ICE CREAM REVIEWS

**GOALS**

* Create a dashboard (using R Shiny, Plotly Dash, Streamlit; or a static report using Tableau) which shows review trends and insights for each brand.
* Filter by product or show charts for all products combined
* Use topic modeling/clustering to group [only negative?] reviews and identify complaints. Use sentiment analysis to “refine” the star ratings and show most important reviews.
* Optional: plots for product comparison (within/between brands) & brand comparisons; i.e. flavor shift
* Optional: update data daily
* EDA: examine ingredients list to determine consumer likes. See my Kaggle notebook for EDA ideas.

**KEYWORDS**

Data collection (web scraping), data dashboard (web app), interactive data visualization, product metrics, emerging trends, business analytics, consumer reviews, EDA, clustering, NLP, topic modeling (LDA-esque methods), sentiment analysis

**NATURAL LANGUAGE PROCESSING FOR REVIEWS**

1. Clustering/topic modeling – common themes and complaints
2. Sentiment analysis – star ratings are not always accurate; we may benefit from measuring sentiment ourselves
3. Text summarization – generate a report from all reviews
4. Identify emerging trends – leverage timestamps

**THE DATA**

Products: name, description, ingredients. Average rating & rating count are inferred from reviews.

Reviews: author, date, stars, title, helpful votes, text. Haagen-Dasz only: taste/ingredients/texture/tags.

**CONSIDERATIONS**

Factor in helpfulness of review.

Sometimes the title is informative (ex: “Way Too Salty”) sometimes it’s not (“Worst Ice Cream Ever”).

Most reviews are ~40-50 words. Use models that work well with **short text**.

Only 20k reviews across all brands. Might want to do topic modeling across all brands. Try to limit filtering to preserve data.

How do we find reviews/sentences within reviews that are meaningful/actionable?

**METHODS FOR SENTIMENT ANALYSIS**

VADER/SentimentR

**METHODS FOR TOPIC MODELING**

use unigrams & bigrams. Filter out stop words & punctuation, lemmatize, stemming, etc.

1.1.: traditional clustering methods (i.e. K-means) on TF-IDF or embedding (word2vec) vectors

1.2.: use Word Mover Distance as a better distance metric

2.: use Haagen-Dasz tags for inspiration. Cluster documents based on occurrence of human-selected tags such as: taste, texture, ingredients, flavor, etc. On the same lines, try to identify \*review words\* that signify an important sentence, i.e. “improvement” or “unsatisfactory”. Perhaps consider parts of speech or semantics which indicate an important sentence.

Three: more advanced methods (LDA/GSDMM, neural networks, graph-based, etc.)

**REFERENCES, citations**

[Short Text Topic Modeling Techniques, Applications, and Performance: A Survey](https://arxiv.org/abs/1904.07695) 24

[Concept decompositions for short text clustering by identifying word communities](https://www.sciencedirect.com/science/article/pii/S0031320317303953) 30

[Short text clustering by finding core terms](https://link.springer.com/article/10.1007/s10115-010-0299-7) 62

[Comparative study of clustering techniques for short text documents](https://dl.acm.org/doi/abs/10.1145/1963192.1963249?casa_token=yexcHVSFidQAAAAA:v5VMuFFbPPt8n-Tw0AxRdDUHFv1T_gmTb5vCARABzu6ahTTYzIC-Lw9Ozoga6I-y2zM-PyQuRiV4Xw) 97

[A dirichlet multinomial mixture model-based approach for short text clustering](https://dl.acm.org/doi/abs/10.1145/2623330.2623715?casa_token=gZ8Oty7pzMUAAAAA:NmGPxVYGcPNnBroGJ0ULNOEEevs4SucwiDJHOf4dxLDqew6JJ6w_HUkbAlNYIBJCXVZAaPLeOYLBxg) 272 **(GSDMM)**

[A Unique Approach to Short Text Clustering (Algorithmic Theory)](https://towardsdatascience.com/a-unique-approach-to-short-text-clustering-part-1-algorithmic-theory-4d4fad0882e1)blog post on GSDMM

[Clustering short text using Ncut-weighted non-negative matrix factorization](https://dl.acm.org/doi/abs/10.1145/2396761.2398615?casa_token=_PXfjTZaJsUAAAAA:QVqifS74CD93kCcspbRCaPp0fJ8-0HSNeyW1kFAe9VWH88k6El_olyfMGCpnKiLcvwk_0QGAAl_Now) 47

[Customer Complaints Analysis Using Text Mining and Outcome-Driven Innovation Method for Market-Oriented Product Development](https://www.mdpi.com/2071-1050/11/1/40) 12

[Google scholar results](https://scholar.google.com/scholar?q=clustering+short+text&hl=en&as_sdt=0&as_vis=1&oi=scholart)

[Quora](https://www.quora.com/What-are-some-good-papers-about-topic-modeling-for-short-texts-especially-for-Tweets)

[From Word Embeddings To Document Distances](http://proceedings.mlr.press/v37/kusnerb15.pdf) 1292 **(WMD)**

[Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves](https://arxiv.org/abs/1604.02038) 14

**TEXT PREPROCESSING**

-Remove stopwords

>English stopwords

>product details

-Filter out extremely common/uncommon words in reviews

>either filter by total frequency or filter words that do not appear in enough reviews

-Normalization: stemming vs lemmatizing. stemming is generally better

>this makes word clouds & interpretation harder, though

>remove stopwords & filter before or after stemming?

-Convert to lowercase, remove extra whitespace, remove numbers

-Remove punctuation (including unicode chars)

>may want to be careful about ' and -

-Remove tokens with 1 or 2 characters

Techniques

-ngrams: unigrams vs bigrams

--------------------------------------------

Topic modeling

1. Manual clustering

-Create a few clusters (taste/flavor, texture, value, ingredients quality, change)

-Build them up by computing commonly co-occuring words, inspecting the list, and

adding relevant words

-Can also use word embeddings to identify similar words

-Assign each document to a cluster based on cluster-word-occurences

2. K-means

-Create term-document matrix; with or without tf-idf

-Use K-means and determine appropriate number of clusters

-Use Word-Mover-Distance, which makes use of word embeddings

3. Traditional models (PLSA/LSA/LSI & LDA)

4. short-text models (GSDMM, neural nets, graph based, etc.)