Semantics of Kickstarter Projects Descriptions: Initial Results

Mariya Petukhova

June 2017

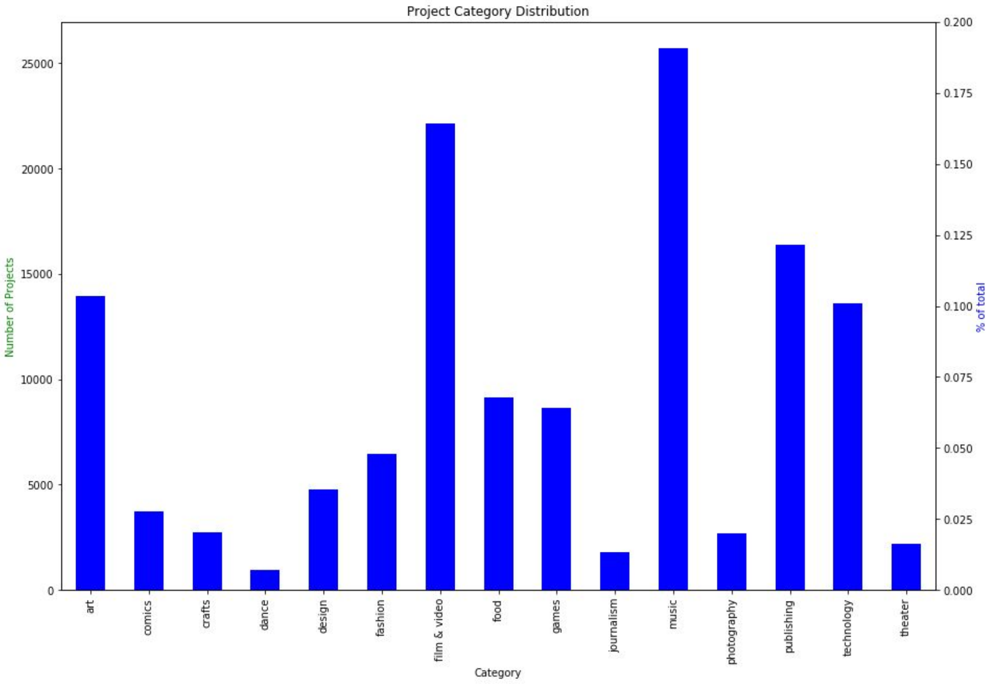
Objective

Conduct NLP analysis of project descriptions on Kickstarter.com to discover an underlying semantic structure

Data

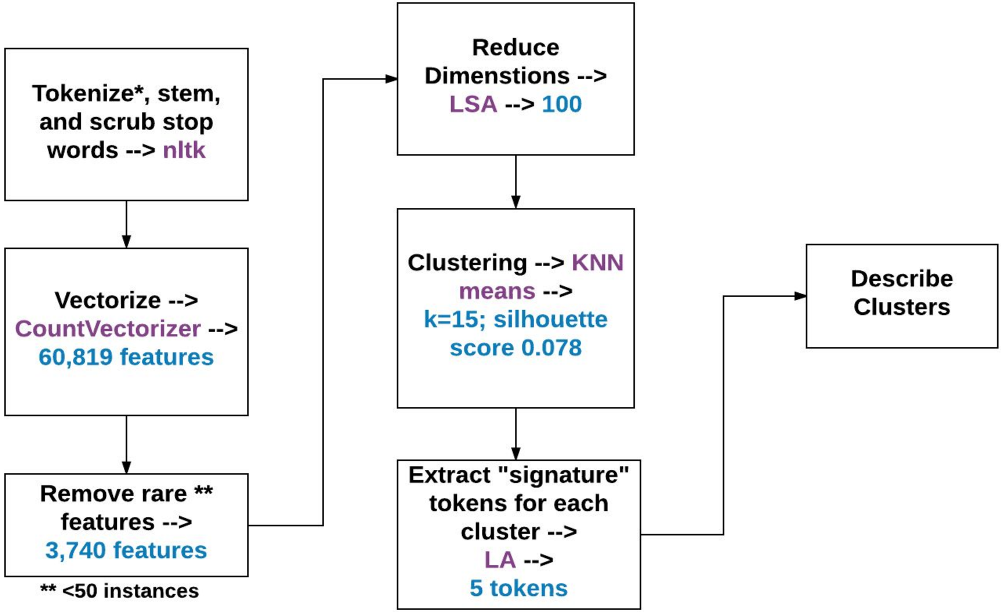
Most recently scraped Kickstarter project data in JSON format available at https://webrobots.io/kickstarter-datasets/ *

