Topic Modeling for Nike's Amazon Reviews

MSDS DSTA 5799 Final Project

About this Presentation

- Very Brief: 5 Minutes
- Main Focus: Deliver Insights
- Technical Part: No time for details but will quickly go through at very top level.
    - For details, please refer to the notebook

Project Overview

Data :

- The database is created by Prof. Julian McAuley at UC-San Diego.
- Picked two smaller datasets that only contain products that are categorized as “**Clothing, Shoes & Jewelry**”“ **from** **Amazon Product Data** ” (in “json.gz” format).

(1) **Meta-data** about products

(2) **Reviews** about products

Project Tasks: 📝

1. Data exaction and insights
2. Text Preprocessing
3. Topic Modeling and Visualizations
4. Data Clustering
5. Insights from the topic models

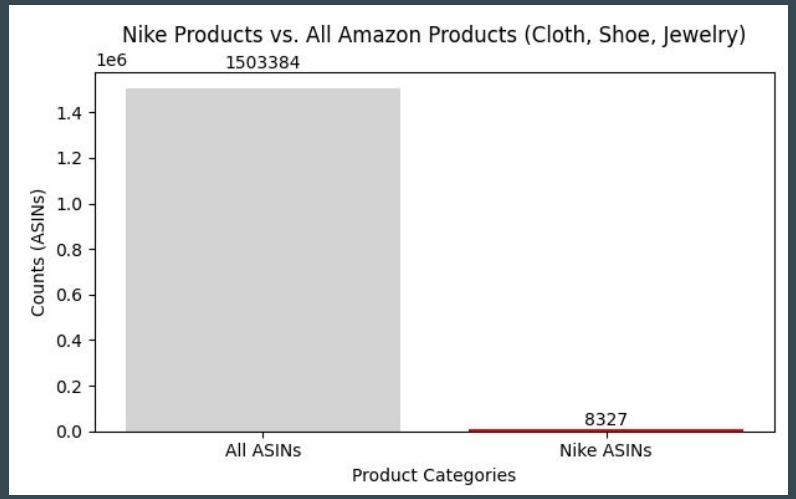
Main Tools & Environment: 🪛

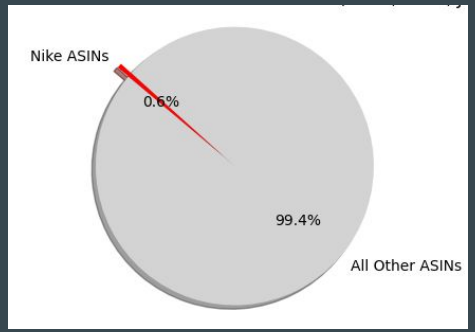
1. Google Colab (and file savings etc)
2. File manipulation packages: gzip, json, pickle, scipy.sparse
3. Compared T&M Toolkit and BERTopic Modeling

1. Data Exaction and Insights

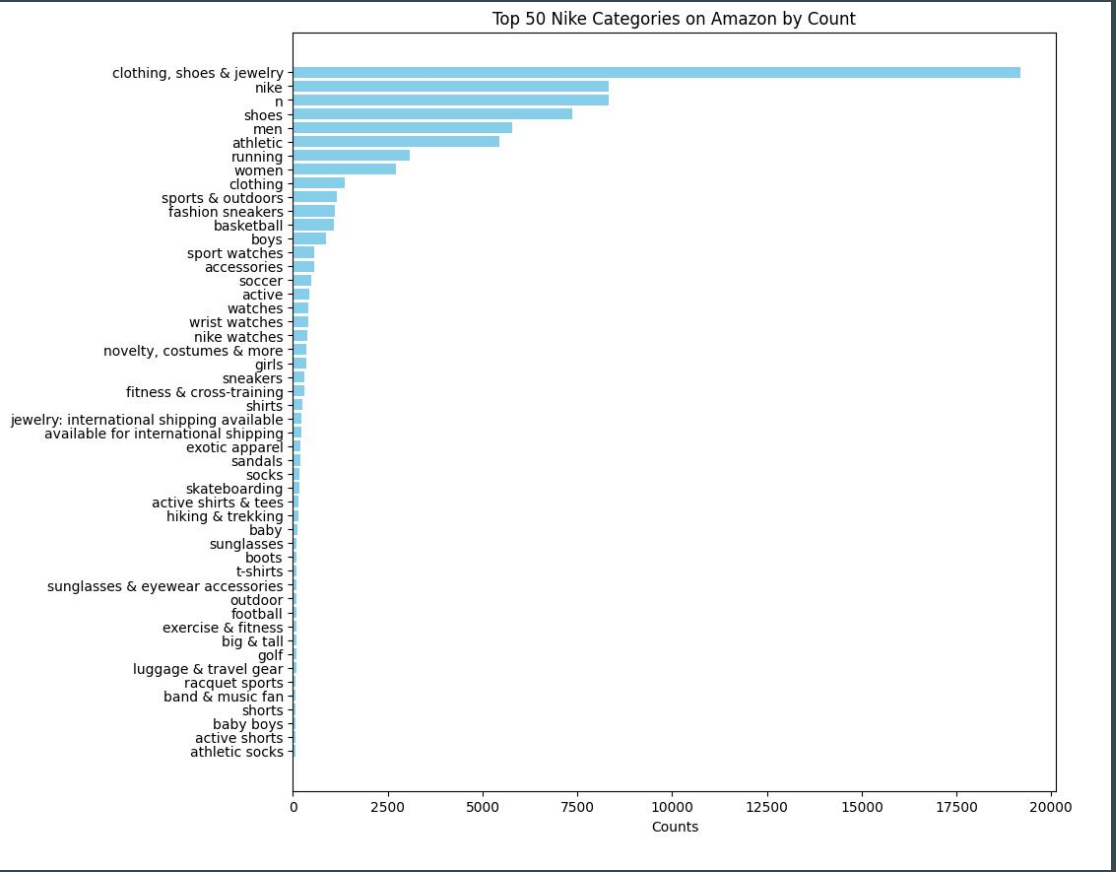
Meta Data:

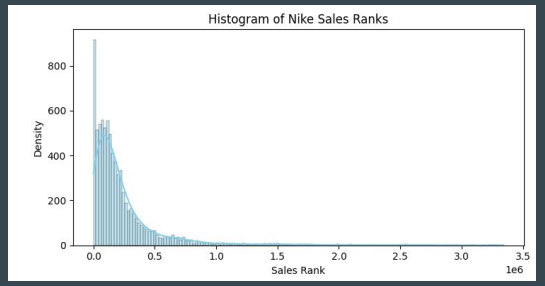
- Exact Nike ASINs from Amazon
- ~ 1,5 million Amazon products (Cloth, Shoes, Jewelry)
- 8,327 Nike
- Top 50 Nike Categories
- Nike's sales rank distributions
- Exact good selling Nike ASINs: Sales rank < 3000 ASINs

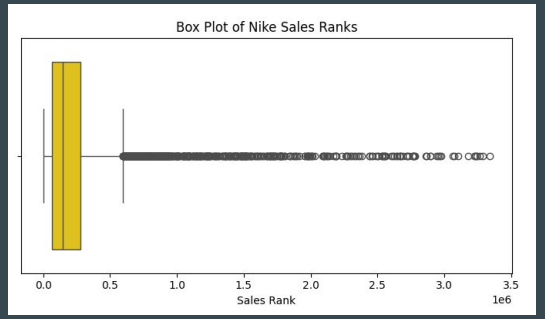


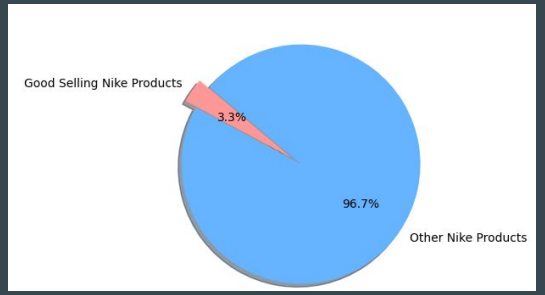


Nike's Top 50 Categories, Sales Ranks, and Good Selling ASINS





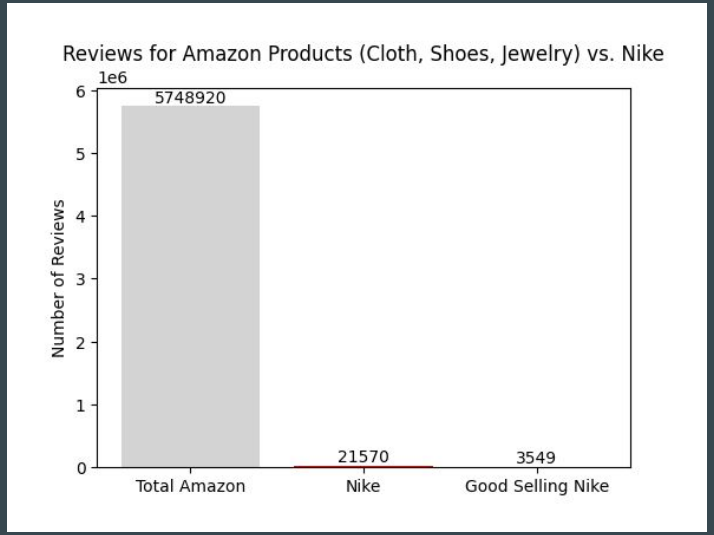


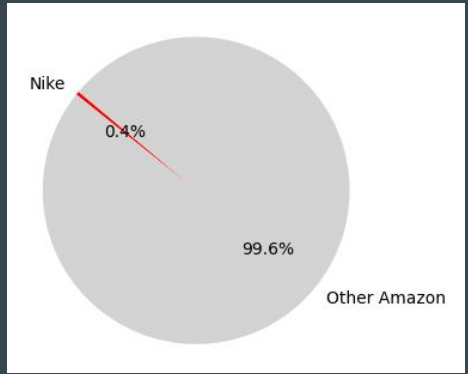


1. Data Exaction and Insights

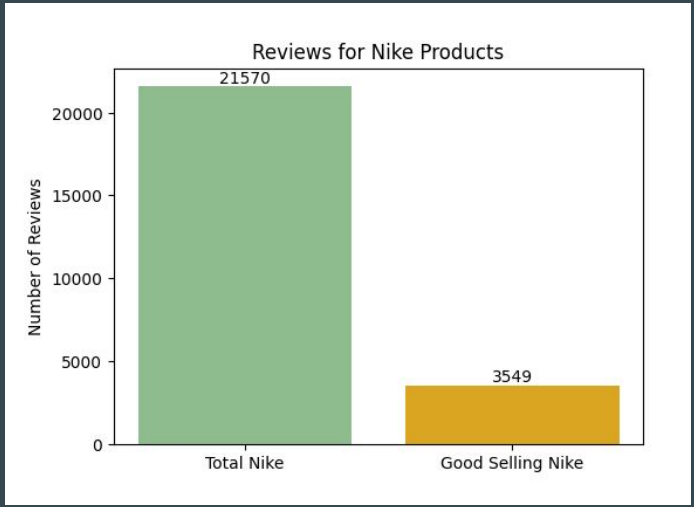
Review Data:

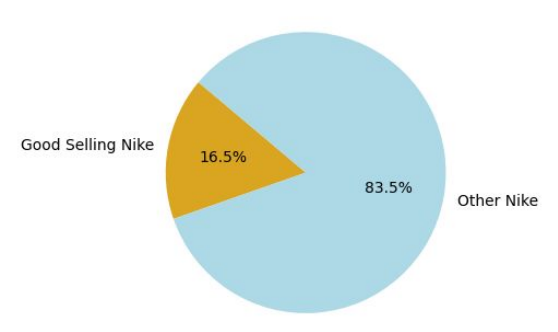
- From Nike ASINs, extract relevant reviews
- ~ 5.7 million Amazon reviews (Cloth, Shoes, Jewelry)
- 21,570 Nike reviews
- Nike reviews about good selling ASINs
- Negative Nike reviews



