CSDS 435: Project 2

John Mays

04/02/23, Dr. Li

1 Dataset Summary

# Num of Tweets	# Num of Words (total)	# Num of Tokens	# Avg Num of Words
4,061	48,233	13,685	11.877

1.1 Top Five Legal Tokens:

1. *"you"*
2. *"rt"*
3. *"your"*
4. *"are"*
5. *"this"*

1.2 Stopwords:

After seeing what common words were in the dataset, I decided to add some of them to the stopwords, but not every common pronoun or common preposition. I decided to leave words like "you" and "are" in because these words are encoded with information. "are" is present tense and might appear on more optimistic tweets more often thant "were", similarly to "you" which could belong to outwardly critical tweets more often. "a" and "an" are pretty neutral, but other common stopwords may not have been. Here are the stopwords I currently use:

"a", "an", "the", "is", "to", "for", "in", "of", "and", "on"

2 Distance Measures

2.1 Normalized Unordered Edit Distance

$$d_1(x, y) = \frac{\sum_k x_k - y_k}{\sum_k x_k + y_k}$$

The intuition behind this was, when I was considering how you compare strings, I was reminded of edit distance (both for words and sentences). Of course, I realized that edit distance required an ordered string, and our bag-of-words model doesn't preserve the order. But I thought, very simply, the difference between two vectors is essentially a version of edit distance that does not pay attention to order (hence "unordered"). But I also thought two four-word sentences with two words in common are about as near as two six-word sentences with three words in common, or at least that this might be a good hypothesis to test since my other distance measure *does not* treat the data this way. Hence, I normalize the distance by dividing by the total number of words in both. This distance measure is $\frac{\text{\# of words not in common}}{\text{\# total number of words in both sentences}}$. Therfore, the minimum value ($d_1(x, x)$) is zero, and the maximum value (for vectors with no words in common) is one.

2.2 Euclidean Distance

$$d_2(x, y) = \sqrt{\sum_k (x_k - y_k)^2}$$

Euclidean distance seemed intuitive. It will, as opposed to my first distance measure, attribute greater distances on average to longer tweets. Two sixty-word tweets will be much farther apart than two ten-word tweets (assuming they each have nothing in common).