Kickstarter Project Write-up

Qingwei Zhang, Flower Yang, Keturah Jones

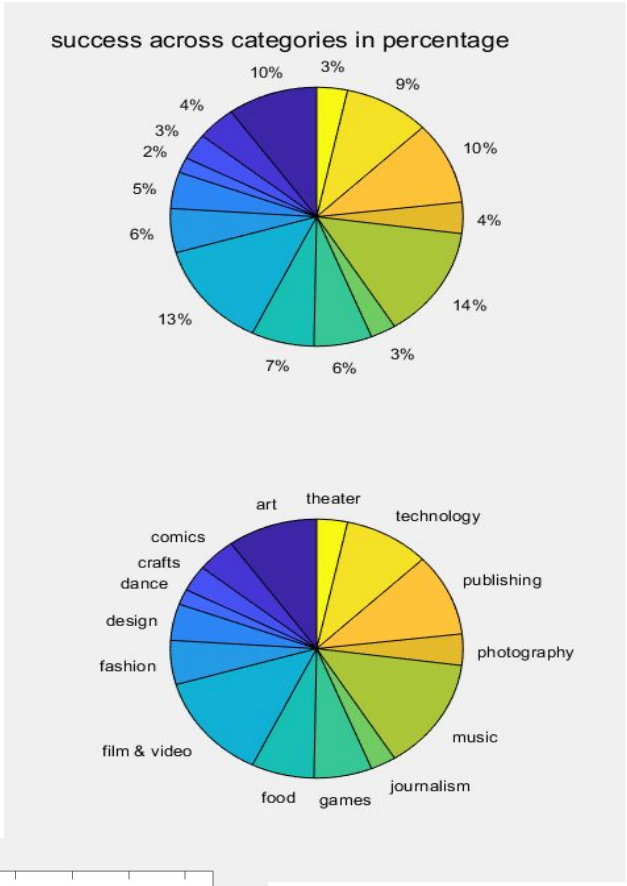
Preprocess

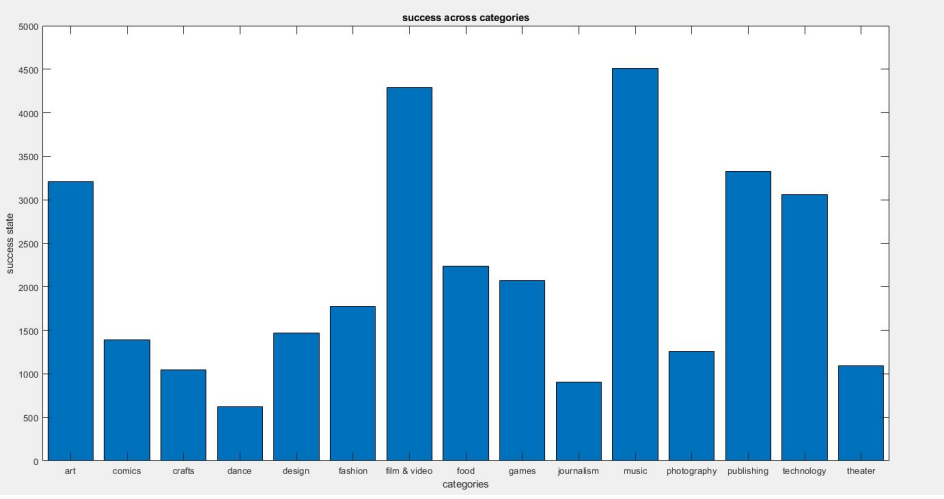
Data from the year 2018 are downloaded. There are 48 to 50 csv files of data for each month. 100 rows are selected randomly from each csv file. 15 out of 37 columns are selected from original data for analysis. The preprocessed data used for analysis are saved in 'final100.mat'.