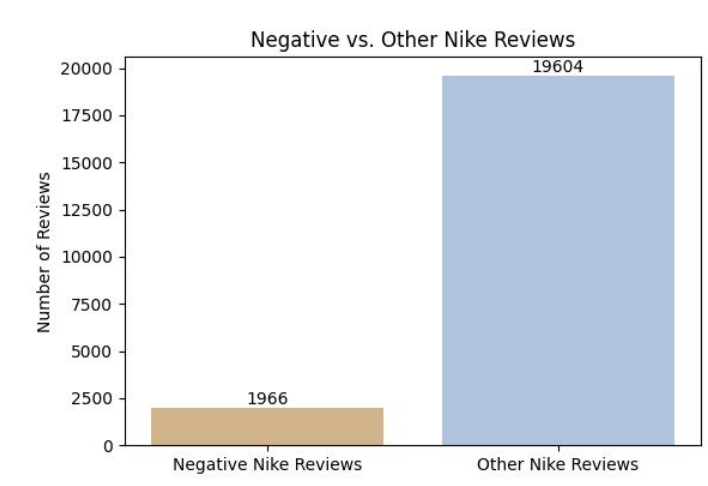


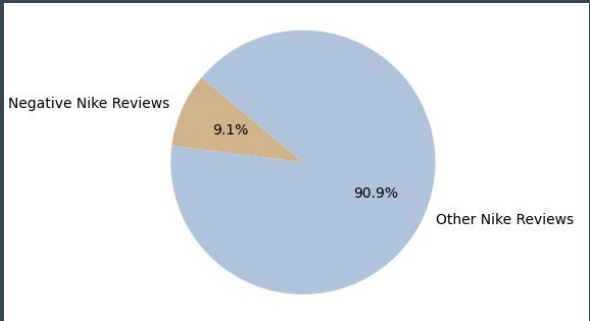
Reviews about Nike's Good Selling ASINs 🥰





Negative Nike Reviews 😡



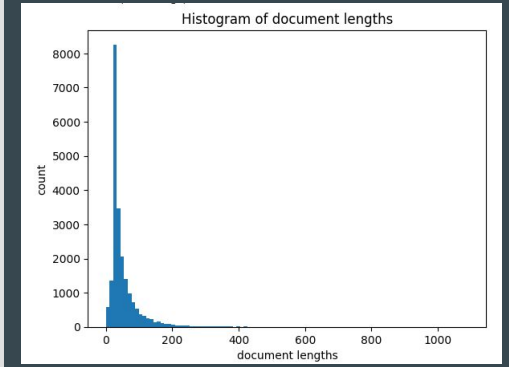


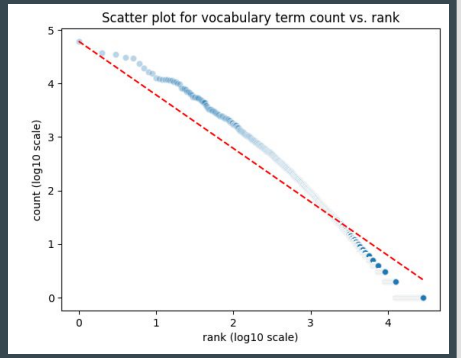
2. Text Preprocessing

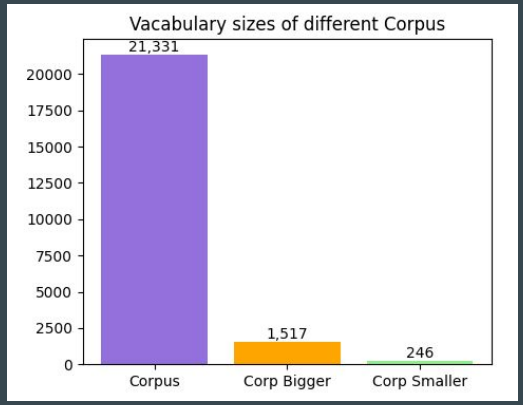
T&M Toolkit

- General
  - *Lowercase*
  - Remove *punctuation, stop words, numbers, words shorter than 3*
  - *Lemmatize*
- Small and Big Corpus
- Big: remove common & uncommon words
- Small: more aggressively remove common & uncommon words, keep only noun, verb, and adj.

Doc. lengths and vocabulary before text preprocessing





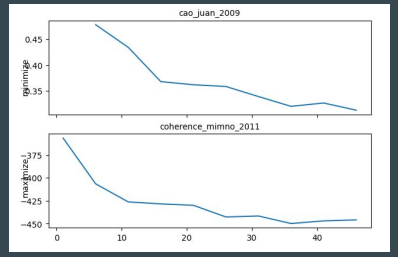


💡Only did "General" text preprocessing for BERTopic modeling.

```
Tokens in corpus:         ['cute', 'work', 'reasonably', 'long', 'cheap', 'gut', 'bad', 'cool', 'look', 'watch']
Tokens in bigger corpus:  ['cute', 'work', 'reasonably', 'long', 'cheap', 'bad', 'cool', 'look', 'watch']
Tokens in smaller corpus: ['cute', 'work', 'cheap', 'bad', 'cool', 'look', 'watch']
```

3. Topic Modeling - LDA from T&M Toolkit

- LDA topic modeling from T&M Toolkit
- Hyper-parameter Tuning
  - Number of topics: range(1, 50,5)
  - Beta: [0.1, 0.05, 0.5]
  - Alpha: (1/k, 50/k)
- T&M Toolkit evaluation statistics
  - cao_jun: how well separated are different topics. It's to minimize.
  - coherence_mino: How meaningful are topics (how often top words in a topic appear together). It's to maximize.
- Best model: **n_topic(31), beta(0.1), alpha(1/k)**



topic_1
> #1. order (0.069922)
> #2. size (0.065014)
> #3. shoe (0.052325)
> #4. return (0.048973)
> #5. receive (0.031495)
topic_2
> #1. size (0.199189)
> #2. small (0.097728)
> #3. shoe (0.082150)
> #4. order (0.068595)
> #5. run (0.056962)
topic_3
> #1. shoe (0.152732)
> #2. play (0.079479)
> #3. basketball (0.051135)
> #4. great (0.045829)
> #5. good (0.042723)
topic_4
> #1. shoe (0.134663)
> #2. run (0.103246)
> #3. nike (0.048931)
> #4. running (0.030643)
> #5. free (0.029393)
topic_5
> #1. great (0.072867)
> #2. shoe (0.050525)
> #3. product (0.043968)
> #4. time (0.039354)
> #5. arrive (0.035711)

