BAIT 508 Group Project: Industry Analysis

Jen-Hsiang (Tony) Yang(Email - ty070101@student.ubc.ca Student ID - 68638030. Responsible for the initial version of the Python script and writeup)
Bhairavi Wadekar (Email - bhwade00@student.ubc.ca, Student ID - 12491247. Responsible for composing the project report)
Sumaita Afifa Ahmed (Email - sahm0019@student.ubc.ca 57330508 Responsible for Python script and writeup)

PROJECT OVERVIEW:

The goal of this project is to conduct an in-depth analysis of public US firms within selected industry sector(s) using various data analyses and natural language processing (NLP) techniques that we learned in BAIT 508. Each team will choose at least one industry sector to investigate and utilize multiple datasets to extract valuable industry insights from the data.

The project will utilize three datasets (located in the data folder):

- public_firms.csv
- major_groups.csv
- 2020_10K_item1_full.csv

DETAILED INSTRUCTIONS:

Part 1. Quantitative Analysis of the Industry Sector

A. **[Industry Sector Selection and Data Filtering; 20 points]**

1. The file "data/major_groups.csv" contains a list of major industry sectors and their corresponding codes (column "major_group"). Your first task is to **choose at least one industry** sector that interests your group. It is okay if multiple groups choose the same industry sector, so you don't need to coordinate with other groups.
2. Next, filter the data in "data/public_firms.csv" to only include the firms belonging to the industry sector(s) you have selected. You can use the "major_group" value, which corresponds to the first two digits of each firm's SIC code,[1] to identify relevant firms. For example, if you are interested in the "Business Service" sector and its "major_group" code is 73, you should retain all firms whose SIC codes start with 73.
3. Now, answer the following questions based on the filtered dataset:
   a. How many unique firm-year ("fyear") observations are there in the filtered dataset?
   b. How many unique firms are there in the filtered dataset?
   c. How many firms in the filtered dataset have records over all 27 years (1994-2020)?

B. **[Preliminary Analysis; 20 points]** Answer the following questions:

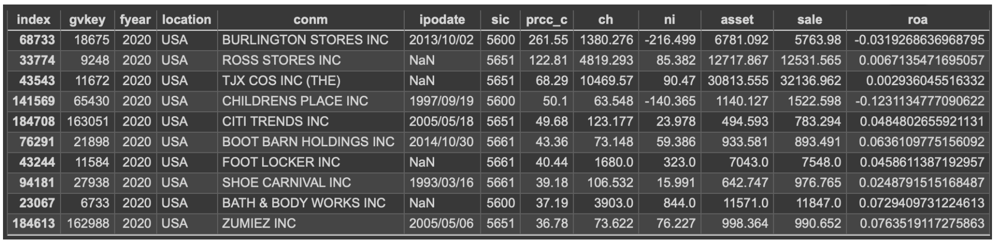
1. What are the top 10 firms with the highest stock price (column "prcc_c") in the year 2020?
2. What are the top 10 firms with the highest sales (column "sale") in the entire history of the dataset?
3. What is the geographical distribution (column "location") of all the firms? In other words, how many firms are there in each location? Please list the top 10 locations.
4. Create a line chart to show the average stock price (column "prcc_c") in the selected sector(s) across the years. If you have selected multiple sectors, draw multiple lines to show them separately.
5. Which firm was affected the most by the 2008 Financial Crisis, as measured by the percentage drop in stock price from 2007 to 2008?
6. Plot the average Return on Assets (ROA) for the firms located in the "USA" across the years. ROA is calculated as ni/asset.

Part 1. ANSWERS

A. The industry chosen is Apparel and Accessory Stores. The corresponding code is 56 and the index is 48. There are 27 unique firm-year observations. There are 105 unique companies in the Apparel and Accessory Stores. 11 companies have records over all the 27 years.

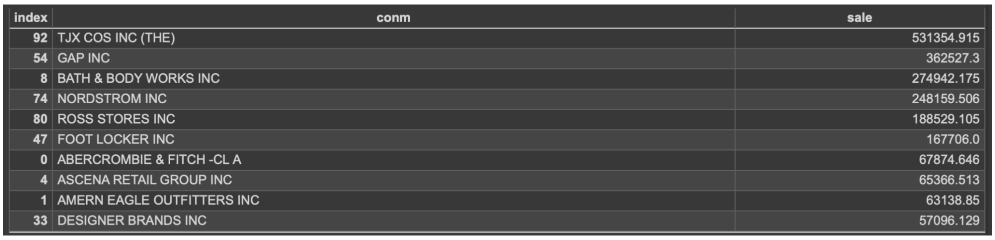
Part 1B. ANSWERS

1. The top 10 firms with the highest stock price are: 'BURLINGTON STORES INC', 'ROSS STORES INC', 'TJX COS INC (THE)', 'CHILDRENS PLACE INC', 'CITI TRENDS INC', 'BOOT BARN HOLDINGS INC', 'FOOT LOCKER INC', 'SHOE CARNIVAL INC', 'BATH & BODY WORKS INC', 'ZUMIEZ INC'



