in the table below, taken from a collection of 806,791 documents.

	df
car	18,165
auto	6723
insurance	19,241

- Compute the cosine similarity between cosmonaut and astronaut given their vector space representations below:

	cosmonaut	astronaut
Soviet	1	0
American	0	1
spacewalking	1	1
red	0	0
full	0	0
old	0	0

- Given the source text “The quick brown fox jumps over the lazy dog”, what would the skipgram with negative sampling training data be for learning the word embedding for the target “*brown*”? Use the parameters from the lecture notes.

3. Information Extraction

Acronym expansion, the process of associating a phrase with an acronym, can be accomplished by a simple form of relational analysis.

- Write a pattern to find three letter acronyms. Compare it to Wikipedia’s TLA page to informally evaluate what you match and what you miss.

4.Semantic Role Labeling

- Discuss the similarities and differences between the FrameNet and PropBank representations of the arguments of the verb “*sell*”.

- Describe appropriate selectional restrictions for this verb. Collect a few corpus examples and analyze how well your selection restrictions worked. Your corpus should be big enough to verify that your selectional restrictions work on several examples, and should also contain examples illustrating where they fall apart. Collect your corpus from any source you like (e.g. online newspapers, the web, online corpora).