Product Reviews Summarizer

SWENG 545 Data Mining

Overview

Companies like Amazon have millions of reviews on their website, from customers all over the world. Given the nature of the site and the fact that their users are looking for the best product to buy, having to sift through hundreds of reviews to find a product can be a real turn off.

Text Analysis and Clustering can be used here to build tools that can summarize multiple properties in 1-2 sentences. Instead of scrolling through a list of reviews, you could simply say “Excellent product but has a history of bad packaging”.

Target audience

Anyone who wishes to use online platforms to shop items can use this tool to quickly find out the summary of the thousands of product reviews.

Data

Collection

- Amazon.com
- Splash container
- BeautifulSoup & html.parser
- Variety of products

In [57]:
```
def get_soup(url):
    r = requests.get('http://localhost:8050/render.html', params={'url': url, 'wait': 2})
    soup = BeautifulSoup(r.text, 'html.parser')
    return soup


def get_reviews(soup):
    reviews = soup.find_all('div', {'data-hook': 'review'})
    try:
        for item in reviews:
            review = {
            'product': soup.title.text.replace('Amazon.com: Customer reviews:', '').strip(),
            'title': item.find('a', {'data-hook': 'review-title'}).text.strip(),
            'rating':  float(item.find('i', {'data-hook': 'review-star-rating'}).text.replace('out of 5 stars', '').strip()),
            'body': item.find('span', {'data-hook': 'review-body'}).text.strip(),
            }
            reviewlist.append(review)
    except:
        pass
```

In [58]:
```
for x in range(1,999):
    soup = get_soup(f'https://www.amazon.com/JBL-Bluetooth-Portable-Stereo-Speaker/product-reviews/B01N1RZC7S/ref=cm_cr_arp_d_pag
    print(f'Getting page: {x}')
    get_reviews(soup)
    print(len(reviewlist))
    if not soup.find('li', {'class': 'a-disabled a-last'}):
        pass
    else:
        break
```

	A	B	C	D
1	product	title	rating	body
12	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	Not what we expected	1	They make this look like the mousse actually coats the seed and comes out like mousse, however it sprays out just like colored water and does not look hose is not sealed so water leaked the whole time we were seeding. We just did this today, I will post later if the grass does/doesn't come up.
13	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	Doesn't work on "any surface" and takes longer than a week to show any improvement.	3	Purchased for Central Florida in JuneIt has been 4 weeks since I used this on a 20 x 50 section of yard (mostly on bare spots).Step 1: Carefully read and r Create a plan on how and where you want to spray the seedsStep 2: Closely mowed the area where we intended to use the sprayStep 3: Follow the dire starting with the worst grassless patch and work down in size. The water color should be green letting you know where you sprayed.Step 3: Wait 1, 2, 3, grass now but not thick and luxurious like the commercial after one week.There was no thick mousse while spraying but the coloring did make it easier t were poking through at 1 week, at two weeks, even more blades were poking through.We did not water during this time frame as we get afternoon and inch and a half of rain during the first week and two inches the second week.There is one spot where were a tree had been removed and we have had n now.The lawn has been mowed twice and there is definitely a thicker feel and look to where the thin spots were except for the one where the tree was to see where you have been but you need about quadruple the recommended patience to see real results.
14	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	Didn't work	2	Didn't work, did nothing and followed the instructions to the T. I even bought another bag to put with it and still nothing. Expected better quality and got
15	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	Garbage, junk, mess, no grass, waste of $$	1	Seller got my green cash! I got green hands, side walk, driveway, clothes, and gym shoes! Oh yes! I got NO green grass! LOL! Amazon should be ashamed NOT PURCHASE THIS PRODUCT!