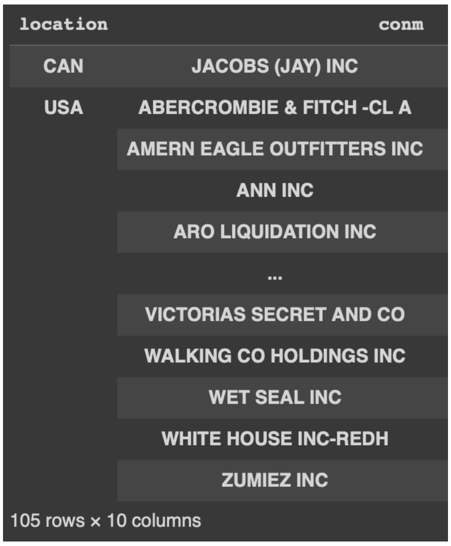
index	gvkey	fyear	location	conm	ipodate	sic	prcc_c	ch	ni	asset	sale	roa
68733	18675	2020	USA	BURLINGTON STORES INC	2013/10/02	5600	261.55	1380.276	-216.499	6781.092	5763.98	-0.0319268636968795
33774	9248	2020	USA	ROSS STORES INC	NaN	5651	122.81	4819.293	85.382	12717.867	12531.565	0.0067135471695057
43543	11672	2020	USA	TJX COS INC (THE)	NaN	5651	68.29	10469.57	90.47	30813.555	32136.962	0.002936045516332
141569	65430	2020	USA	CHILDRENS PLACE INC	1997/09/19	5600	50.1	63.548	-140.365	1140.127	1522.598	-0.1231134777090622
184708	163051	2020	USA	CITI TRENDS INC	2005/05/18	5651	49.68	123.177	23.978	494.593	783.294	0.0484802655921131
76291	21898	2020	USA	BOOT BARN HOLDINGS INC	2014/10/30	5661	43.36	73.148	59.386	933.581	893.491	0.0636109775156092
43244	11584	2020	USA	FOOT LOCKER INC	NaN	5661	40.44	1680.0	323.0	7043.0	7548.0	0.0458611387192957
94181	27938	2020	USA	SHOE CARNIVAL INC	1993/03/16	5661	39.18	106.532	15.991	642.747	976.765	0.0248791515168487
23067	6733	2020	USA	BATH & BODY WORKS INC	NaN	5600	37.19	3903.0	844.0	11571.0	11847.0	0.0729409731224613
184613	162988	2020	USA	ZUMIEZ INC	2005/05/06	5651	36.78	73.622	76.227	998.364	990.652	0.0763519117275863

2. Top 10 firms with the highest sales in the entire history are: 'TJX COS INC (THE)', 'GAP INC', 'BATH & BODY WORKS INC', 'NORDSTROM INC', 'ROSS STORES INC', 'FOOT LOCKER INC', 'ABERCROMBIE & FITCH -CL A', 'ASCENA RETAIL GROUP INC,' 'AMERN EAGLE OUTFITTERS INC', 'DESIGNER BRANDS INC'



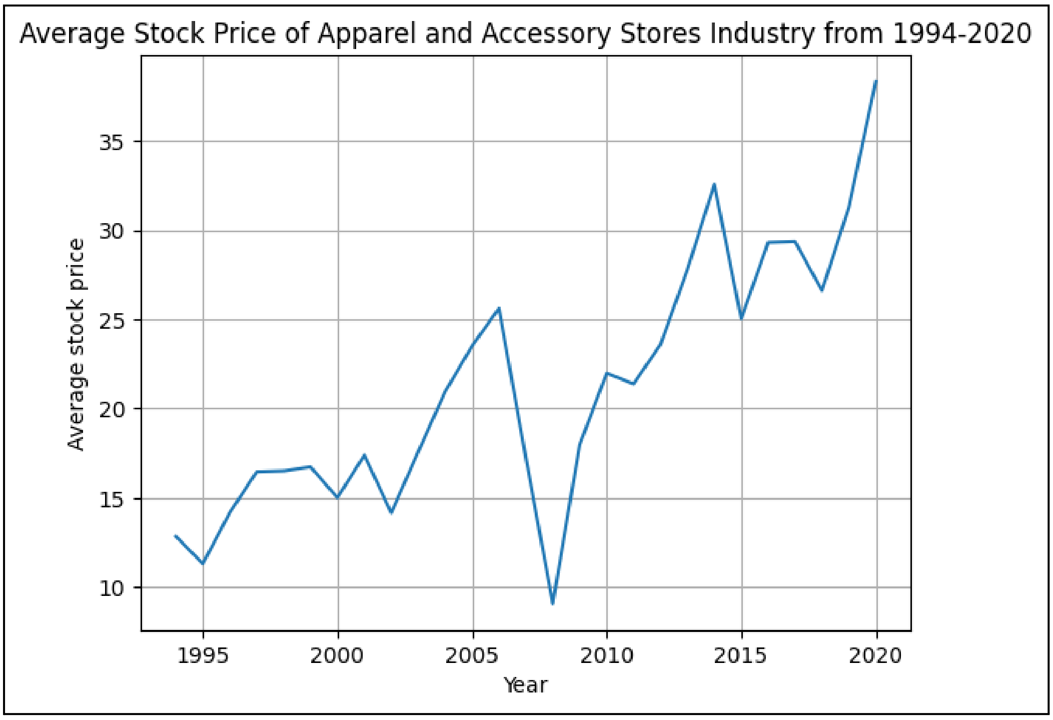
index	conm	sale
92	TJX COS INC (THE)	531354.915
54	GAP INC	362527.3
8	BATH & BODY WORKS INC	274942.175
74	NORDSTROM INC	248159.506
80	ROSS STORES INC	188529.105
47	FOOT LOCKER INC	167706.0
0	ABERCROMBIE & FITCH -CL A	67874.646
4	ASCENA RETAIL GROUP INC	65366.513
1	AMERN EAGLE OUTFITTERS INC	63138.85
33	DESIGNER BRANDS INC	57096.129

3. All the firms in the Apparel and Accessory Store industry is located in USA and CAN. Hence, there can only be top 2 locations. There is only 1 firm that is located in Canada (JACOBS (JAY) INC) and the rest of the 104 firms are all in the USA.

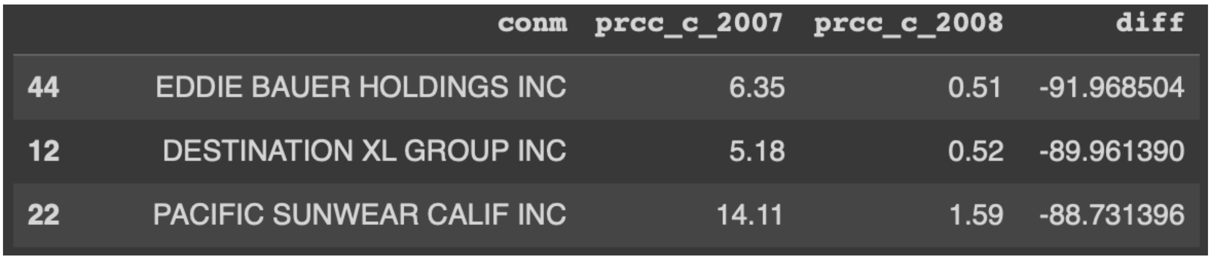


location	conm
CAN	JACOBS (JAY) INC
USA	ABERCROMBIE & FITCH -CL A
	AMERN EAGLE OUTFITTERS INC
	ANN INC
	ARO LIQUIDATION INC
	...
	VICTORIAS SECRET AND CO
	WALKING CO HOLDINGS INC
	WET SEAL INC
	WHITE HOUSE INC-REDH
	ZUMIEZ INC

