APPAREL RECOMMENDATION PROJECT

[PYTHON]

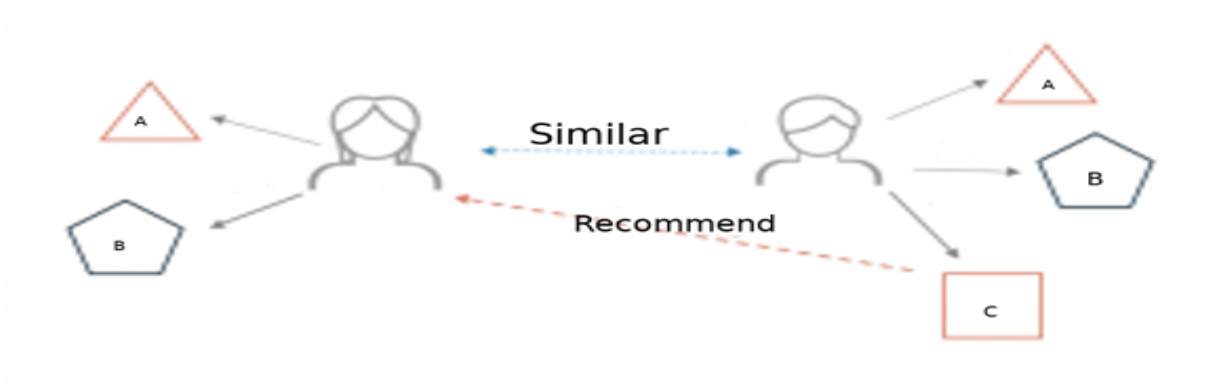
Megha Singh | Natural Language Processing and Machine Learning

PROJECT REPORT

What is a recommender system?

A recommender system is a simple algorithm whose aim is to provide the most relevant information to a user by discovering patterns in a dataset. The algorithm rates the items and shows the user the items that they would rate highly. An example of recommendation in action is when you visit Amazon and you notice that some items are being recommended to you or when Netflix recommends certain movies to you. They are also used by Music streaming applications such as Spotify and Deezer to recommend music that you might like.

Below is a very simple illustration of how recommender systems work in the context of an e-commerce site.



Two users buy the same items A and B from an e-commerce store. As a result, the similarity index between these two users is computed. Depending on the score the system can recommend item C to the other user because it detects that those two users are similar in terms of the items they purchase.

Different types of recommendation engines

Content-based and collaborative filtering recommender systems are the most common types of recommendation systems. In collaborative filtering, users' behaviors are used to provide recommendations to others. The recommendation is based on the preference of other users. A simple example would be providing a user with a movie recommendation based on a friend's recommendation. Memory-based methods and Model-based methods are two types of collaborative models. The advantage of memory-based techniques is that they are simple to implement and the resulting recommendations are often easy to explain. They are divided into two:

- **User-based collaborative filtering**: In this model, products are recommended to a user based on the fact that the products have been liked by users similar to the user. For example, if Derrick and Dennis like the same movies and a new movie come out that Derick like, then we can recommend that movie to Dennis because Derrick and Dennis seem to like the same movies.
- **Item-based collaborative filtering**: These systems identify similar items based on users' previous ratings. For example, if users A, B, and C gave a 5-star rating to books X and Y then when a user D buys book Y they also get a recommendation to purchase book X because the system identifies book X and Y as similar based on the ratings of users A, B, and C.

Model-based methods are based on Matrix Factorization and are better at dealing with sparsity. They are developed using data mining, machine learning algorithms to predict users' rating of unrated items. In this approach techniques such as dimensionality reduction are used to improve accuracy. Examples of such model-based methods include Decision trees, Rule-based Model, Bayesian Model, and latent factor models.

Content-based systems use metadata such as genre, producer, actor, musician to recommend items say movies or music. Such a recommendation would be for instance recommending Infinity War that featured Vin Diesel because someone watched and liked The Fate of the Furious. Similarly, you can get music recommendations from certain artists because you liked their music. Content-based systems are based on the idea that if you liked a certain item you are most likely to like something that is similar to it.

