Capstone Project — NLP Applications

Ian Haggerty • 26.02.2024

Overview

Expected delivery

4th March, 2024

- External amazon product reviews analysed with pandas, spaCy & TextBlob.
- Sentiment and subjectivity ratings for all reviews.
- Similarity ratings across pairs of reviews.
- Wordcloud comparison of the most similar and dissimilar reviews.
- Similarity matrix visualised with Matplotlib & Seaborn.
- Evaluation and further work.

Dataset Description

Reviews of Amazon Products From External Websites (*Bestbuy*, *Newegg*, etc.)

- Dataset contains the review *text*, *title* as well as the *rating*.
- Maintains the original user *id*.
- Temporal data for *date added* and *date updated*.
- The product *category* and *name* are provided.
- External *source* of the review made available via a URL.

Processing Steps

Steps One

- Download dataset from Kaggle.
- Import the CSV file using the pandas library.
- Parse into a dataframe.
- Drop data points where a review hasn’t been provided.

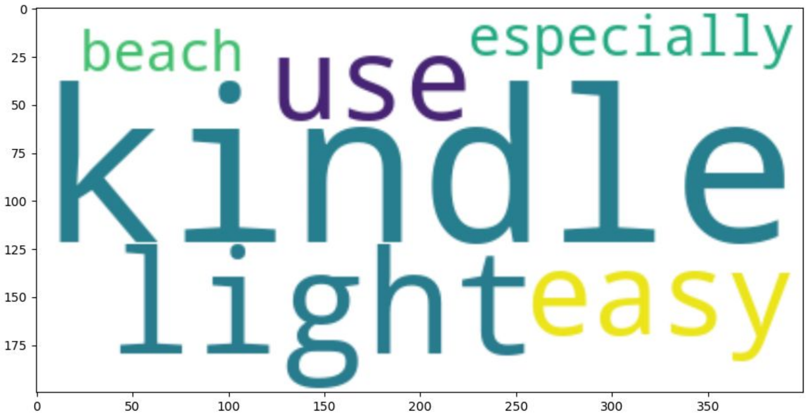
Steps Two

- Perform NLP analysis on review text using the spaCy library.
- Strip out stop-words and lemmatize the remaining tokens.
- Conduct sentiment analysis using the TextBlob library.

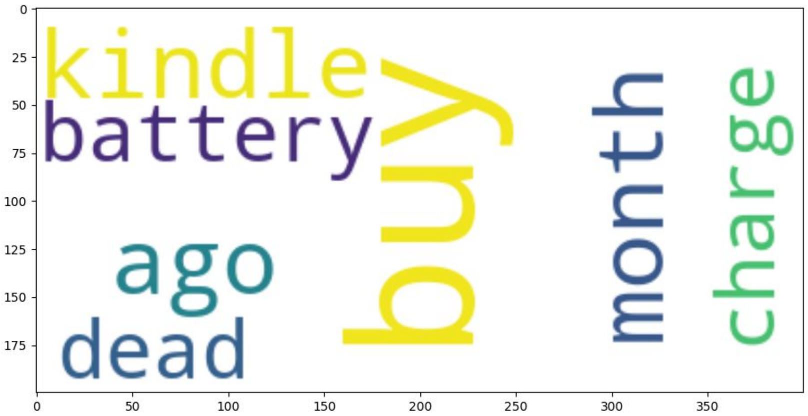
Most Dissimilar Reviews

Similarity = 0.23

"Great light reader. Easy to use at the beach"



"not good quality"



Evaluation

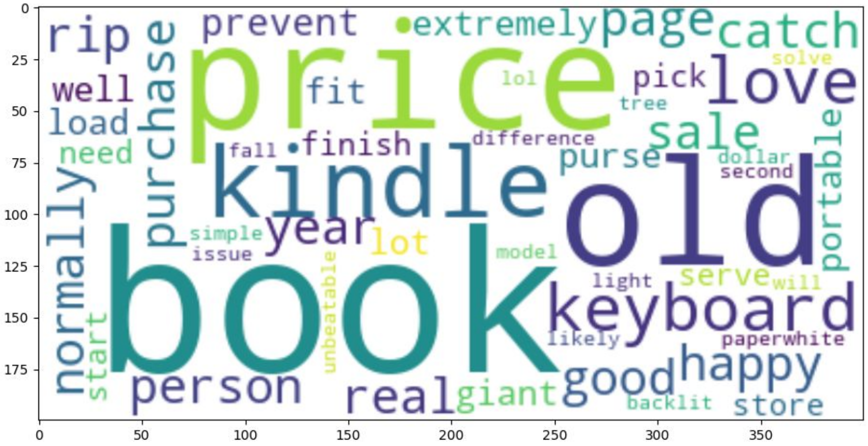
Lowest Similarity

- Reviews which were most dissimilar happened to be some of the shortest, suggesting that the `en_core_web_sm` model doesn't correctly normalize against this feature.
- These reviews were of opposing polarity—as one might expect for syntactically "dissimilar" reviews.
- They were both focused on the amazon kindle.
- The negative reviewer expressed concern about battery life.
- The positive reviewer commended the weight of the kindle and it's ease of use at the beach.

