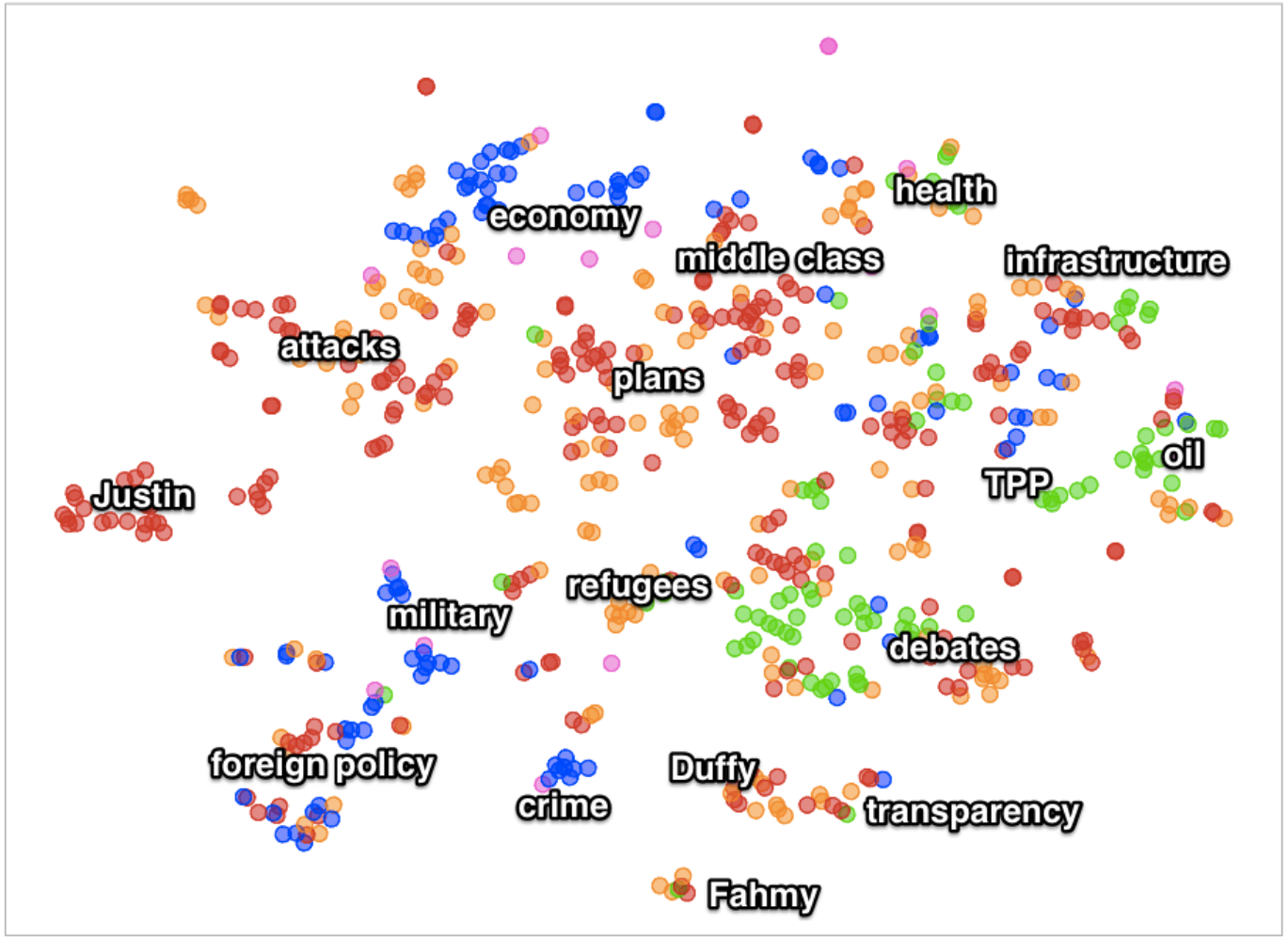
Unsupervised Machine Learning for NLP

K – means clustering for topic classification



E-commerce Search Insights

Customer Intent?

