Homework 4

The code is submitted at github

https://github.com/mas-dse/w9yan/blob/master/DSE220/homeworks/homework_4/Embedding_Clustering.ipynb

1. Data processing

a. Vocabulary V consisting of 5000 is selected of the most commonly-occurring words from original collection of text after removing stopwords and punctuation. And a shorter list C of at most 1000 of the most commonly-occurring words are used as context words C.
In my code, result is saved to object 'vocabulary', and 'context'

b. For each w in vocabulary, compute the probability distribution Pr(c|w) of context words around, as well as the overall distribution Pr(c) of context words.
In code, result is saved to object 'prob_c_under_w' and 'prob_context'

c. Represent each vocabulary item w by a |C|-dimensional vector of (positive) pointwise mutual information, so I got initial data set of 5000x1000.
In code, result saved to dataframe object 'X'

2. 100-dimensional embedding.

SVD(n_components=100) is used to reduce the 5000x1000 dimension dataset to 5000x100, each word in vocabulary is expressed with a new vector of 100. The reason I prefer SVD over PCA is that SVD is more used for latent semantic approximation while PCA is more used when covariance is considered.
Well, in my code I tried both SVD and PCA.

3. Nearest neighbor results

Pick a collection of 25 words from vocabulary, find nearest neighbor for each word with cosine distance.

```
nearest neighbor of communism - losing
nearest neighbor of autumn - tournament
nearest neighbor of cigarette - pitch
nearest neighbor of pulmonary - disk
nearest neighbor of mankind - defeat
nearest neighbor of africa - carolina
nearest neighbor of chicago - carleton
nearest neighbor of revolution - resumed
nearest neighbor of september - january
nearest neighbor of chemical - description
nearest neighbor of detergent - mustard
nearest neighbor of dictionary - text
nearest neighbor of storm - aristotle
nearest neighbor of worship - produces
nearest neighbor of war - free
nearest neighbor of problem - subject
nearest neighbor of wall - block
nearest neighbor of department - secretary
nearest neighbor of society - politics
nearest neighbor of company - letters
nearest neighbor of college - university
nearest neighbor of public - education
nearest neighbor of money - drink
nearest neighbor of wife - husband
nearest neighbor of wanted - maybe
```

Looking at above results, some pairs are quite correlated like the yellow ones, while some others are not obvious correlated as highlighted as red above.

4. Clustering.

I tried both KMeans++, GaussianMixture(EM) to cluster the 5000 words into 100 clusters. GaussianMixture was my choice since it's more generalized model than KMeans regarding the covariance matrix, KMeans restricts covariance to be identity while GaussianMixture suppose it can be any eclipsical shape. And I used covariance_type='full' for this model.

Below listed several clusters with word number between 3 and 30 from result of GaussianMixture.

```
['john', 'god', 'name', 'death', 'word', 'words', 'miss', 'heard', 'boy',
'love', 'wife', 'voice', 'woman', 'girl', 'mother', 'mean', 'alone', 'gone
', 'father', 'dead', 'son', 'police', 'hear', 'yes', 'remember', 'oh']
['also', 'used', 'use', 'find']
['since', 'american', 'however', 'part', 'number', 'less', 'given', 'order
', 'form', 'thus']
['say', 'told', 'nothing', 'knew', 'give', 'want', 'anything', 'really', '
tell', 'sure']
['second', 'early', 'half', 'period', 'close', 'short', 'million', 'third
', 'spent']
['good', 'still', 'come', 'thought', 'think']
['great', 'something', 'look', 'things', 'thing']
['later', 'several', 'four', 'five', 'past', 'six', 'recent', 'hundred', '
ten']
['put', 'head', 'eyes', 'toward', 'room', 'turned', 'open', 'feet', 'acros
s', 'car', 'behind', 'street', 'front', 'stood', 'moved', 'walked', 'opene
d']
['law', 'federal', 'department', 'secretary']
['general', 'public', 'system', 'program', 'business', 'group', 'national
', 'within', 'development', 'interest', 'area', 'service', 'political', 'e
conomic', 'community']
['without', 'left', 'course', 'away', 'hand', 'far', 'side', 'big', 'best
', 'ever', 'least']
['days', 'minutes', 'months', 'hours', 'weeks']
['city', 'members', 'england', 'orleans']
['become', 'help', 'whether', 'act', 'probably', 'result', 'turn', 'cost',
'seem', 'provide', 'run', 'plan', 'leave', 'believe', 'soon', 'increase',
'longer', 'call', 'expected', 'return', 'hope', 'pay', 'bring', 'certainly
', 'decided', 'appear', 'expect']
['found', 'always', 'almost', 'enough', 'took', 'yet', 'end', 'asked', 'lo
oked', 'felt', 'saw', 'seemed', 'seen']
['came', 'right', 'around', 'went', 'got', 'going']
['small', 'water', 'set', 'called', 'president', 'face', 'large', 'childre
n', 'church', 'light', 'family', 'mind', 'country', 'taken', 'body', 'alre
ady', 'moment', 'clear', 'morning']
```

Most of them shows some sense, while the highlighted ones make most sense to me.