	A	B	C	D
1	product	title	rating	body
11	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	Worth the money	4	Your browser does not support HTML5 video. *I was fully reimbursed for this iron* I will be honest with you guys, I thought I got a broken iron. I couldn't get it hot enough to iron a simple cotton dress or get the steam option to work. Took a good 20 minutes before I realized I was aligning the dial in the wrong spot (I noted where it should be aligned). Then things started working correctly. I could hear the steam getting ready and it was able to get wrinkles out almost immediately with no steam needed. This will most likely replace my 20$ Wal-Mart iron I bought 4 years ago. I used to covet Rowenta irons, still wouldn't mind one but this one will do fantastically until then. The auto shut off function does work in the horizontal potistion but I didn't test vertical becasue I have a curious toddler at home. It's got the option to vertically steam (which is so cool but it scared me the first time I tried), uses tap water only (woo no more jugs of distilled water!!) and the cord winds up in the base *this does make it a bit heavier than a standard iron. * It has an anti drip system but it will drip if you, like me, test out the steam button on the low setting. It needs to be on high. All said and done I would recommend buying for an "I just moved out and I need an iron" or "I need a replaced workhorse to complete my sewing projects". I already own a MUELLER slicer and this iron is just as good quality as that is.
12	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	Am I Just Stupid?	3	It seems like a good iron. It's heavy and much larger than I thought it would be (wish it were smaller). I don't do a lot of ironing even though I've been using one for 50 years. I dropped my old iron and it broke so I purchased this one. To be frank, this iron is a bit too complicated for my needs and since I don't use it that often, I have to pull out the manual each time I use it because I don't remember what the buttons and dials are for. There's no ON/OFF switch and the dial is behind the handle which makes it difficult to read and there's no easy to read arrow marking the dial setting. I had to take some nail polish and paint the raised arrow. The only easy to read label is the Mueller brand on the side of the iron!The cord doesn't retract like it should. I have to feed the cord into the hole for storage.I will keep this and use it until it dies but I would have preferred a much smaller, simpler iron with easy to read buttons and dials.
13	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	Big improvement over my last iron	5	I iron 5 days a week,and had a series of irons from a well know brand that starts with "sun". They were just OK and the last one quit on my with a puff of smoke after 2 years. Supposedly it was under warranty but the process was so cumbersome, I gave up and went in search of a new and different model. I like having a retractable cord and not winding it around a hot iron.So far, so great with this new Mueller! The water tank is much larger, the tip has a perfect shape to get in and around the buttons and tight spots on my shirts. My old iron was listed as "large" but this one has the same heated surface area and feels better in my hand.I would recommend to a friend

Data

Pre-processing

- Converting to string datatype
- Tokenizing
- Replacing non-alphabetical characters with blank character
- Converting to lower-case
- Removing stopwords

In [125]:
```
# Removing Stopwords like : 'i', 'me', 'my', 'myself', 'we', 'our', 'ours'
def remove_stopwords(sen):
    sen_new = " ".join([i for i in sen if i not in stop_words])
    return sen_new
```

In [126]:
```
def my_summarizer(dataframe):

    sentences = []
    for s in dataframe['body']:
        sentences.append(sent_tokenize(s))    #Tokeninzing the sentences

    sentences = [y for x in sentences for y in x] # Flattening the sentences list

    #Replacing Non-Alphabetical characters with empty string
    clean_sentences = pd.Series(sentences).str.replace("[^a-zA-Z]", " ", regex=True)
    clean_sentences = [s.lower() for s in clean_sentences] # Converting sentences to Lower-Case

    clean_sentences = [remove_stopwords(r.split()) for r in clean_sentences]
```