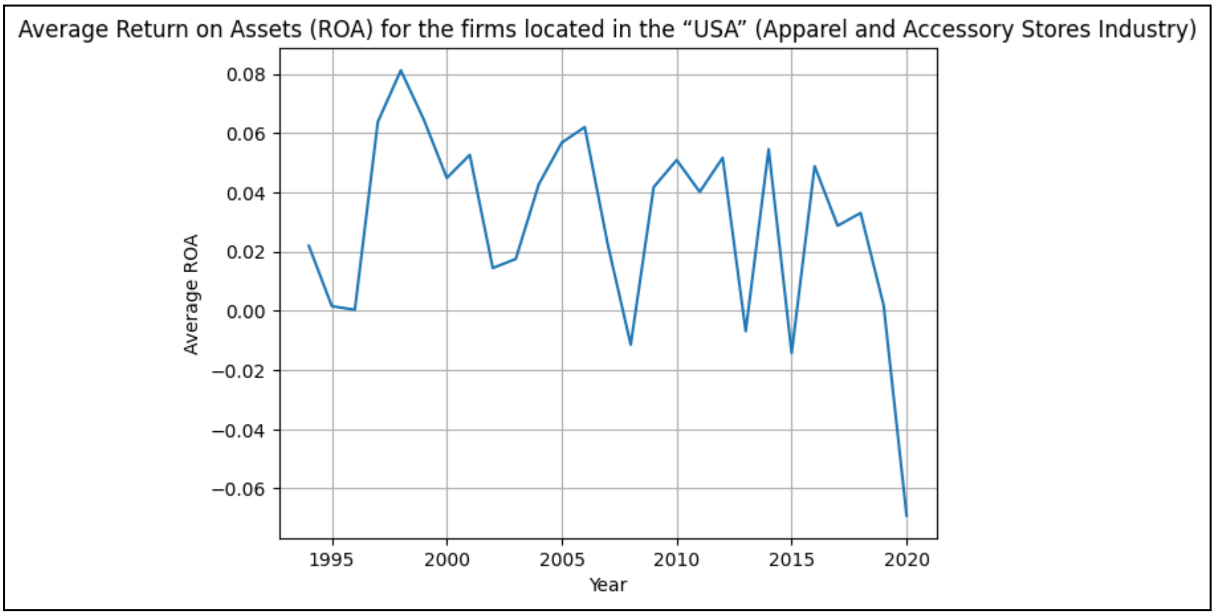
105 rows × 10 columns

4. 

5. EDDIE BAUER HOLDINGS INC was most affected the most by the 2008 Financial Crisis, as measured by the percentage drop of 91.97% in stock price from 2007 to 2008.



	conm	prcc_c_2007	prcc_c_2008	diff
44	EDDIE BAUER HOLDINGS INC	6.35	0.51	-91.968504
12	DESTINATION XL GROUP INC	5.18	0.52	-89.961390
22	PACIFIC SUNWEAR CALIF INC	14.11	1.59	-88.731396

6. 

Part 2. Text Analysis on the Industry Sector

C. **[Text Cleaning; 10 points]** The file "data/2020_10K_item1_full.csv" contains a sample of 5,988 firms and their "item 1" content in their 10-K reports in the year 2020.[2] Load the dataset as a DataFrame and create a new column containing the cleaned text for each "item1" content. Follow the steps below to clean the text:

1. Convert all words to lowercase.
2. Remove punctuations.
3. Remove stop words based on the list of English stop words in NLTK.

D. **[Keyword Analysis; 20 points]** Conduct keywords analysis on your selected industry sector(s). Follow the steps below to complete the analysis:

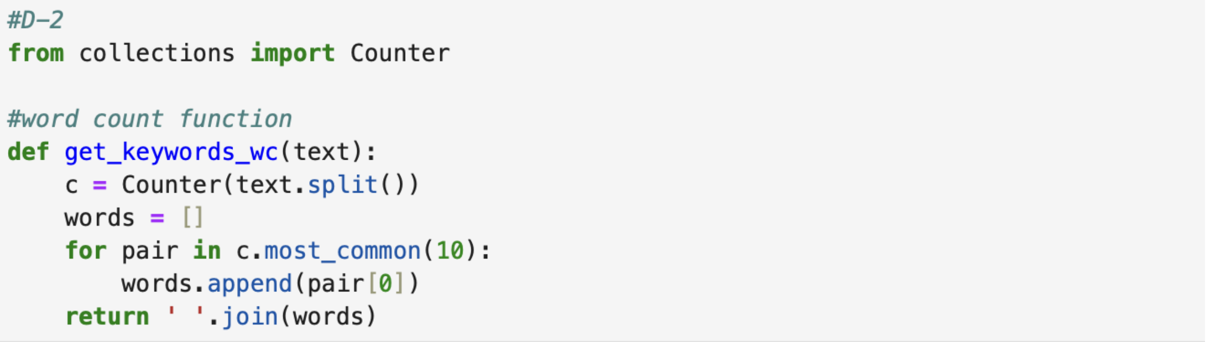
1. Create a new DataFrame that includes only firms in your selected industry sectors. Ensure that you merge the 10-K data with the previous "public_firm.csv" data using an inner join.
2. Generate the top 10 keywords for each firm based on two different methods: word counts and TF-IDF score.
3. Create two wordclouds to visualize the keywords across all firms in the selected sector: one based on the word count of keywords and another based on the TF-IDF score.

