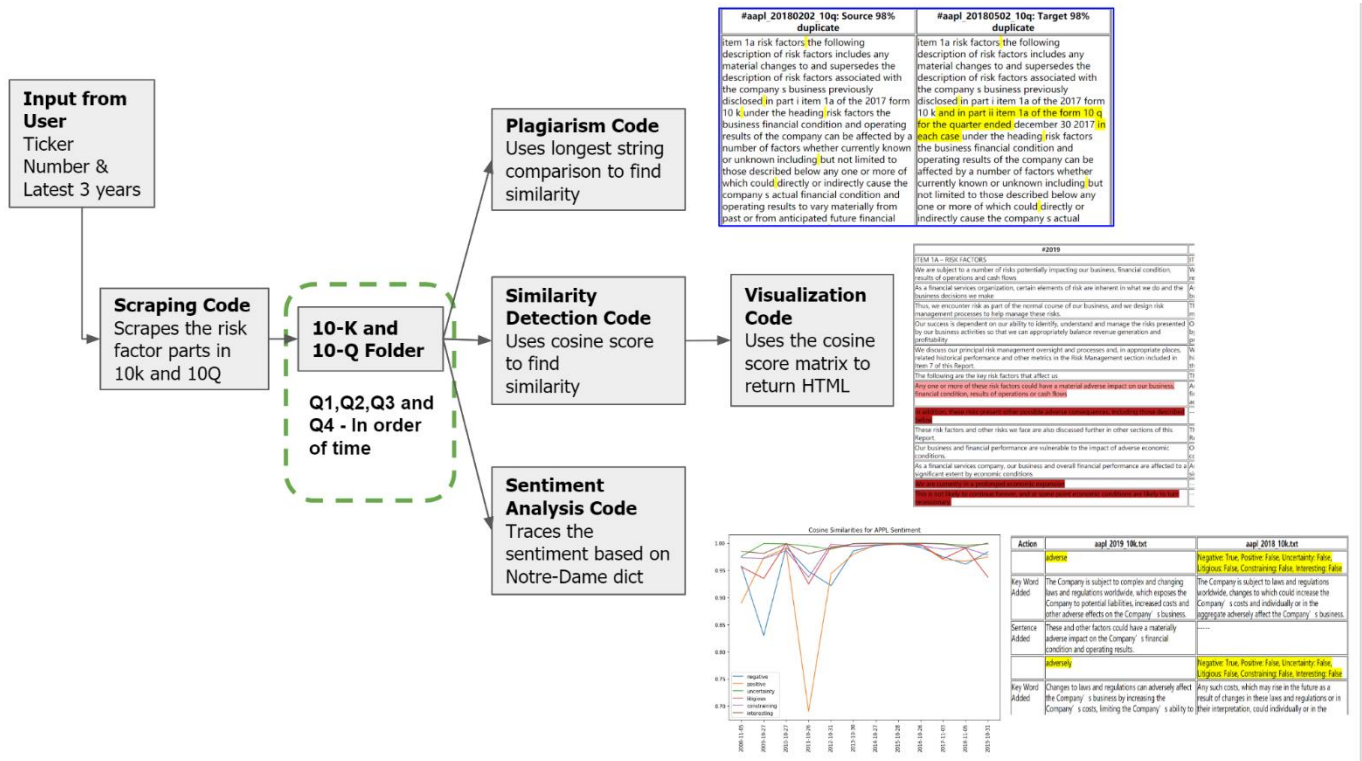


Walkthrough of the code:

Following is the flow of the code:



Part 1: The Web Scraping code

This code reads the ticker.txt file containing all companies in the EDGAR database that is provided by SEC for developers (<https://www.sec.gov/about/webmaster-faq.htm#developers>), then scrapes the 'risk factor' section of the filings made by these companies in the past 3 years in chronological order. However, this step might take around 11 hrs to scrape all the files from the EDGAR database.



















