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 Identification:

Do the Questions Asked by Customers in the Amazon Food Category Receive Correct Answers?

HTTPS://WWW.AMAZON.COM/FMC/LEARN-MORE?REF_=PRIMENOW

A Capstone project:

- Applying NLP modeling to Amazon Question and Answer banks
- Analyzing Unsupervised and Unstructured Text Fields
- Data: 2018 Data Extraction for Amazon category: Grocery and Fine Food
- Predicting the probability of a question being answered correctly
- Identifying the Categories of questions based on Topic Modeling

Data

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

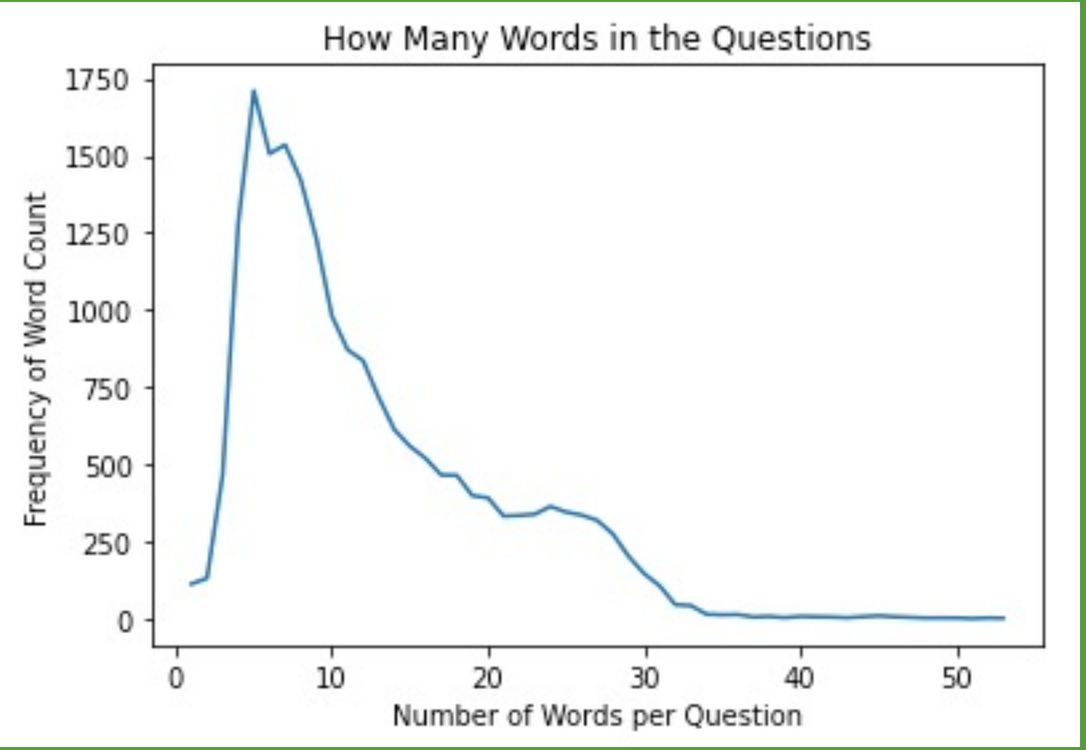
Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems
Mengting Wan, Julian McAuley International Conference on Data Mining (ICDM), 2016

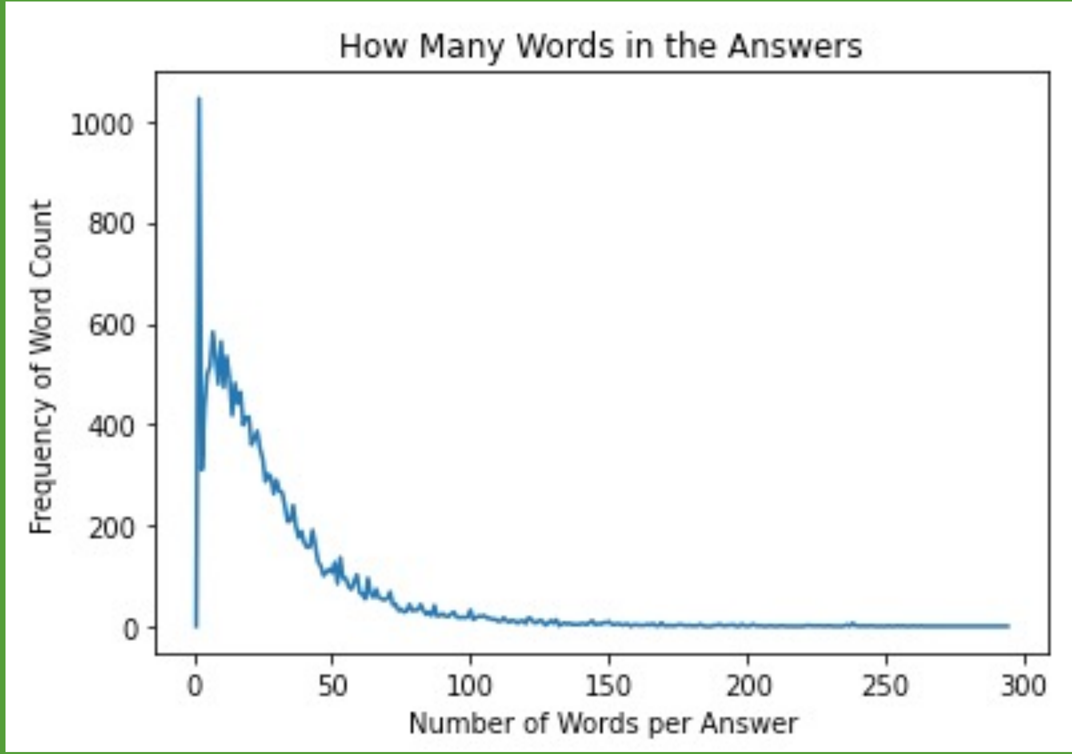
Addressing complex and subjective product-related queries with customer reviews Julian McAuley, Alex Yang World Wide Web (WWW), 2016

