NLP Analysis of Amazon Question and Answer banks

CAPSTONE PROJECT

BOB SPOONMORE

OCT 2021

HTTPS://WWW.BIZJOURNALS.COM/COLUMBUS/NEWS/2017/08/03
/REVIEWING-AMAZON-RESTAURANTS.HTML

Problem Statement

Do the Questions Asked by Customers in the Amazon Food Category Get Answered Correctly?

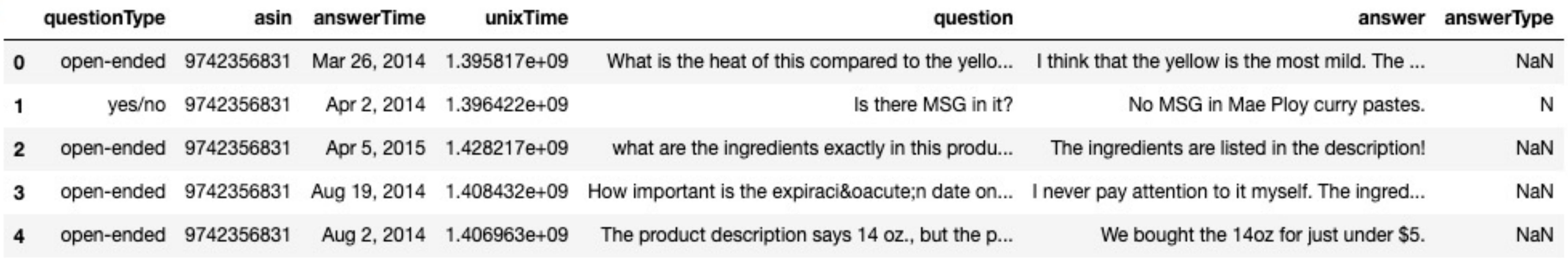
A Capstone project applying NLP modeling to Amazon Question and Answer banks

Data

Source Dataset: https://jmcauley.ucsd.edu/data/amazon/qa/qa_Grocery_and_Gourmet_Food.json.gz:

Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems
Mengting Wan, Julian McAuley International Conference on Data Mining (ICDM), 2016

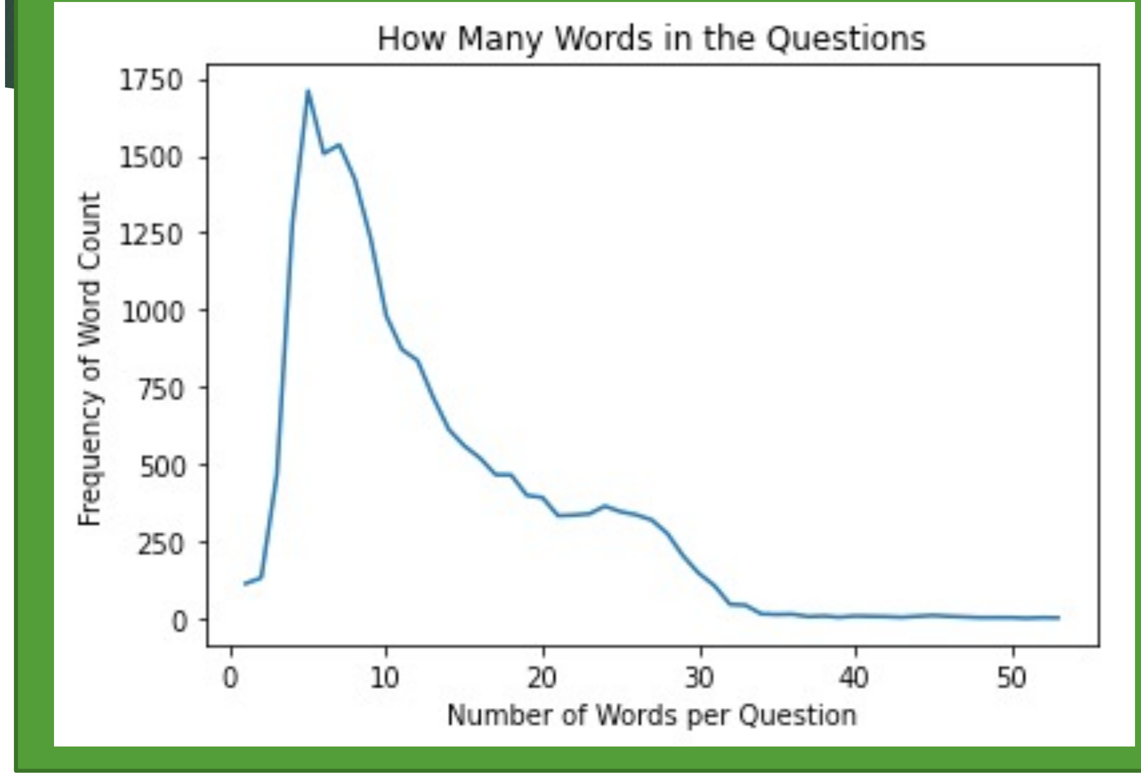
Addressing complex and subjective product-related queries with customer reviews Julian McAuley, Alex Yang World Wide Web (WWW), 2016