Internet #304225557

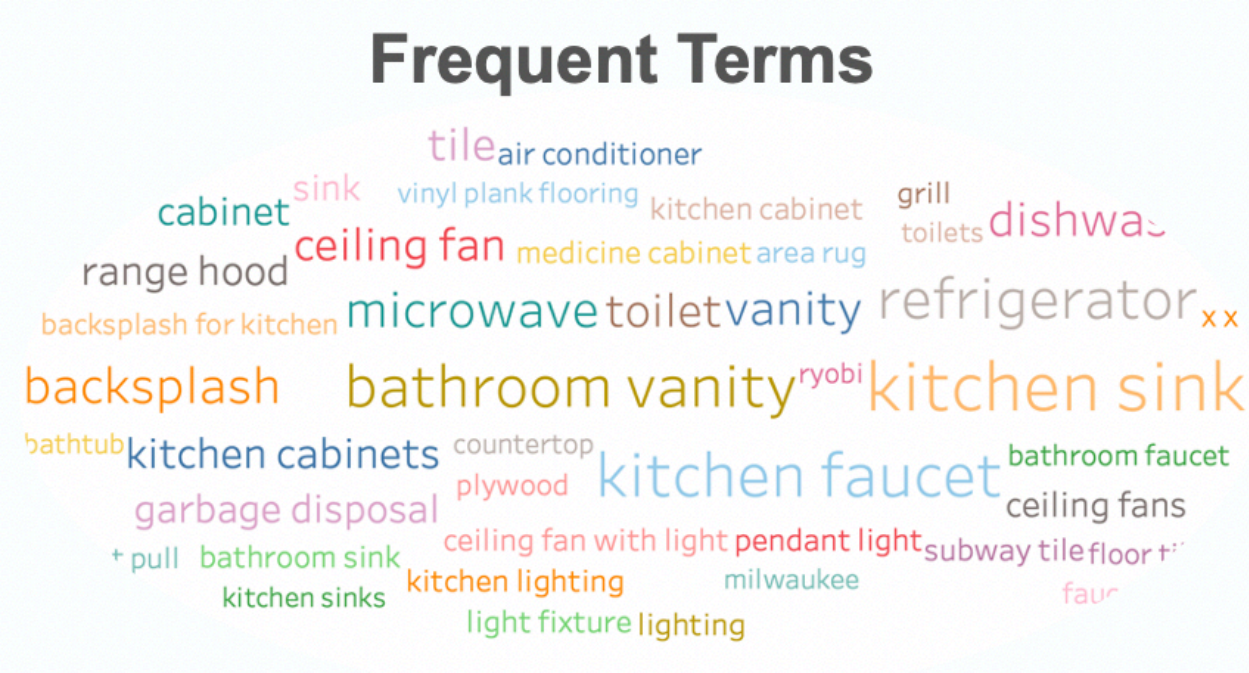
avorites Print

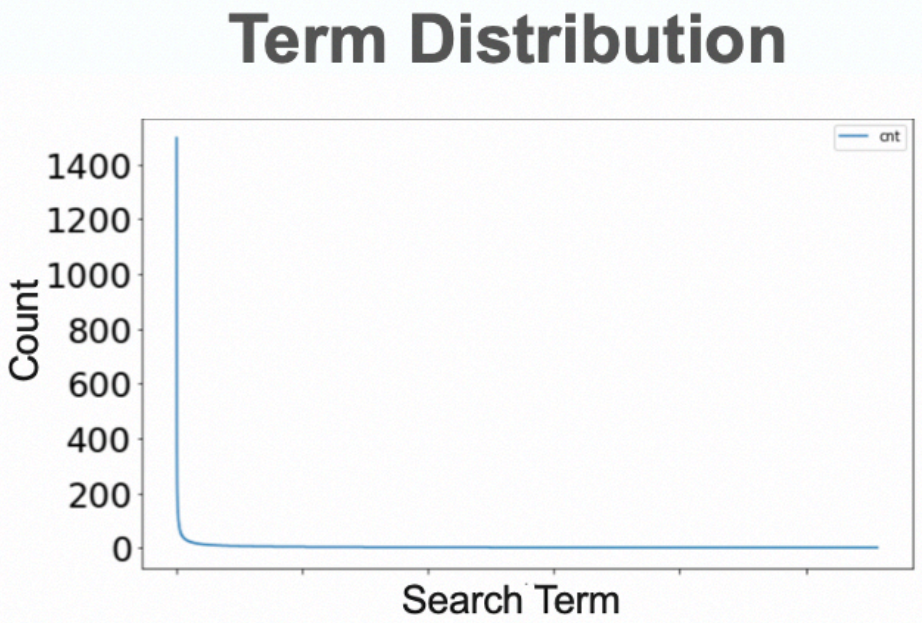
T-FCFS30 Cape 30" L x 18" V
Sink with Sink Grid