topic_6
> #1. shoe (0.145185)
> #2. run (0.073830)
> #3. comfortable (0.052081)
> #4. good (0.051759)
> #5. great (0.042867)
topic_7
> #1. foot (0.068740)
> #2. boot (0.067592)
> #3. wear (0.065425)
> #4. work (0.064660)
> #5. day (0.059687)
topic_8
> #1. black (0.081040)
> #2. shoe (0.079127)
> #3. white (0.074345)
> #4. color (0.070382)
> #5. look (0.054259)
topic_9
> #1. sandal (0.072230)
> #2. comfortable (0.068203)
> #3. wear (0.063897)
> #4. foot (0.061675)
> #5. slide (0.032511)
topic_10
> #1. pair (0.129044)
> #2. love (0.079965)
> #3. buy (0.071917)
> #4. shoe (0.070597)
> #5. color (0.045266)

topic_11
> #1. good (0.119852)
> #2. product (0.086262)
> #3. quality (0.076875)
> #4. nike (0.047539)
> #5. nice (0.047539)
topic_12
> #1. air (0.102736)
> #2. sneaker (0.091323)
> #3. nike (0.074275)
> #4. max (0.057082)
> #5. love (0.047113)
topic_13
> #1. sock (0.156203)
> #2. fit (0.051693)
> #3. wear (0.042888)
> #4. shirt (0.039571)
> #5. great (0.037401)
topic_14
> #1. shoe (0.117898)
> #2. foot (0.087368)
> #3. wear (0.039084)
> #4. walk (0.036263)
> #5. run (0.031534)
topic_15
> #1. bag (0.111262)
> #2. gym (0.059402)
> #3. perfect (0.040261)
> #4. need (0.034208)
> #5. small (0.033881)

3. Topic Modeling - T&M Toolkit (Topic Names and Classification)

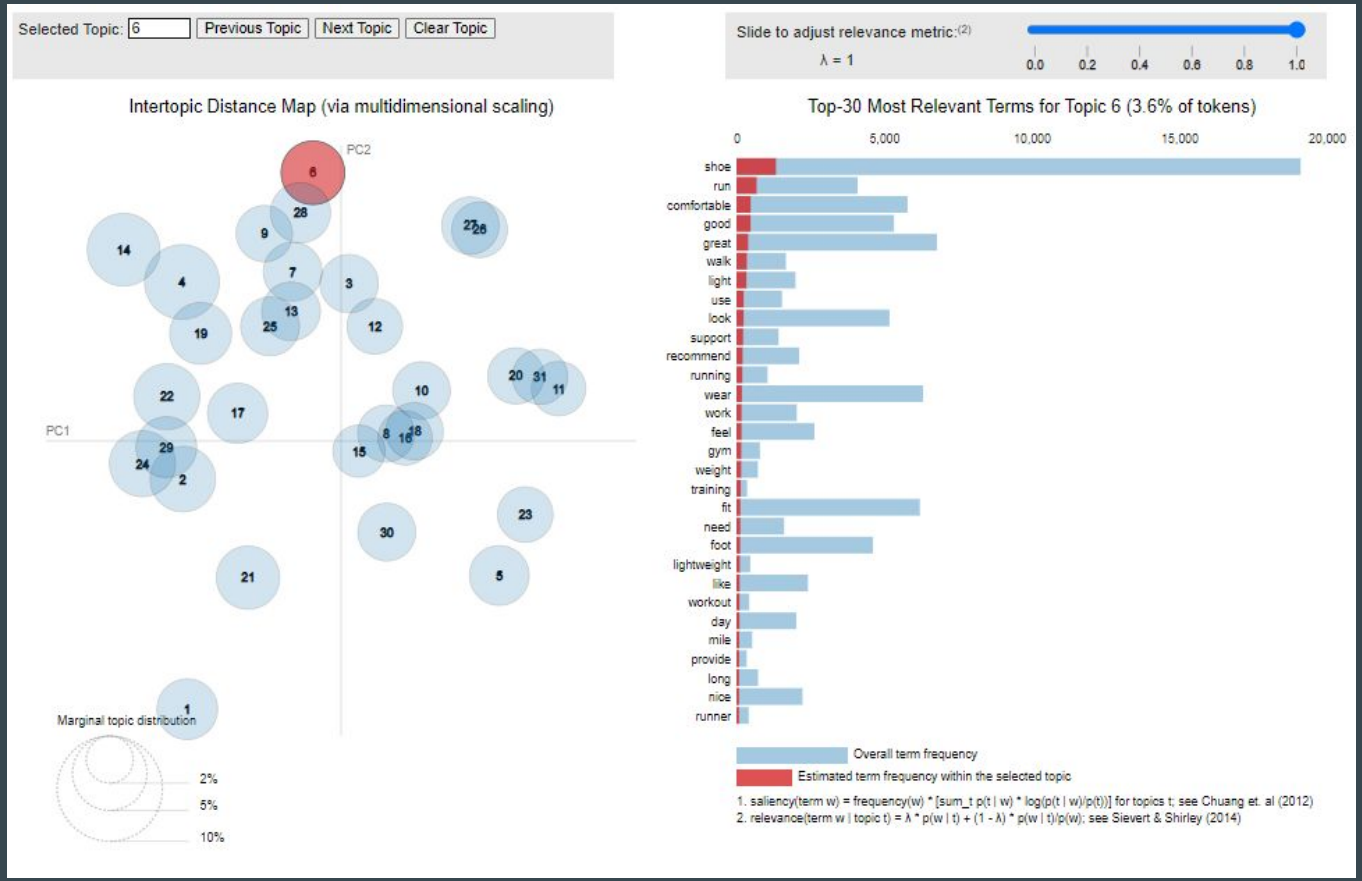
- **Topic labels** (small corpus, tuned lambda_ to be 0.8)

```
array(['1_order_return', '2_size_small', '3_shoe_play', '4_run_shoe',
       '5_great_product', '6_shoe_run', '7_boot_work', '8_black_white',
       '9_sandal_comfortable', '10_pair_love', '11_good_product',
       '12_air_sneaker', '13_sock_shirt', '14_shoe_foot', '15_bag_gym',
       '16_son_old', '17_watch_wrist', '18_size_fit', '19_foot_wide',
       '20_great_look', '21_shoe_buy', '22_shoe_pair', '23_love_gift',
       '24_watch_time', '25_little_bit', '26_light_shoe', '27_shoe_great',
       '28_support_shoe', '29_shoe_month', '30_find_size',
       '31_color_love'], dtype='<U20')
```

