Clustering Startups Based on Customer-Value Proposition milestone

Daniel Semeniuta (dsemeniu), Meeran Ismail (meeran)

November 21, 2017

1 Motivation

Countless new startups are born every single day, and venture capitalists are always on the lookout to find which one will be the next big thing. To do this, they must learn an incredible amount about the startups they want to invest in. One piece of information that is especially valuable to investors is the industry that a startup is in and the industry competition it faces. As such, classifying startups by industry function is an important tool in investing; however, doing this for the many thousands of startups that are formed every day is impossible by hand. We thus want to use machine learning to cluster companies by customer value proposition, given nothing more than short one to two lines describing what the company does.

We believe that clustering startups by their industry allows users interested in the startup landscape to see the competition within specific industries and to see whether certain problems are already being tackled. Having defined industry clusters will allow for a next step of allowing users to input their own descriptions, and the system outputting a classification, along with similarly classified startups from our training set. Our end goal is integration into our group's project within CS 221 as well, which will realize this full potential.

2 Method

Our primary method of classification will be via k-means clustering. Our feature space is an n-dimensional space, where n is the total amount of unique words found in the corpus of our dataset. Currently, each input x of our dataset is represented using a numpy sparse vector via sklearn, where x_i is an indicator of whether word i appears in company x's description. Traditionally understood stop words have been cleaned from the dataset prior to running the learning algorithm. We use the built in methods of sklearn to cluster our dataset into different groupings representing industry or similar purpose of startup.

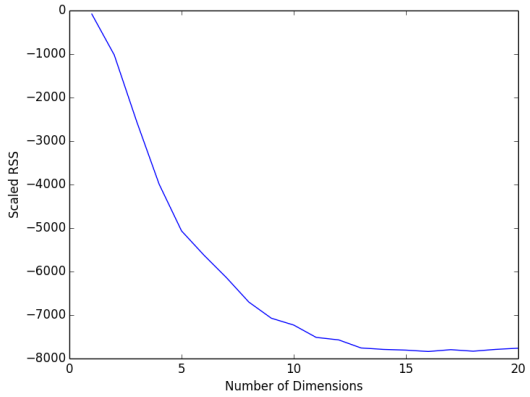
As a feature space the size of the corpus is large (over 90,000 unique words), especially for a clustering algorithm, we also begin preliminary methods of feature reduction. In addition to removing stopwords, we'll use singular value decomposition to reduce the dimensionality, because it is known that reducing to d dimensions by only keeping the d largest singular values creates the closest projection of X onto a d dimensional subspace.

Our methods will be evaluated through the standard means of evaluating clustering algorithms. Currently, numerical success is simply measured by standard reconstruction loss. However, part of the experimentation of our project is discovering what dimensionality we need to reduce the corpus to in order to obtain ideal clustering; thus, because dimensionality can change, our evaluation metric will be K-means reconstruction loss scaled by $\frac{1}{d_{reduced}}$. Additionally, it will be useful to conduct an eye test on the clustering of our data to make sure conceptually that startups in the same cluster are meaningfully related in purpose or industry.

3 Preliminary experiments

Our original baseline approach was a K-means on the entire dataset, with no feature reduction at all.

However, running that turned out to be impossible and extremely expensive, due to the fact that there were over 90,000 features and over 70,000 documents. More importantly, however, even if we were able to efficiently run the algorithm, it wouldn't be very useful; in extremely high dimensional spaces such as ours, almost all points become equidistant. This meant that we had to use dimensionality reduction. A quick trial with the dimensionality in the (inclusive) range of 1-20 gave us the following reconstruction *training* score:

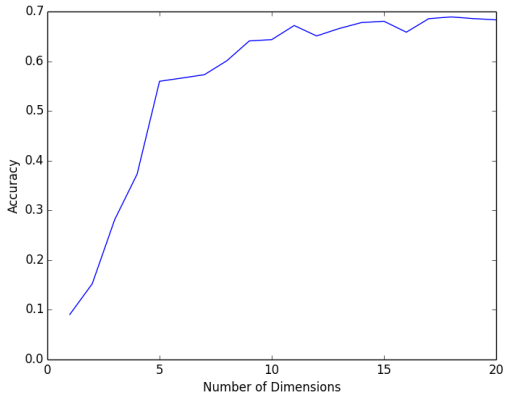


4 Next steps

Our next steps are conducted with the intention of improving our results in clustering. These next steps include further reduction of our feature space. Words like "health" and "healthcare" are obviously interrelated. There is nothing gained from representing these as independent features in the feature matrix. Grouping word stems into the same feature simplifies our dataset and also adds greater weight to a combined health umbrella, for example, than the individual words would. Additionally, methods such as principled component analysis would help us in eliminating features which are correlated with each other, again simplifying our feature matrix.

Choice of language is also a determining factor. Our method's functionality is obviously dependent on the word choice that founders use to describe their startup. However, it is apparent that certain words are more indicative than others of the direction of a startup, i.e. "healthcare," "AI." Startups that are involved in these industries may be highly likely to include these terms in their description. It may help the accuracy of our clustering algorithm by giving more weight to predetermined buzzwords in the feature set, increasing separation between input startups which do contain and do not contain these buzzwords. Further analysis could be done in seeing which sets of predetermined buzzwords decrease error in clustering.

Moreover, there is a strong probability that the evaluation metric we've defined for clustering doesn't necessarily accurately scale the reconstruction loss correctly; given that the CS229 portion of this project is an input into the CS221 portion (where we use the clusters created in this portion and then try to classify startups into them), when testing the accuracy of this classifier on a validation set, we found that the larger the dimensionality of the subspace the corpus was reduced to, the better the accuracy was (while keeping everything else, including the classification algorithm, constant):



This challenges our evaluation metric because if classification becomes easier without changing our classifier, it's a good sign that our classes (in this case clusters) are more specifically designated with respect to certain features and are thus more well-defined, which intuitively would suggest that the clusters themselves must be more tightly defined and thus have a smaller reconstruction error.

5 Contributions

Data cleaning was done by Meeran (as he had worked with similar data before). Initial attempts at clustering and dimensionality reduction were also done by Meeran. Exploration/further analysis was brought up by Daniel.