See More by Nantucket Sinks

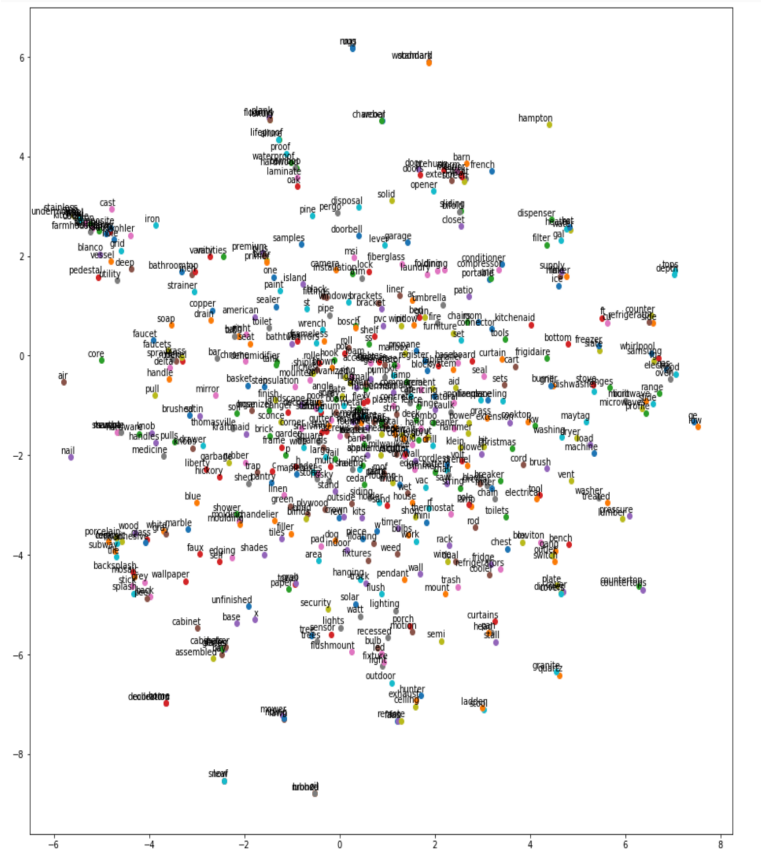
622

Search Term Insights





Keywords Clusters



Kitchen sink – stainless steel – single – double - undermount - farmhouse – apron - drop

Cabinets – X – hardware - drawer – knobs – medicine – shaker - unfinished

Countertop – quartz – granite – marble – island

Backsplash – glass - mosaic – stick - peel

Python Demo

Dataset:
https://www.kaggle.com/zynicide/wine-reviews

NLP Techniques

- Text data pre-processing
  - Regular Expression

    https://www.w3schools.com/python/python_regex.asp
  - Data Cleansing

    Punctuation

    Special characters

    Extra white space

    Lower case
  - More advanced

    Stop words

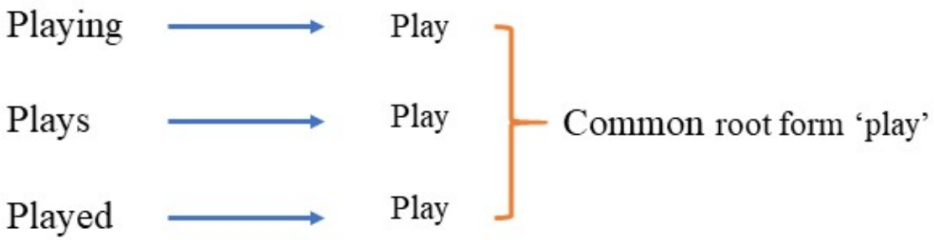
    Stemming

    Tokenizer

    Vectorizer

Stemming

Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.



am, are, is → be

Car cars, car's, cars' → car

Using above mapping a sentence could be normalized as follows:

the boy's cars are different colors → the boy car be differ color

Tokenization

Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or sub words. Hence, tokenization can be broadly classified into 3 types – word, character, and sub word (n-gram characters) tokenization.

Vectorization

- Vectorization is transforming text into a meaningful vector (or array) of numbers.
  - word2vec
  - TF-IDF

    TF-IDF stands for Term Frequency-Inverse Document Frequency which basically tells importance of the word in the corpus or dataset. TF-IDF contain two concept Term Frequency(TF) and Inverse Document Frequency(IDF)

$$TF = \frac{No\ of\ time\ word\ appear\ in\ the\ document}{Total\ no\ of\ word\ in\ the\ document}$$

$$IDF = log_{10} \frac{Number\ of\ Document}{Number\ of\ document\ in\ which\ word\ appear}$$

	The	TFIDF	Vectorization	Process	Is	Beautiful	Concept
Documen1	0	0.9704	0.33	0.096	0	0	0
Document2	0	0.97	0	0	0	0	0.698
Document3	0	0.698	0	0	0	0	0
Document4	0	0.66	0.397	0	0	0	0
Document5	0	0.21	0	0	0	0	0

T-SNE

- t-distributed stochastic neighbor embedding is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for **visualization** in a low-dimensional space of two or three dimensions.
- How to use t-SNE effectively

  https://distill.pub/2016/misread-tsne/