Descriptive Statistics

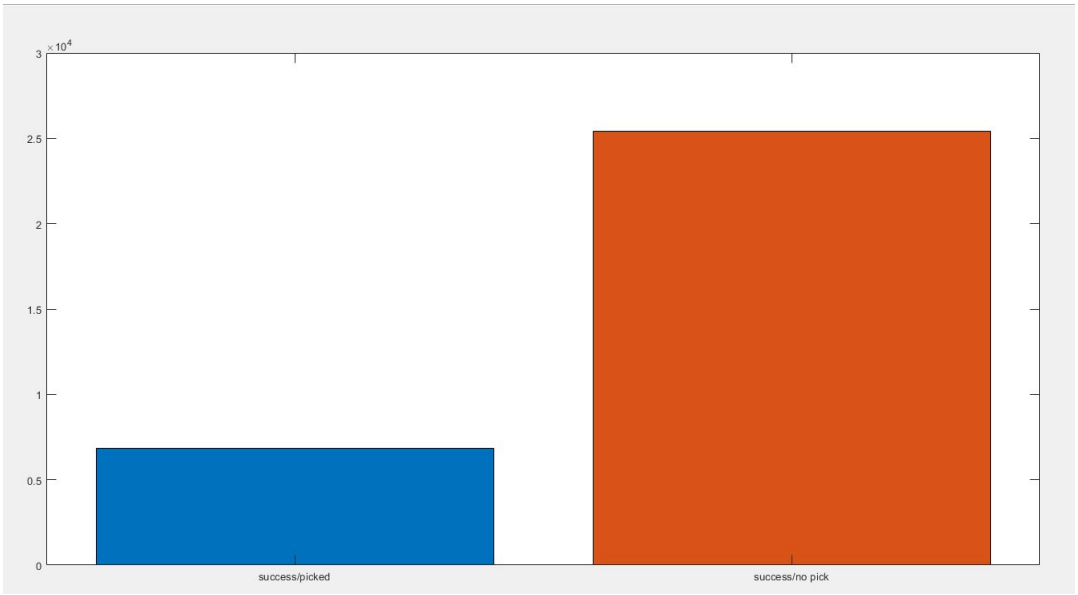
Number of successful projects across categories

The outcome of each project is shown in a variable 'state'. This analysis uses rows in final100.mat that are labeled "success", trying to look at the the number and percentage of successful projects belonging to each category. A pie graph is used to visualize the percentage of successful projects belonging to all the categories. A histogram is used to visualize the actual number of successful projects. We can see from these two graphs, that the four most successful categories are music, film & video, art and publishing with equal weight percentage of 10%. The least successful category is dance, with a weight of 2%.



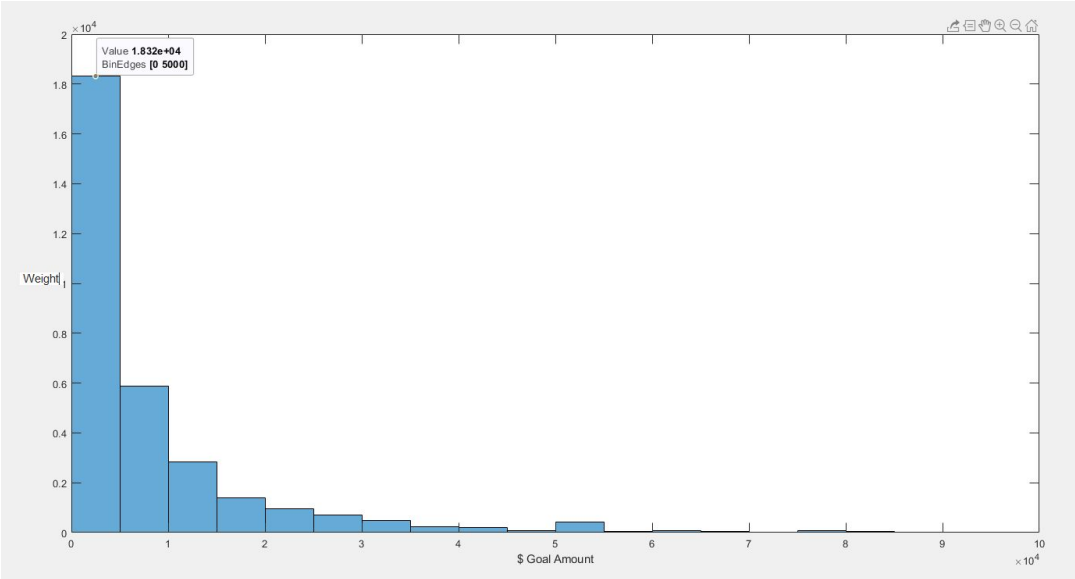


Relationship between "spotlighted"/"staff picked" and being successful



In this graph, the the successful rate with “no staff picked” is higher than the successful rate with “staff picked”. A possible explanation for this might be that people are inclined to be cautious on staff picked with suspicion that the project might be sponsored.

