r-datascience-fundamentals

Groups 2,7

November 29, 2017

Goal

Analyze reddit news posts because they give us a lot of information about what is important to society at a particular time. We use reddit because it is one of the most popular websites in the world and has a very active user community, with 1.6 billion users. The news subreddit alone has 70,000 posts in a month. ## Data

We decided to use one month of Reddit News data. We picked December 2016 for this as it was just after the US elections and that enough time has passed for us to (possibly) notice the effects of news that came out at that time.

Reddit news data usually comes in the form of a title and a link, the title is written by the user and links to the news website's page. It sometimes has thumbnails.

Reddit post data is available on google's bigquery database, it is saved in `news_2016_12.csv`. Following are the most relevant columns for our analysis:

- `time_created` (UTC timestamp) - when was the post created
- `author` - username of user that posted
- `domain` - which domain did the news come from?
- `url` - Specific URL of the news post
- `score` : upvotes - downvotes
- `upvotes` : how many "likes" the post received
- `downvotes` : howmany "dislikes" the post received
- `title` : user-created title of the post

Furthermore, we thought it would be interesting to compare the difference between the user-generated title and the Actual title posted by the news agency. For this, a (scrapy)[https://scrapy.org/] spider was created to crawl all the (cleaned) URLs and retrieve the title. We think this was rather successful since it retreived `24,045` titles form about `31,713` cleaned posts. This is saved as `titles.csv`.

Task

From the data above, we want to do the following:

1. Discover the relationships and trends between elements in our dataset
2. Predict upvotes, comments given a certain title
3. Classify posts (NLP, sentiment, closeness)

Due to the size and variety of elements in our dataset, we will apply a variety of methods to extract informamtion out of it, and may use external data to improve our analysis. However, this also implies that the performance measures of our ML algorithms will vary with the specific relationship under examination.

Cleaning `clean.r`

Everything related to cleaning the original dataset is defined in `clean.r`. Essentially, the script:

- saves `utc_created` as a POSIXct variable
- removes special characters from titles

- removes non-english titles, this halves our dataset.

Descriptive statistics and analysis

This is divided into 3 subparts: genenral, distribution, sentiment

General Statistics - general stuff, scatterplots

Distributions - what distributons does the dataset contain, can we categorize them?

How are the upvotes / comments distributed?

Sentiment - What can we learn from NLP