- Topic Description: Overall, the model performs well. For example, topic 1 is about order return. Topic 2 is about the size fits too small. Topic 3 is about running shoes.
- Document Classification: In Review 1, customer could talk about the watch wrist and sock_shirt (not sure what is that) they bought. However, I couldn't see the sentimental part of the review (what do they think of the product, do they love it or not); In Review 2, the customer seems to love the color of a pair of playful shoes they bought.

	reviewerID	asin	reviewText	rank_1	rank_2
allnikereviews-B0000V9K32.ACT5DY536GISV	ACT5DY536GISV	B0000V9K32	the colour i received is not blue as shown but yellow.Couldnt change it because it was a birthday present for my daughter and havent got time.She really didn,t like it	31_color_love (0.629)	3_shoe_play (0.254)
allnikereviews-B0000V9K32.A3BVWMS9I8OH8U	A3BVWMS9I8OH8U	B0000V9K32	Very cute and is really practical. Fits better on smaller wrists which is my case. I wear them everywhere. I really love this watch!	17_watch_wrist (0.754)	13_sock_shirt (0.129)
allnikereviews-B0000V9K3W.A5RZS69KSJH00	A5RZS69KSJH00	B0000V9K3W	The watch was exactly what i ordered and I got it very fast. Unfortunately it was a bit too big for my wrist. I returned it for a refund without any problems.	17_watch_wrist (0.6703)	1_order_return (0.1147)
allnikereviews-B0000V9K46.A3F8O512N9UNVM	A3F8O512N9UNVM	B0000V9K46	This product came promptly and as described, pleasure doing business with them!-d	21_shoe_buy (0.3441)	10_pair_love (0.3441)
allnikereviews-B0000V9KNM.A2EAKTCKFRF7A4	A2EAKTCKFRF7A4	B0000V9KNM	Why isn't Nike making these anymore? I love this watch, and I get a lot of compliments, questions from people who would like to have one as well.	31_color_love (0.448)	17_watch_wrist (0.448)

3. Topic Modeling - pyDavis



3. Topic Modeling - BERTopic Modeling

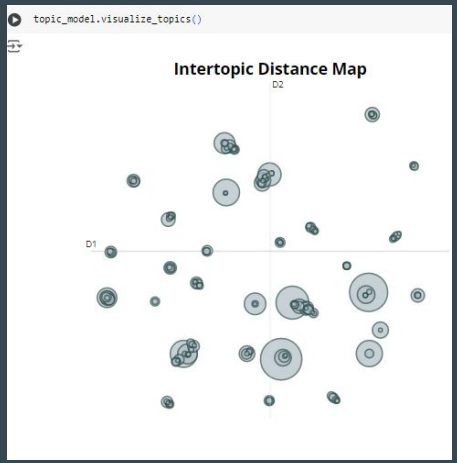
```
freq = topic_model.get_topic_info(); freq.head()
```

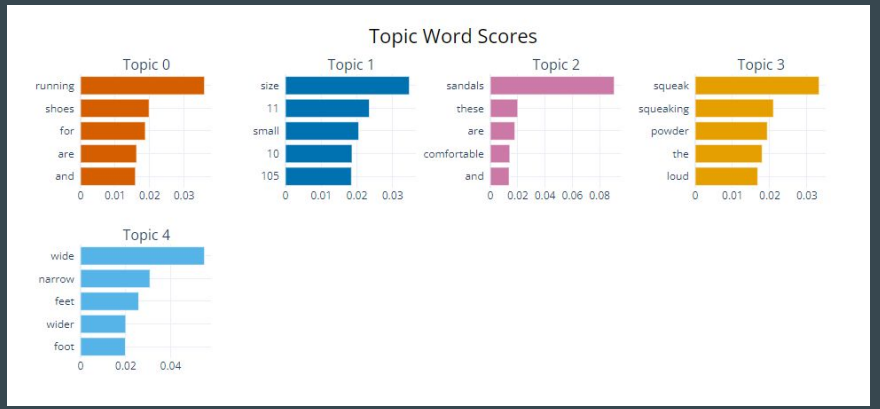
	Topic	Count	Name	Representation	Representative_Docs
0	-1	9273	-1_shoe_nike_size_pair	[shoe, nike, size, pair, foot, comfortable, co...	[bought color style love fit shoe far one comf...
1	0	1145	0_watch_band_wrist_battery	[watch, band, wrist, battery, button, feature,...	[watch little year yesterday wrist band actual...
2	1	972	1_son_grandson_kid_old	[son, grandson, kid, old, boy, school, love, y...	[year old son absolutely love shoe fit well co...
3	2	721	2_sock_crew_dry_foot	[sock, crew, dry, foot, wear, stay, wash, wash...	[sock soft comfortable wear fit fine also like...
4	3	482	3_wide_narrow_foot_width	[wide, narrow, foot, width, tight, wider, toe,...	[shoe nice comfortable however little narrow m...

```
topic_model.get_topic(0)  # Select the
```