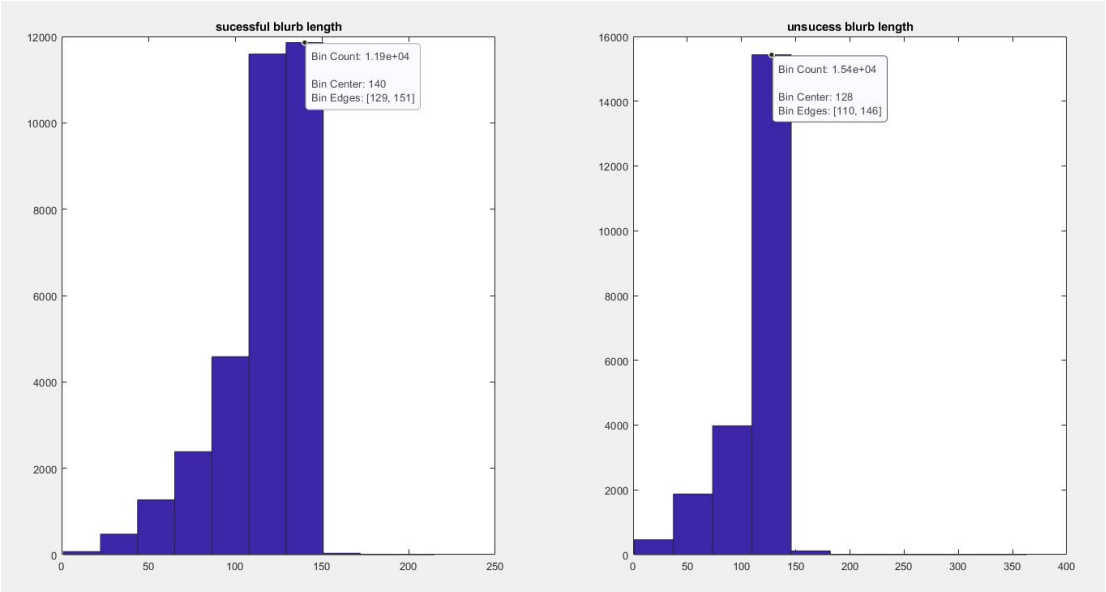
Relationship between goal $$ and being successful



This analysis aims to investigate how the amount of money people expect to earn out of the projects relate to whether they successfully achieve their goal. Out of the successful goal amount,

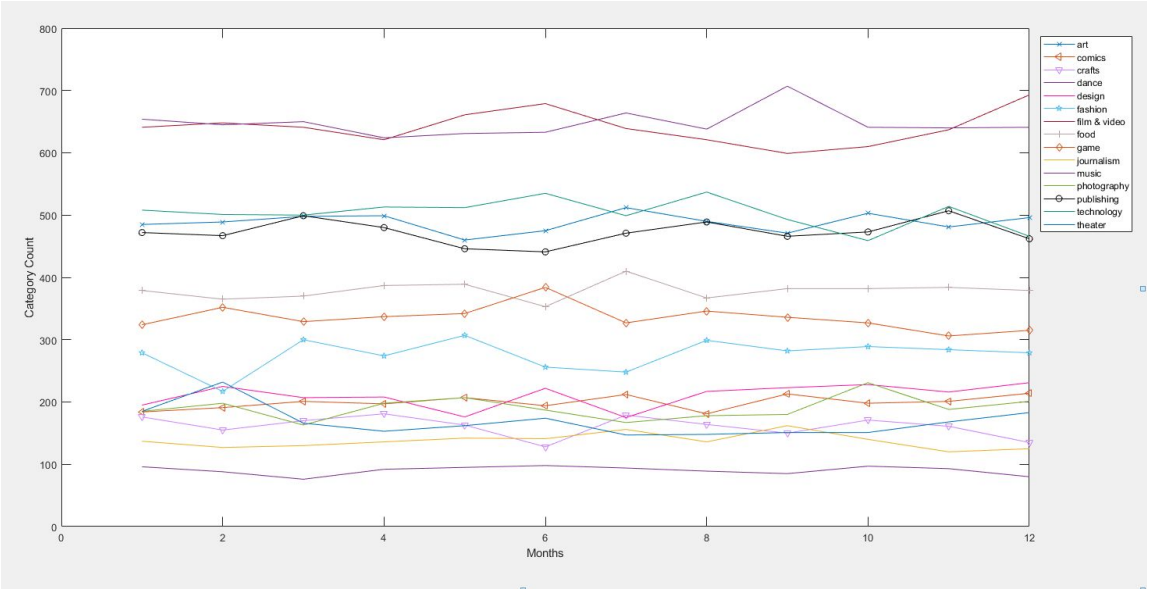
the most successful amount range lies from $0 to $5000. The success rate tend to incline with increasing amount, as shown at the tail of the graph, which makes sense because as the amount of money increases, it becomes harder to achieve it.

