COMP 5790/ 6790/ 6796

Special Topics: Information Retrieval

Instructor: Shubhra ("Santu") Karmaker

Assignment #6: Evaluation + Feedback + Word Embeddings [100 points]

⚠ **Notice:** This assignment is due **Friday, February 21, 2020 at 11:59pm**.

Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a **typeset PDF**. Please do not include scanned or photographed equations as they are difficult for us to grade.

1. Information Retrieval Evaluation [25 pts]

a. **[5 pts]** Why might using raw term frequency counts with a dot product similarity not produce the best ranking results?

b. **[5 pts]** Let d be a document in a corpus. Suppose we add another copy of d to the collection. How does this affect the IDF values for all words in the corpus?

c. **[10 pts]** Suppose there are 16 total relevant documents in a collection. Consider the following result, where $+$ indicates a relevant document and $-$ indicates a non-relevant document.

$$\{+,+,-,+,+,-,-,+,-,-\}$$

d. Calculate the following evaluation measures for this ranked list.
   i. Precision
   ii. Recall
   iii. F_1 score
   iv. Average precision

e. **[5 pts]** Using the same setup as above, assume that the "gain" of a relevant document is 1 and the "gain" of a non-relevant document is 0. Calculate the following:
   i. Cumulative Gain at 7 documents
   ii. Normalized Discounted Cumulative Gain at 7 documents (use $\log_2$ for the discounting function)

2. Word Embeddings (Theory) [35 pts]

a. **[5 pts]** What is the distributional hypothesis and how is it incorporated into Skip-Gram and CBOW?

b. **[10 pts]** In negative sampling we try to learn a model by contrasting actual data from a noise distribution. In the case of word embeddings what is the actual data and what is the noise distribution? How can we obtain samples for them?

c. The following questions relate to subsampling of frequent words: In "Distributed Representations of Words and Phrases and their Compositionality" by Mikolov et al. the probability of a word being discarded is defined as $P(w_i)=1-\sqrt{\frac{t}{f(w_i)}}$, where $f(w_i)$ is the frequency of word w_i and t is a chosen threshold, typically around 10^{-5}.

i. **[10 pts]** How does the subsampling affect very frequent and infrequent words? How does the subsampling affect the ranking of words when ordered by frequency?
ii. **[10 pts]** How does the subsampling affect the size of the context window?

3. Word Embeddings (Practice) [30 pts]

Imagine you trained word embeddings on a corpus and obtained the following vector space with words w_i and dimensions w_{ij}:

wi	*wi*1	*wi*2	*wi*3	*wi*4	*wi*5	*wi*6	*wi*7	*wi*8	*wi*9	*wi*10
the	0.131	0.001	0.023	0.918	0.991	0.912	0.787	0.675	0.787	0.987
cat	0.911	0.891	0.912	0.016	0.099	0.189	0.777	0.776	0.853	0.992
for	0.112	0.009	0.032	0.819	0.971	0.932	0.788	0.677	0.777	0.988
data	0.954	0.919	0.881	0.812	0.901	0.990	0.012	0.002	0.014	0.909
mouse	0.912	0.881	0.922	0.019	0.100	0.199	0.011	0.003	0.016	0.898
it	0.142	0.010	0.026	0.820	0.917	0.923	0.781	0.611	0.722	0.977
dog	0.922	0.882	0.931	0.011	0.101	0.193	0.769	0.762	0.841	0.989
also	0.121	0.004	0.021	0.919	0.981	0.917	0.790	0.617	0.712	0.969
computer	0.912	0.923	0.899	0.853	0.910	0.991	0.022	0.010	0.016	0.912

a. **[5 pts]** What are the three nearest neighbors of 'cat'? List the neighbors with their distances according to cosine similarity.
b. **[5 pts]** What are the three nearest neighbors of 'computer'? List the neighbors with their distances according to cosine similarity.
c. **[5 pts]** What are the three nearest neighbors of 'the'? List the neighbors with their distances according to cosine similarity.
d. **[10 pts]** If you were to cluster the words into three semantically coherent clusters, which words would you cluster together? List each cluster as a set of words and describe their semantics briefly (hint: the clusters don't have to be disjunct). Does this clustering correspond to the clustering that can be obtained from the vector space, e.g. by using an unsupervised clustering method such as k-means?
e. **[5 pts]** Do you see a problem with methods that perform hard clustering (e.g. kNN) as compared to fuzzy clustering (e.g. EM clustering) in the context of word ambiguity?

3. KL-divergence Retrieval Function [10 points]

Show that the KL-divergence retrieval function covers the query likelihood retrieval function as a special case if we set the query language model to the empirical word distribution in the query (i.e., $p(w|\theta_Q)=\frac{c(w,Q)}{|Q|}$, where $c(w,Q)$ is the count of word w in query Q, and $|Q|$ is the length of the query).