	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 expiración 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)

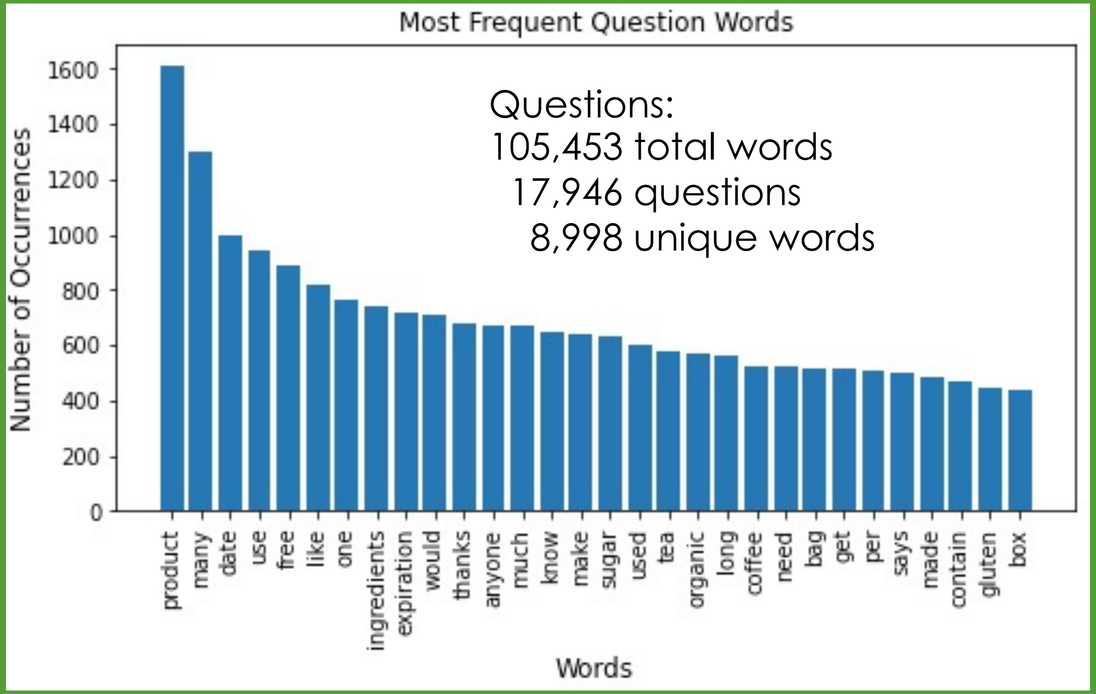
Exploratory Data Analysis

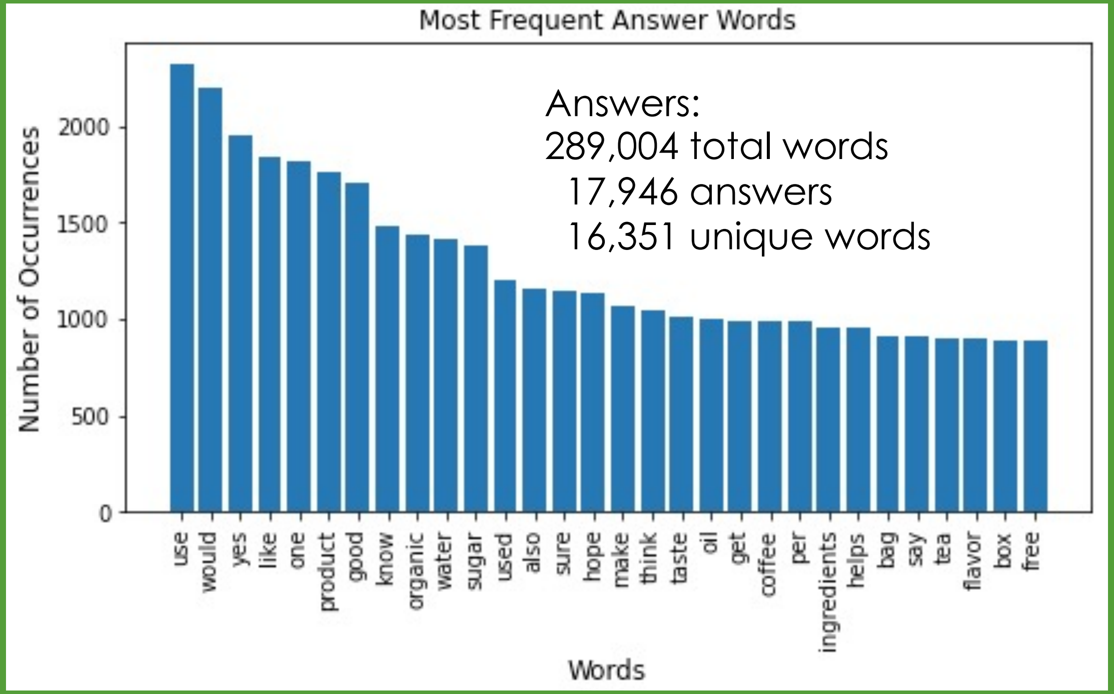




No null values found
Removed all 1 or 2 word questions and answers as these would bias models