- 134,780 projects originated in US and plan for US market
  - Sampled down to 20,000 for
- 22 features, including text blurb with projec description
- Almost 1GB data



*** Actual file name: https://s3.amazonaws.com/weruns/forfun/Kickstarter starter_2017-05-15T22_21_11_300Z.json.gz**

Methodology of Project Blurb Analysis



Tokenize*, stem, and scrub stop words --> nltk

Vectorize --> CountVectorizer --> 60,819 features

Remove rare ** features --> 3,740 features

** <50 instances

Reduce Dimenstions --> LSA --> 100

Clustering --> KNN means --> k=15; silhouette score 0.078

Extract "signature" tokens for each cluster --> LA --> 5 tokens

Describe Clusters

* Unigrams are used per research in "**Using Language to Predict Kickstarter Success**" by K. Sawhney, C. Tran. Stanford University. Department of Computer Science.

Traditional Obscure Cultural Reference

“William Shakespeare's vocabulary has been estimated by the experts at twelve thousand* words. The vocabulary of the Mumbo Jumbo tribe amounts to three hundred words.

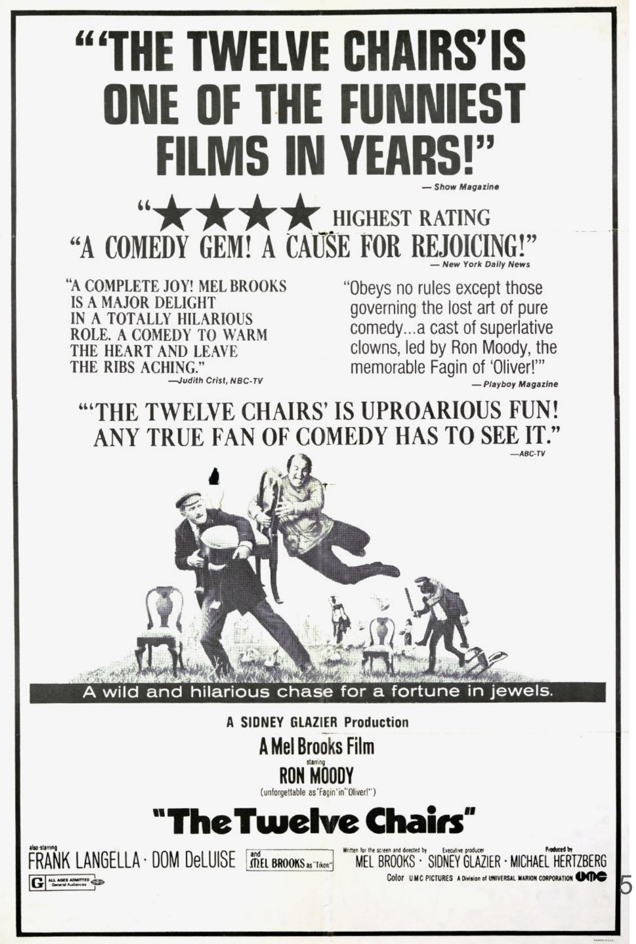
Ellochka Schukin managed easily and fluently on thirty.

Here are the words, phrases and interjections which she fastidiously picked from the great, rich and expressive Russian language:

1. You're being vulgar.
2. Ho-ho!

[…]”

Ilf and Petrov
“The Twelve Chairs



““THE TWELVE CHAIRS' IS ONE OF THE FUNNIEST FILMS IN YEARS!”
— Show Magazine

“★★★★ HIGHEST RATING
“A COMEDY GEM! A CAUSE FOR REJOICING!”
— New York Daily News

“A COMPLETE JOY! MEL BROOKS IS A MAJOR DELIGHT IN A TOTALLY HILARIOUS ROLE. A COMEDY TO WARM THE HEART AND LEAVE THE RIBS ACHING.”
—Judith Crist, NBC-TV

“Obeys no rules except those governing the lost art of pure comedy...a cast of superlative clowns, led by Ron Moody, the memorable Fagin of 'Oliver!'”
— Playboy Magazine

“'THE TWELVE CHAIRS' IS UPROARIOUS FUN! ANY TRUE FAN OF COMEDY HAS TO SEE IT.”
—ABC-TV

A wild and hilarious chase for a fortune in jewels.

