Aditya Chempakasseril

1. While the techniques used to obtain the 100-dimensional embedding were relatively straightforward, a few subtle decisions were made that had a significant effect on the final embedding. First, care had to be taken when cleaning up the raw data. Punctuation had to be removed, but it was important that words with punctuation within them (e.g. "all-natural") remained. If all hyphenated words were removed, for example, there could be significant loss of meaning in some sentences. Furthermore, it was important that the stopwords chosen to be removed were comprehensive but not *too* comprehensive. Again, a fine balance had to be struck between preservation of meaning and removal of syntactic noise.

After cleaning the raw data, subsets V and C, consisting of 5000 and 1000 of the most commonly-occurring words, respectively, were chosen. Initially, subset V (and C) consisted of the *top* 5000 words. However, after analyzing the final clusters, it became apparent that the top two words in V and C contained no useful information. Thus, the elements chosen for V were the top 5000 words starting from the third most commonly-occurring word.

Next, a word co-occurrence matrix between V and C was produced. From this matrix, some basic probabilities were extracted, namely the probability distribution of context words ($c \in \mathbb{C}$) around vocabulary ($w \in \mathbb{V}$) and the overall distribution of context words. In information and statistical theory, there is a probability measurement known as pointwise mutual information (PMI). PMI is a measure of how often two outcomes "coincidentally" occur together. The mathematical formulation follows:[1]

$$pmi(c,w) = log\frac{p(c,w)}{p(c)p(w)} \tag{1}$$

Now, if *p(c)* and *p(w)* are independent, then:

$$pmi(c,w) = log\frac{p(c|w)}{p(c)} \tag{2}$$

In the context of this assignment, the choice of words in a sentence is not completely independent; however, it can be assumed to be mostly independent (barring internal bias and grammatical rules). Thus, for our purposes, Equation 2 was used to calculate PMI. A $|C|$-dimensional vector $\Psi(w)$ could then be produced, with the c'th coordinate of each vector:

$$\psi_c(w) = max\left(0, log\frac{p(c|w)}{p(c)}\right) \tag{3}$$

Only positive PMI were used since a negative PMI indicates that the co-occurrence of *c* and *w* happens less than that dictated by chance, which shouldn't be the case if *c* and *w* are independent. Each of these 5000 vectors represents, in some sense, how closely related different words in *c* and *w* are.

Finally, to reduce Ψ(w) from 1000 dimensions to 100 dimensions, PCA was used. The specific functions used were:

```
pca = PCA(n_components=100)
principal = pca.fit_transform(vectors)
```

where `vectors` is the set of Ψ (w).

2. Table 1 below shows the nearest neighbor for 25 random words $w \in \mathrm{V}$.

Word (w)	Nearest Neighbor
communism	utopian
autumn	late
cigarette	nick
pulmonary	artery
mankind	divine
africa	asia
revolution	english
september	december
chemical	thermal
detergent	fabrics
dictionary	text
storm	saturday
worship	religion
chicago	portland
information	dictionary
society	free
utopian	communism
jail	cook
government	states
effect	seems
timber	fishing
death	life
president	kennedy
order	case
therapist	conversation

Table 1: Nearest Neighbors

The distance metric used in the NN algorithm was the cosine distance, expressed below:

$$1-\frac{\psi(w)\cdot\psi(w')}{\|\psi(w)\|\|\psi(w')\|} \tag{4}$$

This metric will be discussed further in the next section.

As can be seen, the results of the NN search are highly satisfactory. While the nearest neighbors of each w aren't synonyms for the w, they *are* words that convey a similar feeling as the w. This is because each w was embedded using PMI. Thus, words considered similar are words that share similar contexts. Take, for example, “chicago" in Table 1. Its nearest neighbor is “portland.” These are both big cities, which means there is a good chance they will be surrounded by similar words. “Life” and “death” are often spoken of in similar social, political, scientific, and philosophical contexts. “Utopian” and “communism” both refer to different states society can take on. Most of the pairings in Table 1 make sense.

There are a few, however, that are questionable. For example, “storm” and “saturday" seem very weakly related. Storm may appear in many similar contexts to saturday, but the reverse isn’t necessarily true. But, it may also be the case that the Brown corpus includes a major storm or several storms that occur on Saturday(s). Thus, this type of nearest neighbor analysis may be useful in determining how related different words are in reference to *specific* works. For example, the fact that the nearest neighbor of “society” is “free” speaks to the fact that the works in this corpus are probably very politically inclined with major social undertones.

