CSE 250B: Homework 4

Xiyun Liu A53099348

December 2, 2016

1 Description of your 100-dimensional embedding

1. Download the *brown.words*() into *texts*, encode the uni-code words to string using "utf-8", then make each word lower case.

2. Get stopwords from *nltk.corpus* using $stopwords.words('english')$ and punctuation from *string.punctuation*. Remove stopwords and punctuation, and those words contains no character that is alphanumeric, such as "–" and "??" from *texts*.

3. Count the occurrence of each word, save to a dictionary $wordCount$, using $OrderedDict$ to sort $wordCount$ by the occurrence in decreasing order.

4. Select the top 5000 frequent words from the keys of $wordCount$ as *vocabulary* V, and the top 1000 frequent words as *conttext words* C. Record the words and their index in the V and C in to $v_wordToIndex$ and $c_wordToIndex$.

5. Create a 5000×1000 matrix *pr_cw* containing the $Pr(c|w) = \frac{n(w,c)}{n(w,:)}$. Detailedly, for each word in the filtered *texts*, if this word is in $vocabularyV$, for each word in the surrounding window, if it is in *context words* C, increase one to the corresponding value in *pr_cw*. Then, divide each row by the sum of the row.

6. Create a 1×1000 vector *pr_c* and save the overall distribution $Pr(c) = \frac{n(:,c)}{n(:,:)}$ of context words.

7. Calculate the *representation* which is a 5000×1000 matrix using $\Phi_c(w) = max(0, \log \frac{Pr(c|w)}{Pr(c)})$. Each row represents each vocabulary item w by 1000 dimensional vector $\Phi(w)$.

8. Using `sklearn.decomposition.PCA` to do PCA on *representation* to transform the 1000 representation into 100-dimensional representation. Save it to *reducedRepresentation*.

2 Nearest Neighbor Result

Cosine distance neglects absolute frequency difference which is represented by the length of embedding vectors and instead deals with relative difference. Therefore, I use cosine distance to find the nearest neighbor.

word	1^{st} neighbor	2^{nd} neighbor	3^{rd} neighbor	4^{th} neighbor	5^{th} neighbor
communism	phrase	justice	china	museum	located
autumn	storm	wines	winter	trail	fogg
cigarette	shut	lighted	peered	nodded	seated
pulmonary	artery	bronchial	saline	lungs	distributed
mankind	struggle	life	death	belief	history
africa	asia	western	europe	america	germany
chicago	club	board	top	boston	press
revolution	perhaps	lo	guns	known	hope
september	june	december	july	1960	april
chemical	feed	thermal	similar	results	concept
detergent	fabrics	grains	saline	butter	indirect
dictionary	text	occurrence	index	symbolic	stored
storm	autumn	reminded	wedding	eighteenth	clock
worship	shared	conscience	beliefs	life	religion
face	eyes	looked	hair	turned	suddenly
president	kennedy	chairman	conference	director	w.
education	national	public	program	schools	medical
million	billion	approximately	dollars	hundred	year
face	eyes	looked	hair	turned	suddenly
college	university	school	students	brooklyn	student
poet	still	hand	head	carl	sleeping
commission	education	federal	state	agencies	committee
sunday	monday	friday	night	tuesday	saturday
children	women	parents	family	child	girls
business	industry	local	private	public	sales

The result makes sense. The nearest neighbors of the words shares semantic and syntactic meaning with the words.

3 Clustering

K-Means algorithm (`nltk.cluster.KMeansClusterer`) is used to clustering. K-means clusterer starts with k arbitrary chosen means then allocates each vector to the cluster with the closest mean. It then recalculates the means of each cluster as the centroid of the vectors in the cluster. This process repeats until the cluster memberships stabilise. This is a hill-climbing algorithm which may converge to a local maximum. Hence the clustering is often repeated with random initial means and the most commonly occurring output means are chosen. The reason I use K-Means is that this algorithm can cluster words into groups.

The distance function is set to *cosine_distance*. The reason I use cosine distance instead of euclidean distance is because the length of word vectors represents the frequency and cosine distance neglects absolute frequency difference and instead deals with relative difference.