A SIDNEY GLAZIER Production
A Mel Brooks Film
starring
RON MOODY
(unforgettable as 'Fagin' in "Oliver!")

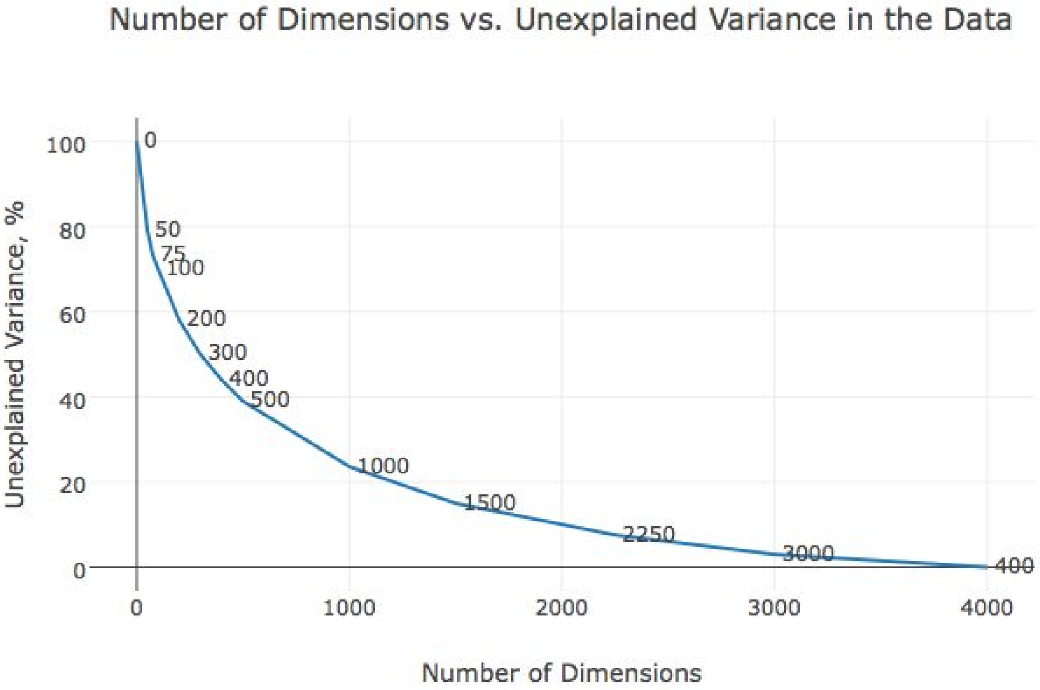
"The Twelve Chairs"

also starring
FRANK LANGELLA · DOM DeLUISE and MEL BROOKS as "Tikon"
Written for the screen and directed by MEL BROOKS · Executive producer SIDNEY GLAZIER · Produced by MICHAEL HERTZBERG
G ALL AGES ADMITTED General Audiences
Color UMC PICTURES A Division of UNIVERSAL MARION CORPORATION UMC