Datasets to use for building recommender systems

We obtained data in a policy compliant manner by using Amazon's Product Advertising API. The data consists of 1,83,000 products. For each product we obtained Image-url, Title, price etc. For this project we primarily focussed on Women's Apparel data. Additionally, Amazon's Product Advertising API can be used to retrieve data for other products. During the development of this project, we will make use of six of the 19 features we have.

1. asin- (Amazon standard identification number)
2. brand- (To which the product belongs to)
3. color- (Color information of apparel, it can contain many colors as a value , example- red and black strips)
4. product_type_name- (Product of apparel , ex- TSHIRT/SHIRT)
5. medium_image_url- (url of the image)
6. Title- (Title of the product)
7. formatted_price- (price of the product)

Data Preprocessing

We remove all those data points in which any of the attribute information is missing.

```
In [18]: # Consider the product which have price information data['formatted_price '].isnull-- Gives the info aboout the dataframe ro
         data = data.loc[~data['formatted_price'].isnull()]
         data.shape[0] # Prints number of data points after eliminating price=NULL.

Out[18]: 28395

In [19]: # Consider the product which have color info data['color'].isnull-- Gives the information about the dataframe row's which ha
         data = data.loc[~data['color'].isnull()]
         data.shape[0]  # # Prints number of data points after eliminating color=NULL.

Out[19]: 28385
```

Remove near duplicate items

These shirts are exactly same except in size (S, M,L,XL)

These shirts exactly same except in color

- Remove all products with very few words in there title.
- Sort the whole data based on title(alphabetically order of title).
- Remove the duplicates which differ at the end.

Text Pre-processing

- Stop words removal

  Removed all the stop words from the title, as hese stop words does not makes any sense.

After Pre-processing our dataset looks like this:

```
In [57]: data = pd.read_pickle('16k_apperal_data_preprocessed')
         data.head()
```

Out[57]:

	asin	brand	color	medium_image_url	product_type_name	title	formatted_price
4	B004GSI2OS	FeatherLite	Onyx Black/ Stone	https://images-na.ssl-images-amazon.com/images...	SHIRT	featherlite ladies long sleeve stain resistant...	$26.26
6	B012YX2ZPI	HX-Kingdom Fashion T-shirts	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	womens unique 100 cotton special olympics wor...	$9.99
15	B003BSRPB0	FeatherLite	White	https://images-na.ssl-images-amazon.com/images...	SHIRT	featherlite ladies moisture free mesh sport sh...	$20.54
27	B014ICEJ1Q	FNC7C	Purple	https://images-na.ssl-images-amazon.com/images...	SHIRT	supernatural chibis sam dean castiel neck tshi...	$7.39
46	B01NACPBG2	Fifth Degree	Black	https://images-na.ssl-images-amazon.com/images...	SHIRT	fifth degree womens gold foil graphic tees jun...	$6.95

Text Based Product Similarity

Bag of words(Bow) on Product titles

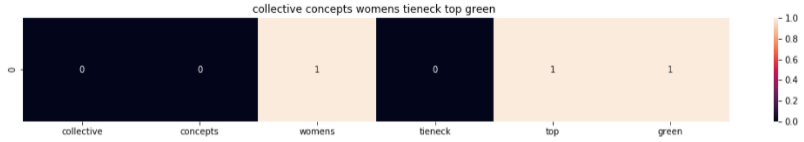
It's a representation used in Natural language Processing and Information Retrieval, using Bag of words we convert our titles to a vector of n-dimension.

We calculate pairwise_distance,the distance from given input apparel to all remaining apparels, the apparels whose Euclidean distance is less from the input apparel those apparel are recommended to the user.

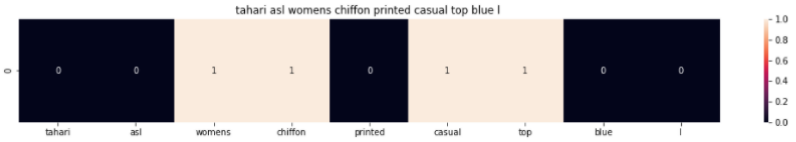
This is the input Apparel:


```
ASIN : B01CDLVBX8
Brand: Bar III
Title: bar iii womens chiffon pleated casual top green
Euclidean similarity with the query image : 0.0
```