E. **[Word embedding; 20 points]** Train a word2vec model and analyze word similarities.

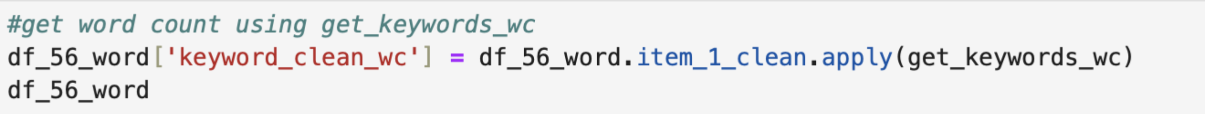
1. Train a word2vec model with the **full 10-K sample** (e.g., "data/2020_10K_item1_full.csv"). Please use the **cleaned text** (e.g., results from Step C) for training.
2. Manually inspect the wordclouds you generated in D.3 and choose three representative keywords that are relevant to the industry sector of your interest. Utilize the trained word2vec model to find the most relevant five words for each of these three keywords.

Part 2. ANSWERS

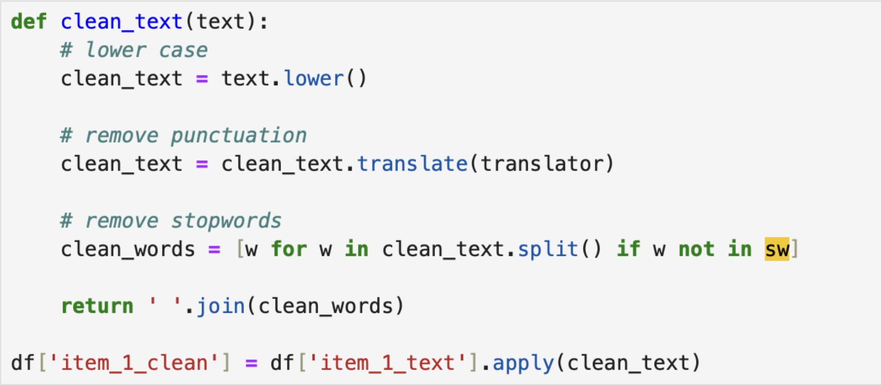
C



```python
#D-2
from collections import Counter

#word count function
def get_keywords_wc(text):
    c = Counter(text.split())
    words = []
    for pair in c.most_common(10):
        words.append(pair[0])
    return ' '.join(words)
```

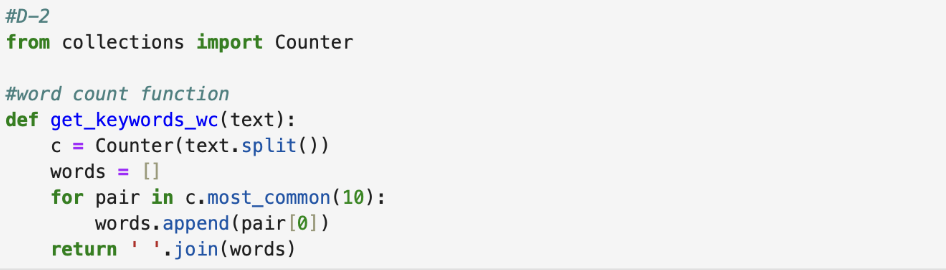


```python
#get word count using get_keywords_wc
df_56_word['keyword_clean_wc'] = df_56_word.item_1_clean.apply(get_keywords_wc)
df_56_word
```

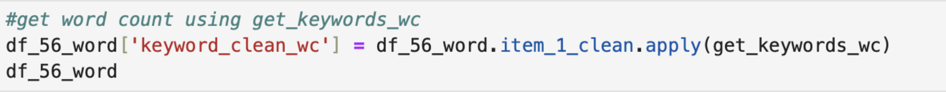


```
def clean_text(text):
    # lower case
    clean_text = text.lower()

    # remove punctuation
    clean_text = clean_text.translate(translator)

    # remove stopwords
    clean_words = [w for w in clean_text.split() if w not in sw]

    return ' '.join(clean_words)

df['item_1_clean'] = df['item_1_text'].apply(clean_text)
```

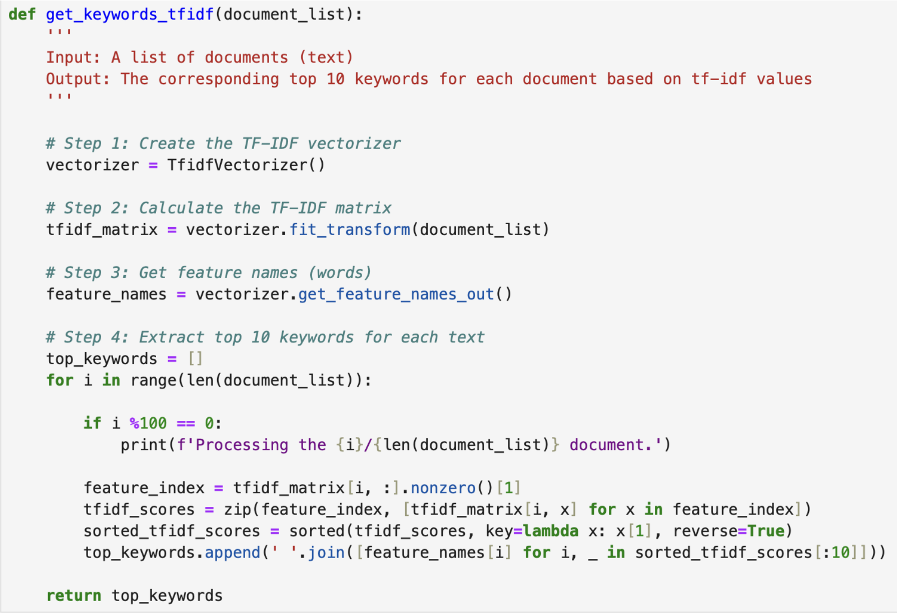
D



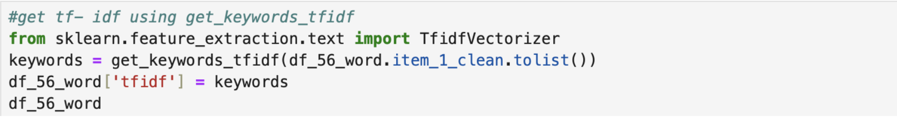
```
#D-2
from collections import Counter

#word count function
def get_keywords_wc(text):
    c = Counter(text.split())
    words = []
    for pair in c.most_common(10):
        words.append(pair[0])
    return ' '.join(words)
```



```
#get word count using get_keywords_wc
df_56_word['keyword_clean_wc'] = df_56_word.item_1_clean.apply(get_keywords_wc)
df_56_word
```