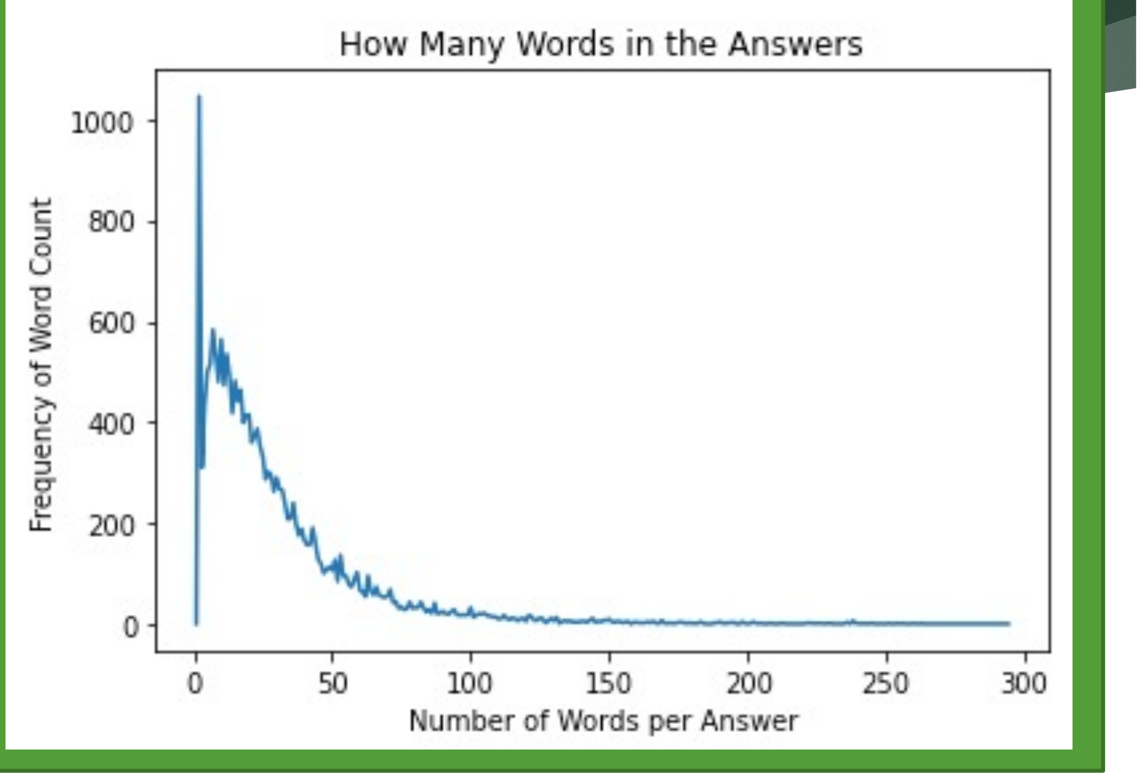


	questionType	asin	answerTime	unixTime	question	answer	answerType
0	open-ended	9742356831	Mar 26, 2014	1.395817e+09	What is the heat of this compared to the yello...	I think that the yellow is the most mild. The ...	NaN
1	yes/no	9742356831	Apr 2, 2014	1.396422e+09	Is there MSG in it?	No MSG in Mae Ploy curry pastes.	N
2	open-ended	9742356831	Apr 5, 2015	1.428217e+09	what are the ingredients exactly in this produ...	The ingredients are listed in the description!	NaN
3	open-ended	9742356831	Aug 19, 2014	1.408432e+09	How important is the expiraci&oacute;n date on...	I never pay attention to it myself. The ingred...	NaN
4	open-ended	9742356831	Aug 2, 2014	1.406963e+09	The product description says 14 oz., but the p...	We bought the 14oz for just under $5.	NaN

shape: (19538, 7)

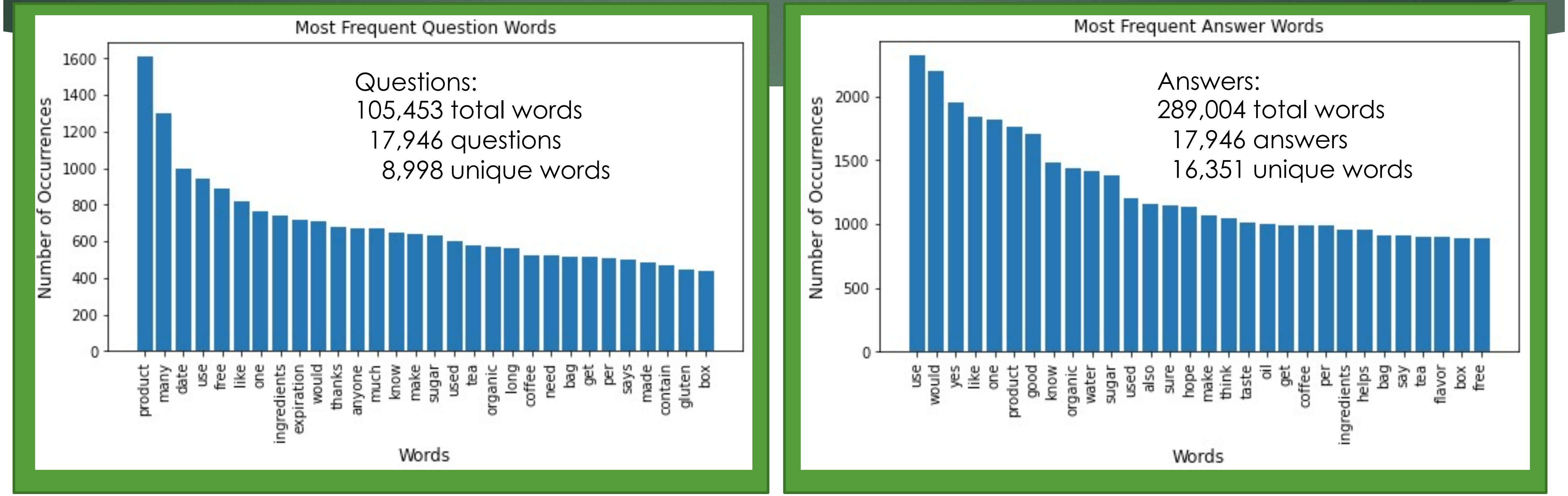
Exploratory Data Analysis





No null values found
Removed all 1 or 2 word questions and answers as these would bias models