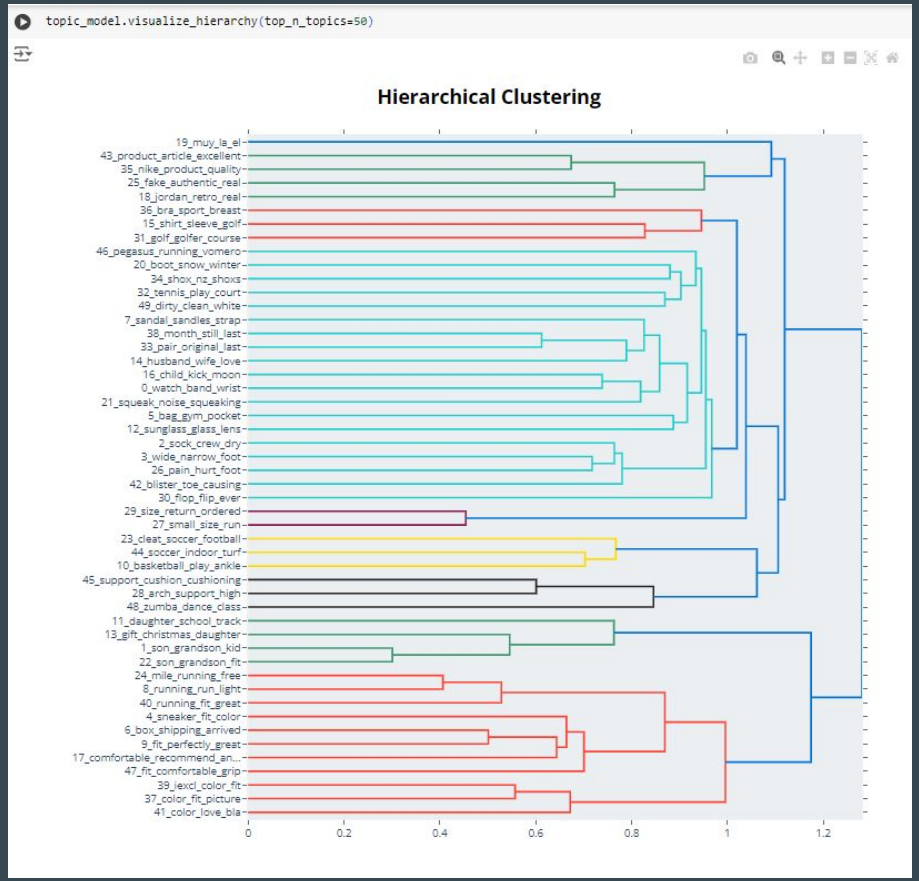
```
[('watch', 0.048902553568546385),
 ('band', 0.019622556165749018),
 ('wrist', 0.01878755306382843),
 ('battery', 0.01545338969609988),
 ('button', 0.012651527256012214),
 ('feature', 0.011710096120367294),
 ('display', 0.011196603332740807),
 ('face', 0.010628347030040518),
 ('easy', 0.010533661659868108),
 ('function', 0.009905576039696561)]
```