Question Words versus Answer 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

Highest Frequency Question Bigrams

- Hamilton_beach
- Dolce_gusto
- Trader_joe
- Agave_nectar
- Genetically_modified

Highest Frequency Answer Bigrams

- Chocolaty_refer
- Possibilities_jimmies
- Aspergillus_oryzae
- Drift_pollinators
- Sri_lanka

TF-IDF Analysis (Sklearn)

Term Frequency - Inverse Document Frequency takes bag of words corpus and down weights words that appear most frequently

<u>Modeling</u>

1) Bag of words from Q & A
2) Fit and Transform to vectorize
3) Combine Corpus, Dictionary to TF-IDF weights
4) Show top values each, similar to Bag of Words frequency but weighting adjusted

Questions

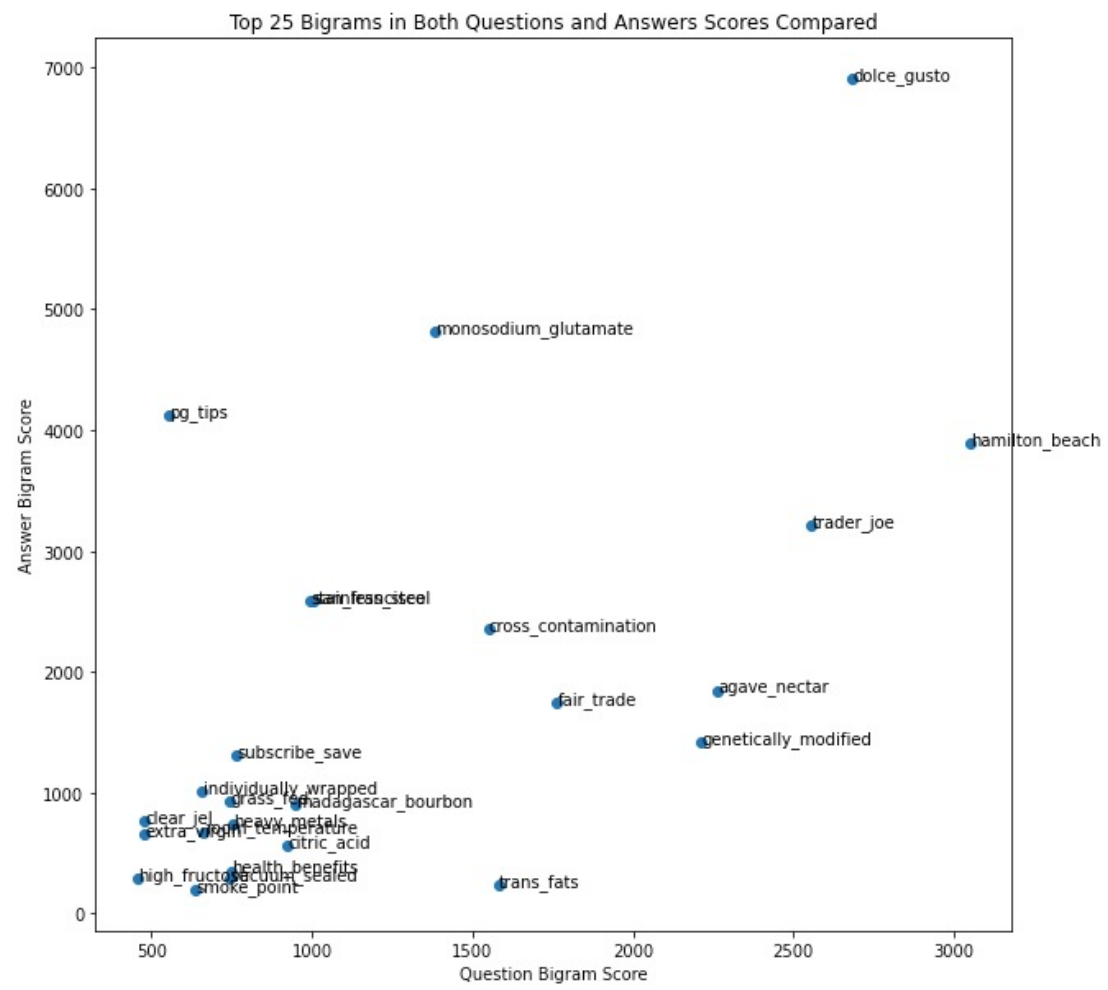
```
Top feature Names:
curry 0.5280719034995475
red 0.37460475940536125
yellow 0.4292527979030419
compared 0.4501747390819909
heat 0.4403363245862047
```

