Queen Mary
University of London

Experimental Search Engine Strategy and Demo

Prepared for Information Retrieval

Prepared by: Danielle Souza Da Silva, Georgia Dean, Adam Toth

14 April 2022

Agenda

- **Overview of Original Proposal**
  - Overview of Experiment
  - Overview of Architecture
- **Data**
  - Data Sources
  - Data Scraping Pipeline
  - Data Structures
- **Tooling and Equations**
  - Tools
  - Indexing Architecture
  - Model Configurations
  - Retrieval Architecture
- **Demo**
- **Results**
- **Conclusion**

Overview of Experiment

Review of goals for experimental search engine

Our experimental search engine uses BM25F model to improve retrieval results of regulatory text.

Why BM25F? *BM25*, considered one of the most effective retrieval models, considers texts as unstructured, undifferentiated in any way. As an extension of this model, *BM25F* considers documents to be composed of several fields (title, abstract, body, etc.) with possibly different degrees of importance, term relevance saturation and length normalisation.

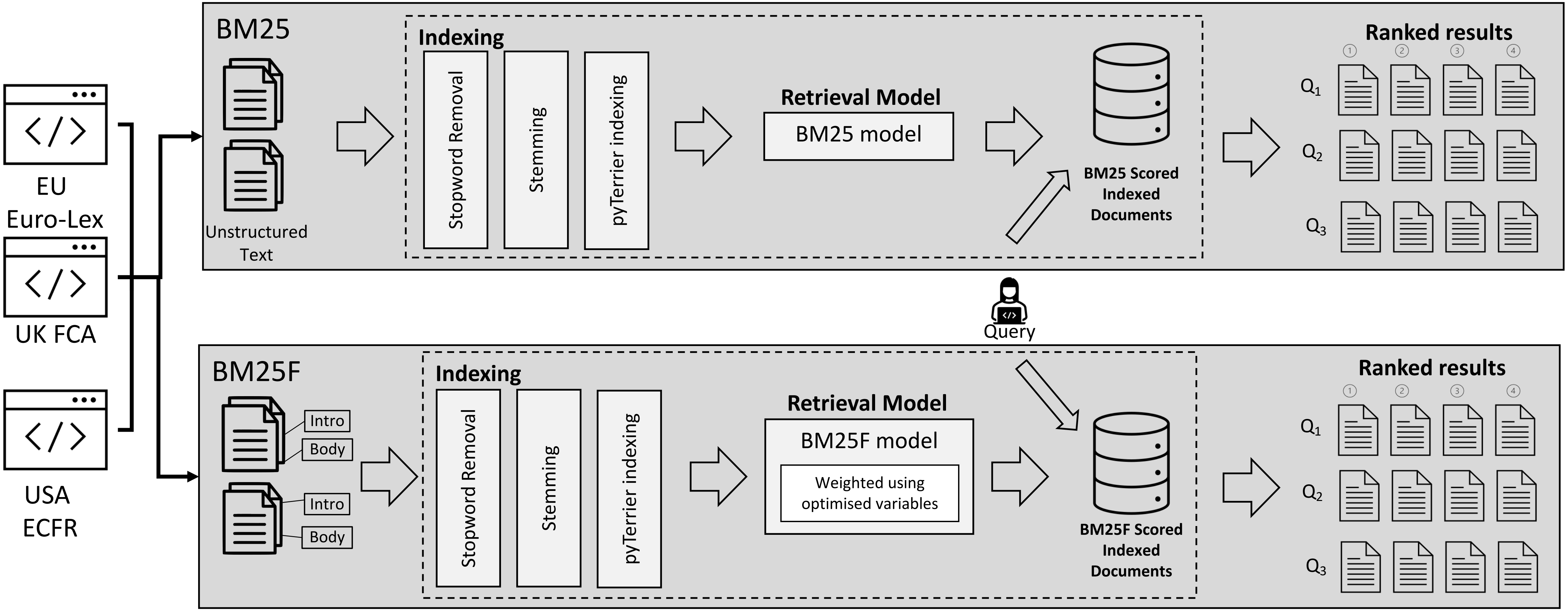
Why regulatory texts? Regulatory texts tend to have consistent structure and unique approach to topic inclusion - highly relevant concepts are mentioned in opening sections, often without repetition in the body. Therefore, some fields may be more predictive of relevance than others.