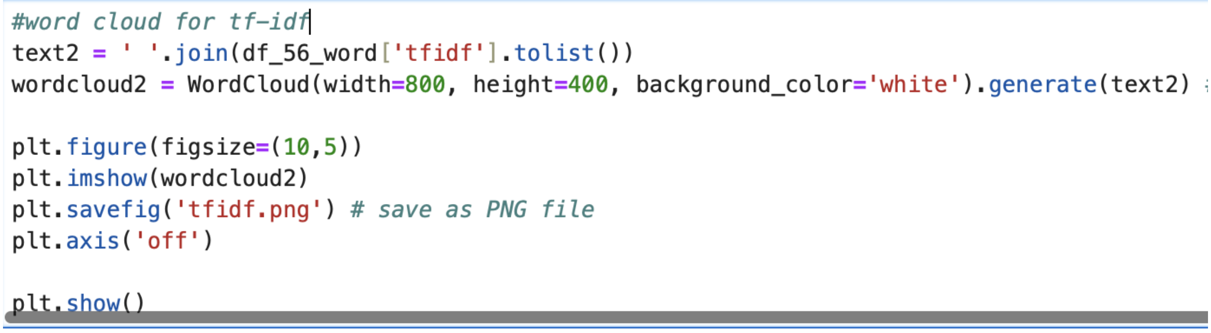


```
def get_keywords_tfidf(document_list):
    '''
    Input: A list of documents (text)
    Output: The corresponding top 10 keywords for each document based on tf-idf values
    '''

    # Step 1: Create the TF-IDF vectorizer
    vectorizer = TfidfVectorizer()

    # Step 2: Calculate the TF-IDF matrix
    tfidf_matrix = vectorizer.fit_transform(document_list)

    # Step 3: Get feature names (words)
    feature_names = vectorizer.get_feature_names_out()

    # Step 4: Extract top 10 keywords for each text
    top_keywords = []
    for i in range(len(document_list)):

        if i %100 == 0:
            print(f'Processing the {i}/{len(document_list)} document.')

        feature_index = tfidf_matrix[i, :].nonzero()[1]
        tfidf_scores = zip(feature_index, [tfidf_matrix[i, x] for x in feature_index])
        sorted_tfidf_scores = sorted(tfidf_scores, key=lambda x: x[1], reverse=True)
        top_keywords.append(' '.join([feature_names[i] for i, _ in sorted_tfidf_scores[:10]]))

    return top_keywords
```

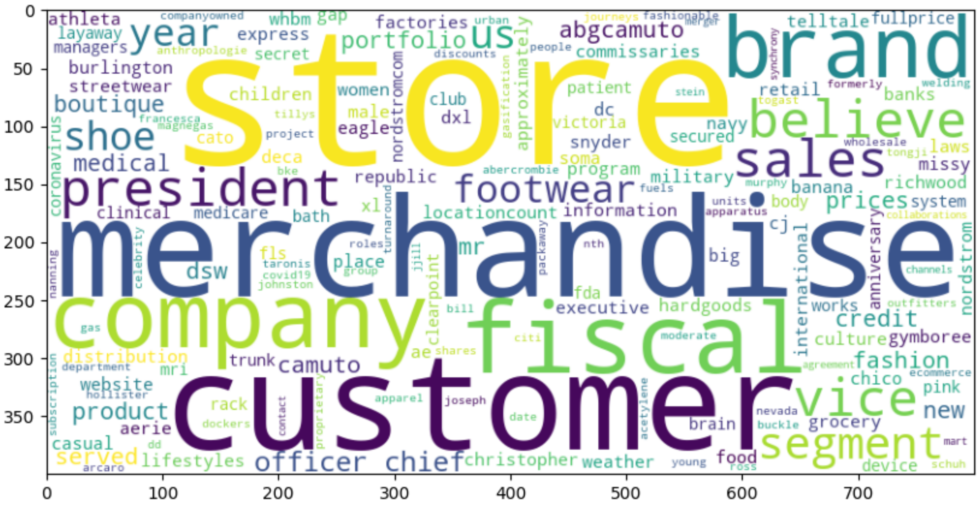


```
#get tf- idf using get_keywords_tfidf
from sklearn.feature_extraction.text import TfidfVectorizer
keywords = get_keywords_tfidf(df_56_word.item_1_clean.tolist())
df_56_word['tfidf'] = keywords
df_56_word
```

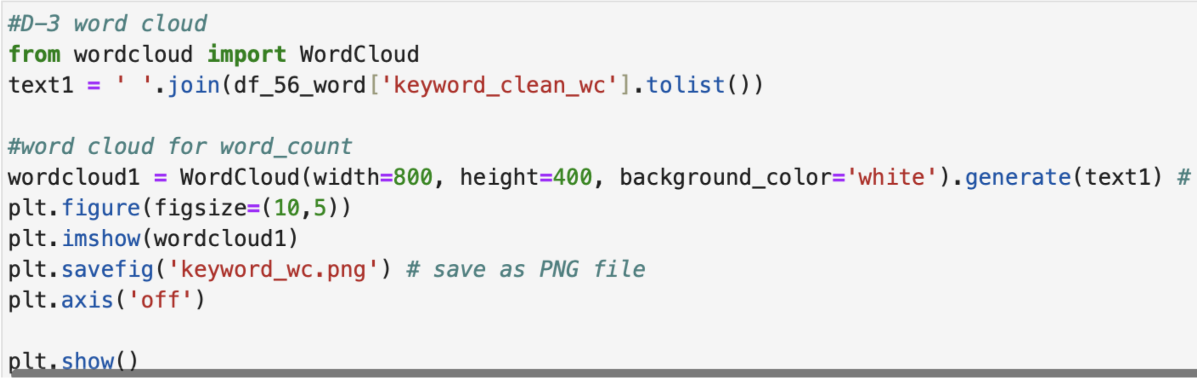
This is the wordcloud based on tfidf.



