William Edgecomb 5/3/16
CS-114 HW5 Write-Up

Part 1.

Counts before PPMI reweighting:

	the	men	feed	dogs	women	bite	like
dogs	91	1	1	1	1	31	11

Counts after PPMI reweighting:

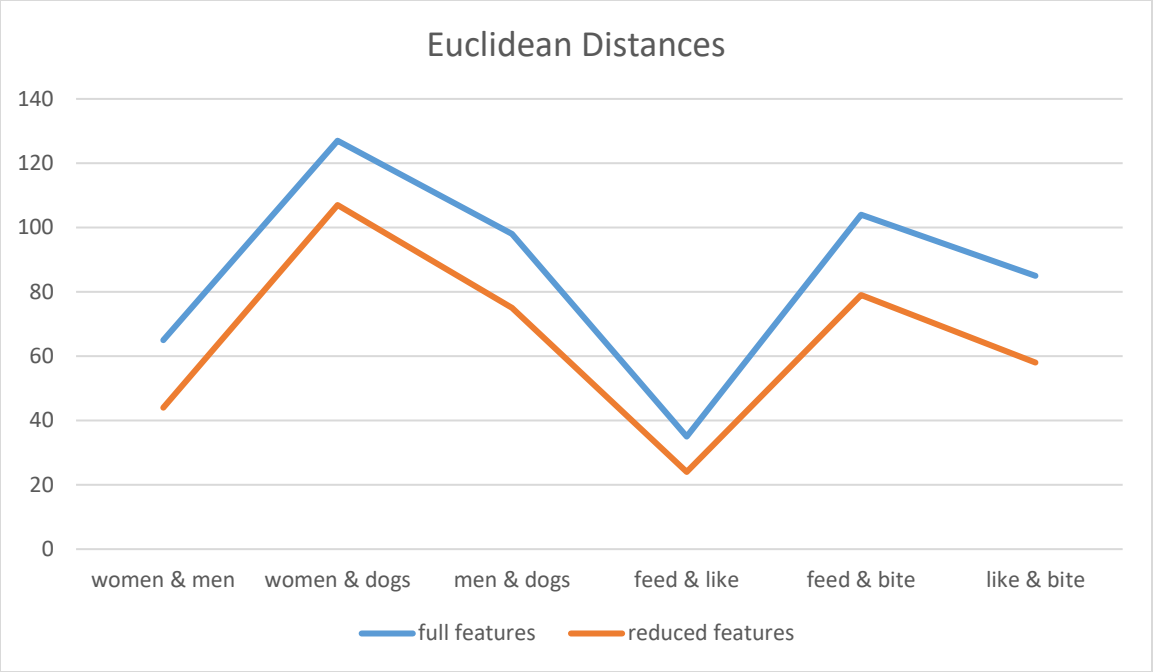
	the	men	feed	dogs	women	bite	like
dogs	169.429	0	0	0	0	81.802	8.439

Reweighting by the PPMI matrix yields the expected result. Features corresponding to promiscuous words, or words that appear next to a lot of different words, get their counts lowered, and vice versa for feature words that are more tightly coupled to the vector word (and by saying counts are lowered or raised, I don't mean in absolute numbers, but in terms of proportion to other counts). So with 'dogs' for example, the ratio of count('bite') to count('the') goes up dramatically. Intuitively this is an improvement. 'bite' is something of an animal-specific word, whereas 'the' is highly promiscuous, so we should probably expect 'bite' to have a stronger semantic relation to 'dog' than 'the'.

Euclidean distances:

	women & men	women & dogs	men & dogs	feed & like	feed & bite	Like & bite
All features	~65	~127	~98	~35	~104	~85
Reduced features	~44	~107	~75	~24	~79	~58

The Euclidean distances indeed confirm naïve intuitions. The distances values suggest (using Euclidean distance as a proxy for semantic relatedness) that human word-human word pairs are more closely related than human word-animal word pairs.



We can see that after the number of features is reduced to three, the relationship between Euclidean distances of word pairs is preserved. While the absolute values of the distances are decreased, the distances remain in proportion. The vectors have fewer dimensions, but travel through feature-space in an analogous fashion. Thus all vectors taken together, the reduced feature vectors retain the same essential distributional semantic information.

Part 2.

Synonym test results:

	Questions Correct	Questions Excluded	Percent Correct (correct / (1000 – excluded))
Google, Euclidean	518	111	58%
Composes, Euclidean	510	90	56%
Google, Cosine	572	111	64%
Composes, Cosine	510	90	56%

Analogy test results:

	Questions Correct	Questions Excluded	Percent Correct (correct / (374 – excluded))
Google, Concatenation and Cosine Similarity	132	53	41.12%
Composes, Concatenation and Cosine Similarity	119	55	37.30%
Google, Vector Subtraction and Euclidean Distance	104	53	32.40%
Composes, Vector Subtraction and Euclidean Distance	118	55	37.00%

The first strategy I tried with the analogy test was to choose the word pair with the Euclidean distance between words that best matched the Euclidean distance between the input words. This got me results around chance. Apparently analogical relations have little to do with semantic distance alone. Then I tried an analogous approach with cosine similarity, and this time had percents correct around 27. Since cosine similarity measures similarity in the *angle* between vectors, it appears that direction is a more accurate proxy for analogical relations than distance. I also tried weighting cosine similarity and Euclidean distance into a single metric, but was unable to improve upon the original cosine similarity results.

Next I tried subtracting vectors and choosing the resultant answer vector that had the closest Euclidean distance to the resultant input vector. This strategy was intuitive. Basically, by subtracting the word vectors, we get a measure of the motion through semantic space that connects one word vector to the other. The hypothesis then would be that the vectors for words in analogous word pairs move similarly through distributional semantic space. The results suggest this is at least somewhat true.

I also tried concatenating vectors and choosing the resultant answer vector that had the closest cosine similarity with the resultant input vector. Concatenating the vectors allows information from both original vectors to be represented in a single metric, but unlike vector addition, which would yield the same endpoint, vector concatenation preserves all of the feature information. The results here were good, with a high value of 41%.

I did not try using vector subtraction with cosine similarity or concatenation with Euclidean distance, although these would be interesting to try. By itself, cosine similarity far outperformed Euclidean distance, so for distributional semantics perhaps it is always or quite often the superior metric.