CSE 250 B Programming 3 Swagata Chakraborty

1. **A description 100-dimensional embedding:**

   The steps are as follows:

   - Raw brown.words() have been downloaded from NLTK, which is pre-processed and the stop words/punctuations are removed to give a set of approximately 510K words from the corpus. The unique words are tabulated along with their corresponding frequencies.
   - The set V (vocabulary word), having the top 5000 frequent words, and a subset C (1000 frequently occurring context word set) of context words are generated.
   - A window matrix M (5000*5 dimension) having rows indexed by 5000 V words is created along with w1, w2, w3, w4 picked from C words, occurring alongside the V word w.
     Window Matrix: (**w1, w2, w, w3, w4**)
   - A dictionary of the 5000*1000 V-C pairs is created, and their co-occurrence count is computed. These values are populated for the above created Window Matrix.
   - P(c|w) = N(c,w) / N(w) is computed; P(c) = N(c) / N(total words) is also computed and henceforth, PHI vector is calculated for the brown words.
   - Principal Component Analysis has been used for dimensionality reduction of the above vector.
   - K-Means has been used on the reduced vector to form 100 clusters of these words and the calculated cosine distance is used to find the nearest neighbor from the clusters. The output for a set of test words as shown in the next topic.

2. **Nearest Neighbor Results:**

Nearest Neighbour algorithm was initiated for different seed values. Few of the results are listed below:

Experiment 1:

```
communism is near to tradition
autumn is near to storm
cigarette is near to smelled
pulmonary is near to artery
mankind is near to regiment
africa is near to asia
chicago is near to club
revolution is near to culture
september is near to december
chemical is near to indicated
detergent is near to fabrics
dictionary is near to text
storm is near to weekend
worship is near to theological
husband is near to wife
hospital is near to district
international is near to association
angel is near to businesses
slowly is near to stared
nuclear is near to weapons
justice is near to cause
building is near to office
brown is near to blue
administration is near to policy
equipment is near to production
```

Nearest Neighbour pairs like September-December, Pulmonary-Artery, administration-policy, husband-wife, revolution-culture, building-office, slowly-stared, detergent-fabrics, Communism – Tradition, Justice- cause, equipment-production, dictionary-text, Africa – Asia, Chicago-club, hospital-district, brown-blue, nuclear-weapon, etc. are good estimates and have very close contextual usage.

On the other hand, pairs like angel-businesses is not so good of nearest neighbor example.

However, if we run the algorithm again, it looks for alternate nearest neighbour pairs like:

Communism- Russia, storm-woods, administration-Kennedy, slowly-quickly, international-economic, angel-beam etc. but also compromises on good nearest selection pairs for a few others.

CLUSTERING:

KMeans algorithm has been used for the clustering and the Euclidean distance has been used for the same.
Yes, the clusters seem coherent. However, some of the huge clusters are not as coherent as the ones listed below and span over more diverse set of contextual words.

Example 1: Numbers

```
62: ['1', '2', '3', '4', '10', '5', '6', '8', '7']
```

Example 2: Amount/Quantity

```
36: ['high', 'less', 'result', 'cost', 'level', 'total', 'rate',
'increase', 'greater', 'expected', 'pressure', 'costs', 'amount', 'higher',
'increased', 'due', 'average', 'lower', 'unit', 'rise', 'rates']
```

Example 3: Learning

```
61: ['program', 'development', 'provide', 'schools', 'research', 'training'
, 'programs', 'additional', 'activities'],
```

Example 4: Places

```
58: ['city', 'york', 'south', 'west', 'north', 'america', 'east', 'central'
, 'entire', 'throughout', 'nation', 'southern', 'western', 'europe', 'citie
s'],
```

Example 5: Administration

```
97: ['state', 'states', 'united', 'public', 'government', 'business', 'act'
, 'federal', 'department', 'policy', 'support', 'defense', 'administration'
, 'services'],
```

Example 6: Military

```
43: ['military', 'addition', 'army', 'staff', 'corps', 'division', 'commiss
ion', 'funds', 'project', 'management', 'carry', 'assistance', 'operations'
, 'economy', 'personnel', 'maintenance', 'continuing', 'maintain', 'budget'
, 'establish', 'expenditures'],
```