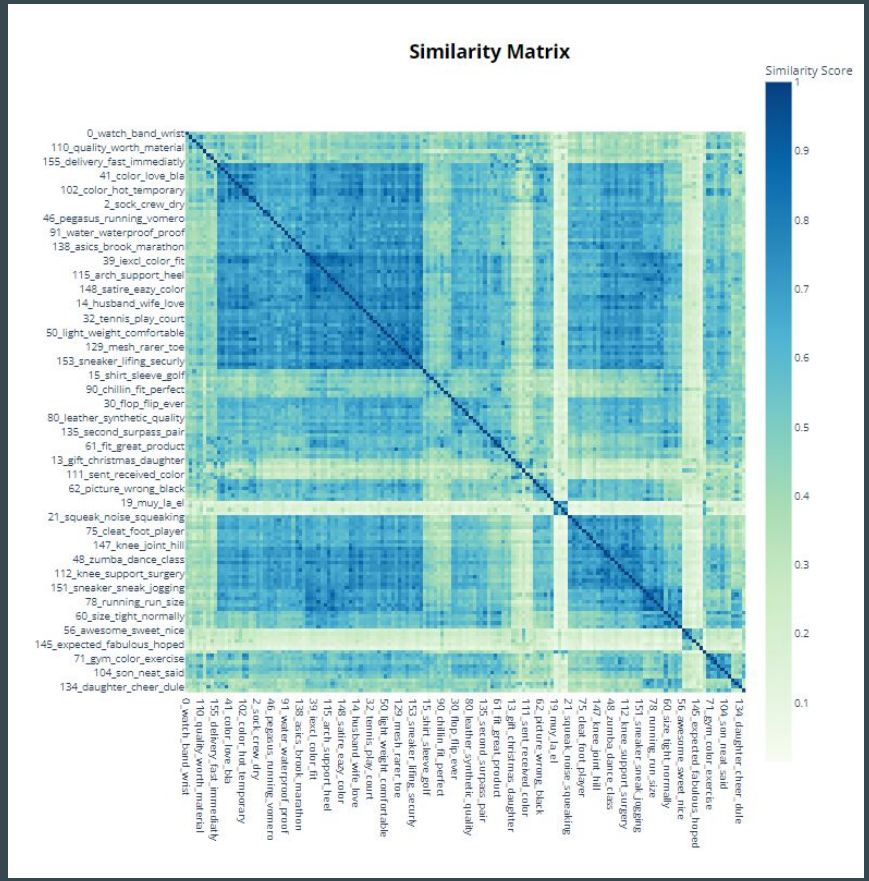
```
topic_model.visualize_topics()
```





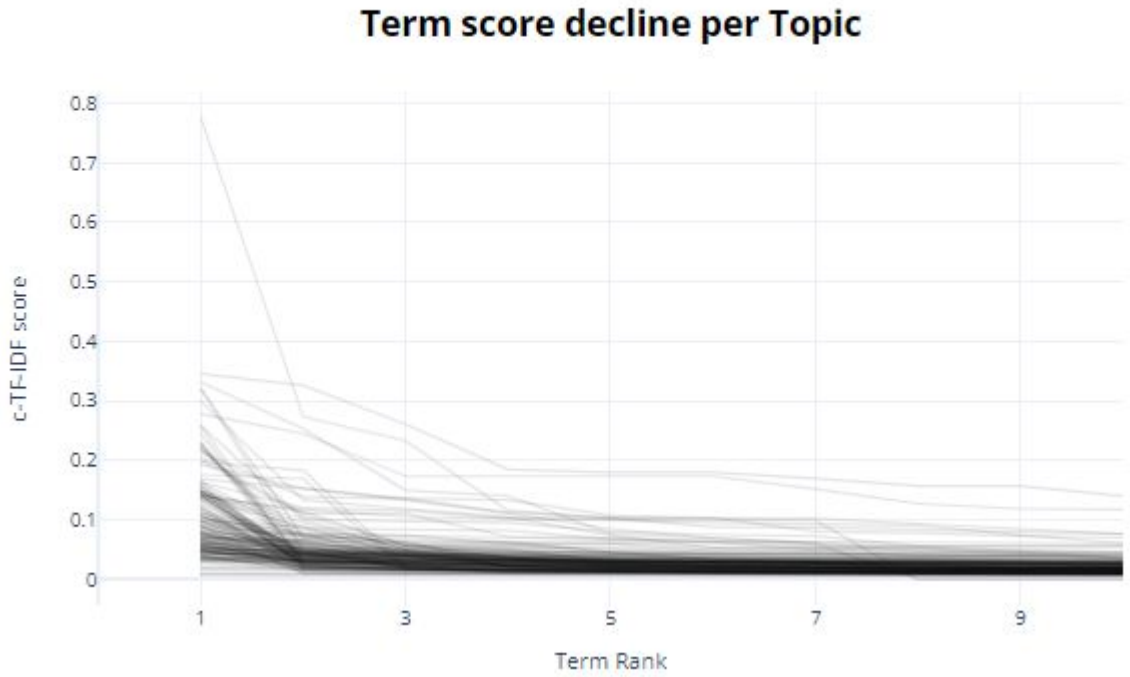
4. Data Clustering and Topic Similarities - BERTopic Modeling





Visualize Term Score Decline

```
topic_model.visualize_term_rank()
```



- Topics are represented by a number of words starting with the best representative word.
- c-TF-IDF score: Each word is represented by the score. The higher the score, the more representative a word to the topic is.
- The c-TF-IDF score slowly decline with each word that is added.
- At some point adding words to the topic representation only marginally increases the total c-TF-IDF score and would not be beneficial for its representation.

5. Actionable Insights after Topic Modeling

Topic Descriptions & Actionable Insights, Attributes people likes & dislikes

Topic 2

- **Description**: Customer dislike running shoes are smaller than expected
- **Action**: Invest if running shoes sizes are correct

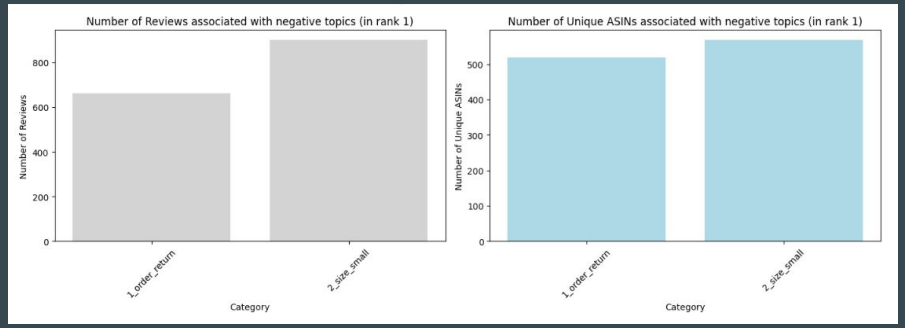
Topic 15

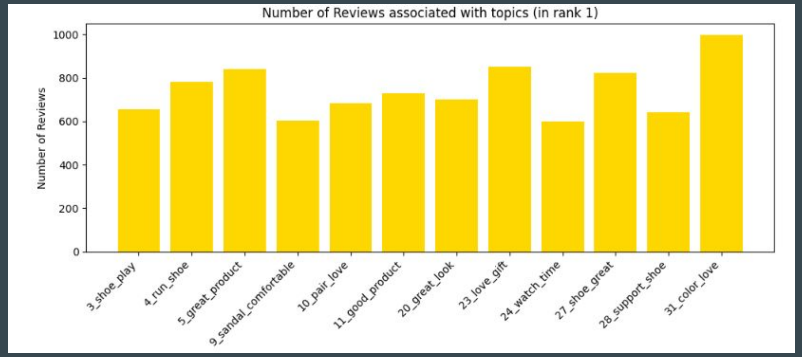
- **Description**: Customers seems to like gym bag in general, but some customers either prefer a small bag size or think the bags are too small (dislike the bag size)
- **Action**: Invest if customers like smaller/bigger gym bags. Accordingly, adjust the bag sizes or add different sizes for different customer needs.

```
topic_1
> #1. order (0.069922)
> #2. size (0.065014)
> #3. shoe (0.052325)
> #4. return (0.048973)
> #5. receive (0.031495)
topic_2
> #1. size (0.199189)
> #2. small (0.097728)
> #3. shoe (0.082150)
> #4. order (0.068595)
> #5. run (0.056962)
topic_3
> #1. shoe (0.152732)
> #2. play (0.079479)
> #3. basketball (0.051135)
> #4. great (0.045829)
> #5. good (0.042723)
topic_4
> #1. shoe (0.134663)
> #2. run (0.103246)
> #3. nike (0.048931)
> #4. running (0.030643)
> #5. free (0.029393)
topic_5
> #1. great (0.072867)
> #2. shoe (0.050525)
> #3. product (0.043968)
> #4. time (0.039354)
> #5. arrive (0.035711)
topic_6
> #1. shoe (0.145185)
> #2. run (0.073830)
> #3. comfortable (0.052081)
> #4. good (0.051759)
> #5. great (0.042867)
topic_7
> #1. foot (0.068740)
> #2. boot (0.067592)
> #3. wear (0.065425)
> #4. work (0.064660)
> #5. day (0.059687)
topic_8
> #1. black (0.081040)
> #2. shoe (0.079127)
> #3. white (0.074345)
> #4. color (0.070382)
> #5. look (0.054259)
topic_9
> #1. sandal (0.072230)
> #2. comfortable (0.068203)
> #3. wear (0.063897)
> #4. foot (0.061675)
> #5. slide (0.032511)
topic_10
> #1. pair (0.129044)
> #2. love (0.079965)
> #3. buy (0.071917)
> #4. shoe (0.070597)
> #5. color (0.045266)
topic_11
> #1. good (0.119852)
> #2. product (0.086262)
> #3. quality (0.076875)
> #4. nike (0.047539)
> #5. nice (0.047539)
topic_12
> #1. air (0.102736)
> #2. sneaker (0.091323)
> #3. nike (0.074275)
> #4. max (0.057082)
> #5. love (0.047113)
topic_13
> #1. sock (0.156203)
> #2. fit (0.051693)
> #3. wear (0.042888)
> #4. shirt (0.039571)
> #5. great (0.037401)
topic_14
> #1. shoe (0.117898)
> #2. foot (0.087368)
> #3. wear (0.039084)
> #4. walk (0.036263)
> #5. run (0.031534)
topic_15
> #1. bag (0.111262)
> #2. gym (0.059402)
> #3. perfect (0.040261)
> #4. need (0.034208)
> #5. small (0.033881)
```