Answers

```
Top feature Names:
red 0.24963019415811116
profile 0.40116616287683743
flavor 0.2004650839356598
deeper 0.42718542776135887
green 0.2503323670052263
mild 0.3294613696529366
yellow 0.5893069023938735
think 0.18863489360450084
```

Bigrams of Text

Common Terms Associated with each other – combined as one



The bigrams shown have some clustering with similarity between Q and A,

However, some outliers show a bias

<u>*Higher Answer Use:*</u>

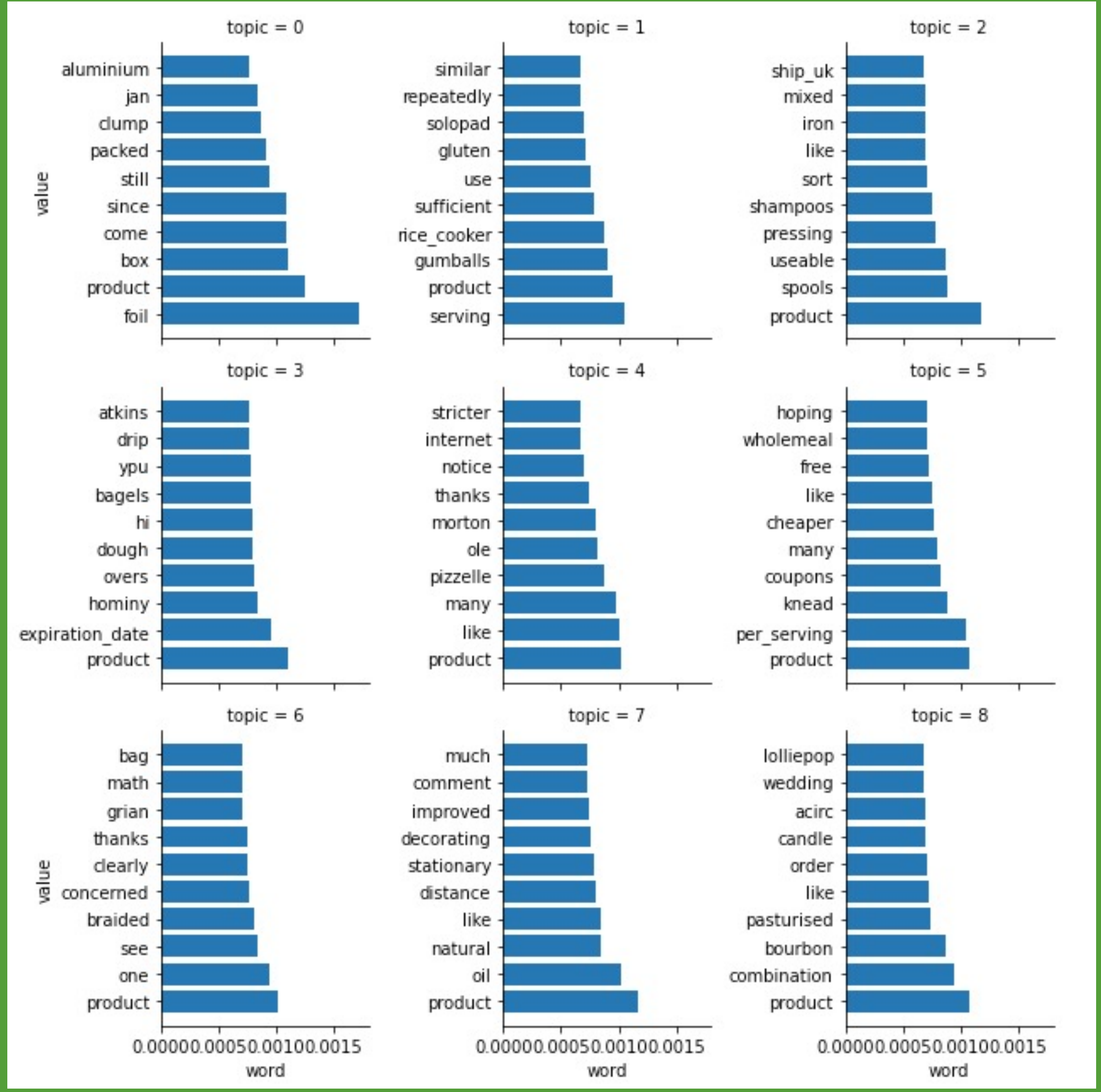
- Dolce_gusto
- Monosodium_glutamate
- Pg_tips
- Hamilton_beach
- Trader_joe

<u>*Higher Question Use:*</u>

- Trans_fats

Topic Modeling (Gensim)