The user can also provide a list of company ticker they want to analyze, for example, 'aapl' or a list of tickers [appl,amzn..] and the years [2017,2018]. The scraper will then scrape only for these tickers and years. (This will take less than a minute)

The final output is kept in a folder named 'risk_factors' (the code creates this folder in case there is none) in form of text files for each combination of company and filing. The naming format of these text files is: "ticker_filingDate_10K/10Q"

Refer to the Web_scraper_README.ipynb for ReadMe file and Risk_Factor_Scraper.ipynb for full code. The outputs will look as follows:

Folder name: risk_factors

Files in the folder (for aapl as input):

	aapl_20180202_10q.txt	
	aapl_20180502_10q.txt	
	aapl_20180801_10q.txt	
	aapl_20181105_10k.txt	
	aapl_20190130_10q.txt	
	aapl_20190501_10q.txt	
	aapl_20190731_10q.txt	
	aapl_20191031_10k.txt	
	aapl_20200129_10q.txt	

Individual File example:

```
>Item 1A.

Risk Factors

The following discussion of risk factors contains forward-looking statements. These risk factors may be important to understanding other statements in this Form 10-K. The following information should be read in conjunction with Part II, Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations" and the consolidated financial statements and accompanying notes in Part II, Item 8, "Financial Statements and Supplementary Data" of this Form 10-K.

The business, financial condition and operating results of the Company can be affected by a number of factors, whether currently known or unknown, including but not limited to those described below, any one or more of which could, directly or indirectly, cause the Company's actual financial condition and operating results to vary materially from past, or from anticipated future, financial condition and operating results. Any of these factors, in whole or in part, could materially and adversely affect the Company's business, financial condition, operating results and stock price.

Because of the following factors, as well as other factors affecting the Company's financial condition and operating results, past financial performance should not be considered to be a reliable indicator of future performance, and investors should not use historical trends to anticipate results or trends in future periods.

Global and regional economic conditions could materially adversely affect the Company's business, results of operations, financial condition and growth
```

Now the files are scraped from the EDGAR database and present in the 'risk_factors' directory.

Part 2: Code for Analysis of reports

To analyze these reports, user can use any of the following three techniques:

- Similarity Detection
- Sentiment Analysis
- Plagiarism Detection

Technique1: Similarity Detection

This technique utilizes Natural Language Processing (NLP) to find the similarity in the reports. It takes the output of the web scraper (in the risk_factors folder) as input and finds the similarity between the documents taking two at a time. These reports are arranged in chronological order and comparisons are made.

To find the similarity, the code uses the Spacy in-built English dictionary: 'en_core_web_lg' and finally outputs a cosine similarity matrix that indicates the similarity between each of the sentences in the two documents. The values of cosine matrix vary from 0 to 1, with 0 being the least similar and 1 being most similar. The code gives three levels of changes:

Small change: Threshold 0.8 ~ 0.99

Moderate change: Threshold 0.5 - 0.8

Big Change: Threshold less than 0.5

These thresholds can be changed as per the requirements. Refer to the 'Similarity, Similarity Visualization, Sentiment Visualization ReadMe.txt' for ReadMe file and 'SimilarityDetection_SimilarityAndSentiment Visualization.ipynb' for full code. The outputs will look as follows along with a cosine matrix.

```
[aapl_2019_3_10q.txt] [aapl_2019_2_10q.txt]
[aapl_2019_3_10q.txt] Apple Inc. | Q2 2019
[aapl_2019_3_10q.txt] Form 10-Q |
[aapl_2019_2_10q.txt] Form 10-Q |

[aapl_2019_3_10q.txt] Global markets for the Company's products and services are highly competitive and subject to rapid technological change, and the Company may be unable to compete effectively in these markets.
[aapl_2019_2_10q.txt] Form 10-Q |
Global markets for the Company's products and services are highly competitive and subject to rapid technological change, and the Company may be unable to compete effectively in these markets.

[aapl_2019_3_10q.txt] The Company's products and services are offered in highly competitive global markets characterized by aggressive price competition and resulting downward pressure on gross margins, frequent introduction
[aapl_2019_2_10q.txt] The Company's products and services are offered in highly competitive global markets characterized by aggressive price competition and resulting downward pressure on gross margins, frequent introduction

[aapl_2019_3_10q.txt] The Company's ability to compete successfully depends heavily on its ability to ensure a continuing and timely introduction of innovative new products, services and technologies to the marketplace.
[aapl_2019_2_10q.txt] The Company's ability to compete successfully depends heavily on its ability to ensure a continuing and timely introduction of innovative new products, services and technologies to the marketplace.

[aapl_2019_3_10q.txt] The Company believes it is unique in that it designs and develops nearly the entire solution for its products, including the hardware, operating system, numerous software applications and related services
[aapl_2019_2_10q.txt] The Company believes it is unique in that it designs and develops nearly the entire solution for its products, including the hardware, operating system, numerous software applications and related services

[aapl_2019_3_10q.txt] As a result, the Company must make significant investments in
[aapl_2019_2_10q.txt] As a result, the Company must make significant investments in

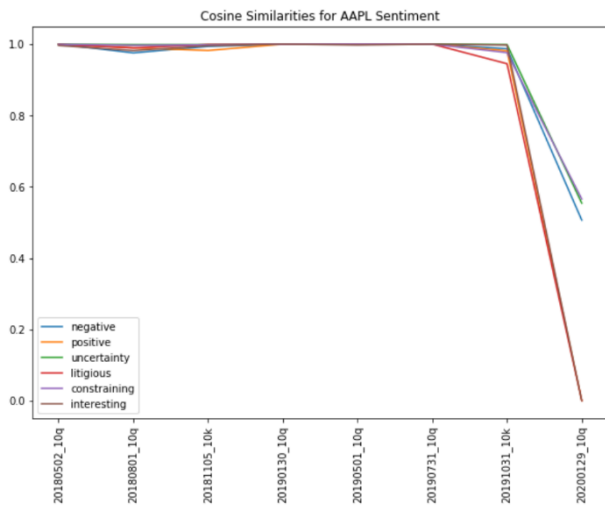
[aapl_2019_3_10q.txt] Item 1A.
[aapl_2019_2_10q.txt] Item 1A.
```

Technique2: Sentiment Analysis

This technique utilizes Natural Language Processing (NLP) and the financial sentiment dictionary from University of Notre Dame 'LoughranMcDonald_MasterDictionary_2018.csv' to find the sentiment in the given reports. The sentiment is divided into 6 categories:

- Negative
- Positive
- Uncertainty
- Litigious
- Constraining
- Interesting

It takes the output of the web scraper (in the risk_factors folder) as input and plots the sentiment values for each of categories for the reports. Refer to the 'SentimentCodeReadMe.txt' for ReadMe file and 'Sentiment analysis Apr 2020' for full code. The outputs will look as follows



Technique3: Plagiarism Detection

This technique utilizes longest string comparison to find the similarity in the reports. The code compares documents word by word to detect plagiarism. The length of the common string can be changed, currently it is min 2 words and maximum 25.

It takes the output of the web scraper (in the risk_factors folder) as input and outputs HTML files as visualization. It also prints the common expression found, its frequency of occurrence and the percentage of similarity between documents.

HTML files that show comparisons between two consecutive documents at a time. The HTML files are linked and can be moved 'Next' or 'Previous' to go to next time frame or previous one respectively

Refer to the 'ReadMe_Plagiarism_Code.txt' for ReadMe file and 'PlagiarismCode' for full code. The outputs will look as follows

Console Output:

```
98 % of document # aapl_20180202_10q is same as document # aapl_20180502_10q
98 % of document # aapl_20180502_10q is same as document # aapl_20180202_10q
Common expressions are:
    services telecommunications mobile communications and media television intellectual property ownership and infringement
    ent tax import and export requirements anti corruption foreign exchange controls and cash repatriation restrictions ( 1 t
    imes )
    releases of confidential information including personally identifiable information that could subject the company to
    significant reputational financial legal and operational consequences the company s business requires ( 1 times )
    business strategies or acquisitions such endeavors may involve significant risks and uncertainties including distrac
    tion of management from current operations greater than expected liabilities and expenses inadequate ( 1 times )
    frequent introduction of new products short product life cycles evolving industry standards continual improvement in
    product price performance characteristics rapid adoption of technological and product advancements ( 1 times )
    jurisdictions are considering imposing additional restrictions these laws continue to develop and may be inconsisten
    t from jurisdiction to jurisdiction complying with emerging and changing international requirements ( 1 times )
    components for its products and build inventory in advance of product announcements and shipments manufacturing purc
    hase obligations typically cover forecasted component and manufacturing requirements for periods ( 1 times )
    activities its corporate headquarters information technology systems and other critical business operations includi
```

