Word2Vec Assignment

Jean Mark Gawron

April 30, 2020

1 PMI

w\c	global_JJ	classic_JJ	ancient_JJ	liberal_JJ
politician_NN	0	5	0	3
agenda_NN	1	1	1	4
conservative_NN	0	4	0	1
liberal_NN	1	6	0	1
cabal_NN	5	1	4	2

Find the following:, using the formulae in vectors1.pdf, slides slides 20-24.

1.1. $p(w = \text{politician\_NN}, c = \text{liberal\_JJ})$

1.2. $p(w = \text{politician\_NN})$

1.3. $p(c = \text{liberal\_JJ})$

1.4. Compute the PPMI score for word *politician_NN* and context *liberal_JJ*.

1.5. Suppose P(w = i | c = j) is equal to P(w = i). What can we say about PPMI(w=i,c=j)? If you don't remember the discussion of this case in class, use the chain rule to turn this into a fact about P(i,j).

1.6. Is it possible for the PMI value of target word i and context word j to be negative?

1.7. When is PPMI undefined?

2 Cosine similarity

Use slides 33-37 to help with the following.

2.1. Using the **counts** (rather than the PPMI values), compute the cosine similarity of target words *conservative* and *politician*. Note: It should be a number between 0 and 1.

2.2. Same two words: Now compute the cosine similarity using vectors with PPMI values. Note: For this problem, think of the log of a probability of 0 as a negative number with a **very** high absolute value. So for the purposes of PPMI any 0 probabilities are going to yield a PPMI of 0. Show at least this much of your work: What are the two PPMI vectors? What are the vectors after they are divided by their length? (We say the have been **normalized**).