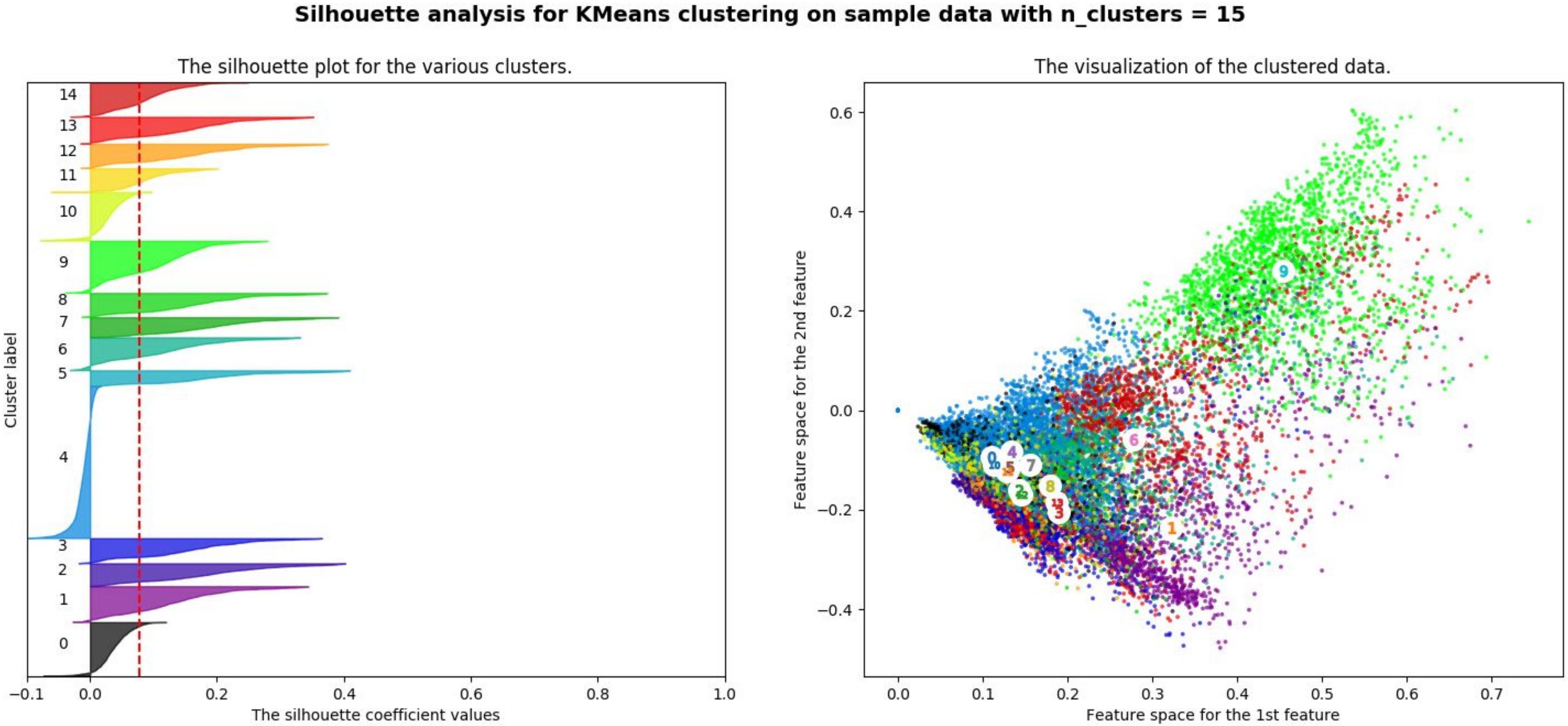
* This makes 60,819 tokens obtained from vectorization extremely high

Dimension Reduction

- 1,500 dimensions explain 85% of variation in the data
  - Curse of Dimensionality in action - very low silhouette scores
  -
  -
- Proceeded with 100 dimensions and 30% explainability