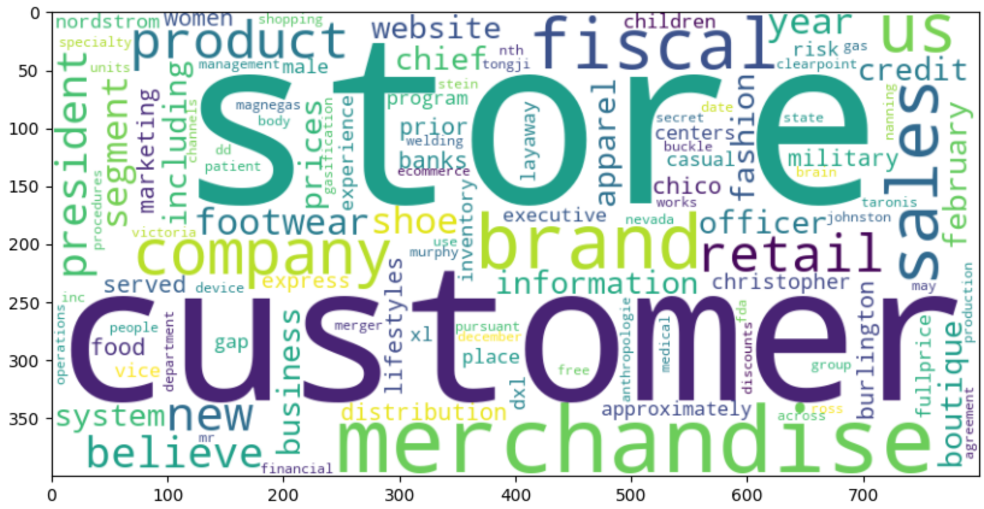
```
#word cloud for tf-idf
text2 = ' '.join(df_56_word['tfidf'].tolist())
wordcloud2 = WordCloud(width=800, height=400, background_color='white').generate(text2)

plt.figure(figsize=(10,5))
plt.imshow(wordcloud2)
plt.savefig('tfidf.png') # save as PNG file
plt.axis('off')

plt.show()
```



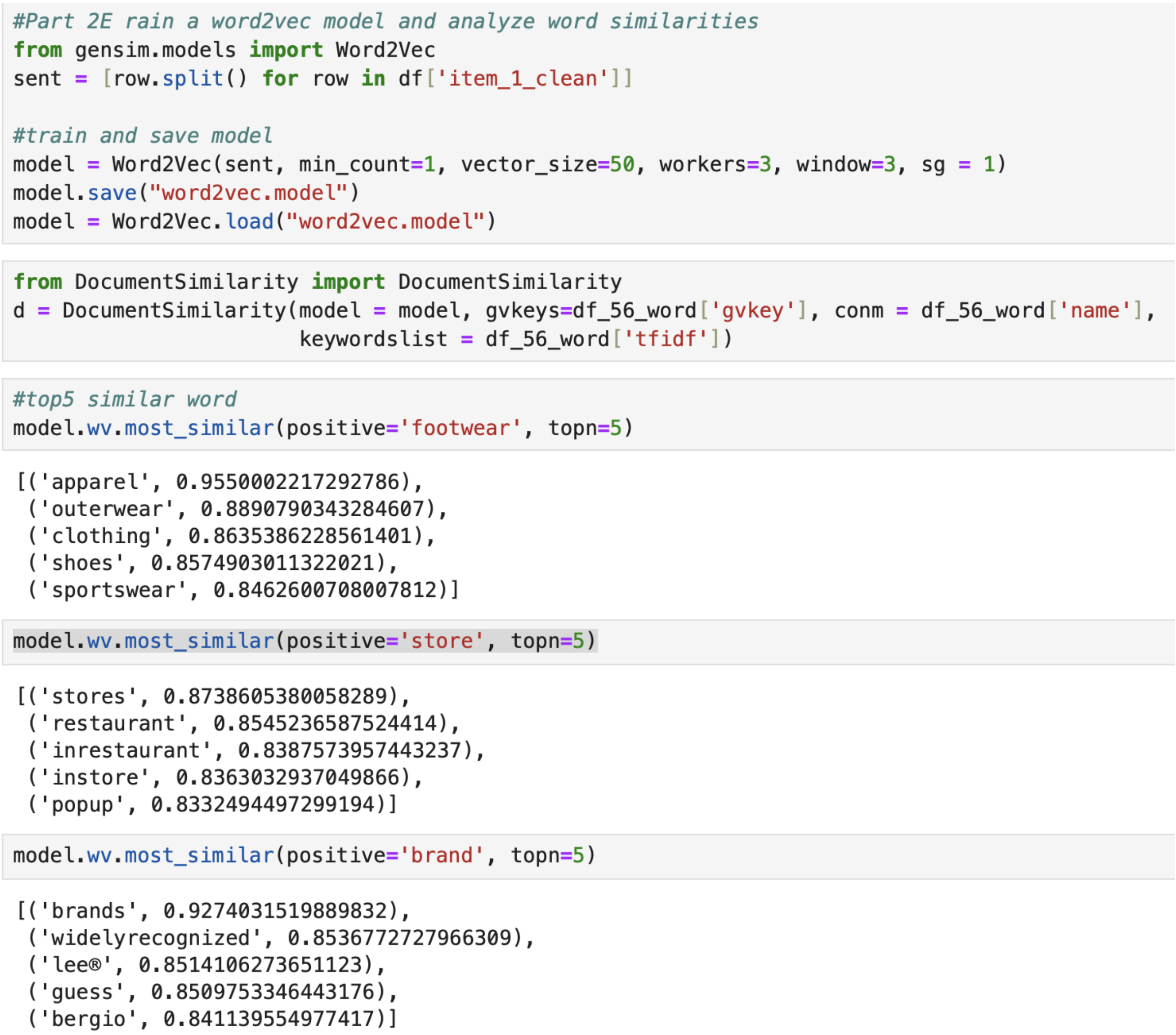
This is the wordcloud based on word count



```
#D-3 word cloud
from wordcloud import WordCloud
text1 = ' '.join(df_56_word['keyword_clean_wc'].tolist())

#word cloud for word_count
wordcloud1 = WordCloud(width=800, height=400, background_color='white').generate(text1) #
plt.figure(figsize=(10,5))
plt.imshow(wordcloud1)
plt.savefig('keyword_wc.png') # save as PNG file
plt.axis('off')

plt.show()
```



