

Information Retrieval on 10-K Filings



ABSTRACT

The financial statements of a company provide a lot of valuable insights about it. However, extracting crucial information from these statements is time-consuming and often quite cumbersome. We propose an interactive information retrieval system that can query the financial information of a company, summarise SEC 10-K filings and visualise a company's performance over the years.

The Problem



PROBLEM

Traders and Investors sometimes act on erroneous information which may lead to loss of capital. Not having access to correct information before executing a trade is the same as shooting in the dark.

EXISTING SYSTEM

Securities Exchange Commission in the US maintains a database: EDGAR- Electronic Data Gathering, Analysis, and Retrieval system. This database contains all the documents filed by companies in the US such as financial reports and other disclosures. However, it is complex and requires a deep understanding of system, to retrieve the correct document.

WEAKNESS OF EXISTING SYSTEM

The data is packed as a mixture of long text documents and XBRLs (extensible Business Reporting Language). Studying long texts can be time consuming and not easy for a retail investor

Who is impacted by this problem?



RETAIL INVESTOR

These are small investors who don't want to or don't have the time to research and thus, are at maximum risk of capital loss.



INSTITUTIONAL ANALYSTS

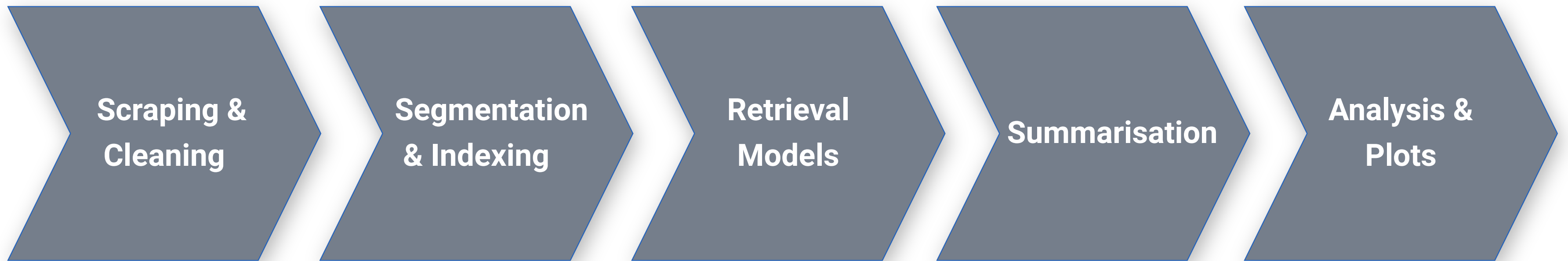
These are the people who do the analysis of a company. Our system extracts crucial data such as balance sheets and income statements which can simplify analysis.



