Kickstarter

RWHD Final Project

Flower Yang, Qingwei Zhang, Keturah Jones
04/30/2019

Kickstarter Datasets

We have a scraper robot which crawls all Kickstarter projects and collects data in CSV and JSON formats. From March 2016 we run this data crawl once a month. Datasets are available from the following scrape dates:

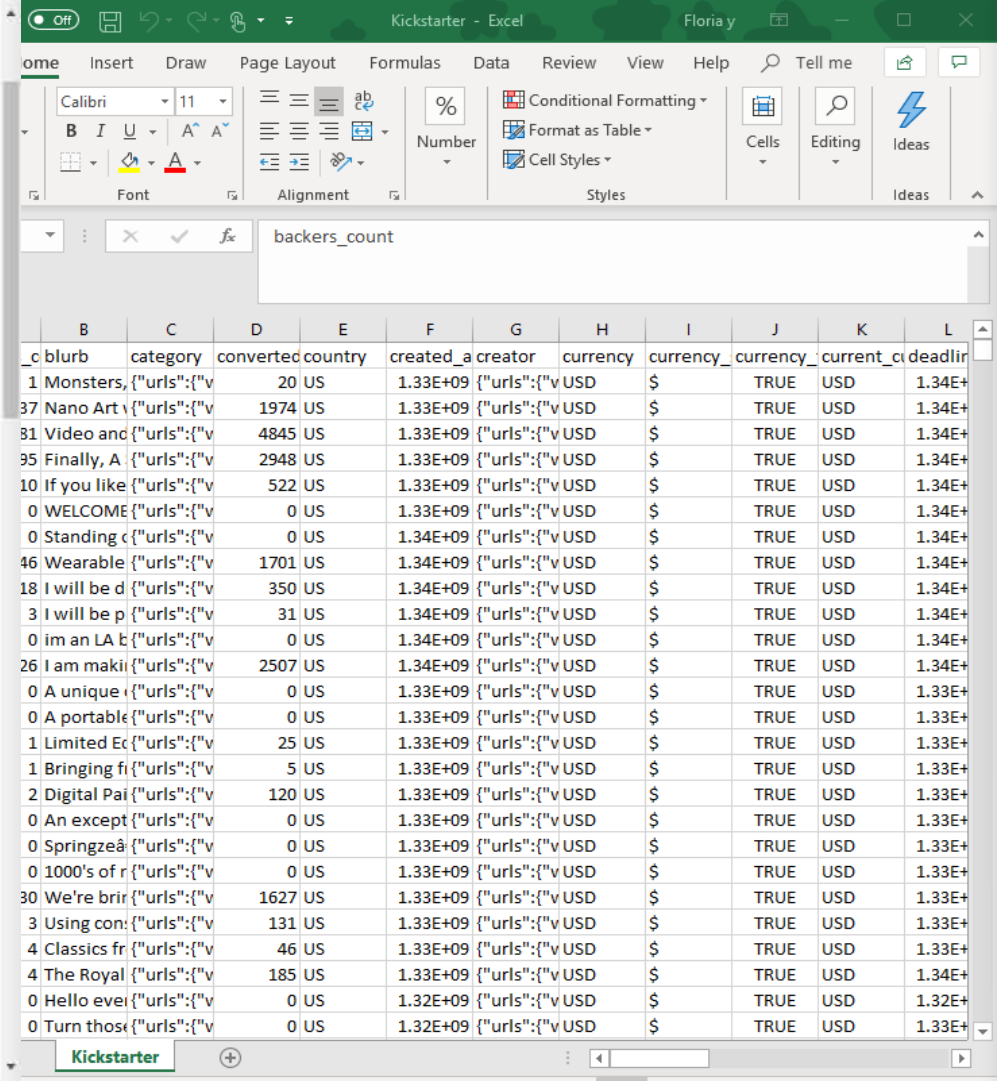
2019

- 2019-03-14 [JSON] – [CSV]
- 2019-02-14 [JSON] – [CSV]
- 2019-01-17 [JSON] – [CSV]

2018

- 2018-12-13 [JSON] – [CSV]
- 2018-11-15 [JSON] – [CSV]
- 2018-10-18 [JSON] – [CSV]
- 2018-09-13 [JSON] – [CSV]
- 2018-08-16 [JSON] – [CSV]
- 2018-07-12 [JSON] – [CSV]
- 2018-06-14 [JSON] – [CSV]
- 2018-05-17 [JSON] – [CSV]
- 2018-04-12 [JSON] – [CSV]
- 2018-03-15 [JSON] – [CSV]
- 2018-02-15 [JSON] – [CSV]
- 2018-01-12 [JSON] – [CSV]

2017



Data Description

Data from 2018 is used in this project. Each month contains approximately 48 to 50 csv files.

Examples of Variables:

1. ID : project ID, consist of numbers
2. Name: Project name
3. Blurb: Project Description
4. Goal: amount of the money goal to succeed the project
5. USD_Pledged : Actual amount in USD reached from the goal
6. State: success, failed, canceled, or live
7. Country & location
8. Deadline: Time of the deadline
9. Launching time
10. Backers: Number of supporters
11. Category: 15 categories of project

Data Pre-processing

37 columns (variables) in original dataset, of which we keep for analysis: backers count; blurb; category; country; deadline; goal; id; launched at; location; name; spotlight; staff pick; state; state changed at; usd pledged

Adding numeric labels (1-15) to represent the 15 categories of projects: art; comics; crafts; dance; design; fashion; film & video; food; games; journalism; music; photography; publishing; technology; theater

