Text Analytics HW2
Safia Khouja sck9826

Question 1

I investigated 2 sets of word2vec model parameters — specifically, I sought to compare the CBOW and skip-gram models. When comparing these two models, I kept the other parameters fixed with the following values: size= 20, workers=20, window =20. I tested the same 9 words on both the CBOW and skip-gram models. I show the results for the top 4 closest neighbors in terms of cosine similarity for each word.

	Skip Gram				CBOW			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th
King	prince	nazareth	domino	rachel	prince	holme	stone	messier
Woman	tomb	her	jurors	herself	mother	son	dead	her
Car	moving	dollar	trips	putting	cars	metzler	getting	bike
Money	spend	cops	pay	taxes	pay	dollars	paying	leasing
Run	replace	sprites	running	machine	slip	switch	plug	replace
Live	bikers	suspects	nets	assholes	stay	go	stop	talk
Drink	liberally	scream	coffee	waving	watch	ride	coffee	laugh
Understand	what	takers	inclined	mean	agree	believe	prove	suggest
The	of	in	and	repeated	and	closest	which	table

The two models perform similarly. Some of the closest neighbors make sense and some of them do not. I believe that my training data may not have been big enough or sufficiently diverse to yield closest neighbors that always make sense. It is most likely that a larger corpus would produce better results with these models. If I have time, I would like to try this exercise with Yelp review data to see if the results improve.

Both the models perform well for the word *money*, meaning that most of the closest neighbors make sense. On the other hand, both models for the word *drink* produce nonsensical results most of the time. Perhaps *money* produces better results because it appears more often and in more diverse situations than *drink* in my training data.

I found it interesting to look at the results depending on the part of speech. I looked at 4 nouns, 4 verbs, and one article. It is notable that the nouns seemed to produce output that made more sense compared to the verbs. The article tends to be similar to articles and prepositions.

Link to code:
https://github.com/SafiaKhouja/sck9826_msia_text_analytics_2020/blob/homework2/Hw2/Homework2_Safia.py

Text Analytics HW2
Safia Khouja sck9826

Question 2

	Word2vec	**Bert**
Learning model details	Skip gram model learns from the training data find word representations that are useful for predicting the surrounding words in a sentence	Pre trained language representation applied to down-stream task
Word context approaches	Relies on word context. Does not take into account word order.	Takes into account word position and order (in addition to context) within a sentence
Corpus size requirements	Large corpus is best (see homework question 1 above). The more words and diverse usage of words, the better.	Small corpus is best. It is critical to use a document-level corpus rather than a shuffled sentence-level corpus such as the Billion Word Benchmark
Ease of installation/use of source code	Word2Vec contained within gensim package in Python. Install the package using pip. Very easy to implement — only	It is available within the keras_bert python Package.
Date of publication	2013	2019
Number of google scholar citations	22889	10047