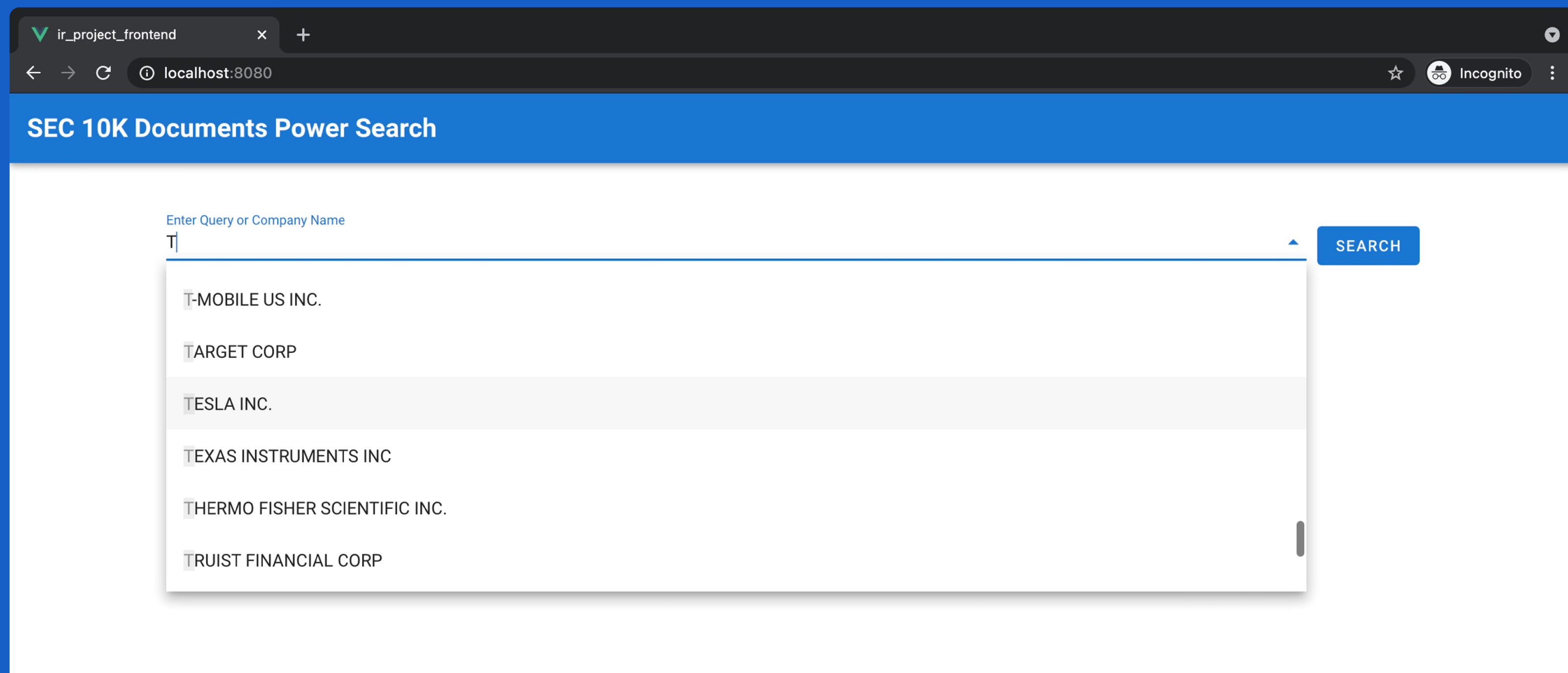
AUDITORS AND LAWYERS

Upcoming features such as extracting policies and disclosures can simplify the task of attorneys and corporate lawyers

Methodology



SEC 10K Documents Power Search



Natural Language Search

User can give natural language queries to search for relevant financial documents.

ir_project_frontend

localhost:8080

SEC 10K Documents Power Search

Enter Query or Company Name

Macbook Sales

DOCUMENT: 461, SCORE: 1.6, DATE OF PUBLISH: 2018-11-05

APPLE INC

and MacBook Pro with Touch Bar™. In October 2018, the Company introduced a new MacBook Air featuring a Retina display and Touch ID, and a new Mac mini with upgraded performance.... was a result of ... was a result of ...

DOCUMENT: 472, SCORE: 1.5, DATE OF PUBLISH: 2019-10-31

APPLE INC

The Company delivers a variety of other services available in certain countries, including Apple Arcade™, a game subscription service; Apple Card™, a co-branded credit card; Apple News+, a subscription news and magazine service; and Apple Pay, a cashless payment service.... shares delivered under a ... was a result of ...

DOCUMENT: 442, SCORE: 1.5, DATE OF PUBLISH: 2016-10-26

APPLE INC

On August 24, 2012, a jury returned a verdict awarding the Company ...In management's opinion, there is not at least a reasonable possibility the Company may have incurred a material loss, or a material loss in excess of a recorded accrual, with respect to loss contingencies, including matters related to infringement of intellectual property rights. However, the outcome

APPLE INC

[Complete Document](#)

Item 1

Item 7

Financial Summary

Tables

☆ Incognito

SEARCH

NETFLIX INC

[Complete Document](#)

