CSE 250B: HW4

WEIWEI LI
PID: A53107958
DIS: WED 7PM

1. A description of 100-dimensional embedding. We want to clustering the Brown corpus. We preprocess the dataset first by removing stopwords and punctuation, and making everything lowercase. Then we constructed a vocabulary set V of 5000 of the most commonly-occurring words, and a context words set C of 1000 of the most commonly-occuring words. For each word w in V, and each occurrence of it in the text stream, we count of how often context words from C appear in the surrounding window of w (two words before of w, two words after w). Using these counts, we construct the probability distribution Pr(c|w) and the overall distribution Pr(c). And represent each vocabulary item w by |C|-dimensional vector $\Phi(w)$, whose c'th coordinate is $\Phi_c(w) = max(0, \log \frac{Pr(c|w)}{Pr(c)})$, which is know as the (positive) pointwise mutual information.

We want a 100-dimensional representation. The basic idea of my 100-dimensional embedding method is as follows: I use the principal component analysis (PCA), and use singular value decomposition (SVD) of the data to project it to a 100-dimensional space.

2. Nearest neighbor results. We pick a collection of 25 words from V, which are listed in table 1.

head	business	face	money	air
third	brown	month	seven	parents
happy	chicago	communism	revolution	chemical
dictionary	september	africa	mankind	worship
pulmonary	storm	cigarette	detergent	autumn

TABLE 1. Words selected from V

We want to find their nearest neighbor word by the cosine distance. The returned results are listed in table 2.

As we can see, the nearest neighbor words are not selected completely random. For example, the nearest neighbor word of 'head' is 'hands', which are the same kind. Also, 'autumn' and 'winter' go into the same category. And people will link the words 'pay' and 'money'. So the results make sense.

Date: November 30, 2016.

hands	local	eyes	pay	water
second	hair	last	ten	children
couldnt	kay	reflection	weapons	instances
tex	kay	kay	nation	theological
artery	eighteenth	slid	fabrics	winter

TABLE 2. Neareat neighbor words

3. Clustering. We use K-means algorithm to cluster the words in V into 100 groups by the euclidean distance and cosine distance. We want to compare which distance of K-means algorithm works better for the vocabulary V. So we repeat the process 5 times for each distance, and get several best clusters from them.

points	image	follows	l	platform	angle
curve	parallel	shear	axis	coating	meets
pencil	zg	tangent	ransformed	substrate	convenient
arbitrary	bundle	vector	vertex		

TABLE 3. K-means: Euclidean distance

capacity	economy	initial	dramatic	machinery
critical	developing	advertising	mere	accomplished
varied	examine	extend	complicated	economical
urgent	expanded			

TABLE 4. K-means: Euclidean distance

looks	memory	somehow	understood	bitter	flesh	laughed
surprise	terrible	forgotten	innocent	sad	grave	gesture
remote	remark	invariably	dreams	gay	anxious	gathering
utterly	intelligent	legend	painful	lonely	vague	adults
noble	unexpected	dive	uneasy	ugly	despair	killer
defend	dare	obliged				

TABLE 5. K-means: Cosine distance

son	married	died	strange	daughter	gets	mothers
finds	professor	enjoyed	loved	thoughts	practically	silent
cook	admitted	believes	lawyer	hate	teach	chose
sending	tells	dying	bride	rejected	romantic	troubled
promptly	prefer	unhappy	grateful	relieved	womans	educated
loves	mercy					

TABLE 6. K-means: Cosine distance

From the words in the clustered group, we can see that cosine distance works better. The words in cluster groups of cosine distance is more coherent. For example, most of

the words in table 5 is about people's feelings, and the words in table 6 is about people's profession and their action. The sizes of the groups clustered by euclidean distance are very different, some groups contain hundreds words, and some only contain a couple words. So I choose the cosine distance. And some of the best group clustered by cosine distance is listed above.