Unstructured Data Predicted Categories



<u>Question Topics</u>

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

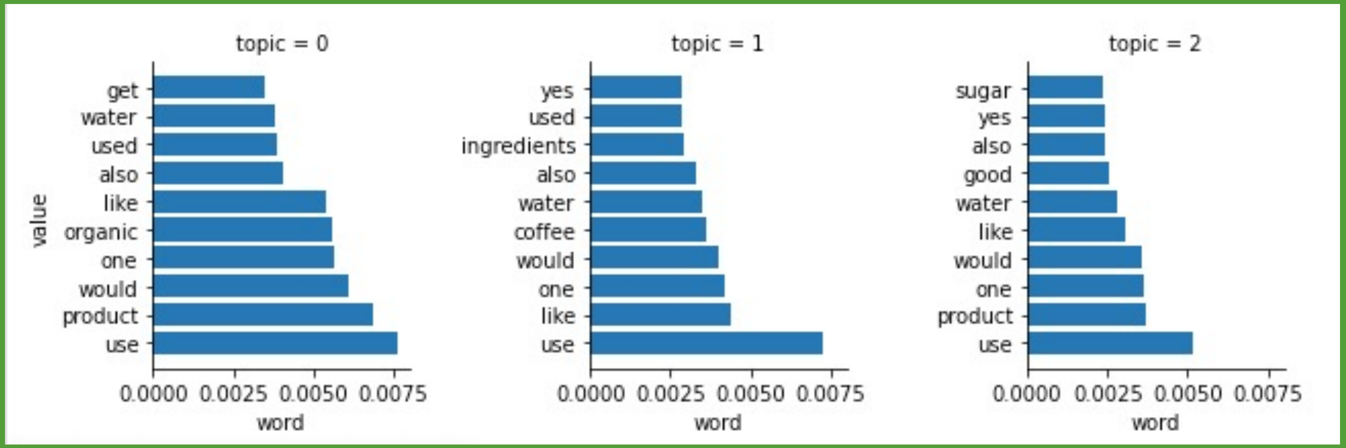
Analysis looked at larger grouping but found significance in these 9 topics for question words

Topic Modeling (Gensim)

Unstructured Data Predicted Categories

Answer Topics

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



Analysis looked at larger grouping but found significance in these 3 topics for answer words

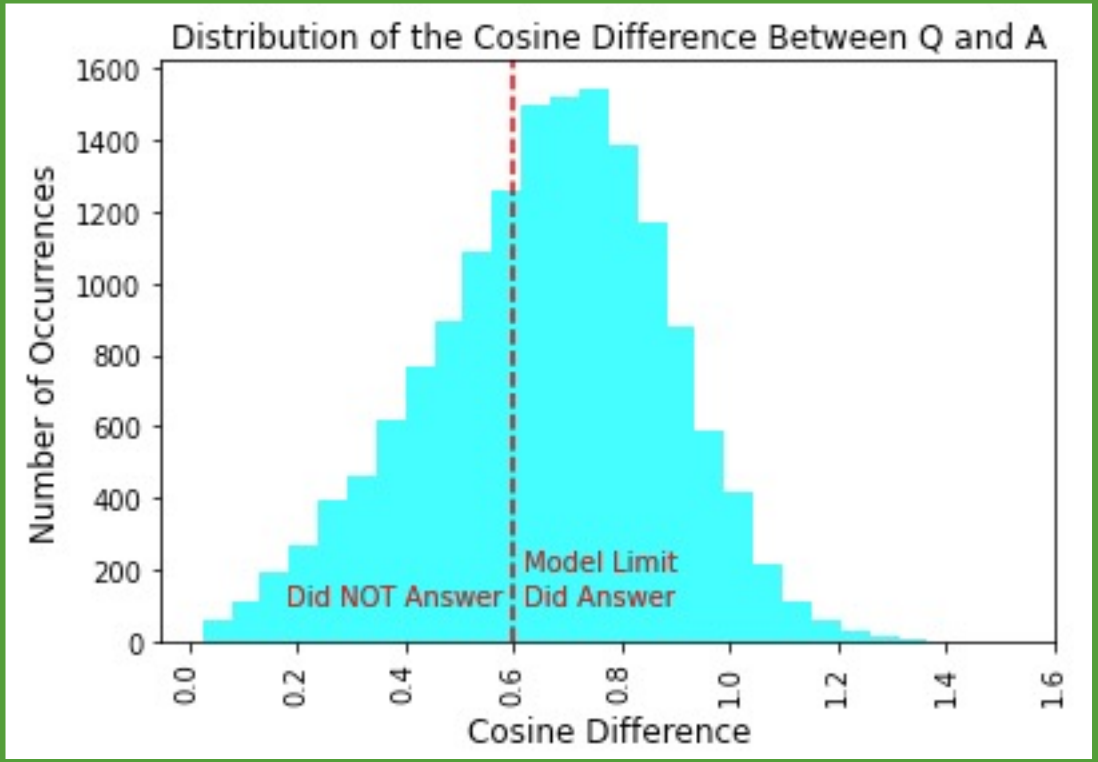
Cosine Difference

Similarity of Q and A vectors measured by Cosine of Angle between them (Highest is Most Similar)

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



Interactive review of questions and associated answers determined model limit of 0.6 or higher best predicted actual answers to questions