Descriptive Statistics

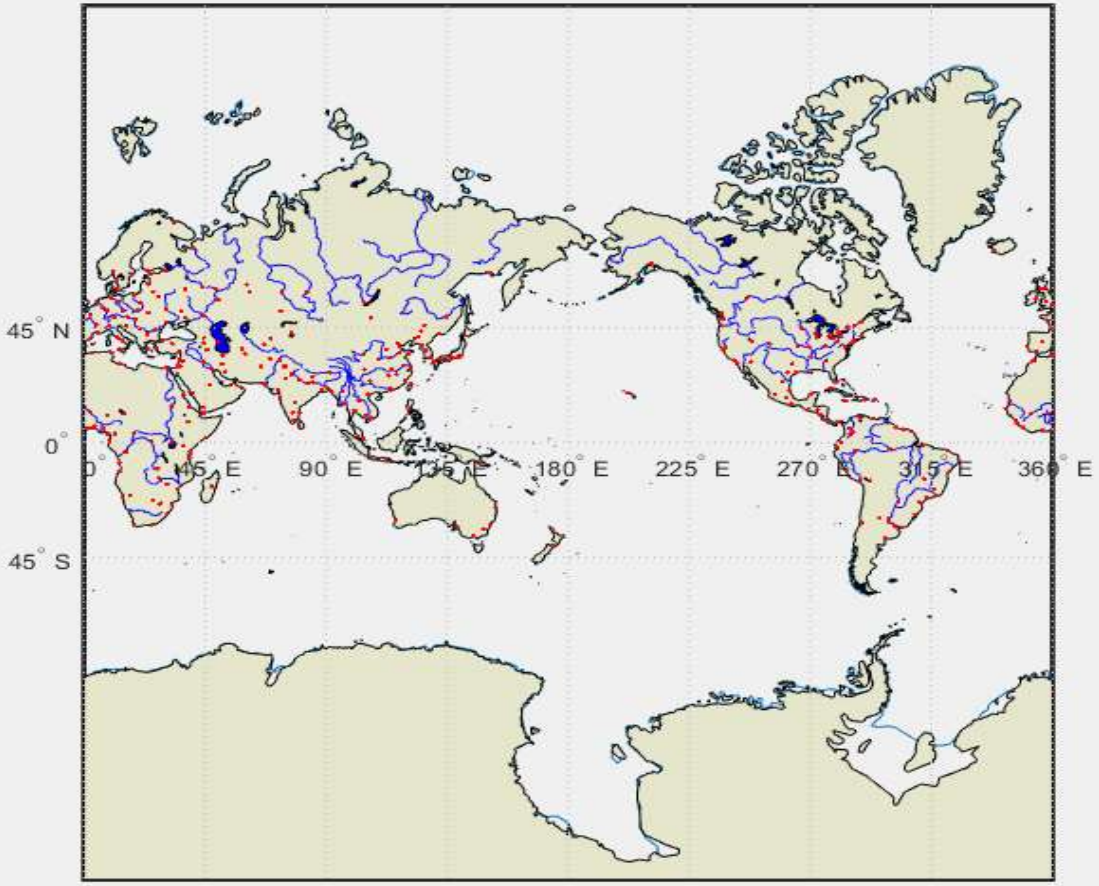
1)Geo analysis -- The whole world

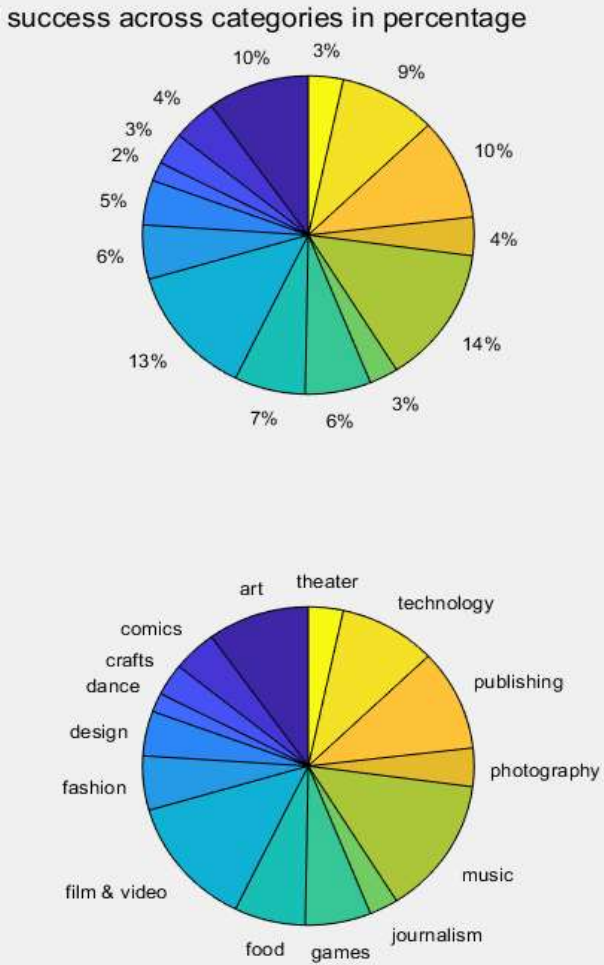
2)Histogram of success across categories(defining success using general (‘state’)