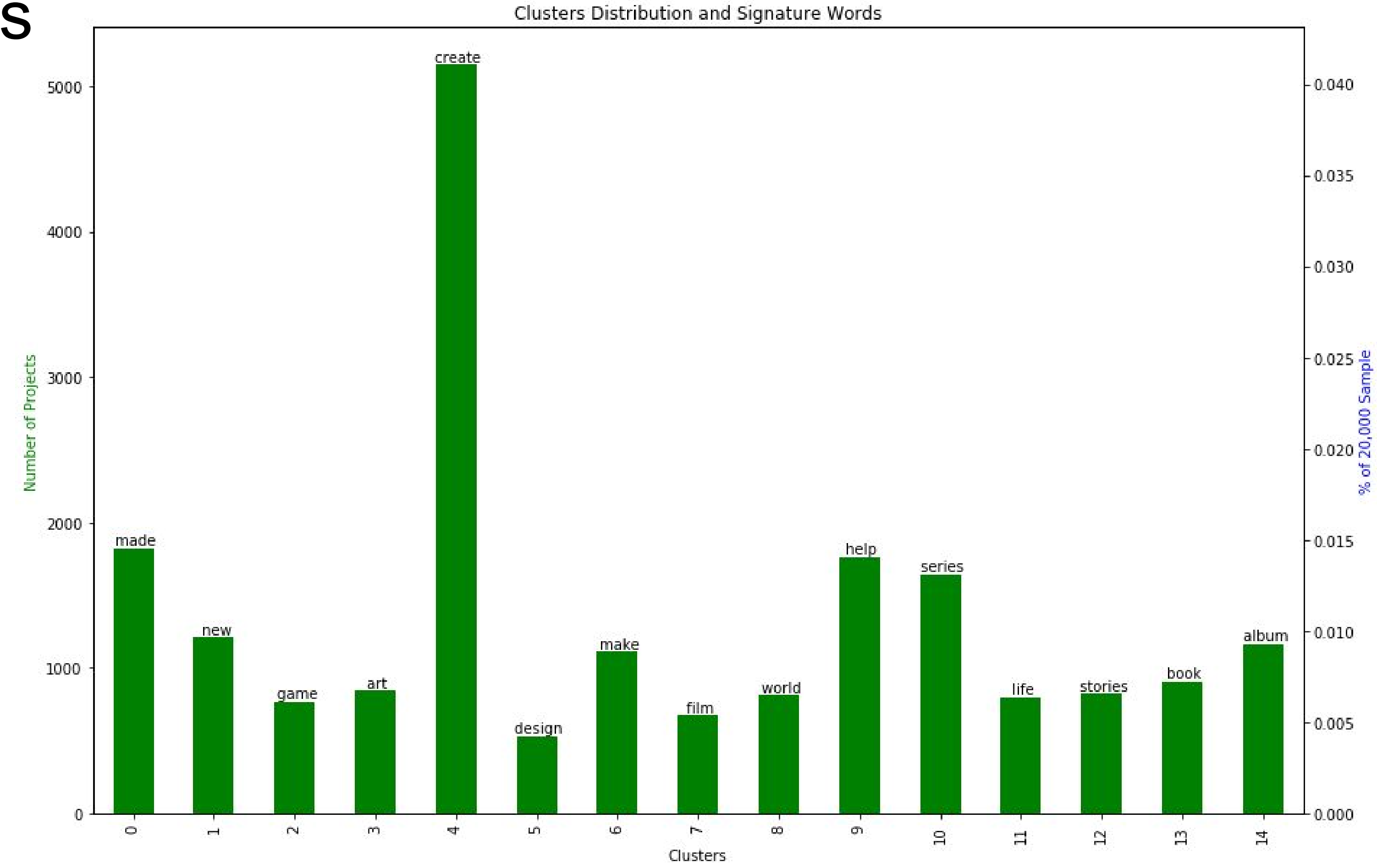


Clustering: KNN Means with k=15

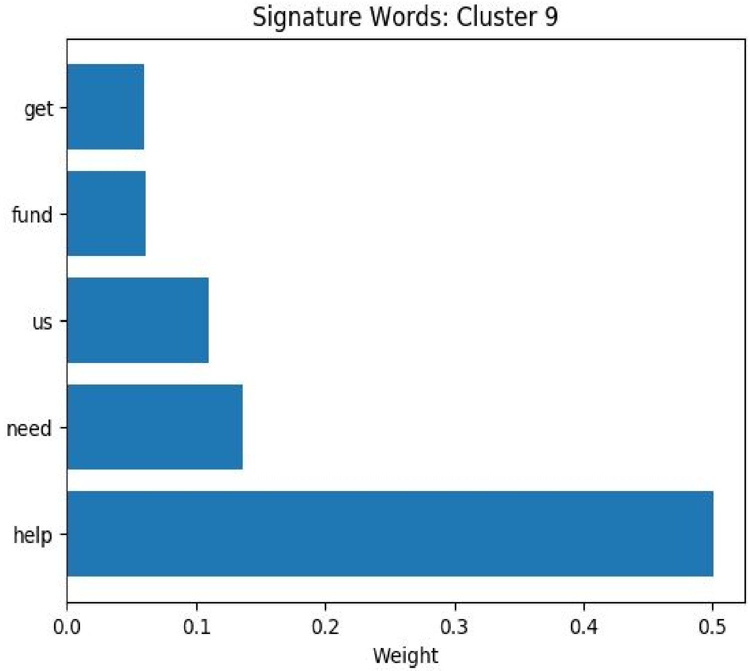


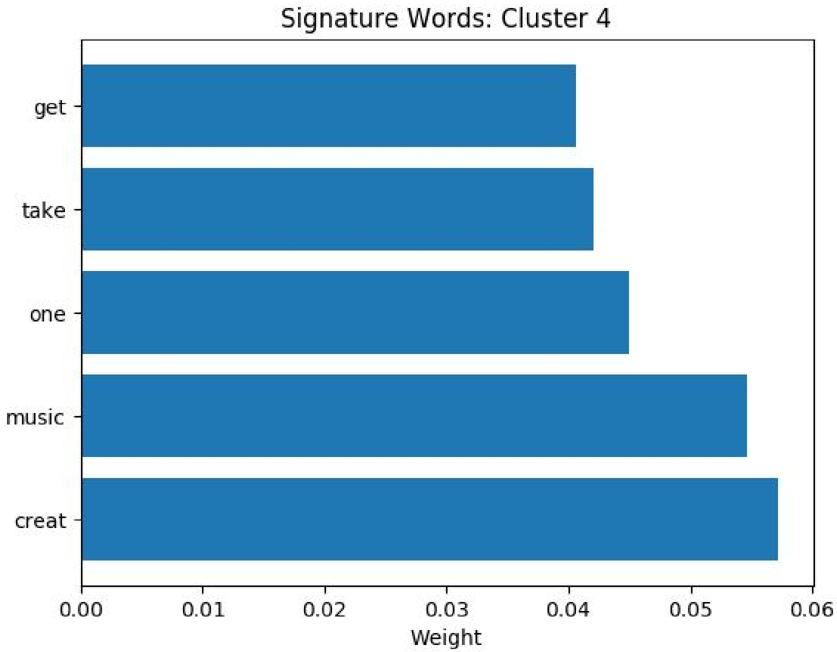
The average silhouette_score is : 0.078

Clusters


Clusters Distribution and Signature Words
Number of Projects
% of 20,000 Sample
Clusters
0
1000
2000
3000
4000
5000
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
made
new
game
art
create
design
make
film
world
help
series
life
stories
book
album
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

The Most Separated (#9) and The Largest (#4) Segments





Clusters Highlights

Cluster ID	Category	Top % in Category	Signature Word	Fun Facts
2	Games	77%	‘game’	The highest pledge, goal, staff picked, and backers numbers
