Cairo University
Faculty of Engineering
Computer Engineering Dept.

Sheet 4: Vector Semantics and Embeddings

1) In a corpus of 10000 documents, document *D* has a total of 250 words and the word "sheet" occurs 20 times. The word "sheet" also occurs in 2500 documents in the corpus. What is the tfidf entry for the term "sheet" in a bag of words vector representation for *D*?

2) We have these two 5-dimensional word vectors for the two words w1 and w2:
w1=[1,0.4,0.3,0.8,0.04] w2=[1,0.2,0.5,0.7,0.6]
Are these words similar or dissimilar? Make only **one** change in the vectors' values to invert the similarity between these two words i.e. if similar becomes dissimilar and viceversa.

3) Consider the following three documents - D1, D2, D3:
D1: Natural language processing is becoming important since soon we will begin talking to our computers.
D2: If computers understand natural language they will become much simpler to use.
D3: Speech recognition is the first step to build computers like us.

   a. Perform text normalization (stop words removal and lemmatization).
   b. What is the document vector for D3?
   c. Using tfidf, what is the document vector for D3?

4) For CBOW model, if we have number of words in vocabulary V and dimension of the embedding d then answer the following:
   a. What are the number of neurons in the hidden layer?
   b. What are the dimensions for the weight matrix between the input and the hidden layer?
   c. What are the dimensions for the weight matrix between the hidden layer and the output?
   d. What is the dimension of the input/output vectors?
   e. For Skip-gram model, mention if the answers to the previous questions change or not?

5) What is the big-O runtime of computing a single probability P(context=c|word = w) for skip-gram model? Express this in terms of the vectors' dimensionality d and the vocabulary size |V |.

6) Alice and Bob have each used the Word2Vec algorithm to obtain word embeddings for the same vocabulary of words V. In particular, Alice has obtained context vectors $\mathbf{u}_w^A$ and center vectors $\mathbf{v}_w^A$ for every w ϵ V, and Bob has obtained context vectors $\mathbf{u}_w^B$ and center vectors $\mathbf{v}_w^B$ for every w ϵ V.
Suppose that, for every pair of words w', w ϵ V, the inner product is the same in both Alice and Bob's model:

$$(\mathbf{u}_w^A)^T \mathbf{v}_{w'}^A = (\mathbf{u}_w^B)^T \mathbf{v}_{w'}^B$$

Does it follow that, for every $w \in V$, $\mathbf{v}_w^A = \mathbf{v}_w^B$? Why or why not?

7) Mention one advantage and one disadvantage for word2vec in comparison with co-occurrence count-based methods.