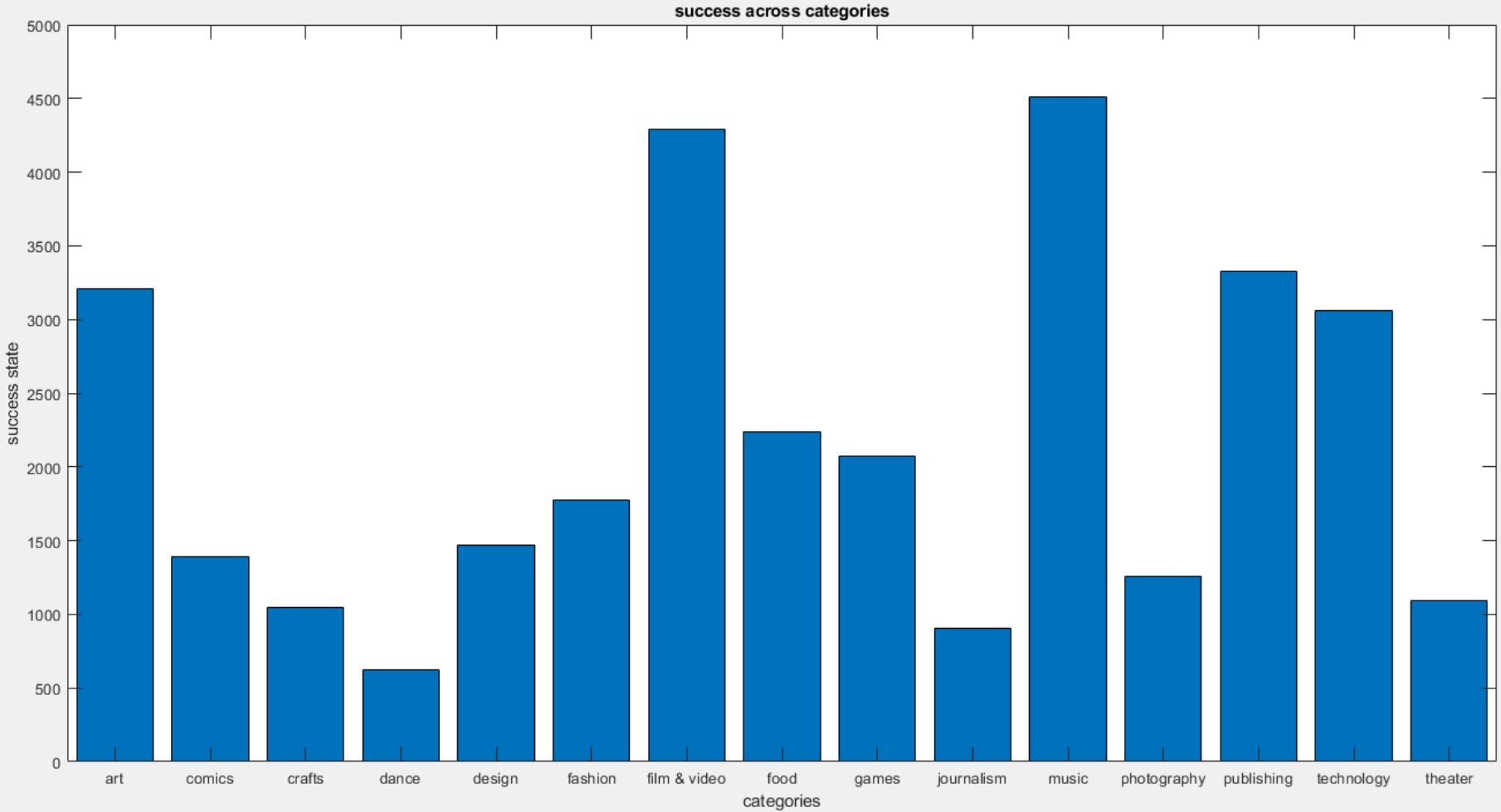
3)relationship between "spotlighted"/"staff picked" and being successful

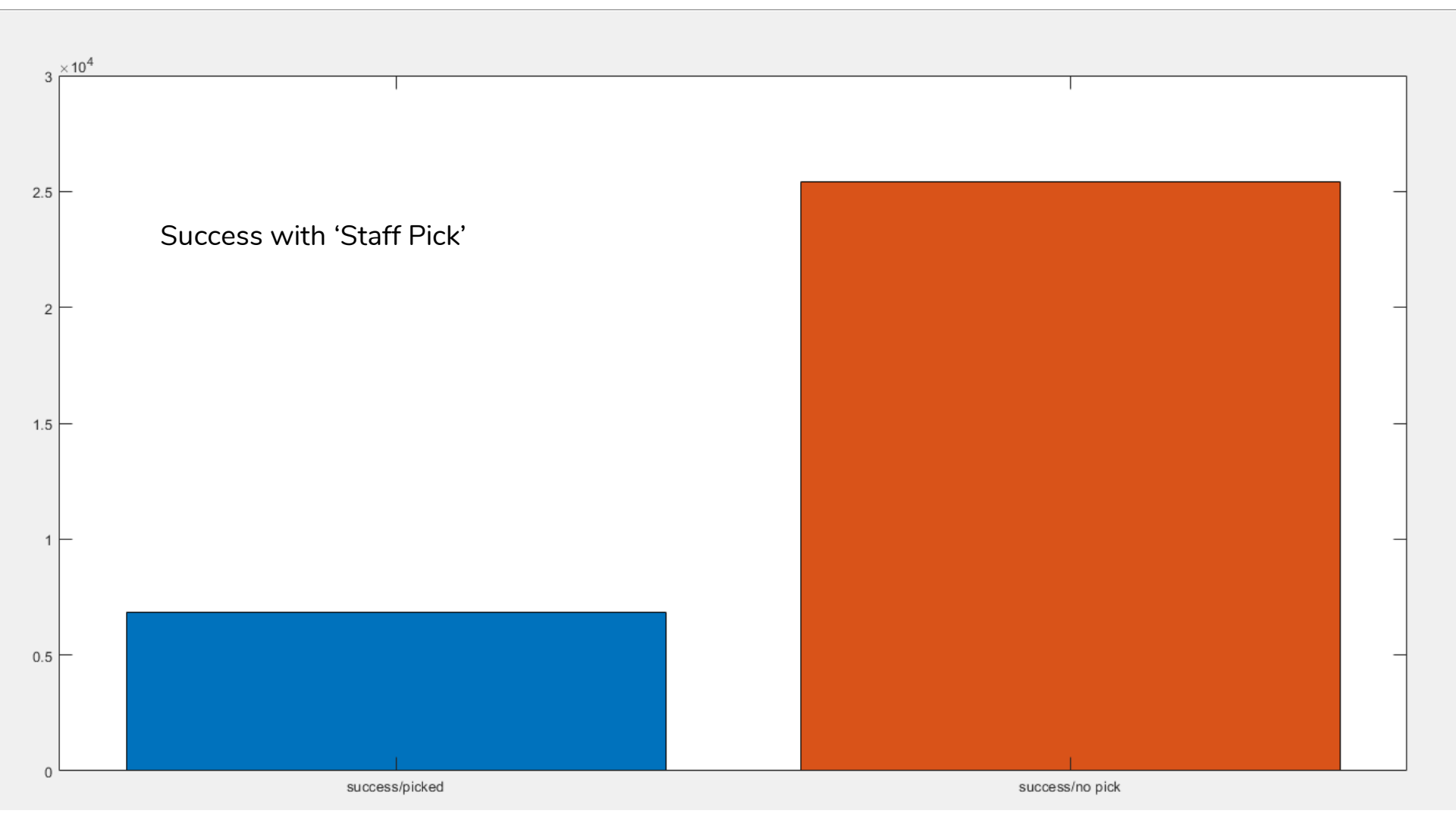
4)relationship between goal $$ and being successful