E



```
#Part 2E rain a word2vec model and analyze word similarities
from gensim.models import Word2Vec
sent = [row.split() for row in df['item_1_clean']]

#train and save model
model = Word2Vec(sent, min_count=1, vector_size=50, workers=3, window=3, sg = 1)
model.save("word2vec.model")
model = Word2Vec.load("word2vec.model")
```

```
from DocumentSimilarity import DocumentSimilarity
d = DocumentSimilarity(model = model, gvkeys=df_56_word['gvkey'], conm = df_56_word['name'],
                       keywordslist = df_56_word['tfidf'])
```

```
#top5 similar word
model.wv.most_similar(positive='footwear', topn=5)
```

```
[('apparel', 0.9550002217292786),
 ('outerwear', 0.8890790343284607),
 ('clothing', 0.8635386228561401),
 ('shoes', 0.8574903011322021),
 ('sportswear', 0.8462600708007812)]
```

```
model.wv.most_similar(positive='store', topn=5)
```

```
[('stores', 0.8738605380058289),
 ('restaurant', 0.8545236587524414),
 ('inrestaurant', 0.8387573957443237),
 ('instore', 0.8363032937049866),
 ('popup', 0.8332494497299194)]
```

```
model.wv.most_similar(positive='brand', topn=5)
```

```
[('brands', 0.9274031519889832),
 ('widelyrecognized', 0.8536772727966309),
 ('lee®', 0.8514106273651123),
 ('guess', 0.8509753346443176),
 ('bergio', 0.841139554977417)]
```

Part 3. Comprehensive Analysis of One Sample Firm

F. **[Firm Analysis and Strategy Suggestion; 10 points]** This is an open question. Pick one firm that you are interested in and try to analyze its market status. The ultimate goal is to provide **one valuable suggestion** to the firm based on your analysis. Some directions you might consider are, but not limited to:

1. Convert the keywords extracted in D.2 into word embeddings with the word2vec model trained in E.1. Add up the embeddings for each firm to create the firm-level embeddings. Use the firm-level embeddings to find the focal firm's competing firms (or, most similar firms).
2. Compare the revenue, market share, and ROA of the focal firm to its competitors and provide suggestions accordingly.
3. Perform an analysis of the historical stock prices, ROA, revenue, and assets of the chosen company. Investigate potential correlations and address noteworthy decreases and increases.

Note: Please focus on **one** direction and provide **one** suggestion to the firm. It is a busy time with many finals and projects. We don't want to overwhelm you with an extensive report 😊.

Project Report

This report inspects the historical performance of GAP Inc, with a focus on its sales, market share and ROA. Based on this, a suggestion is made to Gap Inc. The top 5 most competing firms of GAP Inc as identified through firm-level embeddings are Genesco Inc, Buckle Inc, Shoe Carnival, Amern Eagle Outfitters and Abercrombie and Fitch. This report then focuses on performing an analysis of the historical stock prices, ROA and assets of these companies and investigates potential correlations and noteworthy price changes. It identifies that ROA and stock prices have a moderate positive correlation whereas assets and sales are strongly positively correlated.

As seen in **Figure 1**, Gap Inc has the highest market share throughout the years 1994 to 2020. Gap's greatest competitor in terms of market share is Abercrombie & Fitch and Amern Eagle. However, as seen in **Figure 2**, these two top competitors have better ROA than Gap Inc until 2009 which suggests that they are better in terms of profitability when compared to GAP Inc. 2009 onwards Buckle Inc took over the market in terms of the highest ROA. As the market dominator, Gap should have economies of scale. Gap Inc needs to think about how to improve its efficiency in managing the assets of the company to generate profit. Generating a better profitability will increase Gap Inc's ROA, eventually leading to a higher stock price as ROA and stock price is positively correlated (**Figure 5**).