Questions versus Answers



Questions:
105,453 total words
17,946 questions
8,998 unique words

Answers:
289,004 total words
17,946 answers
16,351 unique words

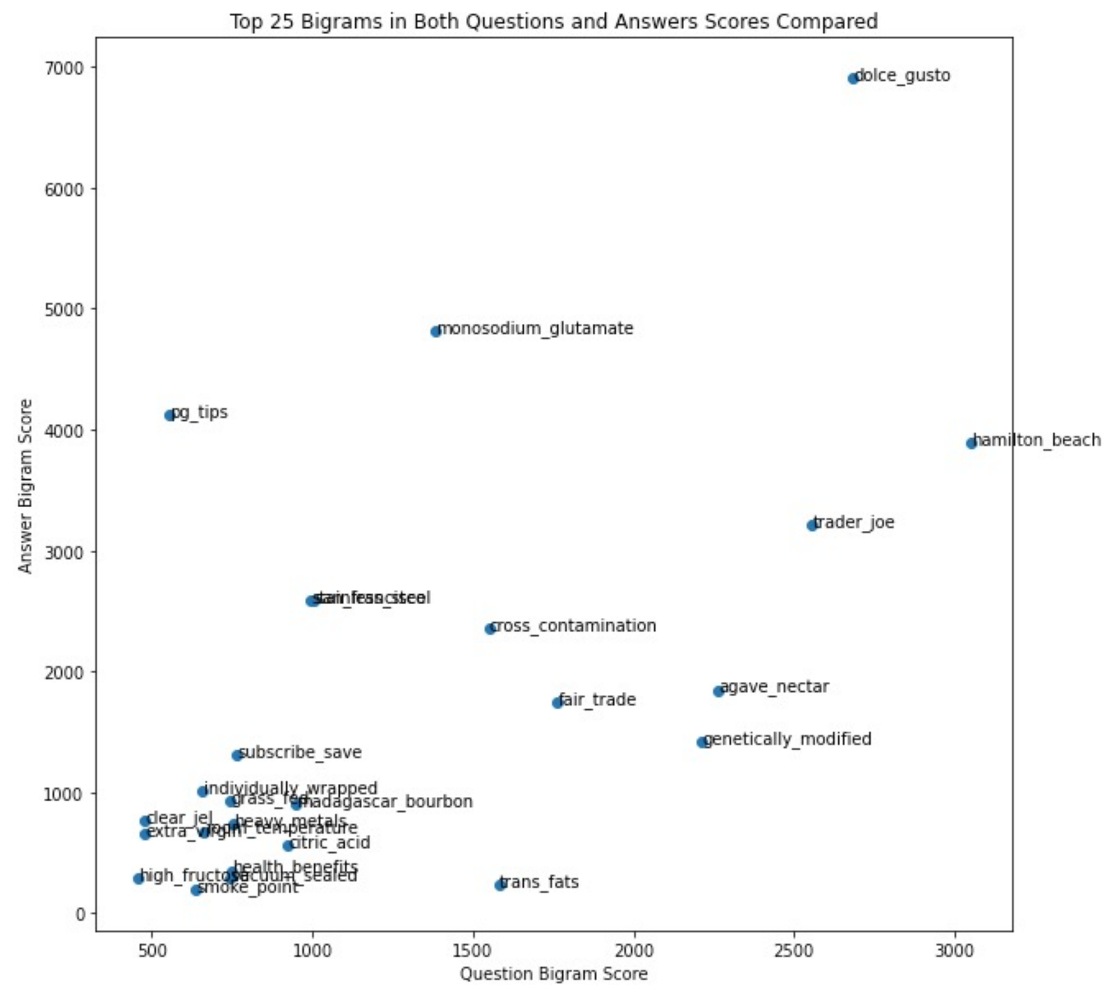
Frequency distribution of words found in questions and answers

Preparing Text for Model

Text Prep on Both Questions and Answers

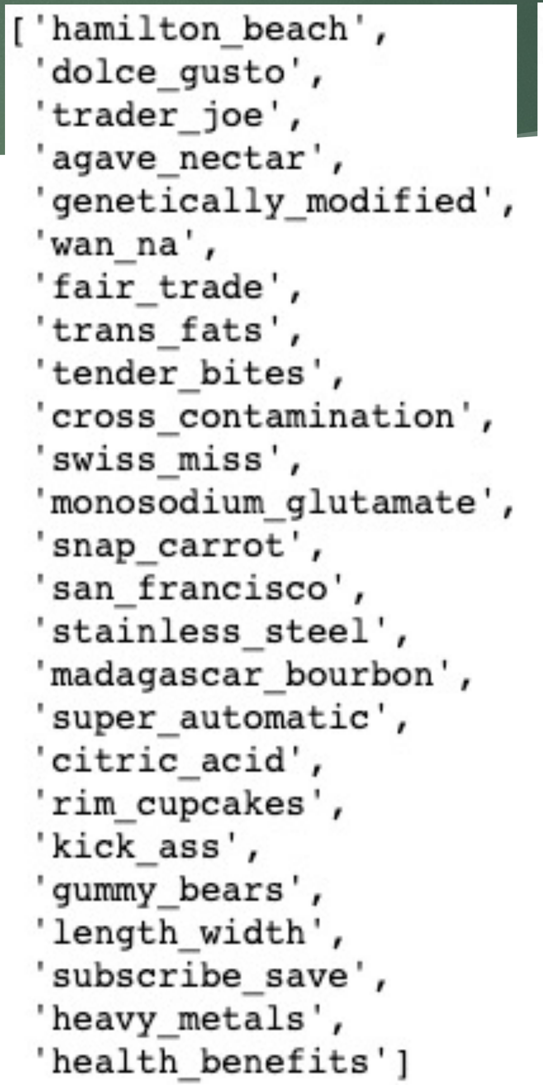
- Preconditioned text: set lower case, remove punctuation, and remove stopwords
- Maintained Integrity of questions as separate lists of words
- Kept Questions and Answers separated, and had total group combined
- Created Bag of Words Dictionary and Corpus based on combined list
- TF-IDF matrix to downgrade most frequent words
- Created Bigrams listing