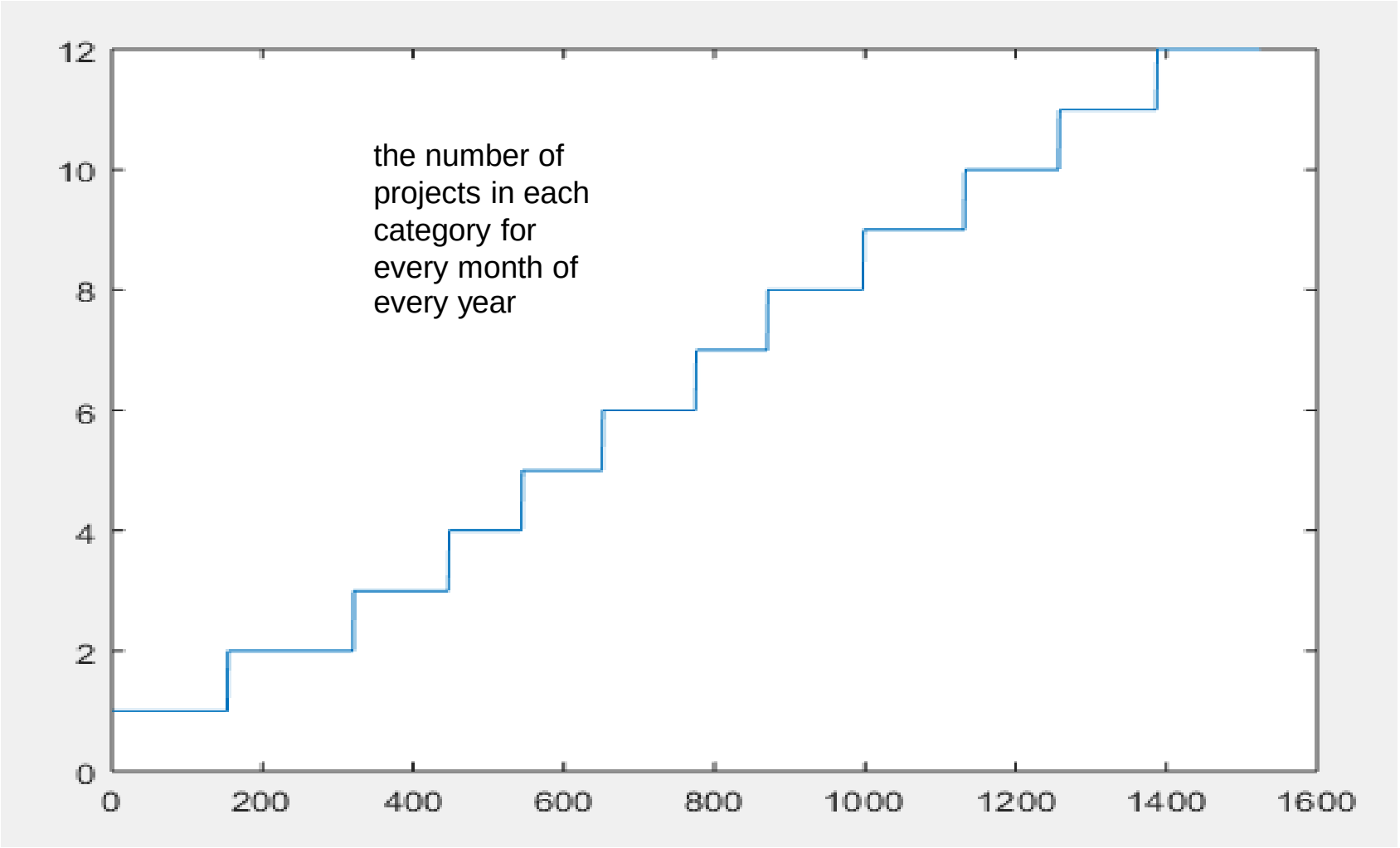
5) the number of projects in each category for every month of every year (a line plot with time on the x-axis).