HTML Output:

#aapl_20180502_10q: Source 1% duplicate	#aapl_2020_q1: Target 81% duplicate
<p>item 1a risk factors the following description of risk factors includes any material changes to and supersedes the description of risk factors associated with the company's business previously disclosed in part i item 1a of the 2017 form 10 k and in part ii item 1a of the form 10 q for the quarter ended december 30 2017 in each case under the heading risk factors the business financial condition and operating results of the company can be affected by a number of factors whether currently known or unknown including but not limited to those described below any one or more of which could directly or indirectly cause the company's actual financial condition and operating results to vary materially from past or from anticipated future financial condition and operating results any of these factors in whole or in part could materially and adversely affect the company's business financial condition operating results and stock price the following discussion of risk factors contains forward looking statements these risk factors may be important to understanding other statements in this form 10 q the following information should be read in conjunction with the condensed consolidated financial statements and related notes in part i item 1 financial statements and part i item 2 management's discussion and analysis of financial condition and results of operations of this form 10 k because of the following factors as well as other</p>	<p>item 1a risk factors the business financial condition and operating results of the company can be affected by a number of factors whether currently known or unknown including but not limited to those described in part i item 1a of the 2019 form 10 k under the heading risk factors any one or more of which could directly or indirectly cause the company's actual financial condition and operating results to vary materially from past or from anticipated future financial condition and operating results any of these factors in whole or in part could materially and adversely affect the company's business financial condition operating results and stock price there have been no material changes to the company's risk factors since the 2019 form 10 k how are you doing today apple inc q1 2020 form 10 q</p>

Part 3: Visualization of Similarity and Sentiment Code

This code produces HTML outputs for the Similarity and Sentiment techniques of the project. For these, the code takes input as the cosine matrix produced by the similarity code and the Notre Dame dictionary.

Refer to the 'Similarity, Similarity Visualization, Sentiment Visualization ReadMe.txt' for ReadMe file and 'SimilarityDetection_SimilarityAndSentiment Visualization.ipynb' for full code.

The outputs will look as follows

Similarity Detection Output:

Overall Similarity	78.1496%
aapl 2019 10k txt	aapl 2018 10k txt
>	>
Item 1A 2019.	Item 1A 2018.
Risk Factors	Risk Factors
The following discussion of risk factors contains forward-looking statements.	The following discussion of risk factors contains forward-looking statements.
These risk factors may be important to understanding other statements in this Form 10-K.	These risk factors may be important to understanding other statements in this Form 10-K.
The following information should be read in conjunction with Part II, Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations" and the consolidated financial statements and accompanying notes in Part II, Item 8, "Financial Statements and Supplementary Data" of this Form 10-K.	The following information should be read in conjunction with Part II, Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations" and the consolidated financial statements and related notes in Part II, Item 8, "Financial Statements and Supplementary Data" of this Form 10-K.
The business, financial condition and operating results of the Company can be affected by a number of factors, whether currently known or unknown, including but not	The business, financial condition and operating results of the Company can be affected by a number of factors, whether currently known or unknown, including but not

Sentiment Analysis Output:

Action	aapl_2019_10k.txt	aapl_2018_10k.txt
	adverse	Negative: True, Positive: False, Uncertainty: False, Litigious: False, Constraining: False, Interesting: False
Key Word Added	The Company is subject to complex and changing laws and regulations worldwide, which exposes the Company to potential liabilities, increased costs and other adverse effects on the Company' s business.	The Company is subject to laws and regulations worldwide, changes to which could increase the Company' s costs and individually or in the aggregate adversely affect the Company' s business.
Sentence Added	These and other factors could have a materially adverse impact on the Company' s financial condition and operating results.	-----
	adversely	Negative: True, Positive: False, Uncertainty: False, Litigious: False, Constraining: False, Interesting: False
Key Word Added	Changes to laws and regulations can adversely affect the Company' s business by increasing the Company' s costs, limiting the Company' s ability to	Any such costs, which may rise in the future as a result of changes in these laws and regulations or in their interpretation, could individually or in the