Relationship between the length of project descriptions/ blurbs and being successful



This analysis tries to compare blurb length with genre successful and unsuccessful state. We can see from the graph that a blurb with length 150 are generally more successful. However, when we cross compare it with the length and unsuccessful, a length with 150 is also considered more unsuccessful. This shows is that success doesn't have a strong correlation the blurb length. A possible bias in the data is, even the non-US country use English blurb and not all non-English blurb comes from non-US country. Thus, Including just US blurb could be biased but including all the country is also not representative.

The number of projects in each category for every month of every year



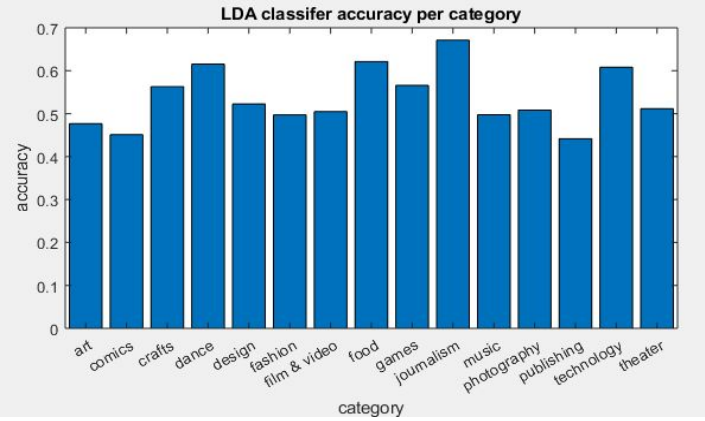
This analysis looks at how the number of projects belonging to each category vary throughout the year 2018. We can see from this graph that the number of project across project is generally consistent in different month. Here, we used data from random 100, so we divided the projects in every 4900 entries. We can see throughout the year, the number of projects in category film & video and music is kept around a constant range around 600, with dance being the lowest across 12 months.

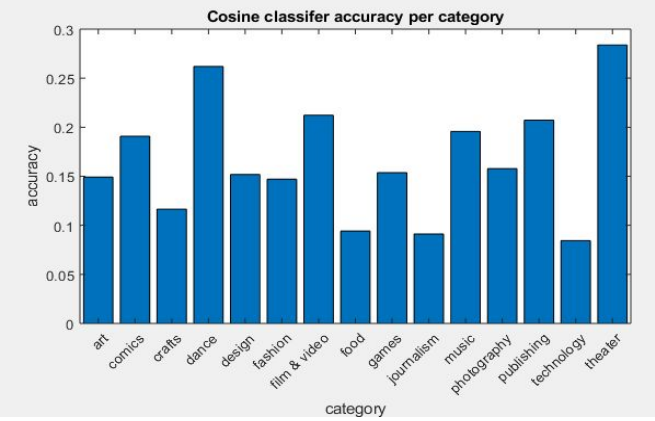
Detailed Analysis #1:

In detail analysis 1, we use average word embedding scores of the project descriptions to classify the success status of the projects. The projects are divided into three status - failed/canceled, successful, wildly successful. The first two are provided in the original dataset. Among the successful ones, those have 20 percent more amount of money pledged than planned are categorized as 'wildly successful'. This analysis aims to investigate whether we can tell (or predict) a project's success status based on its description. First, word clouds are generate to give an overview of how these descriptions of projects belonging to different status look like. From the graph we cannot really eyeball any difference between project descriptions of different status.

