Assignment 3: Kickstarter Projects

Thomas Brambor

2019-03-31

Text Mining Kickstarter Projects

Overview

Kickstarter is an American public-benefit corporation based in Brooklyn, New York, that maintains a global crowd funding platform focused on creativity. The company's stated mission is to "help bring creative projects to life".

Kickstarter has reportedly received more than $4 billion in pledges from 15.5 million backers to fund 257,000 creative projects, such as films, music, stage shows, comics, journalism, video games, technology and food-related projects.

For this assignment, I am asking you to analyze the descriptions of kickstarter projects to identify commonalities of successful (and unsuccessful projects) using the text mining techniques we covered in the past two lectures.

Data

The dataset for this assignment is taken from webroboto.io 's repository (https://webrobots.io/kickstarter-datasets/). They developed a scrapper robot that crawls all Kickstarter projects monthly since 2009. We will just take data from the most recent crawl on 2019-03-14.

To simplify your task, I have downloaded the files and partially cleaned the scraped data. In particular, I converted several JSON columns, corrected some obvious data issues, and removed some variables that are not of interest (or missing frequently), and remove some duplicated project entries. I have also subsetted the data to only contain projects originating in the United States (to have only English language and USD denominated projects).

The data is contained in the file `kickstarter_projects.csv` and contains about 127k projects and about 20 variables.

Tasks for the Assignment

1. Identifying Successful Projects

a) Success by Category

There are several ways to identify success of a project:
- State (`state`): Whether a campaign was successful or not.
- Pledged Amount (`pledged`)
- Achievement Ratio: Create a variable `achievement_ratio` by calculating the percentage of the original monetary `goal` reached by the actual amount `pledged` (that is `pledged` \ `goal` *100).
- Number of backers (`backers_count`)
- How quickly the goal was reached (difference between `launched_at` and `state_changed_at`) for those campaigns that were successful.

Use one or more of these measures to visually summarize which categories were most successful in attracting funding on kickstarter. Briefly summarize your findings.

BONUS ONLY: b) Success by Location

Now, use the location information to calculate the total number of successful projects by state (if you are ambitious, normalize by population). Also, identify the Top 50 “innovative” cities in the U.S. (by whatever measure you find plausible). Provide a leaflet map showing the most innovative states and cities in the U.S. on a single map based on these information.

2. Writing your success story

Each project contains a `blurb` – a short description of the project. While not the full description of the project, the short headline is arguably important for inducing interest in the project (and ultimately popularity and success). Let’s analyze the text.

a) Cleaning the Text and Word Cloud

To reduce the time for analysis, select the 1000 most successful projects and a sample of 1000 unsuccessful projects. Use the cleaning functions introduced in lecture (or write your own in addition) to remove unnecessary words (stop words), syntax, punctuation, numbers, white space etc. Note, that many projects use their own unique brand names in upper cases, so try to remove these fully capitalized words as well (since we are aiming to identify common words across descriptions). Stem the words left over and complete the stems. Create a document-term-matrix.

Provide a word cloud of the most frequent or important words (your choice which frequency measure you choose) among the most successful projects.

b) Success in words

Provide a pyramid plot to show how the words between successful and unsuccessful projects differ in frequency. A selection of 10 - 20 top words is sufficient here.

c) Simplicity as a virtue

These blurbs are short in length (max. 150 characters) but let’s see whether brevity and simplicity still matters. Calculate a readability measure (Flesh Reading Ease, Flesh Kincaid or any other comparable measure) for the texts. Visualize the relationship between the readability measure and one of the measures of success. Briefly comment on your finding.

3. Sentiment

Now, let’s check whether the use of positive / negative words or specific emotions helps a project to be successful.

a) Stay positive

Calculate the tone of each text based on the positive and negative words that are being used. You can rely on the Hu & Liu dictionary provided in lecture or use the Bing dictionary contained in the tidytext package (`tidytext::sentiments`). Visualize the relationship between tone of the document and success. Briefly comment.

b) Positive vs negative

Segregate all 2,000 blurbs into positive and negative texts based on their polarity score calculated in step (a). Now, collapse the positive and negative texts into two larger documents. Create a document-term-matrix based on this collapsed set of two documents. Generate a comparison cloud showing the most-frequent positive and negative words.

c) Get in their mind

Now, use the NRC Word-Emotion Association Lexicon in the tidytext package to identify a larger set of emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust). Again, visualize the relationship between the use of words from these categories and success. What is your finding?

Submission

Please follow the instructions (/Exercises/homework_submission_instructions.md) to submit your homework. The homework is due on Monday, April 8.

Please stay honest!

If you do come across something online that provides part of the analysis / code etc., please no wholesale copying of other ideas. We are trying to evaluate your abilities to visualized data not the ability to do internet searches. Also, this is an individually assigned exercise – please keep your solution to yourself.