45° N
0°
45° S
0°
45° E
90° E
135° E
180° E
225° E
270° E
315° E
360° E


success across categories in percentage
10%
3%
9%
4%
10%
3%
4%
2%
5%
6%
14%
13%
3%
7%
6%
art
theater
technology
comics
crafts
publishing
dance
design
photography
fashion
music
film & video
journalism
food
games


success across categories
success state
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
art
comics
crafts
dance
design
fashion
film & video
food
games
journalism
music
photography
publishing
technology
theater
categories


Success with ‘Staff Pick’
×10^4
3
2.5
2
1.5
1
0.5
0
success/picked
success/no pick


the number of projects in each category for every month of every year
12
10
8
6
4
2
0
0
200
400
600
800
1000
1200
1400
1600

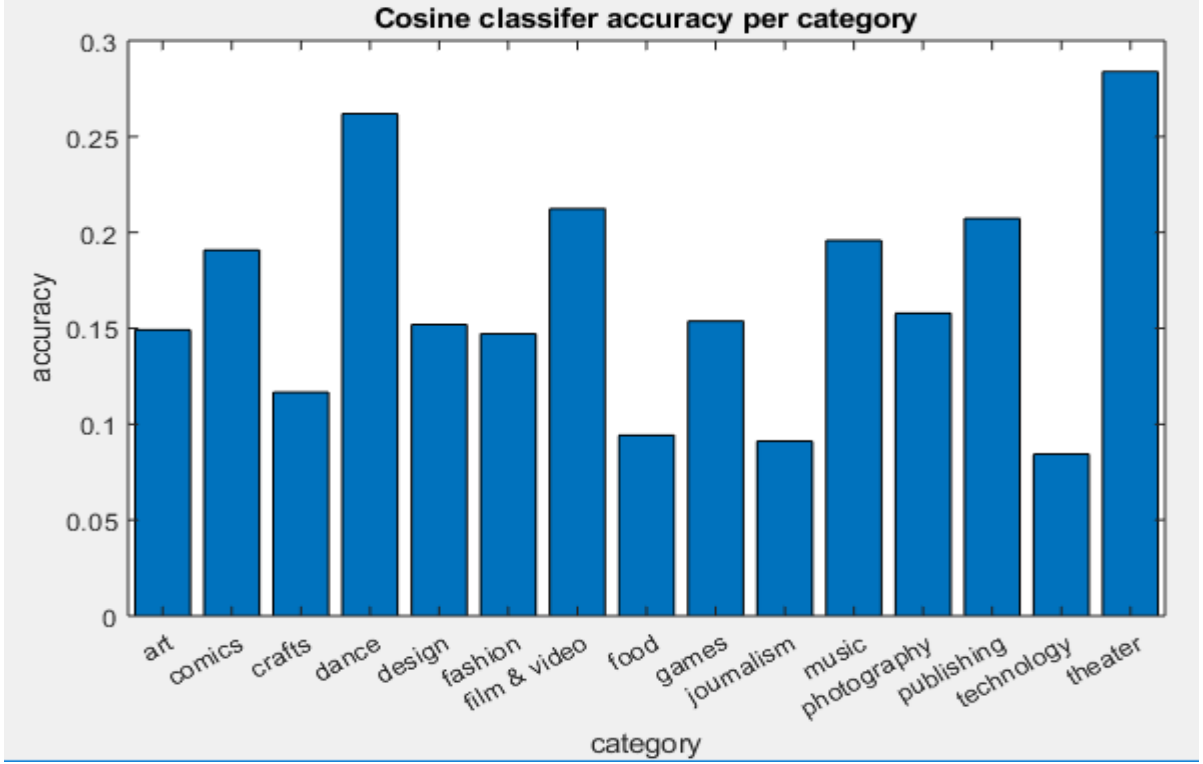
Detailed Analysis 1:
Classification of project success using project description

1) failed/cancelled projects

2) successful projects

3) wildly successful projects: exceed expected outcome by at least 20%

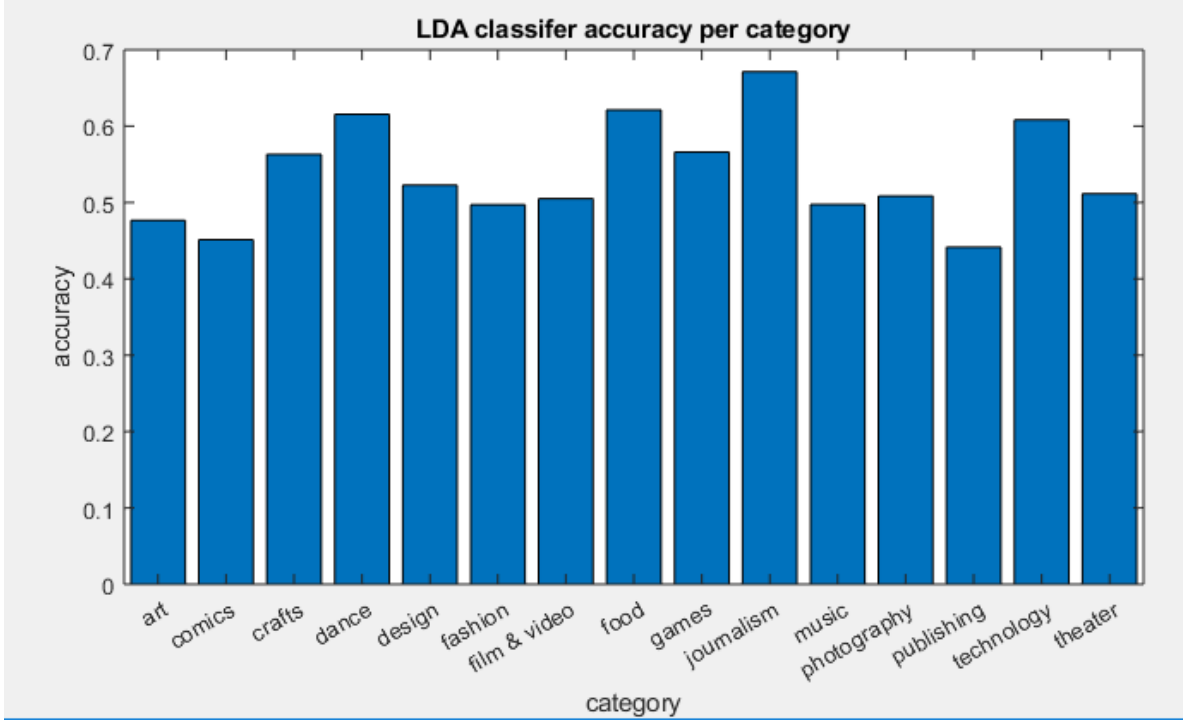
Classification using Cosine Classifier: generate the average word embedding vector for each class. Use 11 months of data as input (training data) to train a classifier, which is then used to predict the class of the remaining month.



