Technical Write Up

The vector space is a compact, compressed representation of a complex collection of vector objects. A vector or vector space is capable of supporting mathematical operators such as being added and subtracted to other vectors, and scaled, or multiplied, by a number. Another way to think of a vector space is as an entity that is represented as a point in a higher dimension, where Euclidean distance between vector spaces, or entities, represents a measure of similarity or dissimilarity. Feature space is a n-dimensional vector of numerical features, where a dimension for either space can be compared to a column or a row of a data frame only in a single direction. An interesting attribute of vector spaces is that they can also be feature spaces, whereas feature spaces can not be vector spaces.

The data frame of tweets seemed to have many words that had a low term frequency, after passing a minimum document frequency of 10% the feature space substantially decreased by 902 dimensions. Contrastingly, the binary and lowercase parameters had little to no affect on the feature space and vector space models. The different parameter settings passed to count vectorizers using a 500-observation data frame had no affect on the number of records, it remained constant at 500 observations, given any combination of parameters. The most useful parameter to use when answering the question of future attendance is most likely the minimum document frequency (min\_df). However, the stop\_words (English) parameter also appears to clean up words that do not provide context to a tweet’s sentiment, on a smaller scale compared to the min\_df parameter (Figure 1).

I think that using weights does not make sense because the 500 documents used during analysis was compared to a population sample size for our question, that addresses changes in attendance at Chiefs games. Using a weight might be useful if we were addressing individual term frequency, for example we could see tweets of “super fans” that have a frequently occurring term in their tweet object text, that does not appear frequently in other documents. For the scenario of determining NFL game attendance, the “super fans” might skew the other objects. Alternatively, weight could be used to determine an overall attitude toward the Kansas City Chiefs organization.

**Figure 1:** Count Vectorizer Parameter and Feature Space Comparison

|  |  |  |
| --- | --- | --- |
| Count Vectorizer | Parameters | Feature Space |
| cv1 | Default | 1009 |
| cv2 | binary = 'False', lowercase = 'True' | 1009 |
| cv3 | binary = 'False', stop\_words = 'english', lowercase = 'True' | 872 |
| cv4 | binary = 'False', stop\_words = 'english', lowercase = 'False' | 872 |
| cv5 | binary = 'False', max\_df = 0.85 | 871 |
| cv6 | binary = 'False', min\_df = 0.10, stop\_words = 'english' | 19 |