14	Music	96%	‘album’ and ‘record’	The highest success rate (65%)
7	Film & Video	89%	‘film’	The 2nd most successful (60%)
13	Publishing	68%	‘book’	The second lowest % live (1.05)
2	Technology Food	30% 25%	‘made’ & ‘food’	The lowest success rate (31%)

Conclusion

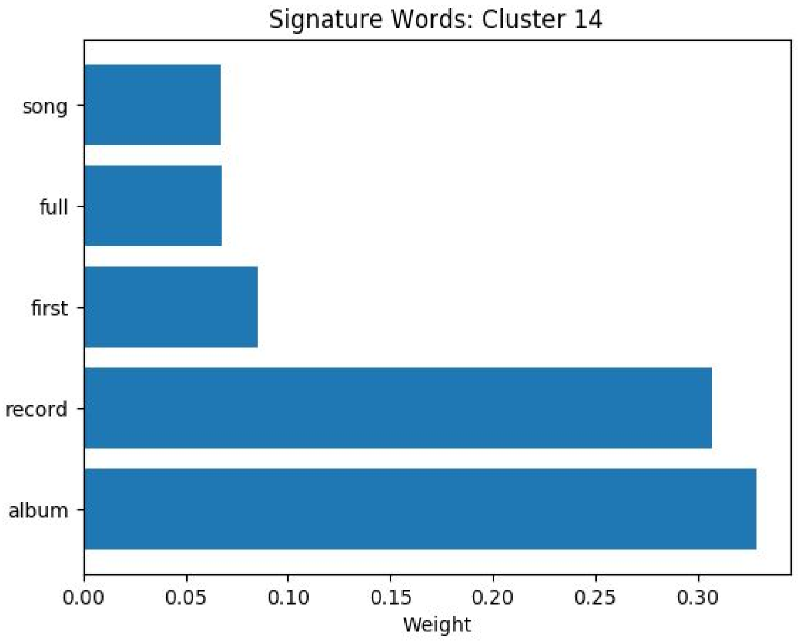
- Only 5 out of 15 clusters closely follow project’s Category, thus, more analysis needed to better understand the semantics behind groupings
- The standalone cluster 9 seems to be just asking for money (top 5 words are ‘help’, ‘need’,’us’, ‘fund’, ‘get’
- Cluster 4 is a catch-it-all - the top words are generic, the size is the largest
- Game is king - the highest pledge and goal amounts, the highest number of staff-picked projects, and the highest backers count