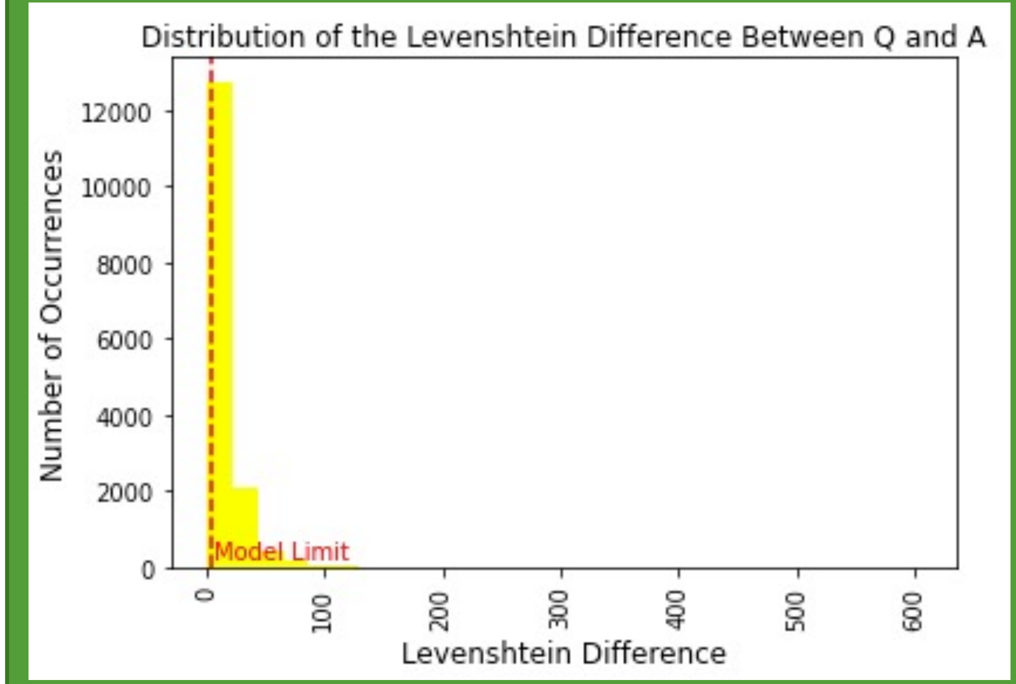
Levenshtein Distance

Similarity of Q & A vectors measured by cumulative change steps from Q words to A words (Lowest is Most Similar)

<u>Modeling</u>

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

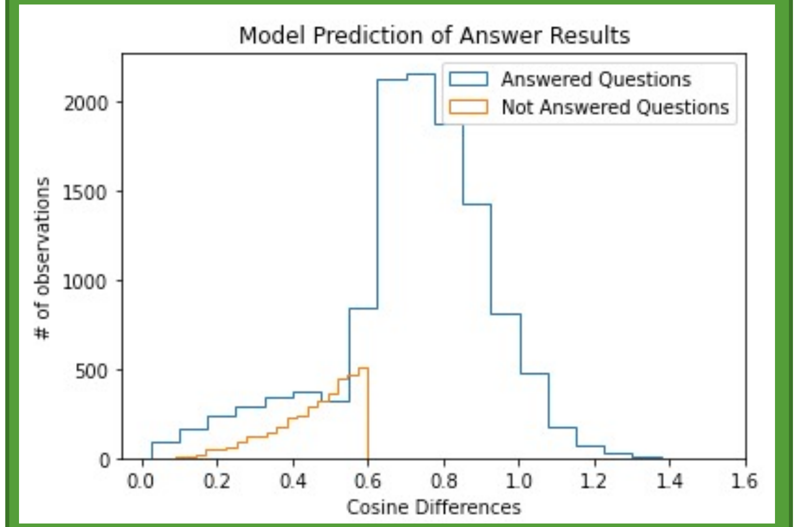
Levenshtein Distant Values:	
count	15571.000000
mean	15.150600
std	18.264013
min	0.000000
25%	6.000000
50%	10.000000
75%	17.000000
max	604.000000