Figure 1- Market share of Gap Inc and its competitors

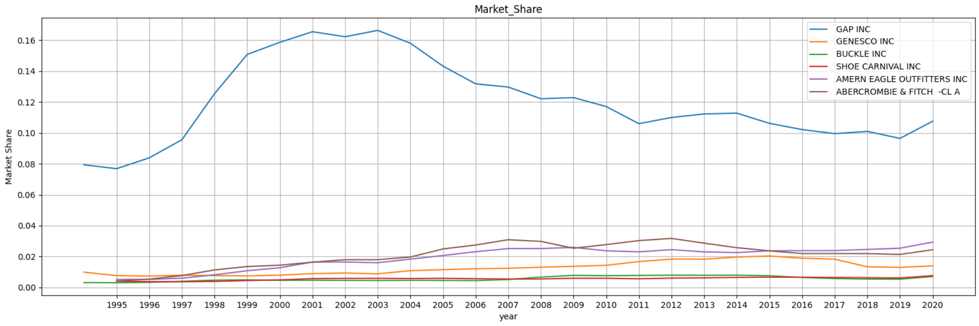


Figure 2- ROA of Gap Inc and its competitors

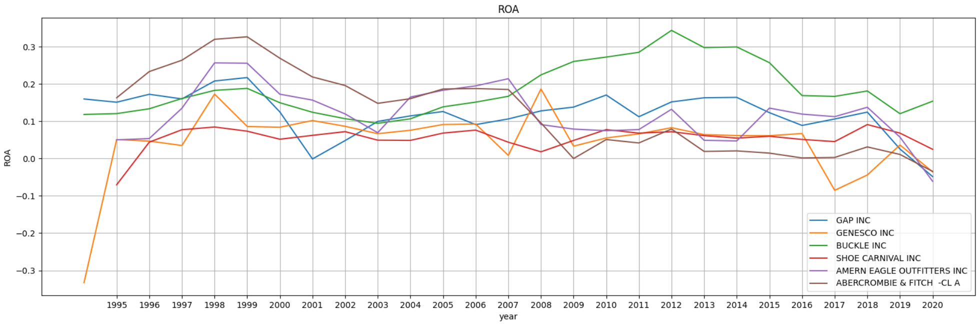
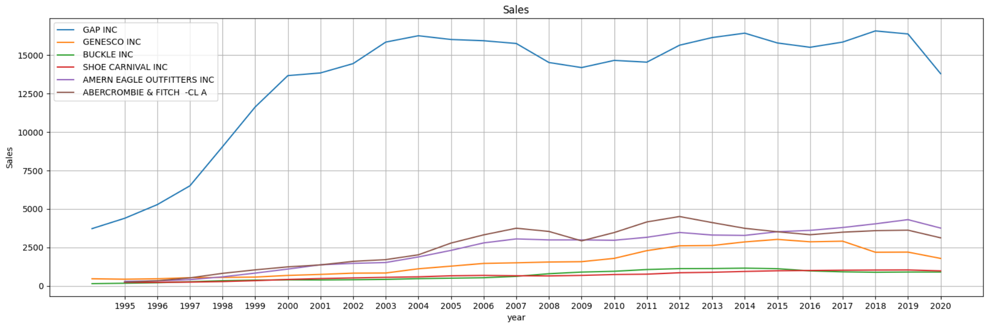
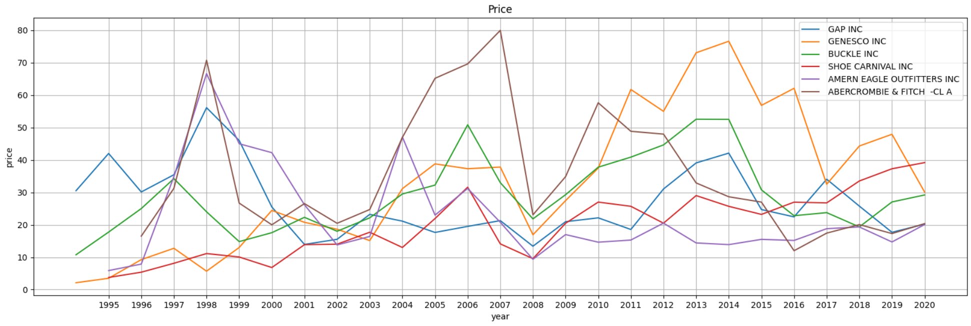


Figure 3- Sales of Gap Inc and its competitors



Looking at the market as a whole shows that Abercrombie & Fitch held a position of considerable popularity in the fashion industry until the year 2012 as reflected by their high stock prices (Figure 4) despite their low market share and sales when compared to GAP Inc. In the early 2000s, Abercrombie & Fitch experienced a notable stock price drop when which could be explained by the possible legal challenges they faced, including lawsuits alleging discrimination within the company. The year 2008 marked the onset of the global financial crisis, leading to a significant downturn in stock prices for all the companies, with Abercrombie and Fitch being hit the most. Post the 2008 financial crisis, there was a noticeable recovery in stock prices across all the companies. Stock values began to ascend during the period from 2008 to 2010, reflecting a broader rebound in the financial markets and greater investor confidence.

Figure 4- Stock prices

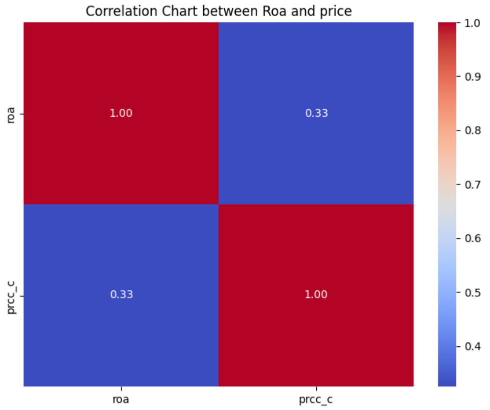


Further analysing reveals key interesting observations regarding the correlations in the market. These industry insights may provide valuable guidance for investors seeking to make informed decisions in the stock market and offer a better understanding of the dynamics within the chosen apparel and accessory stores industry.

1. ROA and Stock Price Correlation:

The correlation charts between Return on Assets (ROA) and stock prices (PRCC) reveal a strong positive correlation. This means that as ROA increases, the stock price tends to rise as well. This finding underscores the significance of ROA as an essential indicator of a company's performance in the stock market. Investors can use ROA as a valuable tool to identify well-performing stocks, as a higher ROA is indicative of a company's ability to generate profitable returns.

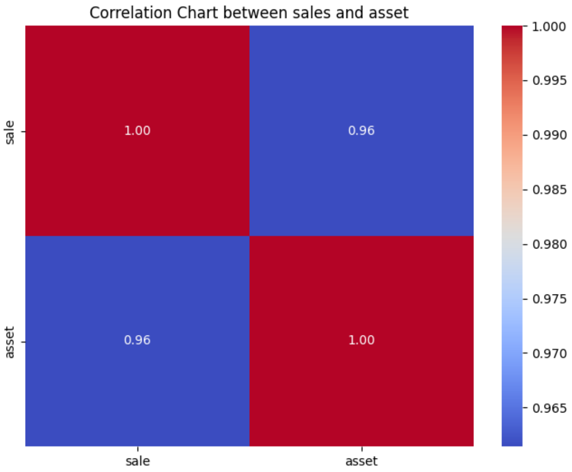
Figure 5- Correlation between ROA and price



2. Asset and Sales Correlation:

The correlation chart between total assets and sales shows a visible positive correlation. This observation suggests that the size of a company, as measured by its total assets, directly impacts its sales performance. This correlation is relevant in the context of the chosen apparel and accessory stores industry. In this capital-intensive and retail-focused sector, companies with more substantial assets, including inventory, cash reserves, and production facilities, are better positioned to generate higher sales volumes. As such, total assets play a crucial role in shaping the revenue-generating capacity of companies within this industry.

Figure 6- Correlation between sales and asset



In conclusion, GAP Inc dominates the market in terms of market share. The rise and fall of its stock prices mirror the cyclical nature of the fashion industry, which often moves in synchrony with the broader economic landscape. This report identifies that while GAP Inc may be the leader in terms of market share, however, its low ROA as compared to its competitors are a cause of concern for the profitability and long term sustenance of the company. This could suggest that GAP Inc has over invested in its assets and are generating lower profit again every dollar invested. Hence, it is suggested that GAP Inc. must improve their ROA by using their assets more efficiently to generate better profitability which will eventually lead to better stock performance as well.