Goal: by computing and assigning relative weights to pre-specified fields of the documents, the goal is to improve upon the retrieval results of the benchmark model BM25.

Evaluation: performance is measured using mean average precision and recall.

Overview of Architecture

Review search engine architecture



Data Sources

Overview of data and how it was scraped for the purpose of this experiment

Dataset

The dataset used for this experimental model was comprised regulatory text that was scraped by the researchers from various regulatory bodies' websites The sources are regulatory texts from the United States, United Kingdom and European Union focusing specifically on financial regulation for a total of **2387** documents.

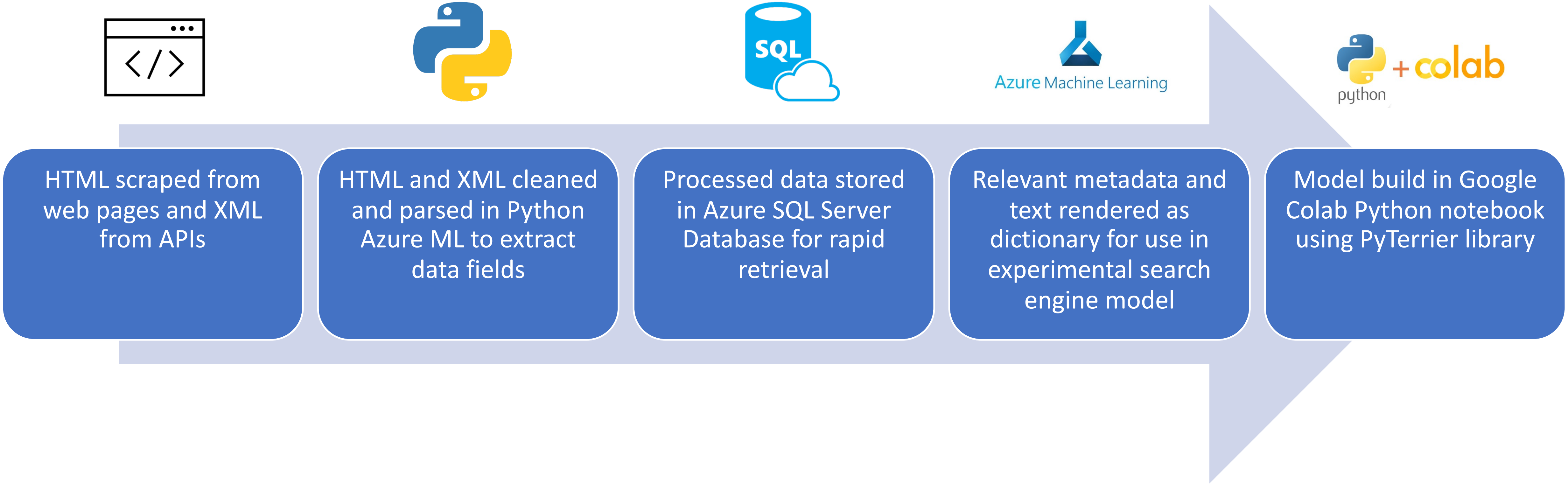
Jurisdiction	Regulations	Number of documents in corpus
United States	E-CFR (Title 12, Title 17)	478
United Kingdom	FCA Handbook	504
European Union	Eur-Lex Financial Regulations	1405

Query relevant document pairs

The query relevant document pairs were completed by domain experts in the regulatory compliance field.
Total number of queries: 12
Range of relevant documents per query: 575