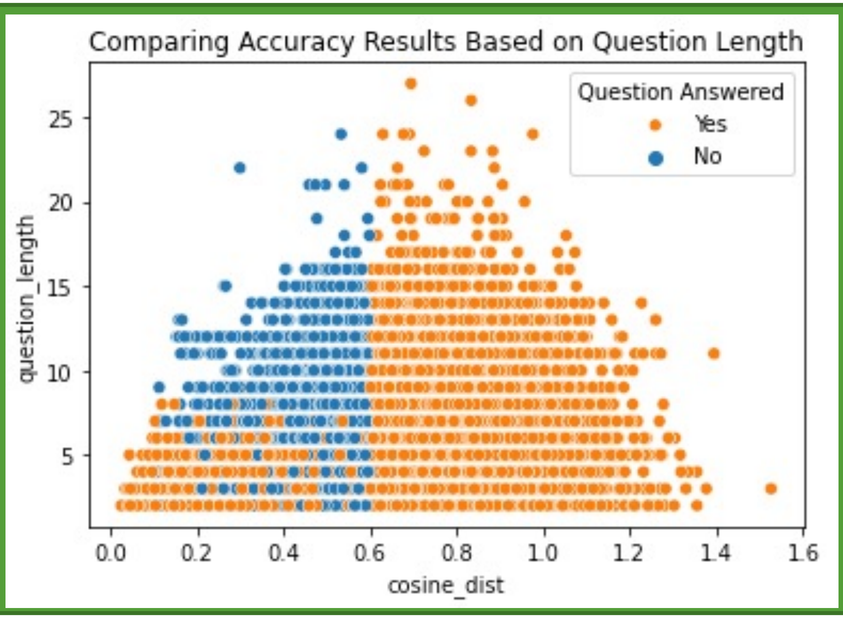


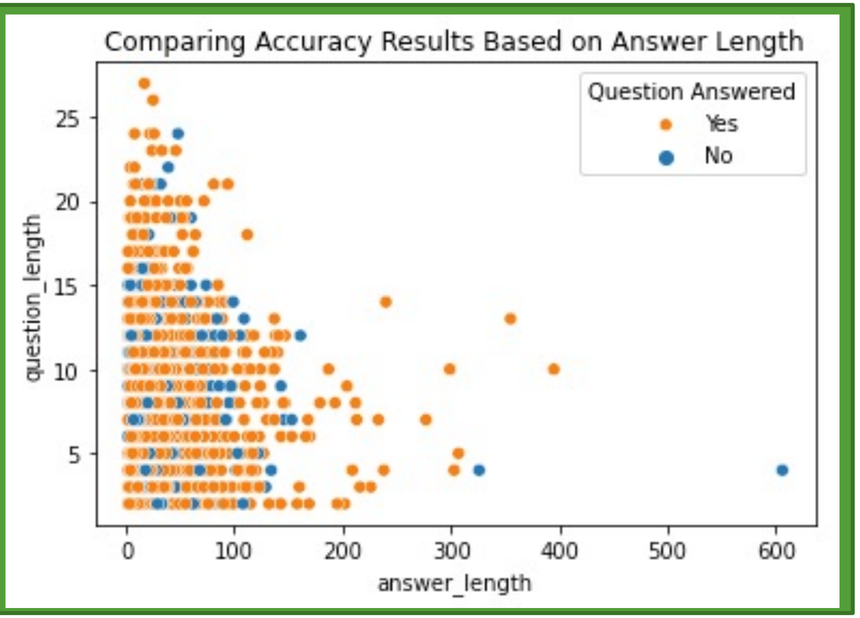
Interactive review of questions and associated answers determined model limit of 5 or lower best predicted actual answers to questions

Final Model Predictions

Combined Cosine and Levenshtein Differences







<u>Combining the measures predicted the following:</u>
11,834 Questions were answered. (76%)
3,737 Questions were NOT answered correctly. (24%)

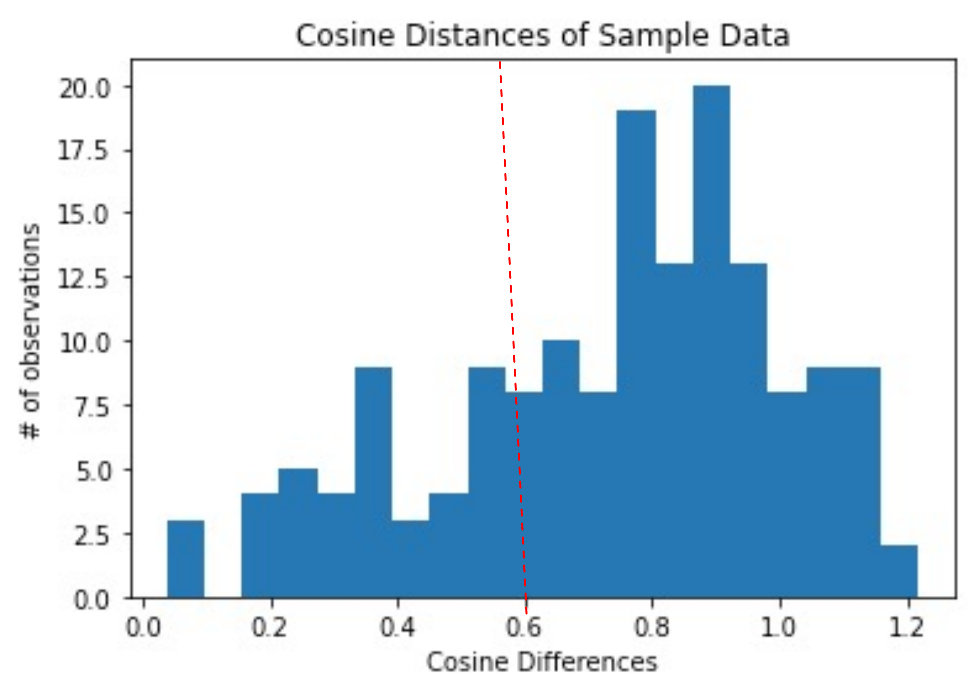
Test Model on New Data

Small sample set of new data (2021 sample of 160 Q & A's) random manual data samples across Amazon food categories

<u>Testing the Model</u>

1) 160 data points
2) Applied to previous corpus and dictionary
3) Results similar (76% correct previous data, 72% new data)

```
If correctness of answer is based on cosine distance less than:  0.6
Potential wrong answer count:  45
Potential right answer count:  115
Potential wrong answer percentage:  28.125
```



Interactive review of questions and associated answers determined model limit of 5 or lower best predicted actual answers to questions

BERT Modeling (torch)

Bidirectional Encoder Representations from Transformers is designed to pre-train from unlabeled text by jointly conditioning on both left and right context

<u>BERT Model</u>

1) Trained on same Q & A main dataset

2) Required to have Supervised Results: Added results column based on Cosine and Levenshtein

3) Results similar (76% correct previous data, 75.8% BERT)

4) Regression Score: 0.757. accuracy

```
LogisticRegression()

lr_clf.score(X_test, y_test)

0.7567428718212176
```