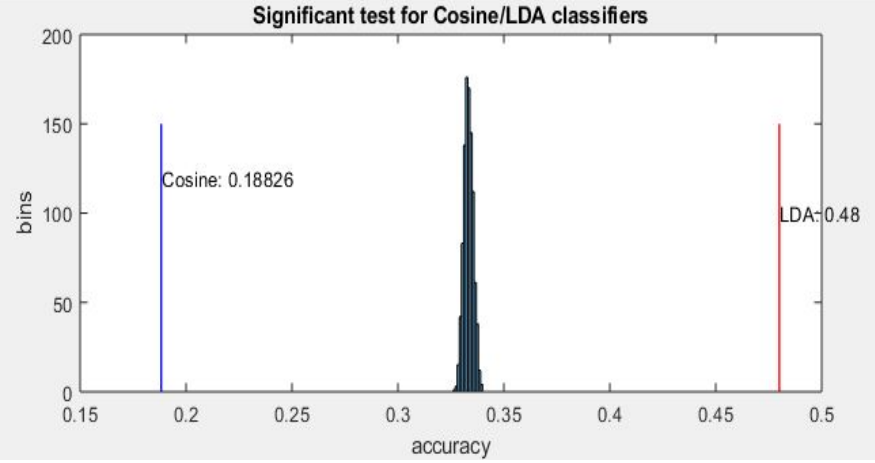
Therefore, a cosine similarity classifier is used to try to discriminate the three status. One month of data is left out for each training trial as test data. The remaining 11 months of data are used as training data. The mean percentage accuracy is calculated by dividing the number of times the classifier correctly predict the success status by the total number of times it makes predictions. The mean accuracy turns out to be 0.188 (p=1), even lower than the chance level 0.333. To improve the prediction accuracy, LDA classifier is used instead, which turns out to be much better. The mean accuracy is 0.48 (p=8.58e-15). Since this strategy seems fairly promising, the classifier is trained again with data divided according to 15 labeled categories (art, technology, etc.) the mean prediction accuracy is calculated and plotted in a bar graph. The rationale is that projects within one category should be more similar to each other in all aspects than to projects in other categories. Thus, classifier trained within categories should have higher prediction

accuracy. As seen from the bar plot, most of them have an accuracy of close to or more than 0.5, which is higher than the overall mean accuracy and definitely much higher than chance level. There are some variations in how well the classifier can predict the outcome of a project that belongs to different categories. We can see that the model can predict some categories, like journalism, technology, food, and dance, better than others. However, when the cosine classifier takes in within-category data instead of across-category data, the percentage accuracy does not seem to increase much.





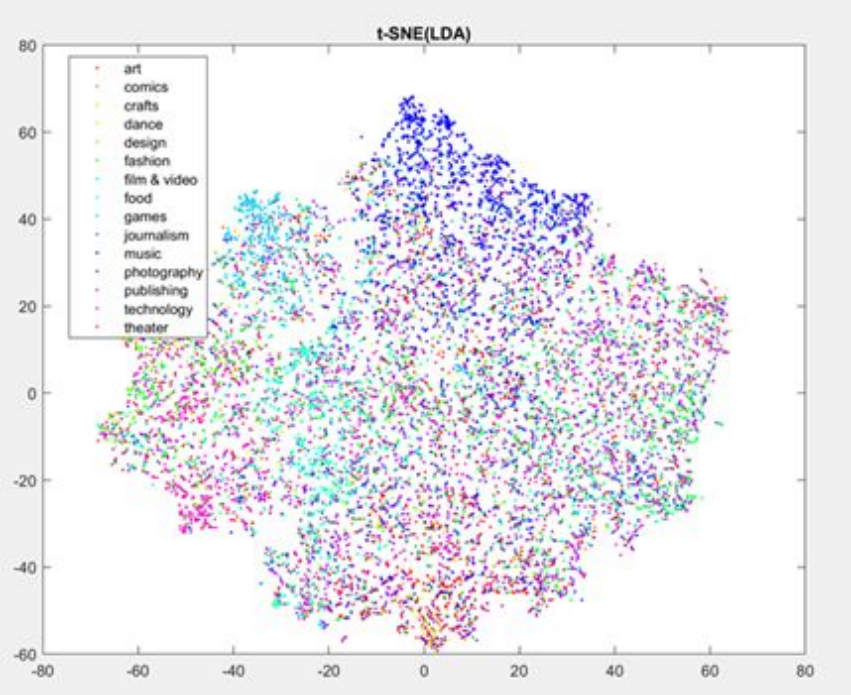
To show the significance level of the two classifiers, a null distribution is generated by randomizing all the labels and repeating the calculation 1000 times. As indicated in the graph, LDA model can reliably make predictions about the outcome of the projects based on project descriptions while the cosine classifier fails to do it.



