Text Visualization

```
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from wordcloud import STOPWORDS
```

```
from google.colab import drive
```

```
drive.mount('/content/drive/')
```

```
Mounted at /content/drive/
```

```
df=pd.read_csv("netflix_titles.csv",usecols=['cast'])
df.head
```

```
pandas.core.generic.NDFrame.head
def head(n: int=5) -> NDFrameT

Return the first `n` rows.

This function returns the first `n` rows for the object based
on position. It is useful for quickly testing if your object
has the right type of data in it.
```

```
ndf=df.dropna()
ndf.head()
```

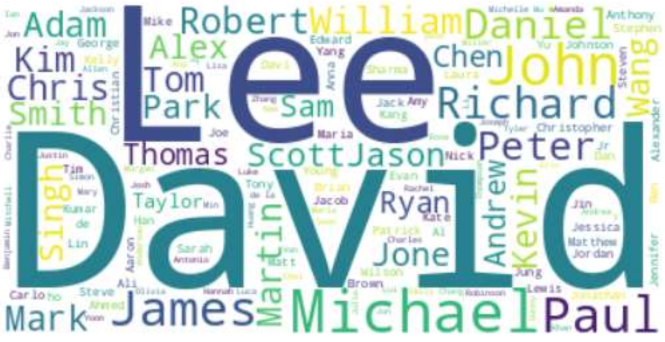
	cast
1	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...
2	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...
4	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...
5	Kate Siegel, Zach Gilford, Hamish Linklater, H...
6	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...

```
text=" ".join(item for item in ndf['cast'])
print(text)
```

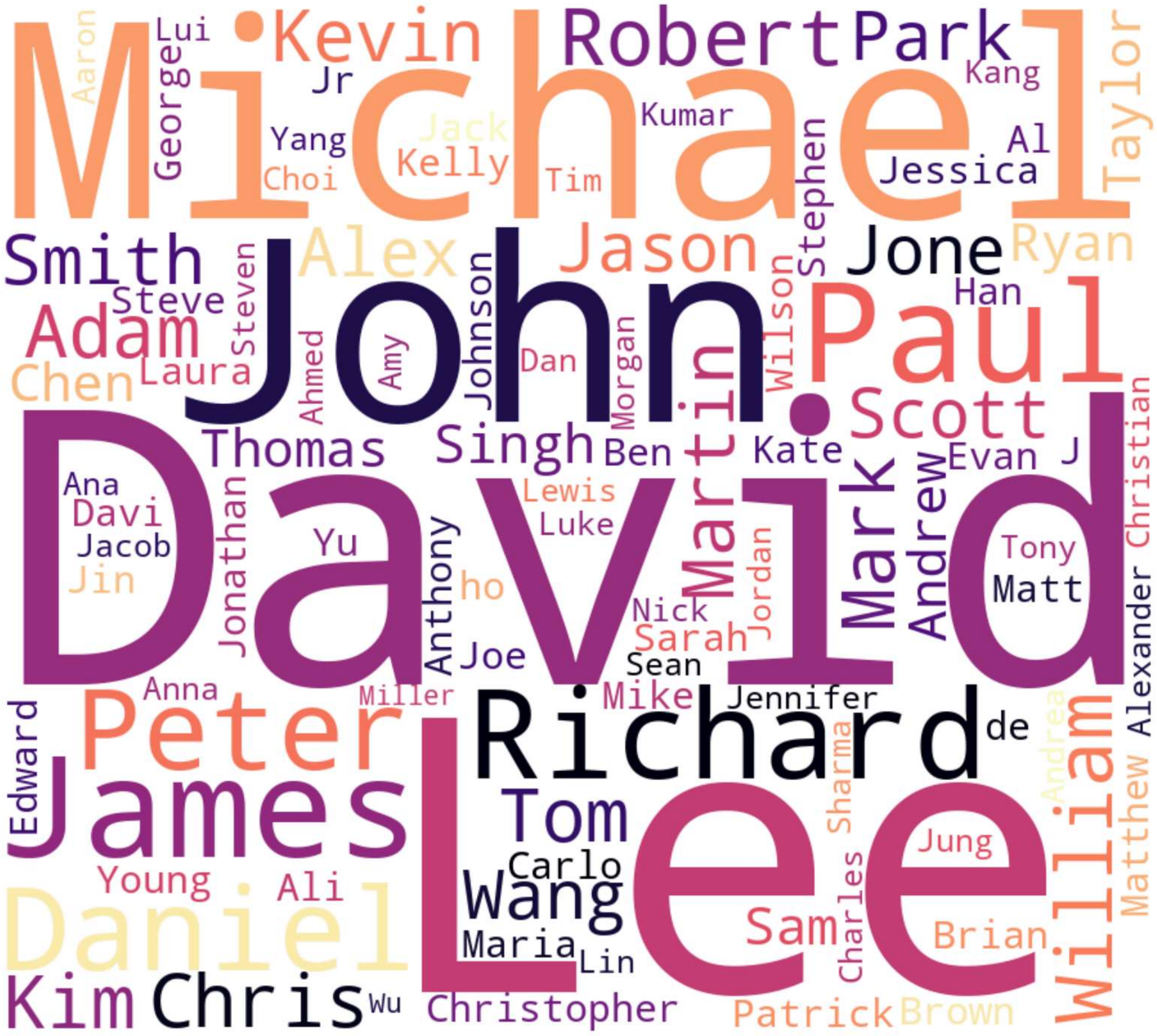
```
Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore
```

```
stopwords=set(STOPWORDS)
```

```
wordcloud=WordCloud(background_color="white").generate(text)
plt.imshow(wordcloud,interpolation="bilinear")
plt.axis("off")
plt.margins(x=0,y=0)
plt.show()
```



```
wordcloud=WordCloud(background_color="white",
                    max_words=100,
                    max_font_size=300,
                    width=900,
                    height=800,
                    colormap="magma").generate(text)
plt.figure(figsize=(20,20))
plt.imshow(wordcloud,interpolation="bilinear")
plt.axis('Off')
plt.margins(x=0,y=0)
plt.savefig("cloud.fig",format="jpg")
plt.show()
```



Market Basket Analysis

Market basket anaylisi is used by companies to identify items that are frequently purchased together.

