NLP Lab 1 Evaluation

Aug 21, 8:15 am to 10:15 am

Objective

In this lab we would build a system based on the principles of Language model and Continuous Bag of Words (CBOW) to determine the similarity between 2 words from the given corpus. We then evaluate the performance by comparing this with the scores produced by word2vec.

Steps

Given the corpus of tweets (C) do the following:

1. Clean the tweets: Replace hashtags, screen names, urls, RT, emoticons with suitable words (token type)
2. Convert the tweets to lowercase
3. (optional): do some minimum preprocessing like lemmatization or stemming
4. Construct the vocabulary (V)
5. Build a unigram counts
6. Choose a subset of V based on a minimum threshold count for the unigrams. Replace the words that have less than the threshold counts with a special symbol
7. From the preprocessed corpus, select a minimum of 10 word pairs for evaluating similarity between them
8. Extract all triples from the preprocessed corpus – let us call this a set T. For each t in T measure the counts c.
9. For each word pair in the list chosen in step 7, w_i and w_j, for each triple in T:
    a. Assign count($w_{i,}$, t) = c if w_i is the center word of t
    b. Assign count($w_{j,}$, t) = c if w_j is the center word of t
    c. Compute delta(t) = abs(count($w_{i,}$, t) - count($w_{j,}$, t))
10. Compute the sum of all deltas generated in step 9 and let this be D
11. Add up all the counts obtained from steps 9(a) and 9(b) and let this be Z
12. Compute and return the score (1 – D/Z)
13. Train the word2vec with the preprocessed corpus
14. For the same word pairs w_i and w_j compute similarity using word2vec
15. Normalize the similarity score obtained in step 14 to have the range (0, 1)
16. Prepare a table and tabulate for each word pairs chosen in step 7 the scores produced by the counting algorithm and the word2vec

17. Inspect the results and draw your conclusions
18. Try different other pairs of words if need be in order to get more insights
19. Post your analysis on the FaceBook

Deliverables

Submit the following by 10:30 am:

1. Source code of tweets cleaning
2. Source of the algorithm implementation
3. Text file containing the cleaned corpus
4. CSV formatted file that show the triples and counts for word pairs
5. CSV formatted file that shows the benchmark between the scores

Do the following by noon 22 Aug 2015:

1. Post your analysis on the Facebook
2. Optionally you can include any graphics, visualization etc