CSE 250B – Word embeddings

Jimmy Zhu (A53054773)

1 Objective

In many language processing applications, it is useful to perform word *embedding*, which clusters words with similar meanings together. In this assignment I apply an example of such an embedding on the `Brown` corpus (from `nltk.corpus`).

2 Approach

After some initial preprocessing of the text (removing stopwords and punctuation), I begin by creating two lists: (1) a vocabulary V of the 5000 most commonly occurring words, and (2) a shorter list C of the 1000 most frequent words.

For each word pairing $(w, c) \in V \times C$, I compute $n(w, c)$, the number of times that context word c appears within a 4-word window of w (two before, two after). Using these counts, I derive the *pointwise mutual information* matrix, defined as follows:

$$\Phi : \Phi_{w,c} = \max\left(0, \log\left(\frac{\Pr(c|w)}{\Pr(c)}\right)\right)$$

where the probabilities are estimated by

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

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

The rows of this matrix Φ embeds each word $w \in V$ as a 1000-dimensional vector.

Dimensionality reduction

Given the 1000-dimensional embedding Φ, I obtained a more compact 100-dimensional embedding Ψ using low-rank approximation via singular value decomposition:

$$\Phi = \begin{pmatrix} \uparrow & & \uparrow \\ u_1 & \cdots & u_p \\ \downarrow & & \downarrow \end{pmatrix} \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_p \end{pmatrix} \begin{pmatrix} \leftarrow & v_1 & \rightarrow \\ & \vdots & \\ \leftarrow & v_p & \rightarrow \end{pmatrix}, \qquad p = 1000$$

Taking the largest 100 singular values σ_i and corresponding basis vectors u_i and v_i, Φ is then approximated as follows:

$$\Phi \approx \Psi = \begin{pmatrix} \uparrow & & \uparrow \\ u_1 & \cdots & u_k \\ \downarrow & & \downarrow \end{pmatrix} \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k \end{pmatrix} \begin{pmatrix} \leftarrow & v_1 & \rightarrow \\ & \vdots & \\ \leftarrow & v_k & \rightarrow \end{pmatrix}, \qquad k = 100$$

The resulting matrix Ψ is the best k-rank approximation of Φ, which ensures that the 100-dimensional embedding is very close to the higher-dimensional embedding.

3 Results

Nearest Neighbor

Using the *cosine distance* metric

$$d(w, w') = 1 - \frac{\Psi(w) \cdot \Psi(w')}{\|\Psi(w)\| \, \|\Psi(w')\|}$$

I performed nearest neighbor classification on the 100-dimensional embedding Ψ to identify the top 3 nearest neighbors for a set of 25 words, as shown in Table 1.

Word	Nearest Neighbors	Word	Nearest Neighbors
`administration`	`policy state support`	`africa`	`asia europe south`
`autumn`	`summer late storm`	`bridge`	`red river old`
`chemical`	`drugs study clinical`	`chicago`	`club portland york`
`cigarette`	`nodded stopped swung`	`college`	`university school students`
`communism`	`century utopian world`	`detergent`	`fabrics soap cleaning`
`dictionary`	`text description list`	`evil`	`fear certainty realize`
`eyes`	`face saw looked`	`horse`	`turned together head`
`jazz`	`music musical art`	`love`	`god knew man`
`metropolitan`	`suburban churches areas`	`pulmonary`	`artery bronchial comparison`
`revolution`	`modern mans nuclear`	`river`	`lake across valley`
`september`	`december week th`	`storm`	`summer saturday night`
`wife`	`mother husband young`	`wine`	`taste always gentlemen`
`worship`	`religion religious theological`		

Table 1: Top 3 nearest neighbors for 25 selected words (listed in order from nearest to farthest).

The results from nearest neighbor with this embedding performs extremely well. It captures related words like `detergent`, `fabrics`, `soap` and even more subtle connections like `evil`, `fear`, `realize`. Of course, some nearest neighbors are not as obvious as others, with groupings like `cigarette`, `nodded`, `swung` and `horse`, `turned`, `together`. This just affirms the effectiveness of the pointwise mutual information embedding.