Methodology

Cosine Similarity w/o tf-idf

- Pretrained GloVe Word Embeddings
- Cosine Similarity Algorithm
- Pagerank Algorithm
- $O(n^2)$

```
# Vector Representation of Sentences
sentence_vectors = []
for i in clean_sentences:
    if len(i) != 0:
        v = sum([word_embeddings.get(w, np.zeros((100,))) for w in i.split()])/(len(i.split())+0.001)
    else:
        v = np.zeros((100,))
    sentence_vectors.append(v)

print(len(sentence_vectors))
# Similarity Matrix Preparation
sim_mat = np.zeros([len(sentences), len(sentences)])

for i in range(len(sentences)):
    for j in range(len(sentences)):
        if i != j:
            sim_mat[i][j] = cosine_similarity(sentence_vectors[i].reshape(1,100), sentence_vectors[j].reshape(1,100))[0,0]

# Applying PageRank Algorithm
nx_graph = nx.from_numpy_array(sim_mat)
scores = nx.pagerank(nx_graph)

ranked_sentences = sorted(((scores[i],s) for i,s in enumerate(sentences)), reverse=True)



sn1 = 10
sn2 = 5
sn3 = 1

try:
    text = ''
    for i in range(sn1):
        text = text + ranked_sentences[i][1]
except:
    try:
        text = ''
        for i in range(sn2):
            text = text + ranked_sentences[i][1]
    except:
        try:
            text = ''
            for i in range(sn3):
                text = text + ranked_sentences[i][1]
        except:
            pass

return text
```

Methodology

Cosine Similarity with tf-idf

- tf-idf measure
- Top 1000 using nlargest module from heapq library
- No loss of important features

```
for word in doc:
    if word.text.lower() not in stop_words:
        if word.text.lower() not in punctuation:
            if word.text not in word_freq.keys():
                word_freq[word.text]= 1
            else:
                word_freq[word.text]+= 1
    #print(word_freq)

x=(word_freq.values())
a=list(x)
a.sort()
max_freq=a[-1]

for word in word_freq.keys():
    word_freq[word]=word_freq[word]/max_freq
#print(word_freq)

sent_score={}
sent_tokens=[sent for sent in doc.sents]
#print(sent_tokens)

for sent in sent_tokens:
    for word in sent:
        if word.text.lower() in word_freq.keys():
            if sent not in sent_score.keys():
                sent_score[sent]=word_freq[word.text.lower()]
            else:
                sent_score[sent]+= word_freq[word.text.lower()]
#print(sent_score)


sentences = nlargest(n=1000,iterable=sent_score,key=sent_score.get)
#print(summary)
```

Methodology

Hugging Face Transformers

- Top 10 from cosine similarity
- Creates an extractive summary
- Min_length = 75 and max_length = 300

```
summarizer = transformers.pipeline("summarization")
```

```
summarized1 = summarizer(text1, min_length=75, max_length=300)
summarized2 = summarizer(text2, min_length=75, max_length=300)
summarized3 = summarizer(text3, min_length=75, max_length=300)
summarized4 = summarizer(text4, min_length=75, max_length=300)
summarized5 = summarizer(text5, min_length=75, max_length=300)
```

```
countList =[]
percentageList = []
i=1
while i<=5:
    countList.append(len(df[df['rating']==i]))
    calculatedValue = (round(float((len(df[df['rating']==i]))/len(df)),3))*100
    percentageList.append(calculatedValue)
    i=i+1
```

```
df_review = pd.DataFrame(columns = ['product', 'rating', 'summary'])
df_review = df_review.append({'product': df['product'][1], 'rating': '1', 'summary':
df_review = df_review.append({'product': df['product'][1], 'rating': '2', 'summary':
df_review = df_review.append({'product': df['product'][1], 'rating': '3', 'summary':
df_review = df_review.append({'product': df['product'][1], 'rating': '4', 'summary':
df_review = df_review.append({'product': df['product'][1], 'rating': '5', 'summary':
```