3. To group the words in *V*, k-means clustering was used on the 5000 vectors $\Psi(w)$ to form 100 clusters. k-means clustering was chosen for several reasons. For one, connectivity-based clustering methods typically run in $O(n^3)$ time, which is quite slow for large datasets. k-means, on the other hand, runs in O(n) time. Furthermore, it is specified that the data should be divided into 100 groups. In general, a major drawback of k-means clustering is that k needs to be specified beforehand. In our case, k *is* specified beforehand, so this drawback is irrelevant. Another drawback of k-means clustering is that the final clusters are sometimes heavily dependent on the choice of the initial k clusters. To combat this, the algorithm was run 20 times, where a different initial set of 100 clusters was chosen each time.

The distance metric used was similar to the cosine distance that was used for the Nearest Neighbor calculations. Because each of the $\Psi(w)$ were formed by calculating the PMI of different word pairs, the cosine distance between any two Ψ are indicative of the degree of similarity of the contexts in which the two corresponding *w* appear. A Euclidean distance metric wasn’t used since, in general, word similarity should be independent of word frequency. For the same reason, the Manhattan distance wasn’t used either.

Because the k-means algorithm in sci-kit learn only supports using Euclidean distance as its distance metric, the Ψ had to be modified such that:

$$sign[d_1(a,b) - d_1(b,c)] = sign[d_2(a,b) - d_2(b,c)] \tag{5}$$

where $d_1(x,y)$ is the modified Euclidean distance between *x* and *y*, and $d_2(x,y)$ is the cosine distance between *x* and *y*. Thus, even if the distances produced by d_1 and d_2 are different, a point *a* that is farther than *b* than from *c* under one variant will be farther from *b* than from *c* in the other variant as well. So how was each Ψ transformed to achieve this equality? All the vectors were simply normalized. The Euclidean distance between two vectors **q** and **p** is:

$$\|\boldsymbol{q} - \boldsymbol{p}\| = \sqrt{\|\boldsymbol{q}\|^2 + \|\boldsymbol{p}\|^2 - 2\boldsymbol{p} \cdot \boldsymbol{q}} \tag{6}$$

However, if **q** and **p** are normalized, then Equation 6 reduces to:

$$\|\boldsymbol{q} - \boldsymbol{p}\| = \sqrt{2}\sqrt{1 - \boldsymbol{p} \cdot \boldsymbol{q}} \tag{7}$$

The cosine distance between **q** and **p** is:

$$cosine\ distance = 1 - \boldsymbol{p} \cdot \boldsymbol{q} \tag{8}$$

It is clear that Equations 7 and 8 satisfy Equation 5.

After normalizing the Ψ and applying sci-kit learn’s k-means clustering algorithm, 100 clusters were formed. Most of the clusters were satisfactory, with the words in each cluster being similar in the sense that they appear in similar contexts. Table 2 gives a few examples. The data is given as a pair of words, where the first word is the cluster label, and the second word is the cluster member. The cluster label is not important in analyzing the clusters.

Cluster 1	Cluster 2	Cluster 3
many,two	school,mr.	mr.,states
many,years	school,mrs.	mr.,united
many,three	school,john	mr.,free
many,later	school,dr.	mr.,america
many,days	school,brown	mr.,nations
many,several	school,william	mr.,countries
many,four	school,j.	mr.,entire
many,times	school,s.	mr.,throughout
many,five	school,a.	mr.,nation
many,ago	school,james	mr.,western
many,six	school,charles	mr.,europe
many,minutes	school,c.	mr.,civil
many,months	school,e.	mr.,friendly
many,hours	school,w.	mr.,european
many,weeks	school,henry	mr.,britain
many,ones	school,robert	mr.,africa
many,o'clock	school,judge	mr.,asia
many,nights	school,b.	mr.,atlantic
many,seventeen	school,jr.	mr.,japan
many,hits	school,richard	mr.,peoples
many,stroke	school,h.	mr.,canada
	school,chairman	mr.,survive
	school,honor	mr.,eastern
	school,d.	mr.,sovereign
	school,fellow	mr.,african
	school,m.	mr.,suspicion
	school,r.	mr.,puerto
	school,f.	mr.,anti-semitism
	school,p.	mr.,rico
	school,l.	mr.,katanga
	school,martin	mr.,treaty
	school,joseph	mr.,alliance
	school,smith	mr.,socialism
	school,arthur	mr.,respective
	school,representative	
	school,edward	
	school,g.	

school,adams
school,vice
school,senator
school,lawrence
school,clark
school,johnson
school,sen.
school,o.
school,albert
school,n.
school,davis
school,hughes
school,t.
school,jefferson
school,mitchell
school,houston
school,gen.
school,brown's
school,victor
school,rev.
school,morris
school,vernon
school,taylor
school,allen
school,notte
school,purchased
school,frederick
school,austin

Table 2: 3 Selected Clusters from 100.

As seen in Table 2, the first cluster comprises words that refer to a quantity of something; the second cluster, names and name prefixes; the third cluster, countries and words related to international affairs.

Works Cited

1. https://en.wikipedia.org/wiki/Pointwise_mutual_information