Topic Modeling and Yelp Reviews

George Dewey

July 17, 2018

What is topic modeling?

A way to understand the structure within a collection of text documents.

The topics created by topic modeling are clusters of related words.

We can try to describe these topics based on the words they contain (but it’s not always easy!)

Definitions

Document - An individual instance of text; for example, a single email or Yelp review.

Corpus - A collection of documents; your entire inbox or all the Yelp reviews for a single business.

An Example Document

“Finally a decent ramen place! It's not the best ramen I've ever had but it is the best in the area. I enjoyed their miso ramen. I'll be back again when I get my ramen cravings”

What topics do you see within this review?

Topic Modeling - Example

Finally a decent ramen place! It's not the best ramen I've ever had but it is the best in the area. I enjoyed their miso ramen. I'll be back again when I get my ramen cravings

Ramen: ramen, ramen place, miso ramen
Like: Decent, best, enjoy, back, again, craving
Dislike: Not the best
Location: in the area
Time: finally, ever

How do we use topic modeling?

The distribution of topics within a document allows us to mathematically describe it!

We assign positive values to the topics we see in a document, and negative values to the topics we don’t see.

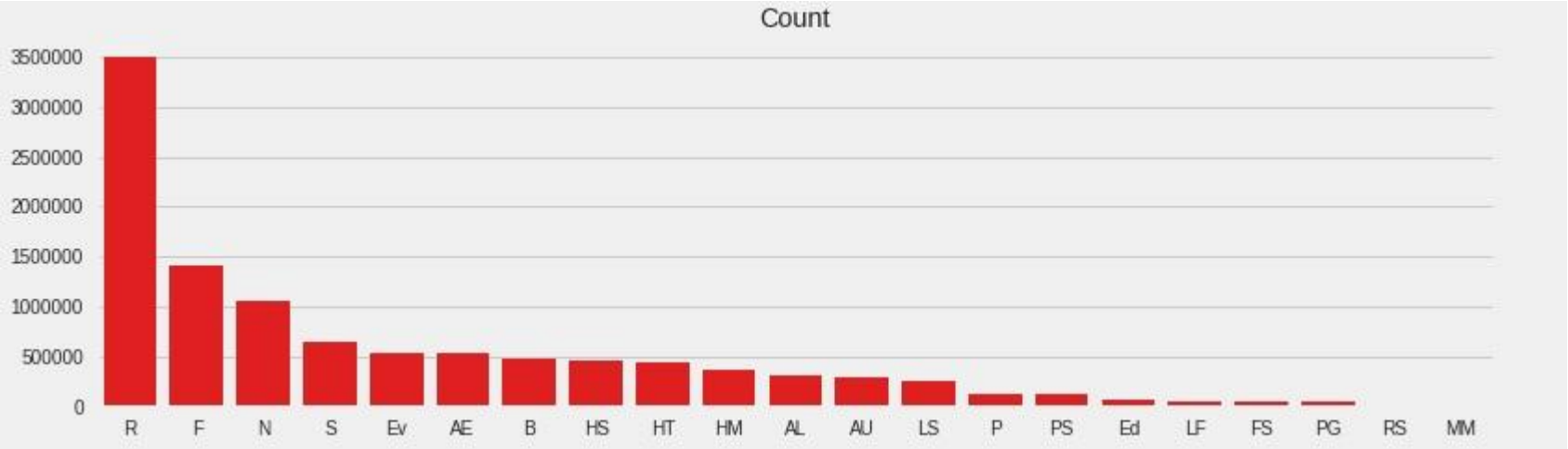
Topic Modeling: An Application

Application: Yelp Reviews

Businesses use “Useful” Yelp reviews to better understand consumer tendencies, inform business strategy, and improve their products and services.

The goal of the project is to use topic modeling to predict if Yelp reviews would be considered ‘Useful’ by other users.

Useful votes by Business Category



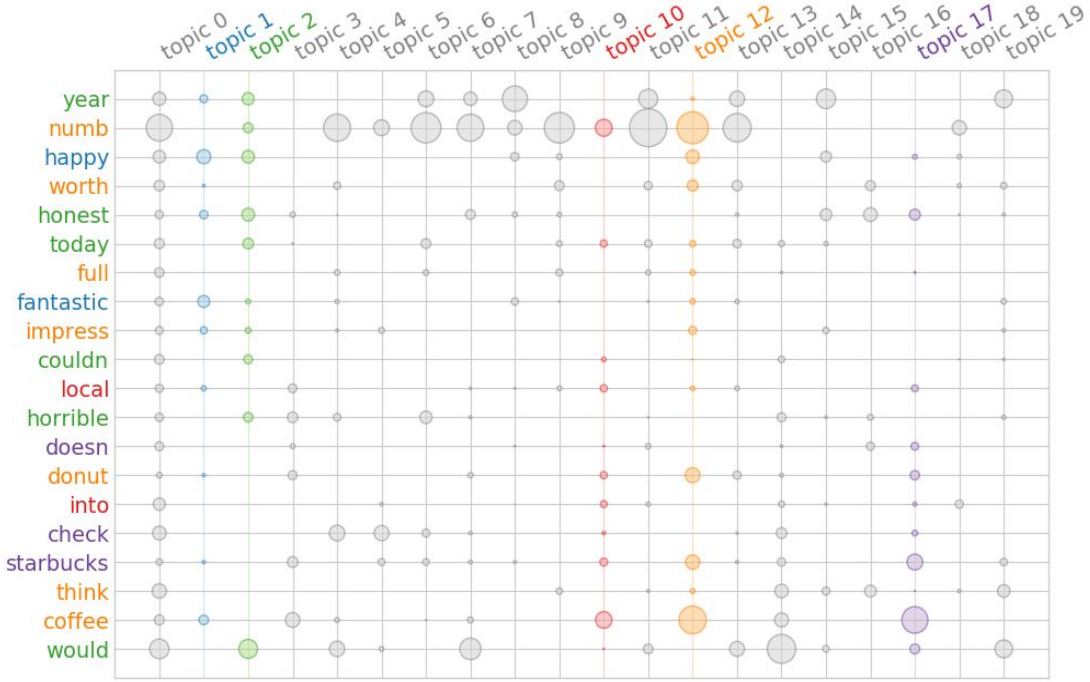
Topic Modeling

Used Latent Semantic Analysis (LSA)

200 topics for restaurant corpus, 325 topics for businesses corpus

Business topics focused on service and staff, while restaurant topics favored types of food and restaurants

Termite Plot - Businesses



Classification

Reviews were considered useful if they had at least 3 useful votes and not useful if they had no useful votes.

In addition to topics, star rating, length of review, and business category were considered in classification.

Results

Business reviews were classified as useful or as not useful with an accuracy of 82%.

Not useful reviews were classified correctly at a higher rate than useful reviews.

Short reviews are hard to classify using topics!

Future Steps

Perform the same analysis with other topic modeling strategies

Limit the analysis to only reviews of a certain length

Potential to use a similar approach for any kind of text-based customer feedback