Detail Analysis #2: How well does t-SNE, word2vec, and LDA predict categories from the project blurbs?

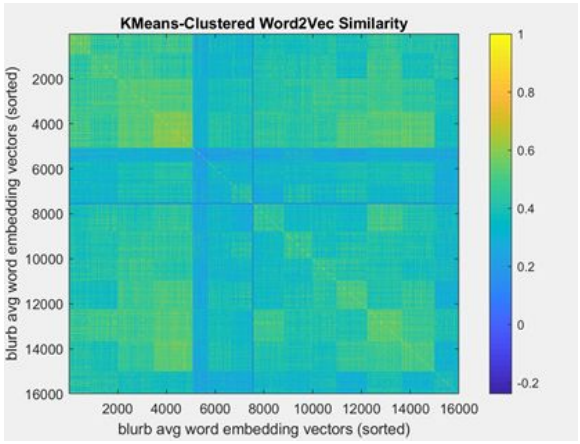
For this analysis, we used a random subset of the dataset to train LDA model and word2vec because the full dataset was too large. We then removed projects that are outside the US from the data, in order to avoid the possibility of project blurbs in non-English languages. Ultimately, we ended up with around 16,000 projects for this analysis. We trained an LDA model on the blurbs, using the same number of topics as there are categories (fifteen overall). For word2vec, we calculated the average word embedding vector for each blurb.

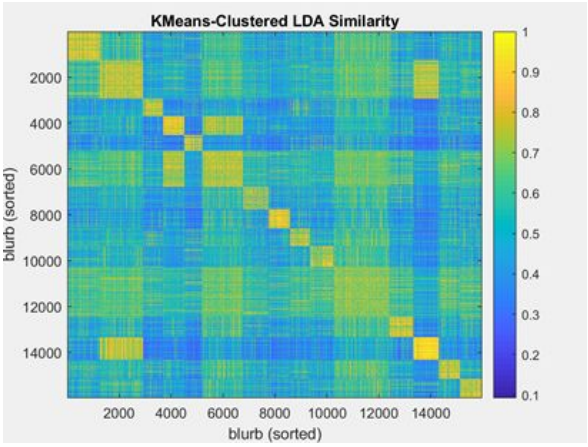
t-SNE



The t-SNE figure above, which reduces the dimensionality of the blurb x topic mixture matrix, shows how the LDA clusters resemble the true categories. It does a decent job of clustering the categories (most notably music, games, and technology), though there is some overlap that makes it difficult to see certain clusters.