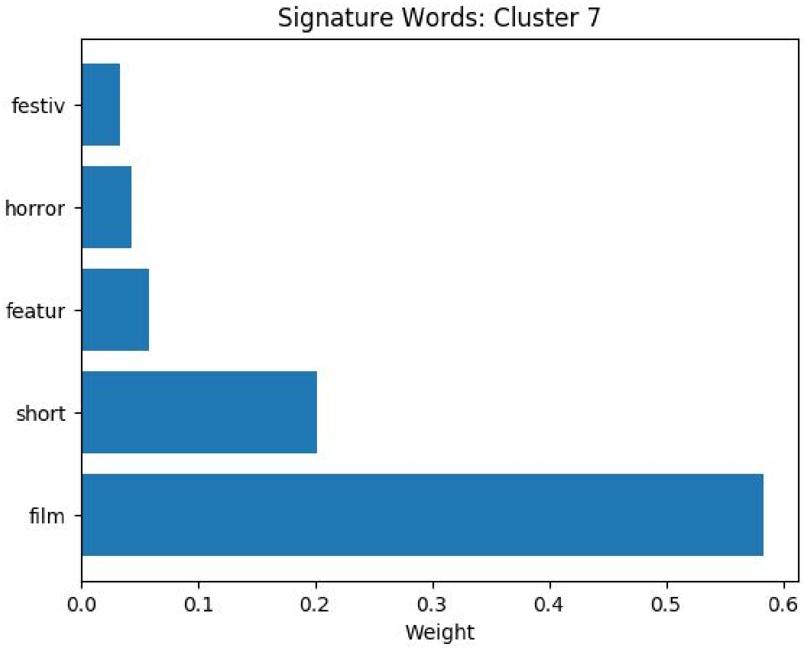
Next Steps

- ❏ Use all 130K + observations
- ❏ Try other clustering techniques
- ❏ Build supervised classification model to see if new clusters or specific signature tokens can predict success of campaign

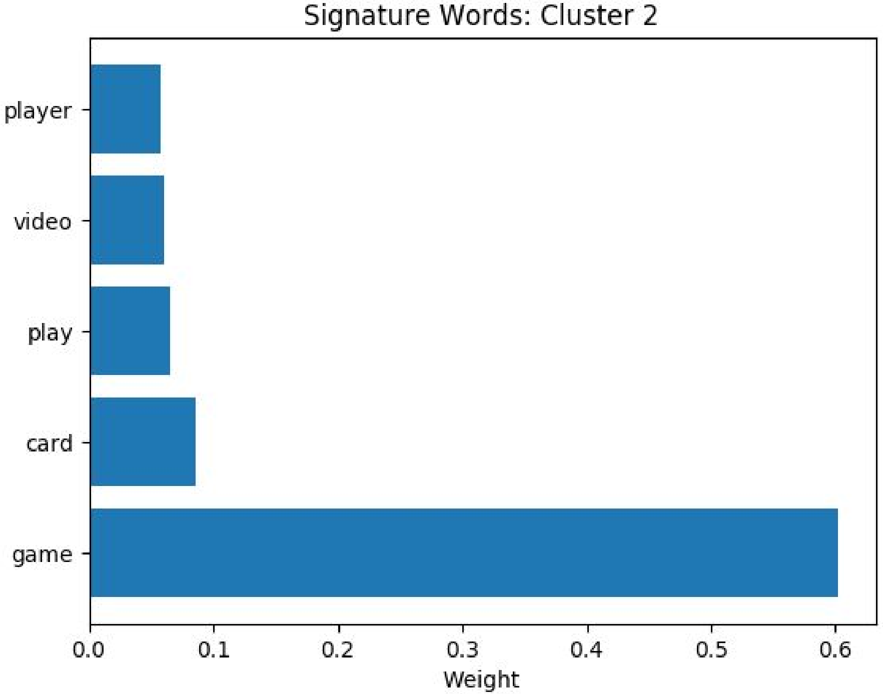
Appendix

Signature Words: Most Successful Clusters





Signature Words: The Highest Pledge and Goal Amounts, Most supported by Staff, and Max Number of Backers Cluster


Signature Words: Cluster 2
player
video
play
card
game
Weight