Clustering

Using the same 100-dimensional embedding Ψ, I then performed *k-means clustering* with 100 clusters, this time just using simple Euclidean distance for the distance metric. This algorithm computes 100 central points in the dataset by iteratively divding and recomputing means for 100 distinct clusters. From the results, ten representative clusters are shown in Table 2.

The clusters obtained from k-means are reasonably coherent in that there is generally a common theme in each cluster (i.e. *travel*, *religion*, etc.). Some clusters are very well identified, like for *religion* (`life`, `god`, `death`). Others have words that mostly agree, but with some outliers, like for *colors* (`white`, `black`, `red`, `hair`, `dress`). But even for outliers, it is fairly straightforward to see the connection to an underlying cluster topic (colors are often used to describe hair and clothing).

4 Conclusion

The pointwise mutual information word embedding is extremely successful in grouping words together, both through nearest neighbor and k-means clustering. Even though I used the *cosine distance* for nearest neighbor and Euclidean distance for k-means, both strategies produced promising results. With nearest neighbors, related words can be found for each specific word in the vocabulary, whereas with k-means, the clusters are not tied to specific words in the vocabulary, but rather to underlying topics that connect the grouped words.

It would be interesting to develop a strategy to use *cosine distance* for k-means clustering, but I did not employ this strategy for the sake of simplicity, and also to avoid any potential problems with convergence that might arise from not using the Euclidean distance in the algorithm.

Cluster	Grouped words
Travel	passed worked visit sea leaving train trip planned rock village traffic drove travel beach mile faced driving angeles vacation lights riding dancing palace japan visitors italy weekend lunch lane journey horizon fogg
Government	state public government national local federal economic policy education plan private support defense administration foreign international labor trade aid responsibility council funds security assistance financial operations governments educational policies agencies housing agency welfare cooperation
Arts	fine color performance audience playing popular film dance musical treated fashion plays excellent arts imagination dramatic songs novel painting truly spite pure sounds television opera singing ballet drama favorite concert ladies characters folk performances films dancers recording dancer sang lively delightful academy dances novels speeches
Quantity	high less often matter seems usually strong greater low material difficult subject likely particularly somewhat gives quality choice rise remain obvious frequently produce atmosphere presence reasonable sensitive fairly details becoming foods
Time	home night morning late meeting hour evening hotel returned afternoon sunday doctor dinner yesterday stayed monday saturday pm tomorrow cool friday tuesday fathers oclock supper wednesday thursday noon
Names	de governor jones honor v joseph martin co remarks david arthur eisenhower speaker author alexander edward vice dean walter lawrence mayor colonel questionnaire baker assistant jackson clark atlanta lee johnson samuel rev van howard sen davis hughes albert k sergeant appointment jefferson kennedys mitchell houston ford memorial victor gen douglas browns ambassador morris francis vernon taylor purchased allen donald commissioner gov frederick
Religion	life god death love spirit st mans born faith knows christ lord condition universe unity birth jesus sin holy soul wisdom heaven conscience virgin salvation eternal kingdom inspired belongs glory
War	nuclear attack greatest possibility aircraft capable greatly weapons task advance apparent begins theme device generation risk easier passage millions missile concrete precisely destroy instrument valuable prime target extensive warfare increasingly weapon bound worlds emergency furthermore striking introduction turns camera automatically intention meanwhile accurate worst abstract naval missiles remote fallout starts consequence proportion committed african submarine overcome exposure accurately bombers unexpected targets variations strategic registration valid earliest strategy certainty resolved drawings indication negotiations sailing submarines
Colors	white black red hair blue green deep thin bright gray thick dress wore nose sky yellow tall tiny pair brilliant pink suit skin wearing golden tongue jacket falling bare stained shade shirt beard faint tie shining stiff blonde brushed jungle
Statistical charts	af points image aj plane fixed follows hence q curve meets pencil zg tangent transformed arbitrary bundle vertex

Table 2: Ten representative clusters obtained from k-means clustering ($k = 100$).