Bigrams of Text



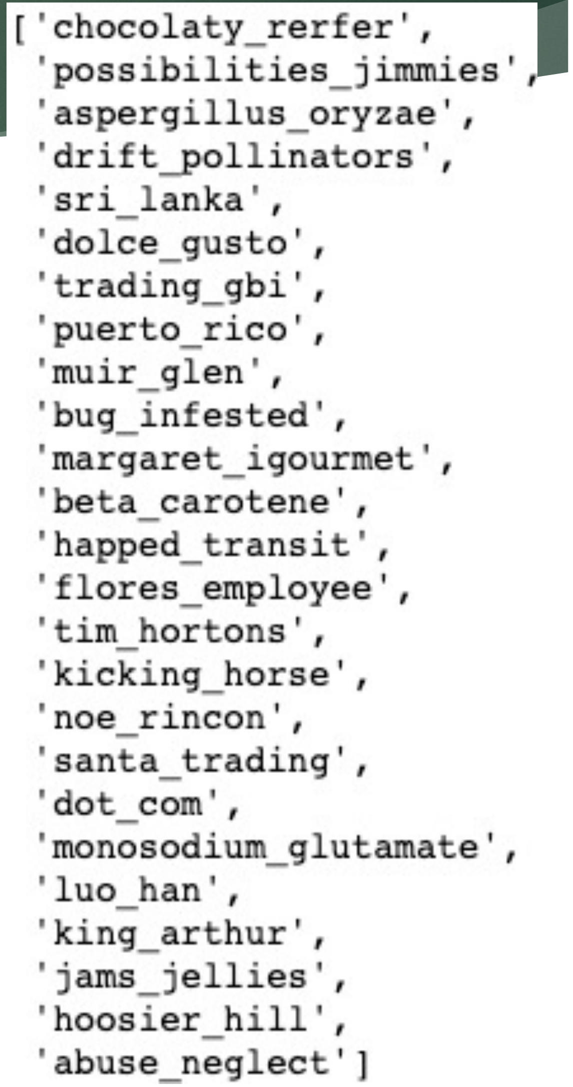
Top 25 Bigrams in Both Questions and Answers Scores Compared

Question Bigrams



```
['hamilton_beach',
 'dolce_gusto',
 'trader_joe',
 'agave_nectar',
 'genetically_modified',
 'wan_na',
 'fair_trade',
 'trans_fats',
 'tender_bites',
 'cross_contamination',
 'swiss_miss',
 'monosodium_glutamate',
 'snap_carrot',
 'san_francisco',
 'stainless_steel',
 'madagascar_bourbon',
 'super_automatic',
 'citric_acid',
 'rim_cupcakes',
 'kick_ass',
 'gummy_bears',
 'length_width',
 'subscribe_save',
 'heavy_metals',
 'health_benefits']
```

Answer Bigrams

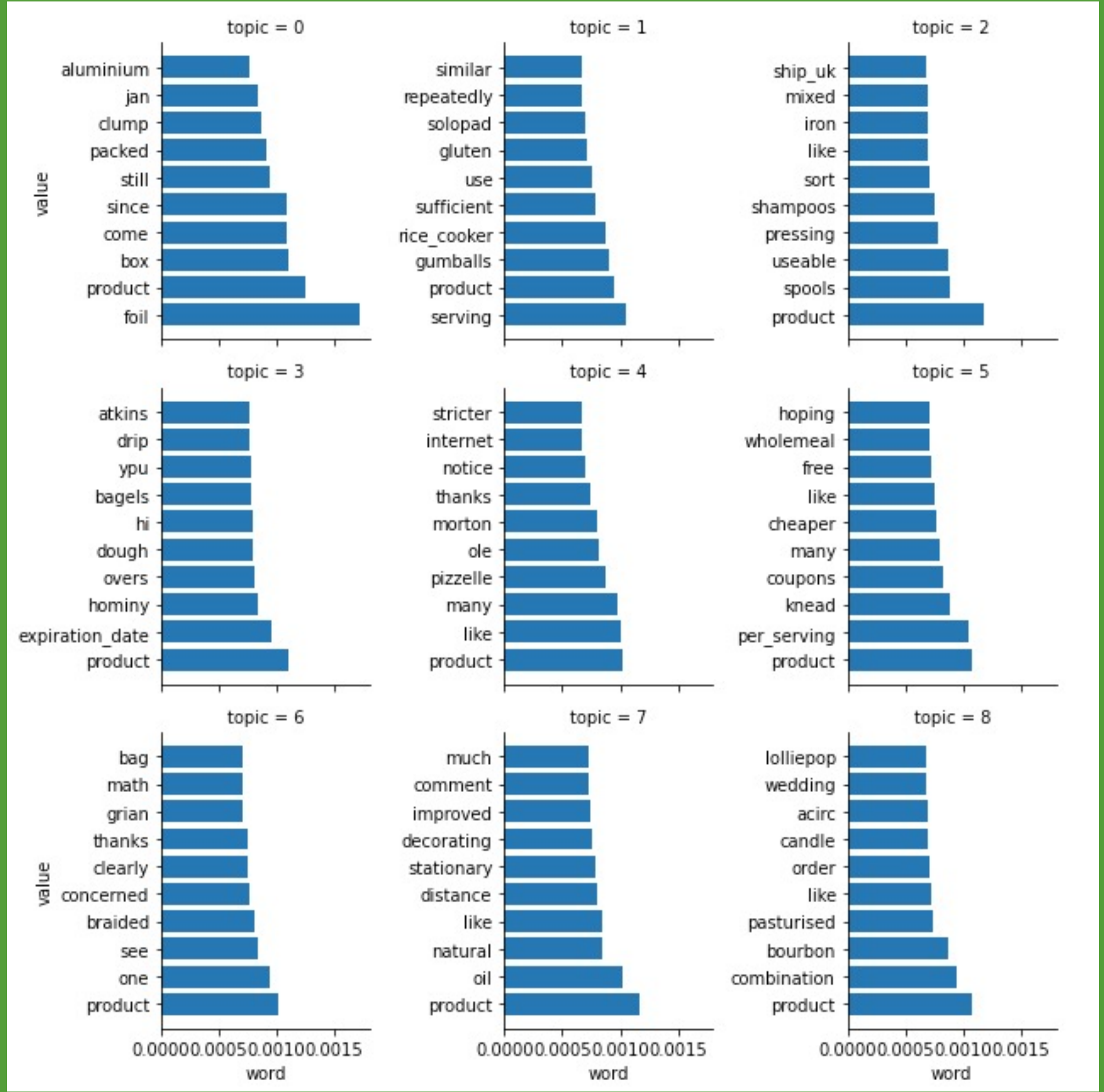


```
['chocolaty_rerfer',
 'possibilities_jimmies',
 'aspergillus_oryzae',
 'drift_pollinators',
 'sri_lanka',
 'dolce_gusto',
 'trading_gbi',
 'puerto_rico',
 'muir_glen',
 'bug_infested',
 'margaret_igourmet',
 'beta_carotene',
 'happed_transit',
 'flores_employee',
 'tim_hortons',
 'kicking_horse',
 'noe_rincon',
 'santa_trading',
 'dot_com',
 'monosodium_glutamate',
 'luo_han',
 'king_arthur',
 'jams_jellies',
 'hoosier_hill',
 'abuse_neglect']
```

