Programming Assignment 3

1. **Description of 100-dimensonal embedding:**

Step 1: Preparing the data - The words were loaded by calling brown.words(). Stopwords.words('english') was used from nltk package to remove the stopwords from the word corpus. After removing the stopwords, the Counter function was used to get the count of each words in the corpus. The result is then sorted to get 5000 common words and 1000 common words which forms our set V and C respectively. Now, create a dictionary which will contain information about number of times each word in V and C occurs in the entire corpus.

Step 2: Counting the occurrence of neighbors - For each word in V, I created a vector with 1000 values. Each value in the vector corresponds to the number of occurrence of each word in C as neighbor within the sliding window of 4 neighbors around w (thus it's a count). We now have a 5000 X 1000 matrix, we call it as cw_count.

Step 3: Getting the Probabilities – From the matrix cw_count,

n(w,c) is each element in cw_count.

n(w,:) is a vector obtained by summing each row in c_w to produce 5000 X 1 vector. It is a measure of total number of neighbors for w that belongs to C.

Now,

$$Pr(c|w) = \frac{n(w,c)}{n(w,:)}$$

$$Pr(c) = \frac{number\ of\ repeats\ of\ c\ in\ corpus\ that\ belongs\ to\ set\ C}{total\ number\ of\ repeats\ of\ all\ words\ in\ corpus\ that\ belongs\ to\ set\ C}$$

Step 4: Positive Mutual Information

From the probabilities calculated, we can create |C| dimensional vector by the formula,

$$\phi_c(w) = \max\left(0, log\left(\frac{pr(c|w)}{pr(c)}\right)\right), which\ is\ again\ a\ 5000\ X\ 1000\ matrix$$

Step 5: Dimensionality reduction

I used PCA from sklearn.decomposition.PCA to transform the Positive Mutual information into a 100-dimensional representation for each word in w ($\psi(w)\ \epsilon\ R^{100}, for\ w\ \epsilon\ V$). PCA transforms the data in such a way that the components with maximum variance are kept to keep as much as information as possible with the constraint that the newly added component is orthogonal to preceding components.

2. **Nearest neighbor results:**

These are the collection of words and their neighbors,

SNo	Word	Neighbor
1	Pulmonary	artery
2	Africa	Asia
3	Chicago	Portland
4	chemical	milligrams

5	September	December
6	Detergent	Fabrics
7	Worship	Christian
8	Dictionary	Text
9	State	Federal
10	World	Nations
11	Men	Women
12	Day	Night
13	Work	Job
14	General	Public
15	God	Christ
16	Hands	Head
17	History	Literature
18	College	University
19	South	North
20	Six	Five
21	Land	Area
22	Washington	President
23	Tax	Income
24	Million	Billion
25	Book	Fiction

3. **Clustering:**
   The distance between two words was calculated by using cosine formaula. The clustering is done by kmeans clustering algorithm (nltk.cluster.kMeansClusterer, cosine_distance). I chose cosine similarity because it is large for two vectors that point in the same direction. This is important in understanding the semantics between the words. Kmeans is the logical thing to do once you have the distance because distance is the only parameter that defines the words in the space we built.

   Here are some clusters that I got and I have mentioned why I think they are related,

Writing/literature: [88, ['word', 'words', 'music', 'book', 'english', 'read', 'england', 'written', 'story', 'reading', 'french', 'note', 'writing', 'letters', 'write', 'poet', 'books', 'frequently', 'names', 'famous', 'poems', 'writer', 'master', 'song', 'naturally', 'message', 'speaking', 'bible', 'stories', 'roman', 'papers', 'poem', 'italian', 'numerous', 'hardy', 'writes']]

Time/year:[98, ['last', 'year', 'early', 'week', '1960', 'island', '1961', 'continued', 'during', 'series', 'march', 'fiscal', 'season', 'date', '1959', 'june', '1958', 'previous', 'mark', 'session', 'november', 'april', 'july', 'december', 'boston', 'ended', 'september', 'january', 'august', 'october', '1957', 'prior', '1954', 'february', 'journal', 'editorial', 'tv', 'scheduled', '31', 'schedule', '1962', 'ending', '1953']]

Expense/logistics: [54, ['work', 'money', 'job', 'worth', 'workers', 'build', 'offered', 'buy', 'reasonable', 'notice', 'advance', 'goods', 'raise', 'easier', 'expense', 'extra', 'strike', 'hopes', 'lots', 'spending', 'supplied', 'equipped', 'waste', 'branches', 'further', 'should']]

Direction/distance: [35, ['long', 'came', 'house', 'around', 'left', 'away', 'far', 'toward', 'side', 'along', 'open', 'line', 'across', 'car', 'behind', 'office', 'street', 'turn', 'center', 'town', 'outside', 'near', 'road', 'started', 'fire', 'moved', 'coming', 'inside', 'followed', 'move', 'reached', 'river', 'building', 'wall', 'floor', 'police', 'stand', 'lay', 'ran', 'running', 'steps', 'main', 'middle', 'corner', 'moving', 'pool', 'ahead', 'closed', 'drive', 'reach', 'built', 'walk', 'bridge', 'sea', 'pass', 'key', 'spread', 'bar', 'opposite', 'camp', 'pointed', 'jones', 'traffic', 'block', 'shore', 'circle', 'somewhere', 'stone', 'truck', 'draw', 'putting']]

Color/property:[55, ['water', 'eyes', 'white', 'light', 'black', 'red', 'dark', 'hair', 'blue', 'color', 'wide', 'blood', 'green', 'sun', 'deep', 'mouth', 'filled', 'thin', 'bright', 'gray', 'dog', 'sharp', 'beauty', 'dust', 'walls', 'dry', 'dress', 'warm', 'bodies', 'grew', 'wind', 'narrow', 'cool', 'soft', 'snow', 'sky', 'double', 'pure', 'colors', 'throat', 'pair', 'hills', 'leaves', 'lights', 'ring', 'spirits', 'ears', 'pleasant', 'fruit', 'warmth', 'shade', 'faint']]