KMeans Graphs

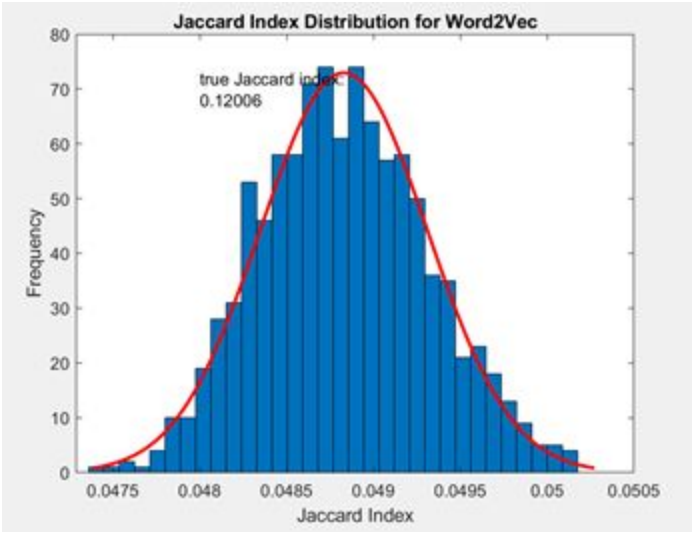


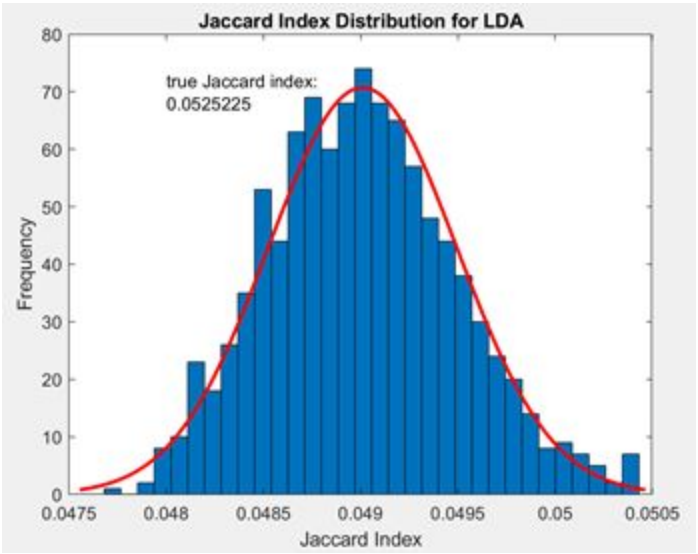


We then used kmeans clustering (k = 15 clusters to mirror the fifteen different categories) on the average word embedding vectors, resulting in the left figure, and on the topic mixtures, resulting in the right figure. There are some clusters, though they remain around 0.5-0.7 correlation and aren't as distinct as the KMeans-Clustered LDA similarity figure. It is clearer to see clusters with the LDA figure, and there is also some highly correlated overlap outside the diagonal.

Jaccard Index Distribution Histograms

To further compare how well word2vec predicted the categories versus LDA, we calculated the Jaccard index, which in this case represents the overlap between the list of projects with predicted category assignments (the fifteen clusters received from kmeans) and the list of projects with true category assignments. We calculated the Jaccard index for one cluster at a time, matching it to the true category that has the maximal Jaccard index. After that, we averaged together those fifteen Jaccard index values to receive the true overall Jaccard index. Then to produce a distribution like in the two figures below, we randomized the list of true category assignments and calculated the overall Jaccard index using the process described. We did this one thousand times to produce a normal distribution histogram for both LDA and word2vec.





These figures show very similar distributions due it to being random. However the true Jaccard index is higher for word2vec than LDA, which is surprising. Most likely, this is due to the difficulty we had in calculating the true Jaccard index for LDA. When running kmeans on the LDA data, the projects' predicted categories list was less than the projects' true categories list. So in order to calculate the overall Jaccard index, we had to shorten the true categories list, which most likely messed up the indices of the list and which categories were lined up with which projects. Possibly non-English languages even in the U.S. interfered with this result. If we were able to correct it, then I would hypothesize that the true Jaccard index for the LDA would be much higher.

Conclusion

Overall, we believe this is a comprehensive analysis of the Kickstarter data. The randomization of 100 Kickstarter projects chosen in each Excel sheet for each month in 2018 allowed for us to look at a representative dataset without having to analyze all of the 2018 data. However, there are some potential biases that could skew the data. An example of this is the non-English projects from the U.S., which possibly produced an erroneous result for the true Jaccard index for the LDA. Another example of a potential bias is that we only looked at a year. Perhaps analyzing over other years would provide different results than the ones we obtained for 2018. If we were to improve upon our analyses, analyzing Kickstarter data from other years would be helpful in seeing if the result we found in the 2018 data would be generally consistent across multiple years. Another idea for improvement would be to find a way to remove the non-English projects from the U.S. so we can properly obtain the true Jaccard index of the LDA-predicted categories and the true categories.