Overall mean accuracy across categories:

0.1883

Classification using LDA classifier: use 11 months of data as training data and the last 1 month as test data.



Overall mean accuracy across categories:

0.48

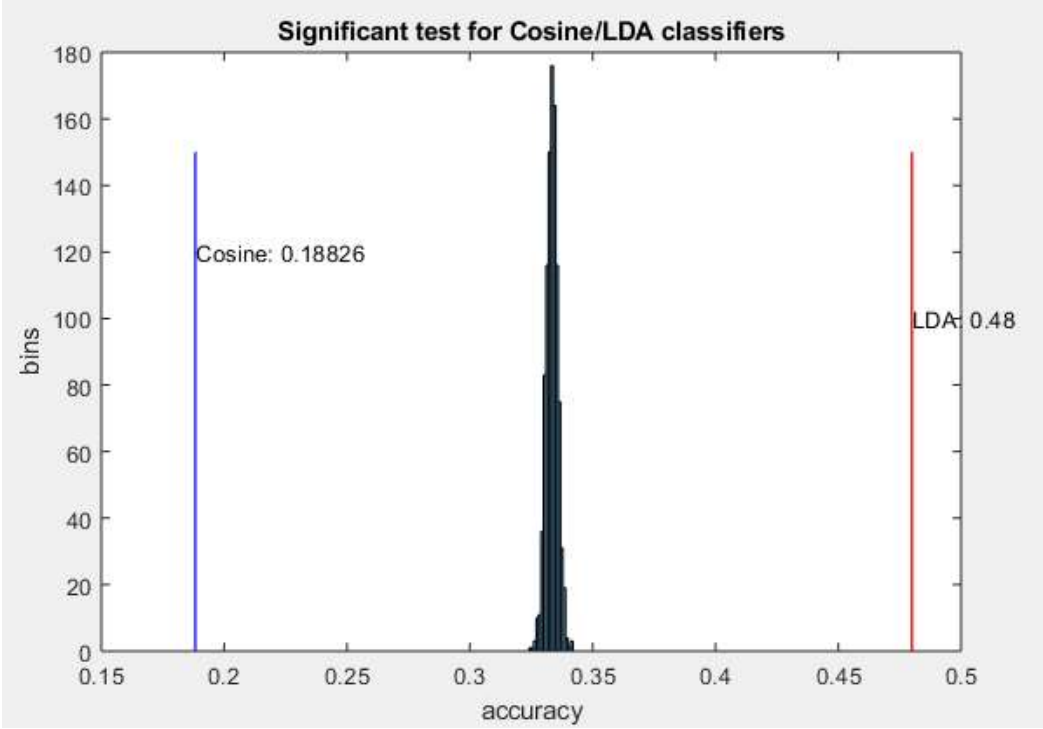
Significance Test

Cosine classifier:

- Mean accuracy: 0.188
- P-value: 1

LDA classifier:

- Mean accuracy: 0.48
- P-value: 8.5822e-15

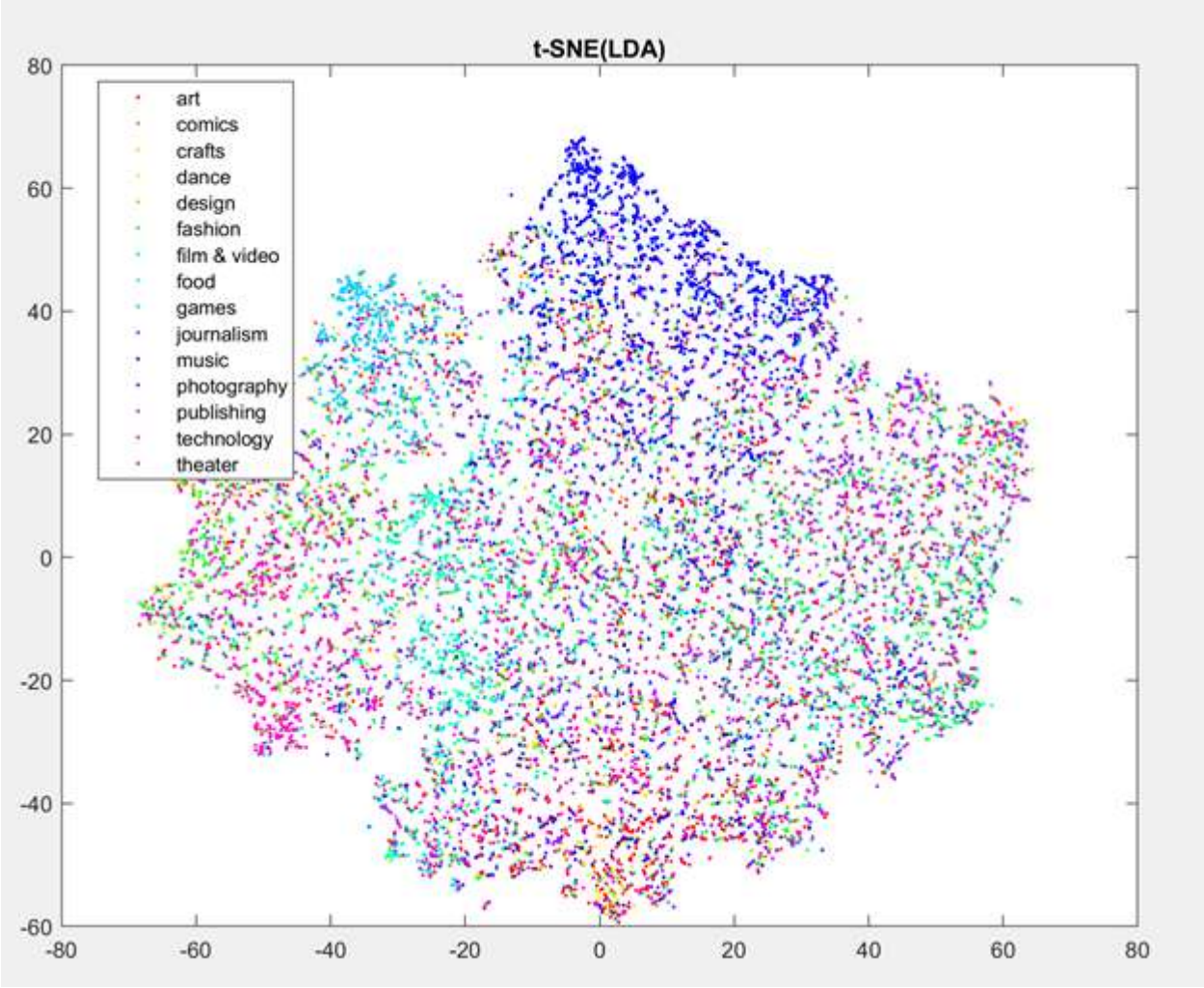


Detailed Analysis 2:
How well can t-SNE, LDA and word2vec recover the categories from the blurbs?

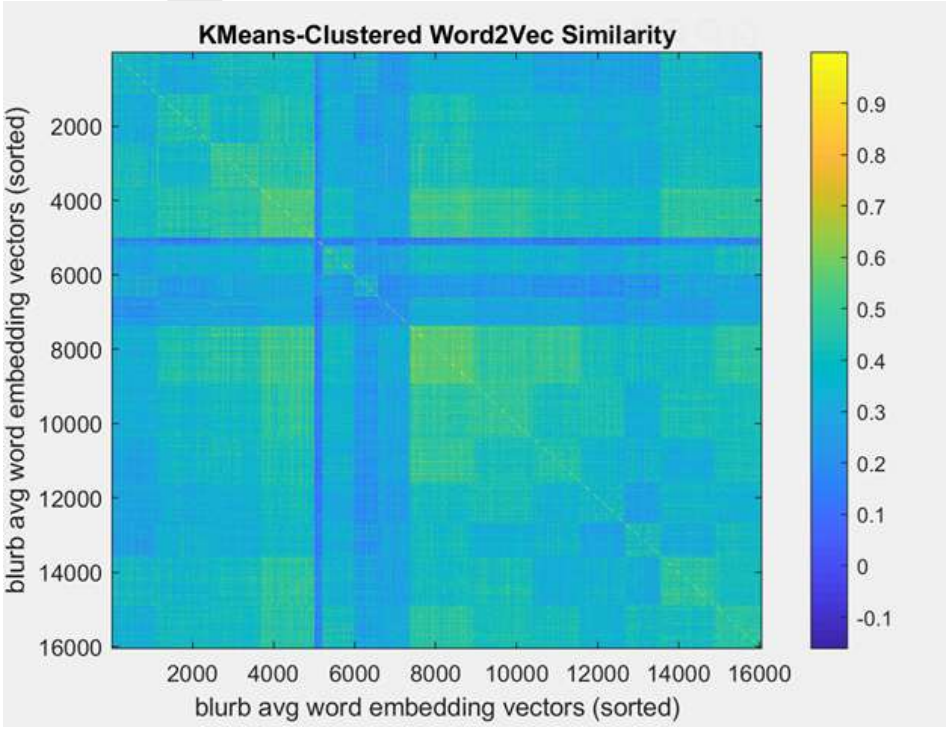
- Trained an LDA model on the blurbs, using the same number of topics as there are categories (fifteen overall).
- For word2vec, calculated the average word embedding vector for each blurb
- Used a subset of the dataset to train LDA model and word2vec because full dataset was too large

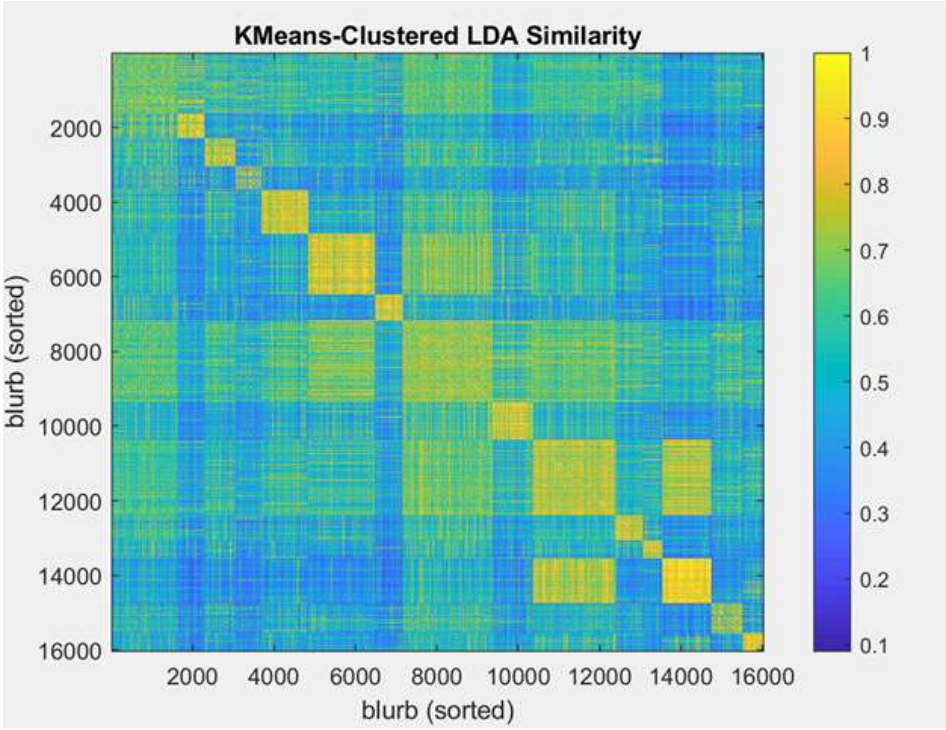
t-SNE

Produced t-SNE to reduce dimensionality of the blurb x topic mixture matrix of LDA