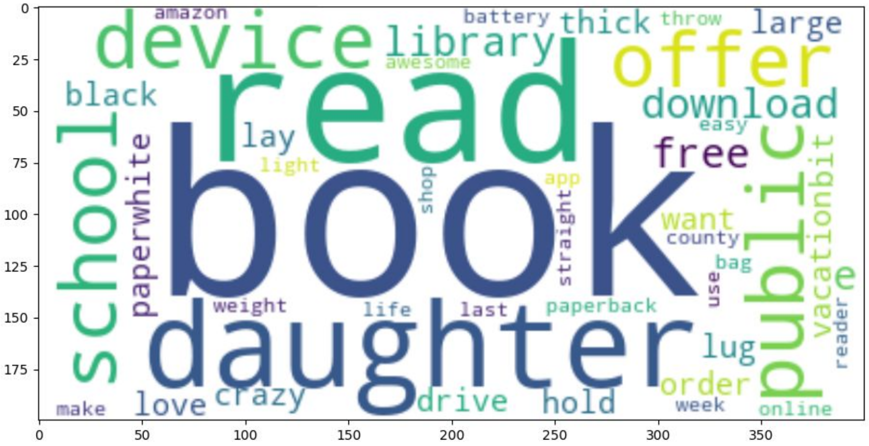
Most Similar Reviews

Similarity = 0.82

"A Great Buy"



"Light Weight - Makes a world of difference when taking books on the go!"



Evaluation

Highest Similarity

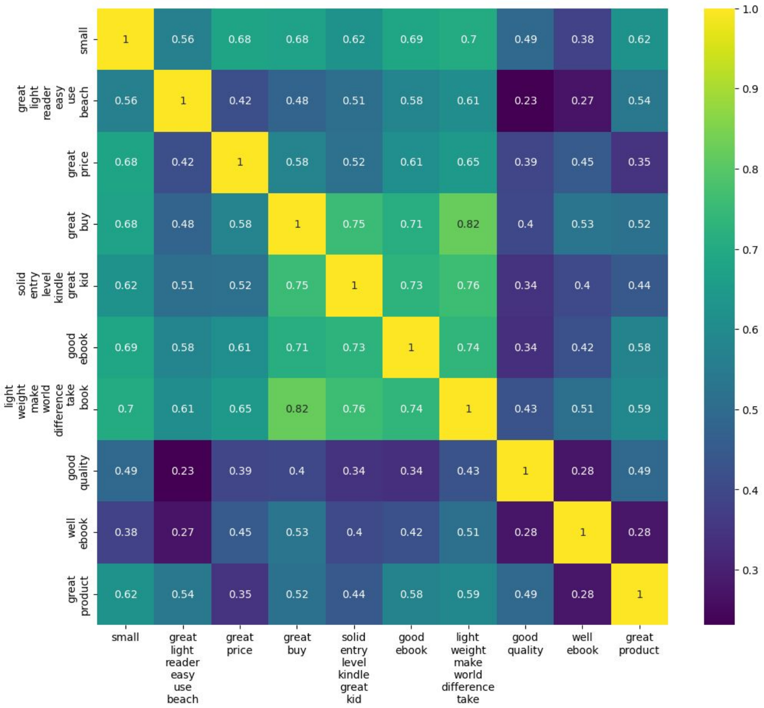
- Reviews which were most similar also happened to be some of the longest, suggesting that the `en_core_web_sm` model doesn’t correctly normalize against this feature.
- The polarity and subjectivity ratings were close, as one might expect for syntactically “similar” reviews.
- Both reviews were on the amazon kindle, although the 2nd reviewer didn’t explicitly mention it—opting to use the word ‘this’.
- The 2nd reviewer bought the kindle for her daughter.

Evaluation

Remarks

- The `en_core_web_sm` model doesn’t use word to vector algorithms, such as word2vec, to evaluate similarity.
- Instead it depends upon part-of-speech tagging and named entity recognition.
- Using the `en_core_web_md` model produces higher similarity ratings (delegating to cosine similarity, internally).
- Using the `en_core_web_trf` on the *entire* dataset would yield more accurate and insightful results.

Review Similarity Matrix



	small	great light reader easy use beach	great price	great buy	solid entry level kindle great kid	good ebook	light weight make world difference take	good quality	well ebook	great product
small	1	0.56	0.68	0.68	0.62	0.69	0.7	0.49	0.38	0.62
great light reader easy use beach	0.56	1	0.42	0.48	0.51	0.58	0.61	0.23	0.27	0.54
great price	0.68	0.42	1	0.58	0.52	0.61	0.65	0.39	0.45	0.35
great buy	0.68	0.48	0.58	1	0.75	0.71	0.82	0.4	0.53	0.52
solid entry level kindle great kid	0.62	0.51	0.52	0.75	1	0.73	0.76	0.34	0.4	0.44
good ebook	0.69	0.58	0.61	0.71	0.73	1	0.74	0.34	0.42	0.58
light weight make world difference take book	0.7	0.61	0.65	0.82	0.76	0.74	1	0.43	0.51	0.59
good quality	0.49	0.23	0.39	0.4	0.34	0.34	0.43	1	0.28	0.49
well ebook	0.38	0.27	0.45	0.53	0.4	0.42	0.51	0.28	1	0.28
great product	0.62	0.54	0.35	0.52	0.44	0.58	0.59	0.49	0.28	1

Evaluation

Review Similarity Matrix

- The similarity matrix gives a visually compelling way to see which reviews were close to one another.

Evaluation

Further Work

- The *rating* in conjunction with *numHelpful* has been not been considered for this analysis.
- The latter of which can be used to gauge the strongest user sentiment.

Evaluation

Further Work

- A lot of customer insights can be derived from the reviews.
    - “How often is the kindle bought as a gift?”,
    - “When are users most likely to give a review?”,
    - “Are negative reviews longer or shorter?”,
    - “Is there a correlation between the use of proper english and sentiment?”
    - “What pairs of products are likely to bought together?”