Item 1

Item 7

Financial Summary

Tables

Cash Flows

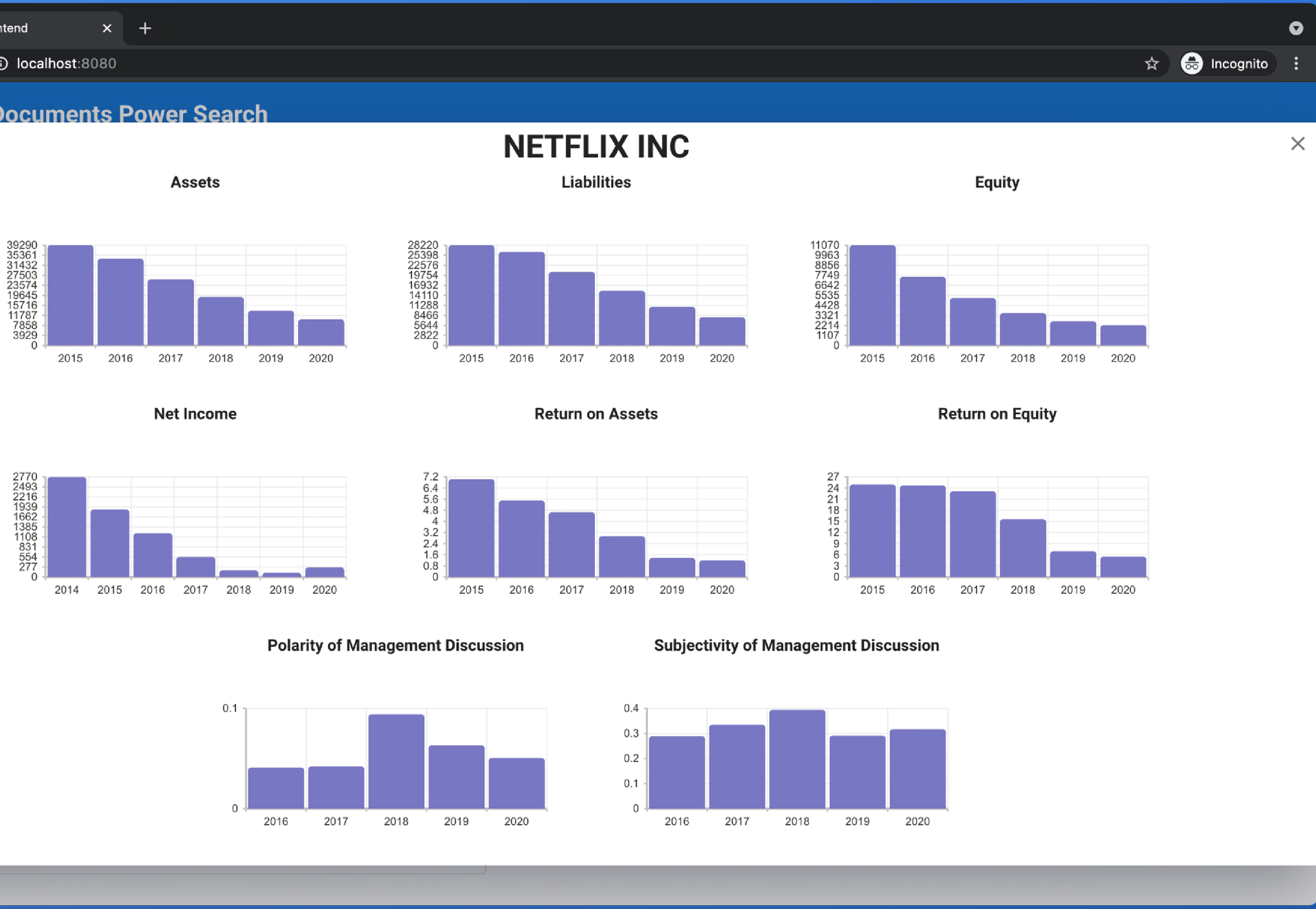
Balance Sheets

Cash and cash equivalents	1467576	1809330
Short-term investments	266206	501385
Current content assets, net	3726307	2905998
Other current assets	260202	215127
Total current assets	5720291	5431840

Financial Summary & Tables

Important parts of 10-K documents of a company are summarised and crucial tables are presented to the user.





Visualitation & Sentiment Analysis

Change in crucial financial indicators of a company and sentiments of management are visualised using plots.

Evaluation

Method	Precision@3	Precision@5	Precision@10
TF-IDF	0.78	0.74	0.54
BM-25	0.66	0.65	0.51
log(TF)-IDF	0.48	0.46	0.41

Handling of Different cases

HANDLING NEW DOCUMENTS

- 10-K documents are filled annually. Around 90% of the companies file their documents during the first quarter, hence our system will have to reindex only during this specific time period. After this time period, our system will be resilient enough to balance less load.
- Each document consists of a fixed format with specific items. Our system passes documents through an automated pipeline, wherein the document is cleaned and preprocessed before indexing, identifying key items of the document, so we can easily incorporate new documents into our system.

HANDLING NEW QUERIES

- We have experimented with different indexing, and preprocessing techniques and our final system use the combination, which gave the best evaluation results for queries whose relevant documents were present in our system.
- We have put in a threshold for the value, such that if for a query if the value is where less than that threshold, then we won't show any document, because in all likelihood there are no relevant documents for that particular query.

Limitations & Challenges

- Currently, our system handles only 10-K form formatted documents. There are several other useful documents present in the database such as quarterly filings, public disclosures, insider trading etc.
- Each document is a combination of text, images, and tables. Each document can be very large in size, and our system may require additional resources for handling a very large dataset of documents.
- Each company may follow a different structure in tables, and the data analysis results may not be accurate always.

The image features a solid blue background. In the top center, there is a curved, segmented shape in white, yellow, and purple. In the bottom left corner, there are several overlapping cylindrical shapes in yellow and purple. In the bottom right corner, there are several overlapping triangular and quadrilateral shapes in white, yellow, purple, and orange.

Thank You