KMeans clustering

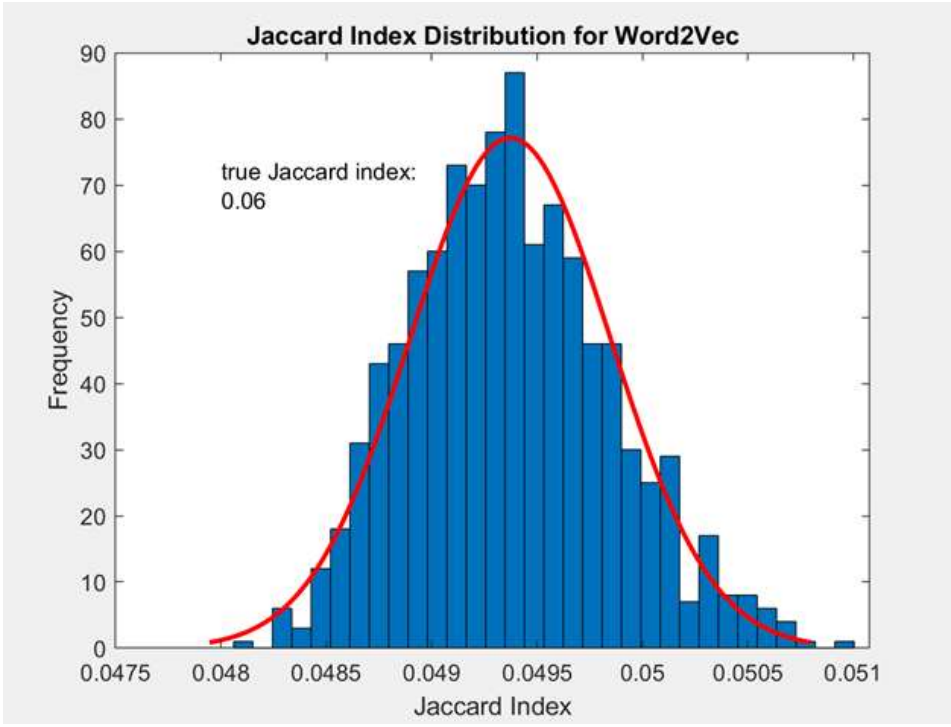


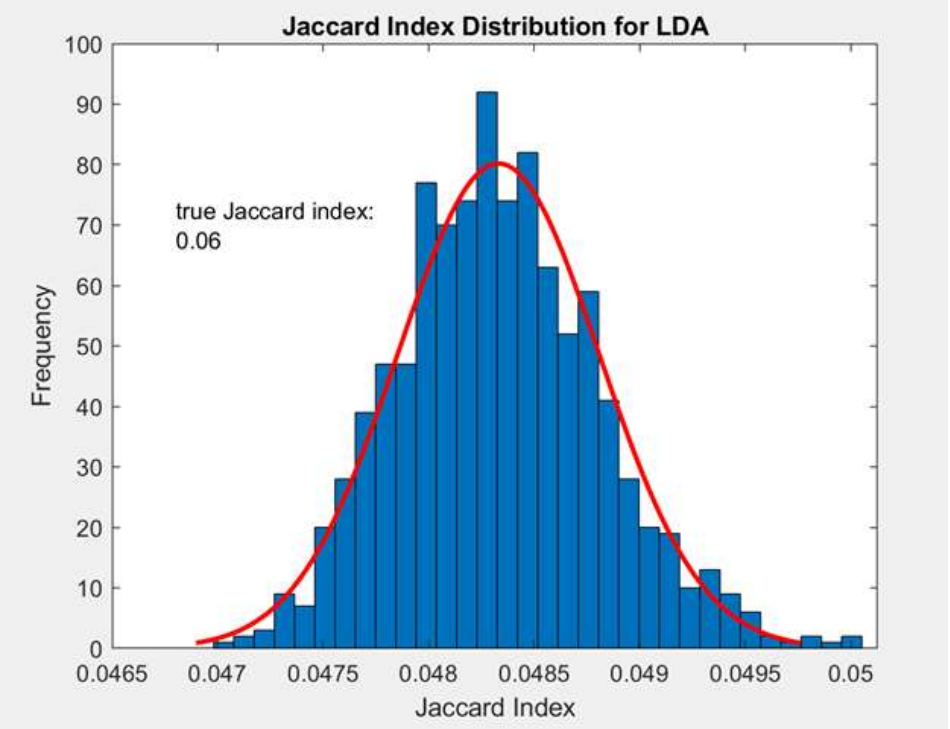


Jaccard Index

- Calculated true overlap of actual categories and predicted categories (from kmeans) for both LDA and word2vec
- Randomized actual categories and calculated overlap of that result and predicted categories 1000 times.

Jaccard Index Distribution Histograms





Solo Analysis: more classifiers

Use the time difference between deadline and launching time to classify success status

1) Calculate the mean time diff. for the three classes;
2) Use 11 months of data to train the LDA classifier;
3) Use the remaining one month of data as test data;
4) Repeat;
5) Calculate accuracy.

Accuracy = 0.4119

P-value = 6.8294e-12

Solo Analysis: more classifiers

Use the number of backers to classify success status

1) Calculate the mean number of backers for the three classes;
2) Use 11 months of data to train the LDA classifier;
3) Use the remaining one month of data as test data;
4) Repeat;
5) Calculate accuracy.

Accuracy = 0.4247

P-value = 6.9781e-12

Solo Analysis: Significance