Data Scraping Pipeline

Process used to scrape data



HTML scraped from web pages and XML from APIs

HTML and XML cleaned and parsed in Python Azure ML to extract data fields

Processed data stored in Azure SQL Server Database for rapid retrieval

Relevant metadata and text rendered as dictionary for use in experimental search engine model

Model build in Google Colab Python notebook using PyTerrier library

https://www.iconspng.com/image/68515/python-language-logo
https://www.clipartmax.com/middle/m2H7K9G6b1G6G6m2_azure-sql-database-icon/
https://towardsdatascience.com/azure-machine-learning-service-part-1-an-introduction-739620d1127b
https://miteshparmar1.medium.com/structure-your-code-better-in-google-colab-with-text-and-code-cells-b6fa73feec20

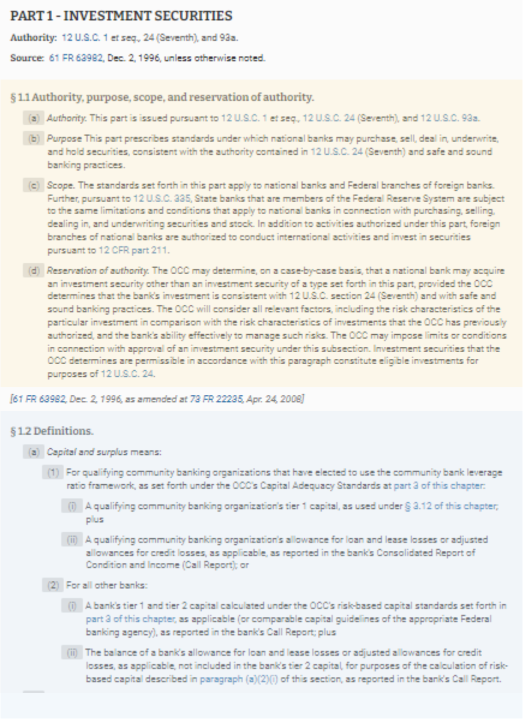
Data Structures

Explanation of how data was structured

ECFR (USA)

Preamble

Text



PART 1 - INVESTMENT SECURITIES

Authority: 12 U.S.C. 1 *et seq.*, 24 (Seventh), and 93a.

Source: 61 FR 63982, Dec. 2, 1996, unless otherwise noted.

§ 1.1 Authority, purpose, scope, and reservation of authority.

(a) *Authority.* This part is issued pursuant to 12 U.S.C. 1 *et seq.*, 12 U.S.C. 24 (Seventh), and 12 U.S.C. 93a.

(b) *Purpose* This part prescribes standards under which national banks may purchase, sell, deal in, underwrite, and hold securities, consistent with the authority contained in 12 U.S.C. 24 (Seventh) and safe and sound banking practices.

(c) *Scope.* The standards set forth in this part apply to national banks and Federal branches of foreign banks. Further, pursuant to 12 U.S.C. 335, State banks that are members of the Federal Reserve System are subject to the same limitations and conditions that apply to national banks in connection with purchasing, selling, dealing in, and underwriting securities and stock. In addition to activities authorized under this part, foreign branches of national banks are authorized to conduct international activities and invest in securities pursuant to 12 CFR part 211.

(d) *Reservation of authority.* The OCC may determine, on a case-by-case basis, that a national bank may acquire an investment security other than an investment security of a type set forth in this part, provided the OCC determines that the bank's investment is consistent with 12 U.S.C. section 24 (Seventh) and with safe and sound banking practices. The OCC will consider all relevant factors, including the risk characteristics of the particular investment in comparison with the risk characteristics of investments that the OCC has previously authorized, and the bank's ability effectively to manage such risks. The OCC may impose limits or conditions in connection with approval of an investment security under this subsection. Investment securities that the OCC determines are permissible in accordance with this paragraph constitute eligible investments for purposes of 12 U.S.C. 24.