It is frequently used by restaurents retail stores,online shopping platforms to encourage customers to make more purchases in a single visit. This is a use-case of data-science in marketing that increases company sales and drives business growth and commonly utilizes the Apriori Algorithm

Apriori Algorithm

It is the most common technique for performing market basket analysis. It is used for association rule mining, which is a rule based process used to identify correlations between items purchased by users

The Apriori algorithm has three main compnents

- Support
- Lift
- Confidence

Lets calculate the support,confidence and lift

Support

The first component of the Apriori Algorithm is support - we use it ti assess overall popularity of a given product with the following formula.

Support(item)=Transactions comprising the item / Total transactions.

A high support value indicates that the item is present in most purchases, therefore the market should focus on it more.

Confidence

It tells us the liklihood of different purchase combinations. We calculate that using the following formula:

Confidence(Bread -> Milk)=transactions compromising bread and milk / Transactions comprimising bread.

Lift

Finally lift refers to the increase in the ratio of the scale of milk when you sell bread: Lif=Confidence(Bread-> Milk)/Support(Bread) This means that customers are 1.3 times more likely to buy milk if you also sell bread

Step 1: Pre-Requistes for performing market basket analysis

Step 2 : Reading the dataset

```
import pandas as pd
df=pd.read_csv('/content/Groceries_dataset.csv')
df.head()
```

	Member_number	Date	itemDescription
0	1808	21-07-2015	tropical fruit
1	2552	05-01-2015	whole milk
2	2300	19-09-2015	pip fruit
3	1187	12-12-2015	other vegetables
4	3037	01-02-2015	whole milk

Next steps: Generate code with df View recommended plots

```
df['single_transaction']=df['Member_number'].astype(str)+'_'+df['Date'].astype(str)
df.head()
```

	Member_number	Date	itemDescription	single_transaction
0	1808	21-07-2015	tropical fruit	1808_21-07-2015
1	2552	05-01-2015	whole milk	2552_05-01-2015
2	2300	19-09-2015	pip fruit	2300_19-09-2015
3	1187	12-12-2015	other vegetables	1187_12-12-2015
4	3037	01-02-2015	whole milk	3037_01-02-2015

Next steps: Generate code with df View recommended plots

```
df2=pd.crosstab(df['single_transaction'],df['itemDescription'])
df.head()
```

	Member_number	Date	itemDescription	single_transaction
0	1808	21-07-2015	tropical fruit	1808_21-07-2015
1	2552	05-01-2015	whole milk	2552_05-01-2015
2	2300	19-09-2015	pip fruit	2300_19-09-2015
3	1187	12-12-2015	other vegetables	1187_12-12-2015
4	3037	01-02-2015	whole milk	3037_01-02-2015

Next steps: Generate code with df View recommended plots

```
def encode(item_freq):
  res=0
  if item_freq>0:
    res=1
  return res
basket_input=df2.applymap(encode)
```

Step 4: Build the Apriori Algorithm for Market Basket Analysis

```
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
frequent_itemsets=apriori(basket_input,min_support=0.001,use_colnames=True)
rules=association_rules(frequent_itemsets,metric='lift')
rules.head()
```

```
/usr/local/lib/python3.10/dist-packages/mlxtend/frequent_patterns/fpcommon.py:110: DeprecationWarning: DataFrames with non-bool typ
  warnings.warn(
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(bottled water)	(UHT-milk)	0.060683	0.021386	0.001069	0.017621	0.823954	-0.000228	0.996168	-0.185312
1	(UHT-milk)	(bottled water)	0.021386	0.060683	0.001069	0.050000	0.823954	-0.000228	0.988755	-0.179204
2	(other vegetables)	(UHT-milk)	0.122101	0.021386	0.002139	0.017515	0.818993	-0.000473	0.996060	-0.201119
3	(UHT-milk)	(other vegetables)	0.021386	0.122101	0.002139	0.100000	0.818993	-0.000473	0.975443	-0.184234

Next steps: Generate code with rules View recommended plots

```
rules.sort_values(['support','confidence','lift'],axis=0,ascending=False).head(8)
```

```
/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `transform_
  and should_run_async(code)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
623	(rolls/buns)	(whole milk)	0.110005	0.157923	0.013968	0.126974	0.804028	-0.003404	0.964550	-0.214986
622	(whole milk)	(rolls/buns)	0.157923	0.110005	0.013968	0.088447	0.804028	-0.003404	0.976350	-0.224474
695	(yogurt)	(whole milk)	0.085879	0.157923	0.011161	0.129961	0.822940	-0.002401	0.967861	-0.190525
694	(whole milk)	(yogurt)	0.157923	0.085879	0.011161	0.070673	0.822940	-0.002401	0.983638	-0.203508
551	(soda)	(other vegetables)	0.097106	0.122101	0.009691	0.099794	0.817302	-0.002166	0.975219	-0.198448
550	(other vegetables)	(soda)	0.122101	0.097106	0.009691	0.079365	0.817302	-0.002166	0.980729	-0.202951
648	(sausage)	(whole milk)	0.060349	0.157923	0.008955	0.148394	0.939663	-0.000575	0.988811	-0.063965