6.

Word embeddings

(a)

$$P(c|w) = \frac{P(w,c)}{P(w)} = \frac{P(w,c)}{\sum_c P(w,c)} = \frac{n(w,c)}{\sum_c n(w,c)}$$

$$P(c) = \frac{\sum_w n(w,c)}{\sum_{c,w} n(w,c)}$$

$$\Phi_c(w) = \max\left(0, \log\frac{P(c|w)}{P(c)}\right)$$

Vocabulary V are 5000 most commonly-occurring words with the most counts

Context C are 1000 most commonly-occurring words with the most counts

Use PCA to reduce the dimension to get the 100-dimension embedding

1. get the covariance matrix

2. do the spectral decomposition for the covariance matrix

3. choose the largest 100 eigenvalues and their corresponding eigenvectors

4. get the mean of these points w;

5. subtracted by the mean then do the projection in the directions of the eigenvectors

Here are the formulas

$$x' = U^T(x - \mu)$$

$$M = U\begin{pmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_m \end{pmatrix}U^T$$

(b)

$$d\big(\psi(w), \psi(w')\big) = 1 - \frac{\psi(w) * \psi(w')}{||\psi(w)||||\psi(w')||}$$

Randomly choose 25 words, get their nearest neighbor

Below are the results:

Apparently, the nearest neighbor is of the similar contexts (class) or is usually used with the words, like

1958 and 1960 are time

First and the are usually used together: “the first”

Lovely and beautiful are synonyms

taxpayers ==> statistics
killing ==> lonely
See ==> The
Little ==> The
First ==> The
sin ==> images
sequence ==> connection
average ==> total
slowly ==> suddenly
mention ==> happen
1958 ==> 1960
external ==> educational
liquor ==> operator
lovely ==> beautiful
lightly ==> shoulder

continuously ==> 20
stayed ==> sat
transfer ==> interpretation
Five ==> The
formula ==> binomial
familiar ==> convinced
lives ==> family
government ==> state
uneasy ==> ugly
chance ==> reason

(c)

I use k-means algorithm and simple Euclidean distance to do the clustering since they are the widest used algorithm and distance. Also, we have do much conversion for the representation. Thus, I would like to make the algorithm and distance function as simple as possible to maintain the properties we have already got in the pointwise mutual information representation and the complex embedding.

The results are amazing, related words are clustered together! some of the best clusters are:

clst 0 :
through , between , under , against , without , upon , within

They are all preposition

clst 3 :
20 , 100 , 25 , 50 , 60 , 200 , 300 , 75 , 500 , 70 , 35 , 80 , 45

They are all numbers from 20-1000

clst 8 :
10 , 15 , 12 , 30 , 14 , 11 , 13 , 40 , 23

They are all numbers under 40

clst 10 :
make , see , get , go , come , take , put , find , give , help , keep , tell , leave , call , bring , talk

They are all commonly-used verb and often only have structure function, not detailed meaning

clst 11 :
children , others , themselves , women , boys , girls

They all refer to plural pronoun

clst 14 :
still , since , always , almost , however , though , become , often , yet , quite , already , became , although , strong , usually , soon , sometimes

They are all adverb and often used in the beginning of the sentence.

clst 16 :
ten , couple , seven , eight , nine , twenty , fifteen , twelve , eleven

They are all numbers in English

clst 35 :
had , have , been , has

all have

clst 40 :
the , a , this , an , any , each , same , another , every ,
clst 41 :
year , day , during , week , month ,
clst 42 :
1 , 2 , 3 , above , 4 , 5 , 6 ,
clst 43 :
out , up , into , over , back , down , off , home , away , along ,