A few meaningful clusters are shown below. I also labeled each cluster with a title.

Cluster 4: Time ['day', 'home', 'week', 'morning', 'st.', 'hour', 'club', 'evening', 'returned', 'sunday', 'dinner', 'died', 'post', 'san', 'monday', 'saturday', 'newspaper', 'tomorrow', 'p.m.', 'guests', 'arrived', 'beach', 'boston', 'friday', 'tuesday', 'theater', 'philadelphia', 'calling', 'suffered', "o'clock", 'restaurant', 'a.m.', 'reception', 'scheduled', 'supper', 'wednesday', 'atlanta', 'thursday', 'funeral', 'wedding', 'weekend', 'noon', 'cocktail', 'announcement', 'workshop', 'arrive', 'luncheon', '29', 'vernon', 'wagner']

Cluster 32: Bible['life', 'church', 'god', 'death', 'love', 'spirit', 'heart', 'neither', 'fear', 'truth', "man's", 'born', 'faith', 'speak', 'knows', 'christ', 'everyone', 'lord', 'condition', 'created', 'secret', 'mission', 'accept', 'universe', 'wonder', 'birth', 'jesus', 'struggle', 'refused', 'bible', 'vision', 'loved', 'sin', 'protestant', 'holy', 'soul', 'deny', 'liberty', 'identified', 'wisdom', 'saved', 'heaven', 'unlike', 'conscience', 'mankind', 'virgin', 'salvation', 'eternal', 'gentle', 'kingdom', 'inspired', 'forgive', 'belongs']

Cluster 36: People['men', 'old', 'almost', 'yet', 'called', 'young', 'children', 'family', 'gave', 'today', 'past', 'seen', 'miss', 'known', 'wife', 'age', 'sometimes', 'child', 'strong', 'alone', 'women', 'living', 'except', 'live', 'person', 'lost', 'son', 'picture', 'friends', 'fine', 'working', 'sent', 'boys', 'girls', 'appeared', 'met', 'husband', 'de', 'learned', 'lived', 'scene', 'interested', 'married', 'playing', 'older', 'americans', 'parents', 'battle', 'finished', 'regular', 'mark', 'remembered', 'rich', 'failed', 'jewish', 'writer', 'independence', 'realized', 'leading', 'join', 'finds', 'informed', 'fought', 'younger', 'musicians']

Cluster 43: Money ['money', 'tax', 'amount', 'pay', 'paid', 'bill', 'income', 'date', 'oil', 'gross', 'fund', 'entitled', 'estate', 'extra', 'passage', 'bonds', 'bills', 'dollar', 'taxes', 'load', 'excess', 'reserve', 'cash', 'revenue', 'adjustment', 'receiving', 'returns', 'tied', 'farmers', 'builder', 'contributions', 'revenues', 'bears', 'monthly', 'taxpayers', 'receives']

Cluster 60: Modern Government ['state', 'president', 'company', 'board', 'department', 'party', 'washington', 'secretary', 'report', 'committee', 'meeting', 'police', 'county', 'congress', 'member', 'district', 'army', 'former', 'press', 'recently', 'reported', 'chief', 'staff', 'plans', 'hospital', 'democratic', 'director', 'officer', 'chicago', 'project', 'manager', 'citizens', 'workers', 'reports', 'officers', 'yesterday', 'campaign', 'election', 'vote', 'official', 'texas', 'jury', 'attorney', 'headquarters', 'california', 'officials', 'senate', 'duty', 'minister', 'joined', 'executive', 'republican', 'co.', 'appointed', 'engineer', 'representatives', 'proposal', 'legislature', 'latest', "president's"]

Cluster 74 : School ['college', 'students', 'professor', 'trained', 'automobile', 'demanded', 'estimate', 'profession', 'similarly', 'scholarship', 'visiting', 'harvard', 'campus', 'elected', 'connected', 'universities', 'seventh', 'dartmouth', 'graduate', 'carleton', 'brooklyn', 'builders', 'anti-trust', 'bearing', 'attracted', 'tended', 'drivers', 'thinks', 'publications', 'furnished', 'expects', 'belgians', 'historians', 'mathematics']