What I Learned from the Process

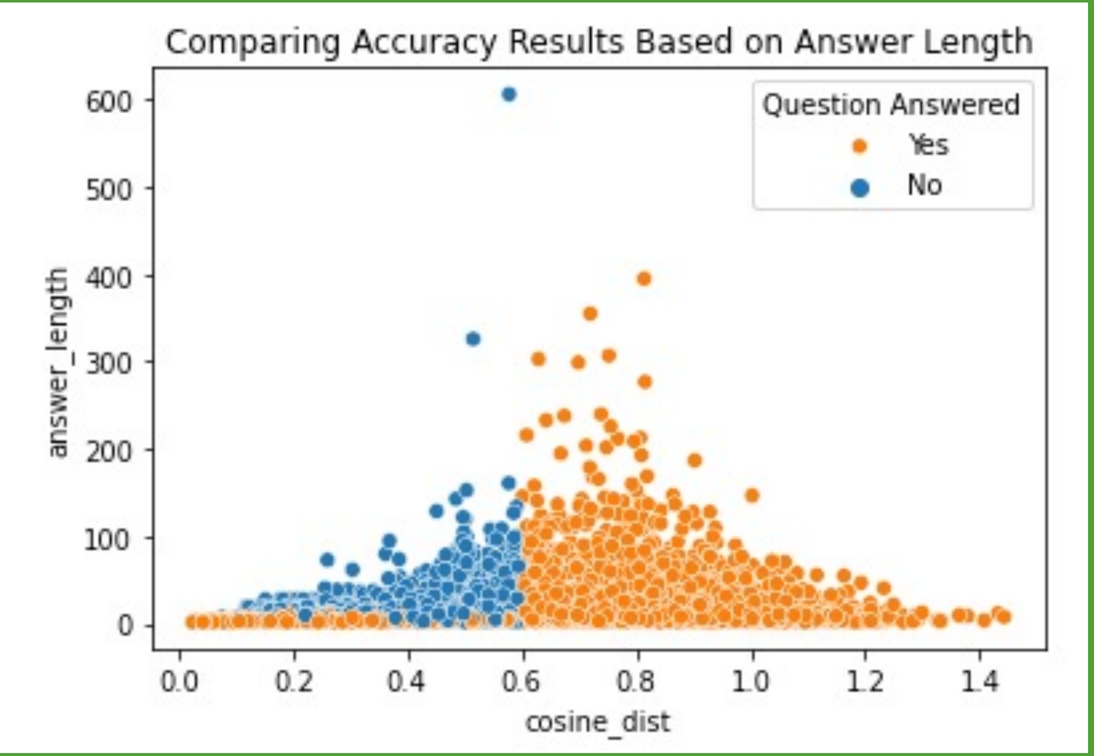
Reflections on Personal Lessons Learned

1) The analysis tools struggle with 1 & 2 word Questions and Answers: Removed from Model
2) Because unsupervised data, Each step required personal data review of Q&A alignments
3) Difference between Word2Vec and Doc2Vec models. Utilized Doc2Vec which required question bag of words to be in form of lists of lists
4) Bag of Words analysis required Questions and Answers to be in a flat file format
5) Maintained processed Question and Answer lists in list format to expedite processing time, as initial approach to build into dataframe memory structure overloaded my computer
6) Corpus is a list of indicies and counts, Dictionary is a list of indicies and word locations. Had to keep clear for each model to align with results to see how it related to each word
7) Visualizations of counts and scatter plots explain relationships much better than tables
8) Topic Modeling very manual: required review of all questions topic words to determine topic
9) There is a difference between Cosine Angle, Soft Cosine, and Cosine difference.
10) Levenshtein is very sensitive to text length, so long answers had big impact on measure
11) Cosine alone did not model dynamics completely
12) Bert Model had heavy system requirements and crashed my computer multiple times

Output Discoveries from the Process

Applied Learning from the Results

1) The Amazon data challenging: Prank questions, non serious answers, jargon, inconsistent slang, inconsistent styles, answers conversation style
2) A pretrained model failed to predict the results. First pass used Google trained model, but over 12,000 unique words in Amazon data were not found. Too many abbreviations and jargon
3) Some Answers just restated the Question with adding YES or NO, thus the levenshtein model was needed.
4) The question topics categories predict areas for improvement to lower number of question types
5) BERT model required heavy computing resources (20 minutes for main step to execute)



Recommendations

Next Steps

1) **<u>Amazon can prevent 2/3 of questions from being asked</u>** initially based on Topic Modeling:
   1) 1/3 of questions tied to <u>understanding product contents, counts, or packaging</u>. Make sure allergy potential and gluten contents is clearly identified for customer. Ensure ingredients listing clearly shown and counts or quantities verified
   2) 2/9 of questions on <u>shipping conditions or sourcing</u>. Make sure it is clear to customer origin of shipment (country) and conditions requirements (refrigeration)
   3) 1/9 of questions on <u>match to equipment</u>. Make sure alignment to equipment is clearly stated (pods for coffee makers, etc.)

<u>Of the remaining questions that will always be asked:</u>

2) 1/9 of questions tied to <u>product understanding</u>. Accuracy can be increased based on length and contents. Ensure answer does not contain jargon and is within a window for word count (3-100 words). Add check for entry size on answer box
3) 2/9 of questions are <u>customer complaints</u>, and should be treated separately than simple answers. Utilize modeling to collect these topics for improved customer appreciation