5. Actionable Insights after Topic Modeling

- Can see how many reviews are associated with each topic of interest
- Take a look at actual review text, asins of specific topics that have high ranks.






```
# sample reviews with "order return" topic
review_and_topicRanks[review_and_topicRanks['rank_1'].str.contains('1_order_return')].head()
```

	reviewerID	asin	reviewText	rank_1	rank_2
allnikereviews-B0001YMVTS.A1HU5FWBPL98EI	A1HU5FWBPL98EI	B0001YMVTS	i had some problems with this order, the bill didn't arrive with the watch to the p.o. box and it couldn't be sent to my country as it was supposed to, when i finally recieved this watch (after sending several emails to solve the situation) it just didn't work, i had to spent money to fix it.	1_order_return (0.586)	24_watch_time (0.336)
allnikereviews-B0001YMVTS.A1AS3XEPV7EYM4	A1AS3XEPV7EYM4	B0001YMVTS	The product never arrived. The vendor told me it was US postal service issue. The US postal service was of no help. i paid for a product I never received. I won't use Amazon ever again and will advise others the same.	1_order_return (0.8486)	13_sock_shirt (0.0794)
allnikereviews-B0006MFAW0.A3TWL3QWHRQ12V	A3TWL3QWHRQ12V	B0006MFAW0	This watch is very cute. It is described as being a petite watch but that is a polite word for saying it is a children's watch. I am not sure why Nike chose to make such a small band width. Even if you are very petite you might want to try to find this in a store to try on before ordering online.	1_order_return (0.502)	17_watch_wrist (0.377)
allnikereviews-B0006NGUE6.A3IMJ28EKNX085	A3IMJ28EKNX085	B0006NGUE6	My husband liked the colors of these shoes, but the width was wrong. Did not have this shoe in a 7 wide, so I ordered him a B width. He said they hurt his feet. Took them to a shoe repair shop and had them stretched twice. Didn't help. Giving them to the Goodwill. It was too late to have them sent back.	1_order_return (0.4137)	19_foot_wide (0.4137)
allnikereviews-B0006NGUE6.AYXKUVV8A1GXG	AYXKUVV8A1GXG	B0006NGUE6	The shoes look good and comfortable. I have worn them for about half a dozen times now. The size is right because I have another Nike Air, so I just ordered the same size as my other Nike Air. The price is right. This is my first pair of golf shoes so I don't want to break the bank.The shipping was slow, though. I purchased it from Amazon. I had to call Amazon's customer service 1 week after I placed the order in order to find out why the item has not been shipped yet. After my call, the item was shipped the next day. I was disappointed with Amazon that the item was not shipped right away. It took about a week to start shipping even though the item was in stock in Amazon's warehouse. It took about 2 weeks to get the item.	1_order_return (0.6882)	12_air_sneaker (0.1465)

```
# sample reviews with "order return" topic
review_and_topicRanks[review_and_topicRanks['rank_1'].str.contains('2_size_small')].head()
```

	reviewerID	asin	reviewText	rank_1	rank_2	ra
allnikereviews-B0002164KC.AYRYZ0458MNTQ	AYRYZ0458MNTQ	B0002164KC	By far, the best pair of shoes I've ever owned. Well, you only need to know that their size is normally smaller than normal. You may need to order them bigger in size than you would normally do, e.g, your size is 12, get them 12.5. Anyways, I've been trying to shop for them Worldwide, but I failed to find them. I'd love to buy them in all colors!	2_size_small (0.5516)	9_sandal_comfortable (0.2516)	
allnikereviews-B0006NGUE6.A22Z9W8U4SUT3A	A22Z9W8U4SUT3A	B0006NGUE6	a good looking shoe but not to be recommended to buy unseen and untried. For the size on the shoe, it is SMALL and therefore a v tight fit. I would not buy this brand again via the internet	2_size_small (0.5409)	31_color_love (0.0794)	
allnikereviews-B0006NGUE6.A1GL2RIOBHC40A	A1GL2RIOBHC40A	B0006NGUE6	I bought this shoe for my husband who is normally a 10 1/2 so I got him an 11 just to be safe and they were still kinda tight so my advice is this shoe runs just a bit smaller especially if u have a wider foot	2_size_small (0.4194)	23_love_gift (0.2527)	
allnikereviews-B0006NGUE6.A3JBMKCMO5PLZL	A3JBMKCMO5PLZL	B0006NGUE6	I only did standard shipping. I ordered on May 1 and they got to the house May 3rd. It’s a Nike Golf shoe can’t go wrong and the price was good. I think they may run about a half size small. I don’t care I will stretch them out. I am not a huge golfer this is more for work functions.	2_size_small (0.5737)	24_watch_time (0.1452)	
allnikereviews-B0006NGUE6.A5SE11MEAPDSX	A5SE11MEAPDSX	B0006NGUE6	After finally learning that Nike shoes are always smaller than the listed size, I ordered 1/2 size larger than normal. They are the most como9frtable golf shoes I have ever owned!!	2_size_small (0.7302)	29_shoe_month (0.09384)	