Topic Modeling (Gensim)

Question Topics

1) Product understanding:
2) Product match to equipment:
3) product quality or clarity?:
4) Product shipment and conditions:
5) Customer complaints on product use:
6) Quantity and Sizing Understanding:
7) Product sourcing conditions:
8) Customer quantity concerns:
9) Customer complaints on product use:

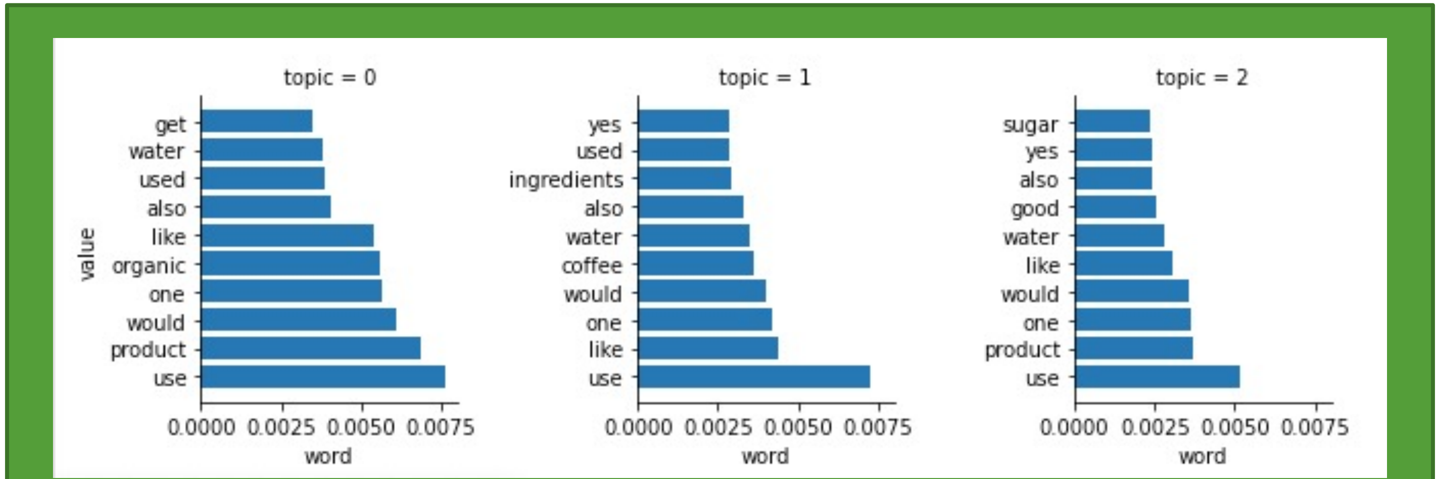


RTSpoonmore 2021

Topic Modeling (Gensim)

Answer Topics

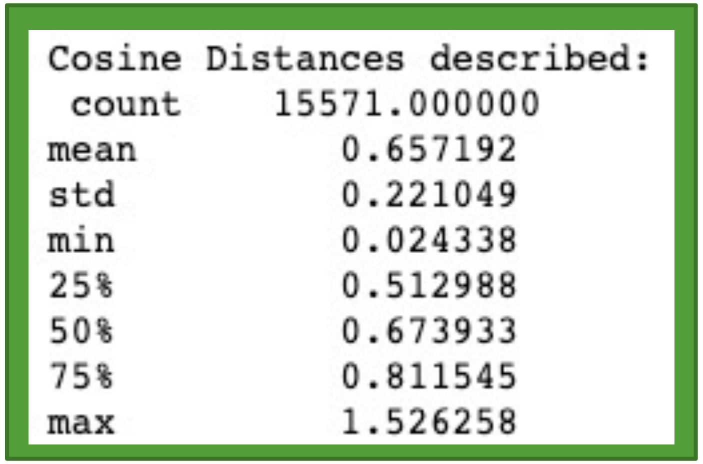
1) Product use instructions:
2) Product Ingredients Detailed listing:
3) Personal descriptions of product or
   competitor use:



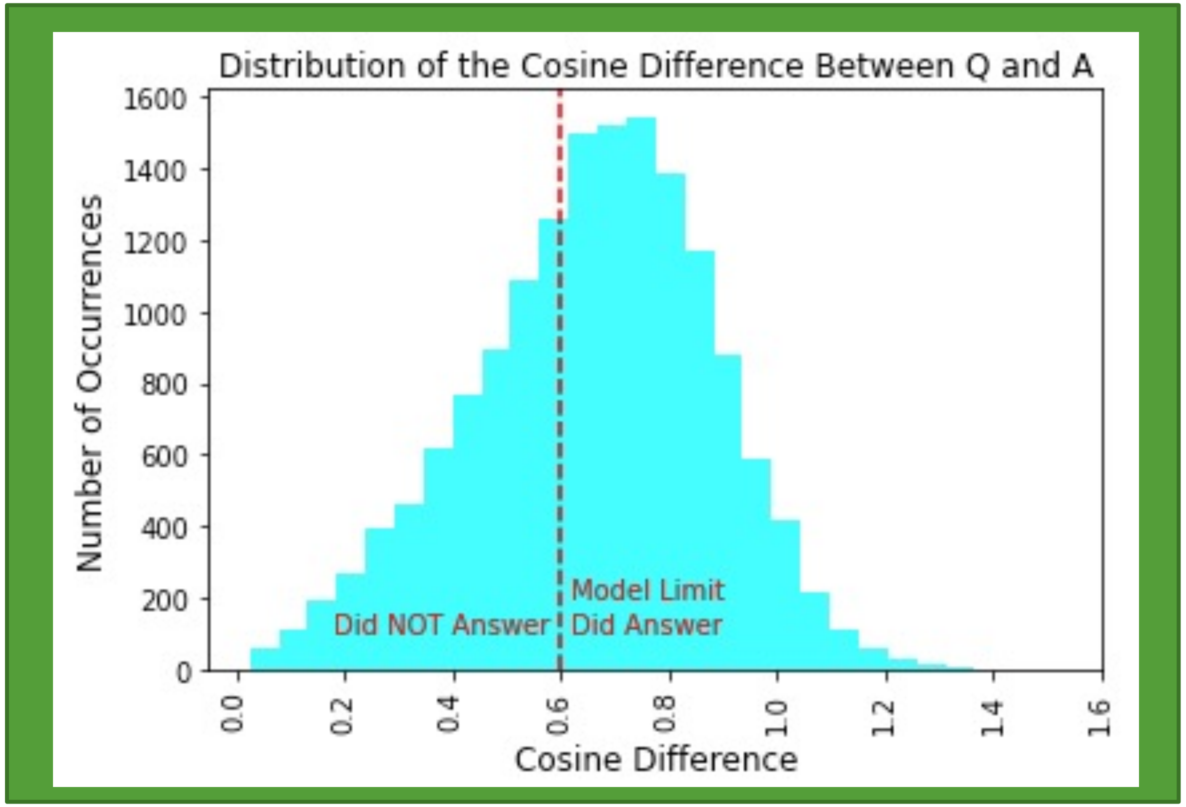
Cosine Difference

Modeling

1) Gensim Doc2Vec
2) Vectorized Questions and Answers
3) Cosine difference between Q & A
4) Set model limit to 0.60



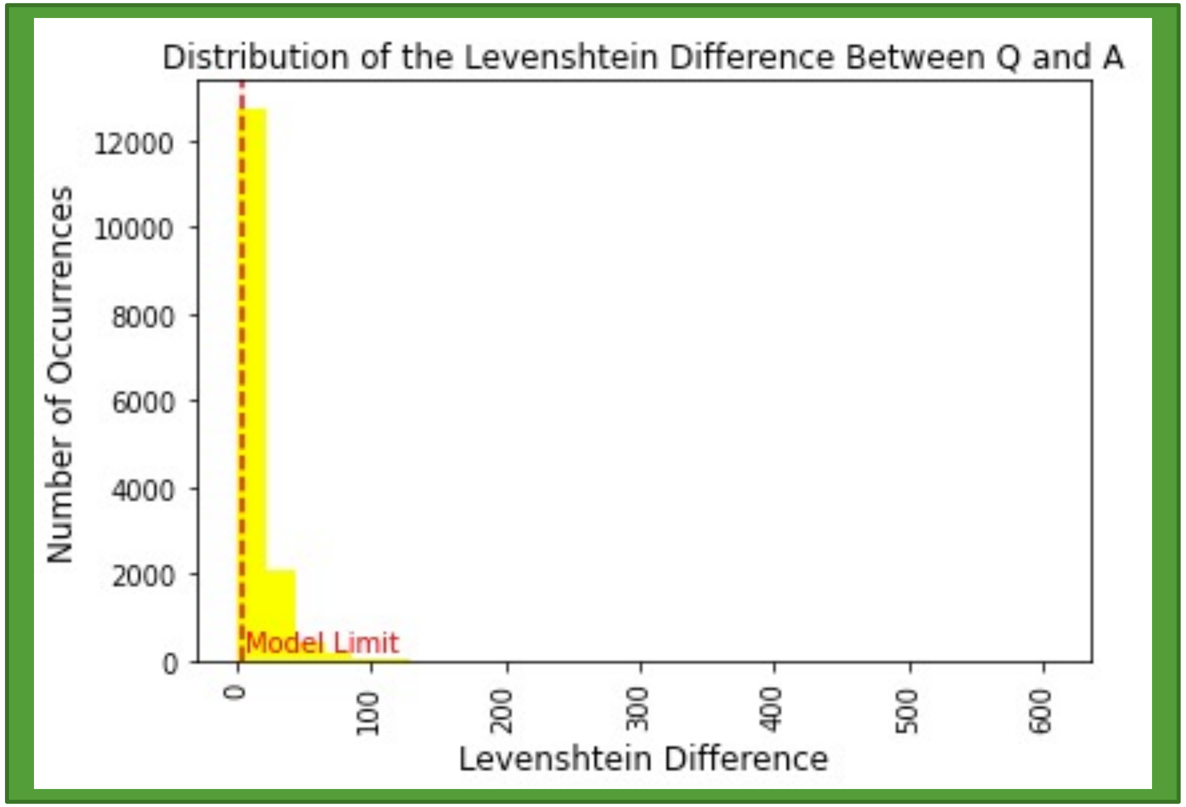
```
Cosine Distances described:
 count     15571.000000
mean          0.657192
std           0.221049
min           0.024338
25%           0.512988
50%           0.673933
75%           0.811545
max           1.526258
```