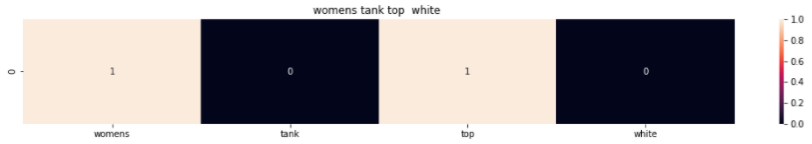
Based on input apparel the recommended apparels are:



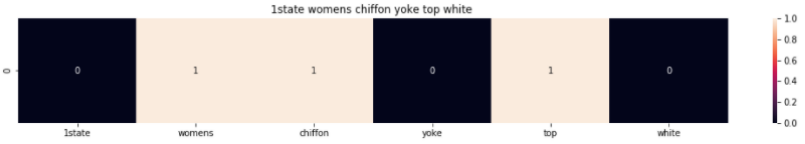
```
ASIN : B06Y15Y3Z7
Brand: Collective Concepts
Title: collective concepts womens tieneck top green
Euclidean similarity with the query image : 2.8284271247461903
============================================================
```



```
ASIN : B01M68JIFB
Brand: Tahari ASL
Title: tahari asl womens chiffon printed casual top blue l
Euclidean similarity with the query image : 2.8284271247461903
============================================================
```



```
ASIN : B00JPOZ9GM
Brand: Sofra
Title: womens tank top  white
Euclidean similarity with the query image : 2.8284271247461903
============================================================
```



```
ASIN : B071YPB9P9
Brand: 1.State
Title: 1state womens chiffon yoke top white
Euclidean similarity with the query image : 2.8284271247461903
============================================================
```

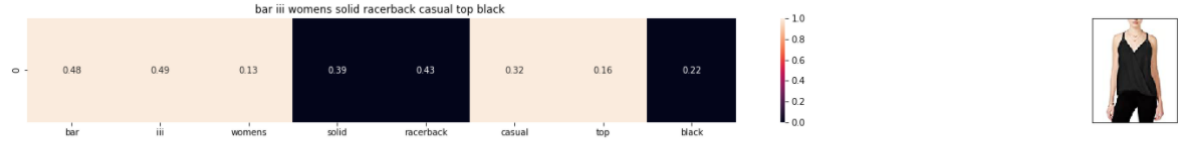
TF-IDF based product similarity

In information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.Using tf-idf we convert our titles to a vector of n-dimension.

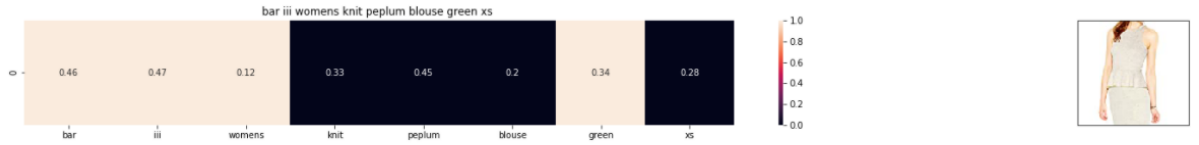
We calculate pairwise_distance,the distance from given input apparel to all remaining apparels, the apparels whose Euclidean distance is less from the input apparel those apparel are recommended to the user.

This is the input Apparel:

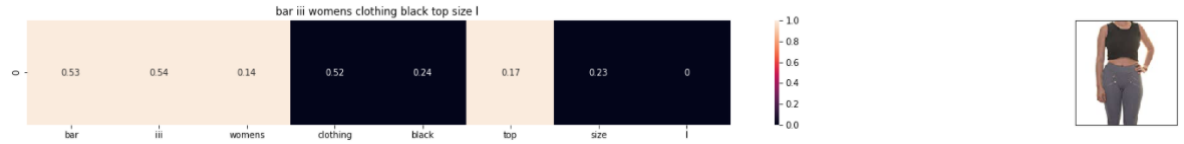
ASIN : B01CDLVBX8
BRAND : Bar III
Eucliden distance from the given image : 0.0



