qualtrics

Identifying Exemplar Survey Responses

The Problem

With large amounts of free response data, it can be nearly impossible to gain any meaningful business insights without spending countless hours reading entries.

We set out to greatly reduce that time by building a model that groups survey responses in a meaningful way and returns the most representative, or "exemplar," response from each group.

Setting Up for Success

We prepared the raw text data for our model by correcting misspelled words and removing unnecessary links, websites, and words that provide little to no meaning.

Spell Correct

~~You have an~~ exclent ~~platform at qualtrics.com.~~
~~We've had a~~ grest ~~experience with~~ customer service!!

excellent platform great experience customer service

Technologies

python™ Natural Language Analyses with NLTK

gensim

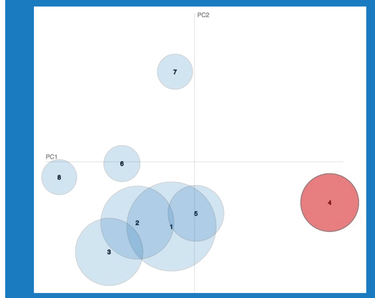
Aspell

LANGID.PY

Two Competing Models

We developed two different solutions to see which would produce better results:

LATENT DIRICHLET ALLOCATION (LDA)

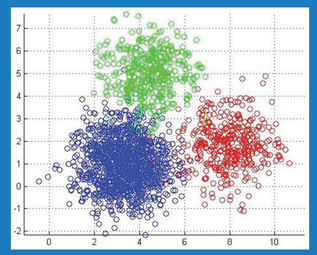


Simply put, the model considers each document (in this case, a survey response) as a multinomial distribution of n topics, where each topic is a multinomial distribution of N words. LDA then estimates the parameters to these distributions. By extracting one document from each topic, the model provides the user with a small subset of documents that are representative of the general trends in a corpus.

VS.

K-MEANS

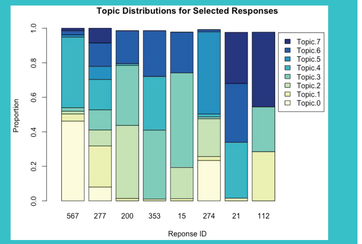
This model uses a pre-trained doc2vec model from Associated Press data to infer vector representations for our survey responses. We then applied k-means clustering with cosine similarity to group the response vectors. The model then compares each response vector to its cluster's centroid, and the response that is the most similar to the centroid is the exemplar response for that cluster.



Results

The LDA model produced better clusters than K-means.

We created a topic model that will automatically extract exemplar survey responses from a corpus. We also produced a visualization using pyLDAvis so that users can interactively explore the topic modeling that our algorithm uses.



Impact for Qualtrics

- ✓ **Enables their clients to make business decisions quickly and effectively based on the data our model returns.**
- ✓ **Saves their clients countless hours of sorting and analyzing data to gain meaning, improving their bottom line.**
- ✓ **Helps clients understand major trends in their survey responses.**
- ✓ **Provides valuable R&D knowledge for Qualtrics, expands the use cases for these two machine learning methods, and opens the way for further investigation.**