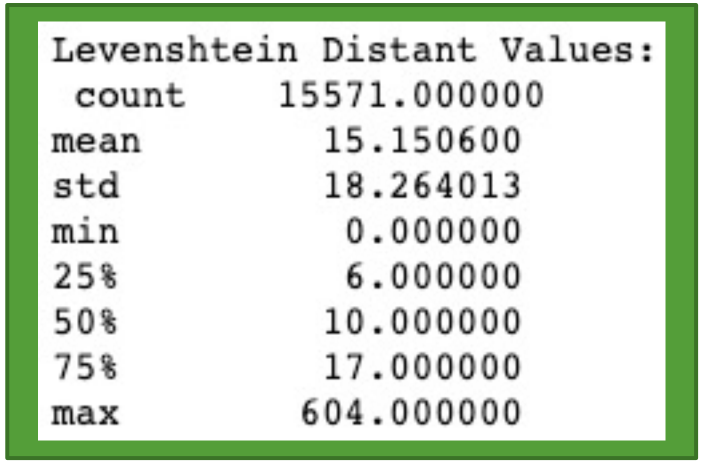


Levenshtein Distance

Modeling

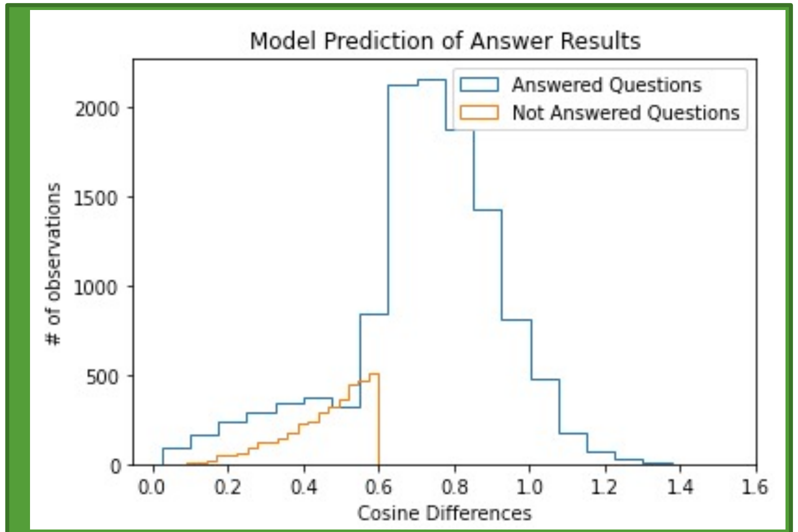
1) Levenshtein change difference
2) Determined word changes between Q & A
3) Noticed some Answers were restatements of Q but had only a few words difference. This was to add YES or NO to the question. These had low Cosine differences
4) Set model limit to 5