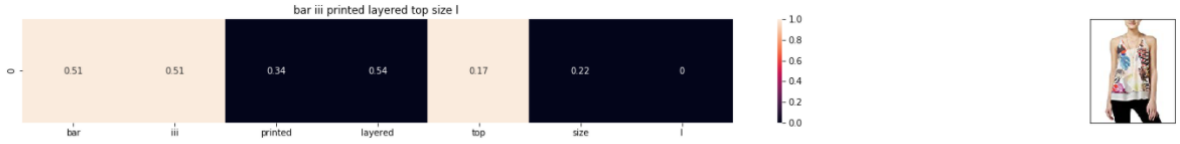
ASIN : B0724ZCX9F
BRAND : Bar III
Eucliden distance from the given image : 0.9093599127727673



ASIN : B01AND54VS
BRAND : Bar III
Eucliden distance from the given image : 0.9368579831412406



ASIN : B01KVR2YZA
BRAND : Bar lll
Eucliden distance from the given image : 0.9645461014433538



ASIN : B06XPYWQ5Y
BRAND : Bar III
Eucliden distance from the given image : 1.006127086981935

Text Semantics based product similarity

Bag of word and Tf-idf technique does not take semantics into consideration.

We used the word 2 vec technique to convert a given title into an n-dimensional vector. Word2vec is a collection of models used to create word embeddings. We used google pre-trained word 2 vec model in this project.

Deep Learning to Recommend Products

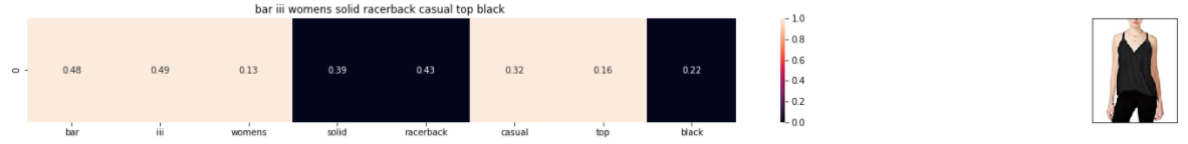
Visual features based product similarity.

We used keras and tensorflow to extract

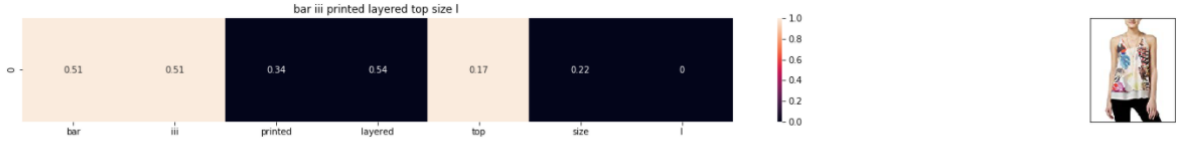
This is the input image

ASIN : B01CDLVBX8
BRAND : Bar III
Eucliden distance from the given image : 0.0

Recommended images

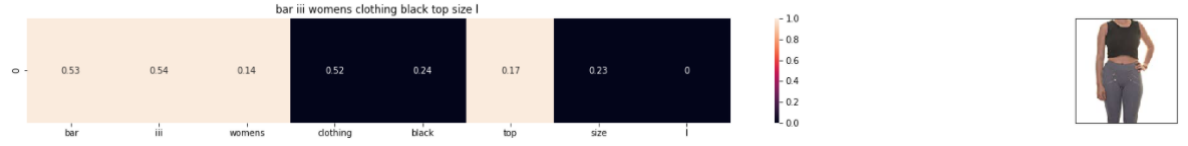


ASIN : B0724ZCX9F
BRAND : Bar III
Eucliden distance from the given image : 0.9093599127727673



ASIN : B06XPYWQ5Y
BRAND : Bar III
Eucliden distance from the given image : 1.006127086981935

ASIN : B01AND54VS
BRAND : Bar III
Eucliden distance from the given image : 0.9368579831412406



ASIN : B01KVR2YZA
BRAND : Bar lll
Eucliden distance from the given image : 0.9645461014433538

<u>Measuring goodness of our Solution:</u>

In order to determine which of the given options is going to provide the best results, we can perform A/B testing. It is best to use TF-IDF models for this project as opposed to Bag of words (Bow).

The recommendation system we are performing is a content-based one, so metrics are not applicable, in case of collaborative filtering we could have used the suitable metric.