	A	B	C	D	E	F
1	product	rating	summary	Counts	Percentage of Total	Most Common Words
2	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	1	this is my first 1 star I think if not its been along time any ways don't buy this product its a waste of you hard earned money its just green paint in a bag with seeds I followed directions and got nothing but covered in green stain that is very hard to remove . this product doesn't work at all like it says it's been several weeks and I still have no grass sprouting up call the water I put on my guess it's just a bad batch of grass.	2631	76.7	('Product', 1003) ('Grass', 854) ('Money', 834) ('Work', 810) ('Seed', 726)
3	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	2	it only grew two small areas and has not grown any higher even if you water it and in the front of the house area I spray one bag and water it like they say to do it . I will never buy it again and tell people about it.One bath and countless hand washings later, I'm still green. we believe this is a scam product, better off buying a bag of dirt, and a bags of seed, old fashioned way.	292	8.5	('Grass', 134) ('Product', 99) ('Seed', 76) ('Work', 72) ('Green', 53)
4	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	3	all the bare spots except 1 have grass now but not thick and luxurious like the commercial after one week . the water color should be green letting you know where you sprayed the seed . if you buy this, be extremely careful of the additive packet, pour it outside away from everything . it's not right time for this product unless you have good water sources you are willing to use.	162	4.7	('Grass', 73) ('Seed', 52) ('Green', 33) ('Work', 26) ('Like', 26)
5	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	4	easy to use and grass is already coming up in just a week . I used this product 12 days ago and have started to see new grass in spots where nothing existed before . it came up in thick and stayed green all year round, even under the snow, decided to do entire yard after having some waterline repairs . by end of last summer my grass had started growing in pretty good and it was very thick and green	96	2.8	('Product', 38) ('Grass', 34) ('Good', 20) ('Seed', 20) ('Well', 18)
6	Hydro Mousse Liquid Lawn System - Grow Grass Where You Spray It - Made in USA	5	the product worked great on my patchy lawn . it grew green grass within a week and looks great . the green is just to blend any already dead grass while new grass grows but more importantly shows you where you sprayed the seed to get an even spread and lastly sticks it to the dirt so it's not just sitting on top.	251	7.3	('Grass', 70) ('Product', 66) ('Great', 58) ('Works', 45) ('Use', 44)

	A	B	C	D	E	F
1	product	rating	summary	Counts	Percentage of Total	Most Common Words
2	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	1	the iron is not new it's used and it doesn't work . after two months the iron started smelling like the plastic was burning . the iron was dirty like a dusty hand grabbed it. this is the second time that I have bought an iron on amazon . it has come showing obvious where in tear from repeated use .	183	8.8	('Iron', 181) ('Water', 78) ('Cord', 58) ('Used', 56) ('Months', 39)
3	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	2	the iron is heavy, it never turns off and its awkward, and I don't care for the retraceable cord . it replaced a 1,500-watt iron on which the power cord had failed, but otherwise did a splendid job . i could deal with a $25 iron that worked ok but this cord messing up the fabric has taught me to look for one not in that position .	68	3.3	('Iron', 68) ('Cord', 45) ('Water', 31) ('Used', 20) ('Like', 20)
4	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	3	the iron works well but the retractable plug wouldn't go back in . the iron was crinkled and stiff like paper that got wet, and then dry . a water chamber was partially filled, and the source of the water damage . I don't do a lot of ironing even though I've been using one for 50 years .	80	3.9	('Iron', 86) ('Cord', 60) ('Water', 20) ('Heavy', 19) ('Steam', 19)
5	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	4	the iron is good the steam it does to remove wrinkles is good but the water comes out and wets the clothes more than normal . if you are using it for sewing I would use one that doesn't automatically turn off, but that's a nice thing about this iron if your forgetful.the iron does not fall and I was taking care of it so that it would not be damaged but it was.	207	10	('Iron', 249) ('Cord', 116) ('Steam', 74) ('Retractable', 64) ('Water', 61)
6	Mueller Professional Grade Steam Iron, Retractable Cord for Easy Storage, Shot of Steam/Vertical Shot, 8 Ft Cord, 3 Way Auto Shut Off, Self Clean	5	great iron,have it for a week and used every day, a very good product for its price,the water capacity is perfect and it makes good amount of steam .I have never had an iron with a retractable cord, it seems like a small detail but definitely helps for an easy clean up especially because I pull out the iron for last-minute situations . this iron is well made, it's light weight enough for easy use, but doesn't feel cheaply made .	1537	74.1	('Iron', 1418) ('Great', 611) ('Love', 520) ('Cord', 506) ('Steam', 442)

Improvements & Future Applications

The Team

Namratha Sri Mateti

PERCEPTION

It's not how you see the problem!

Dominic Thomas

INTELLECTION

It's not how you think about the problem!

Parv Bhatt

EXPRESSION

It's not how you explain it and your solution to others