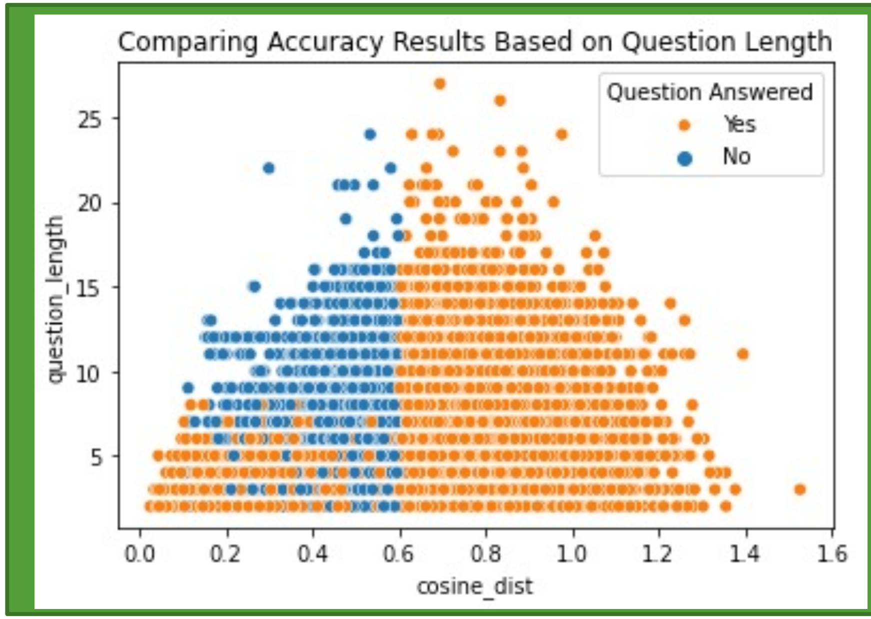


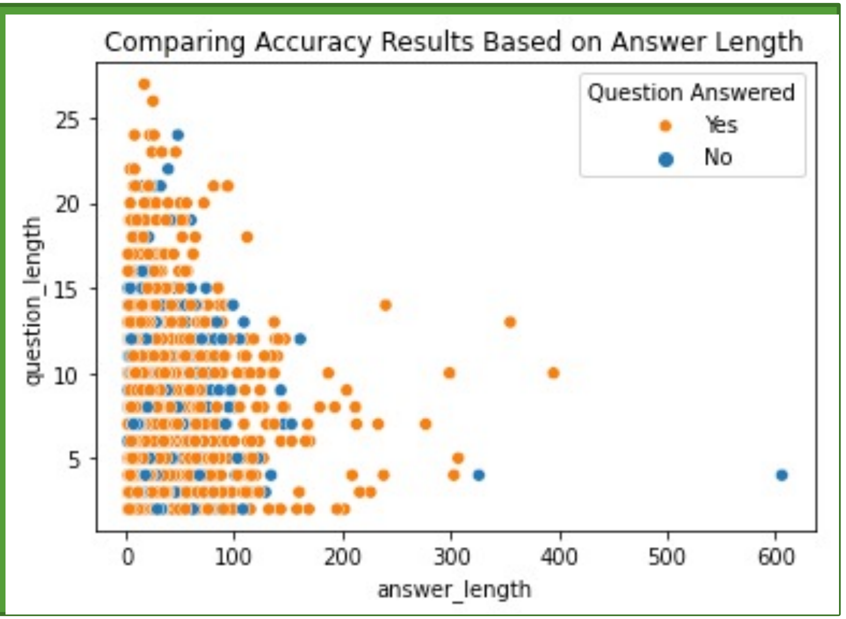


```
Levenshtein Distant Values:
 count     15571.000000
mean          15.150600
std           18.264013
min            0.000000
25%            6.000000
50%           10.000000
75%           17.000000
max          604.000000
```

Final Model Combined Cosine and Levenshtein Differences

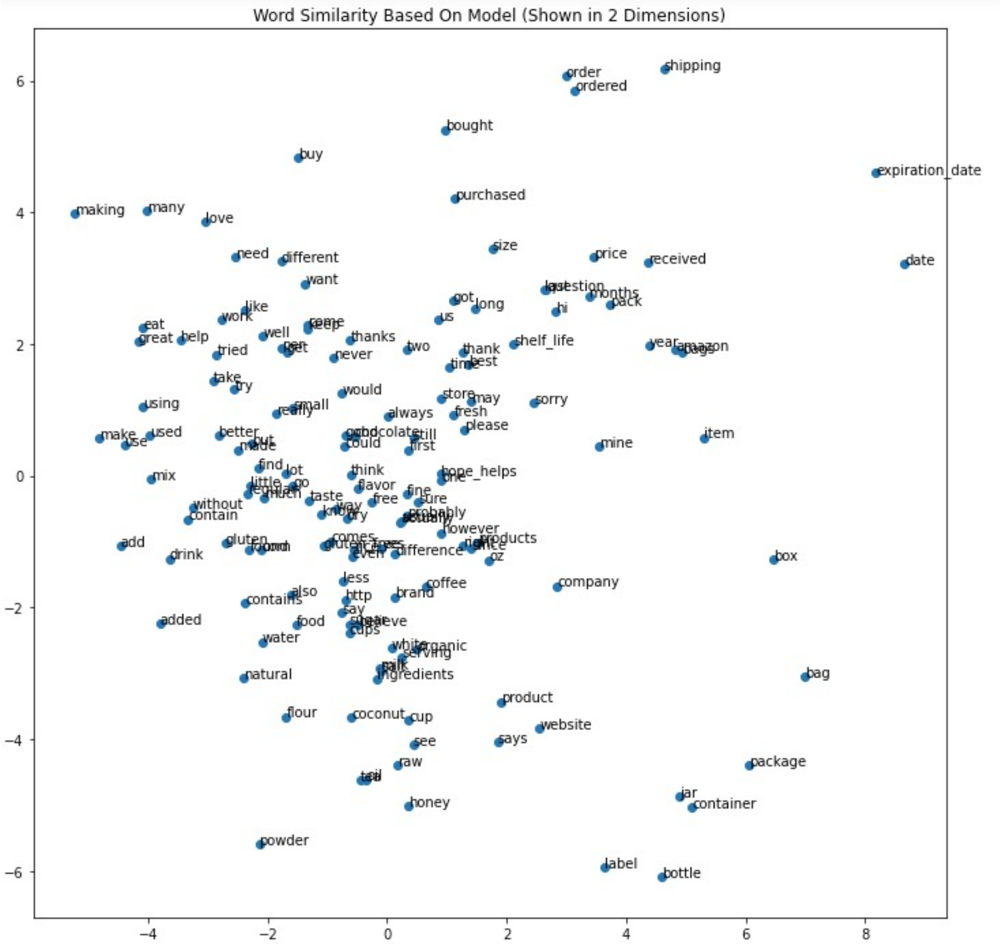

Model Prediction of Answer Results


Comparing Accuracy Results Based on Question Length


Comparing Accuracy Results Based on Answer Length

Combining the measures predicted the following:
11,834 Questions were answered
3,737 Questions were NOT answered correctly

What was learned from the process



Recommendations