[61 FR 63982, Dec. 2, 1996, as amended at 73 FR 22235, Apr. 24, 2008]

§ 1.2 Definitions.

(a) *Capital and surplus* means:

(1) For qualifying community banking organizations that have elected to use the community bank leverage ratio framework, as set forth under the OCC's Capital Adequacy Standards at part 3 of this chapter:

(i) A qualifying community banking organization's tier 1 capital, as used under § 3.12 of this chapter; plus

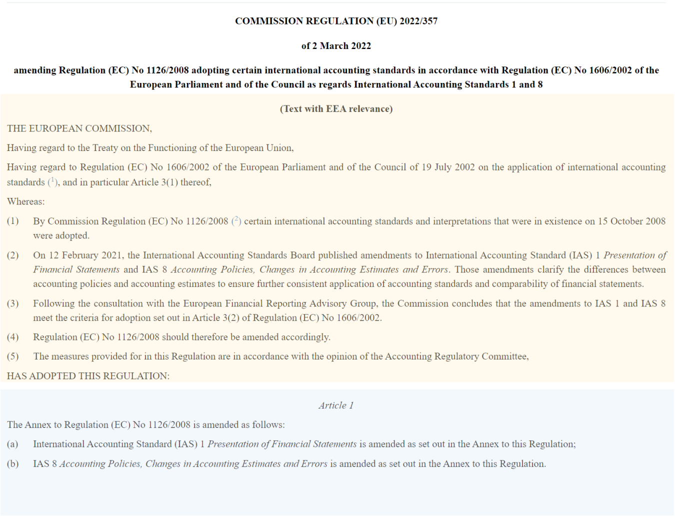
(ii) A qualifying community banking organization's allowance for loan and lease losses or adjusted allowances for credit losses, as applicable, as reported in the bank's Consolidated Report of Condition and Income (Call Report); or

(2) For all other banks:

(i) A bank's tier 1 and tier 2 capital calculated under the OCC's risk-based capital standards set forth in part 3 of this chapter, as applicable (or comparable capital guidelines of the appropriate Federal banking agency), as reported in the bank's Call Report; plus

(ii) The balance of a bank's allowance for loan and lease losses or adjusted allowances for credit losses, as applicable, not included in the bank's tier 2 capital, for purposes of the calculation of risk-based capital described in paragraph (a)(2)(i) of this section, as reported in the bank's Call Report.

Euro-Lex (EU)



COMMISSION REGULATION (EU) 2022/357

of 2 March 2022

amending Regulation (EC) No 1126/2008 adopting certain international accounting standards in accordance with Regulation (EC) No 1606/2002 of the European Parliament and of the Council as regards International Accounting Standards 1 and 8

(Text with EEA relevance)

THE EUROPEAN COMMISSION,

Having regard to the Treaty on the Functioning of the European Union,

Having regard to Regulation (EC) No 1606/2002 of the European Parliament and of the Council of 19 July 2002 on the application of international accounting standards (1), and in particular Article 3(1) thereof,

Whereas:

(1) By Commission Regulation (EC) No 1126/2008 (2) certain international accounting standards and interpretations that were in existence on 15 October 2008 were adopted.

(2) On 12 February 2021, the International Accounting Standards Board published amendments to International Accounting Standard (IAS) 1 *Presentation of Financial Statements* and IAS 8 *Accounting Policies, Changes in Accounting Estimates and Errors*. Those amendments clarify the differences between accounting policies and accounting estimates to ensure further consistent application of accounting standards and comparability of financial statements.

(3) Following the consultation with the European Financial Reporting Advisory Group, the Commission concludes that the amendments to IAS 1 and IAS 8 meet the criteria for adoption set out in Article 3(2) of Regulation (EC) No 1606/2002.

(4) Regulation (EC) No 1126/2008 should therefore be amended accordingly.

(5) The measures provided for in this Regulation are in accordance with the opinion of the Accounting Regulatory Committee,

HAS ADOPTED THIS REGULATION:

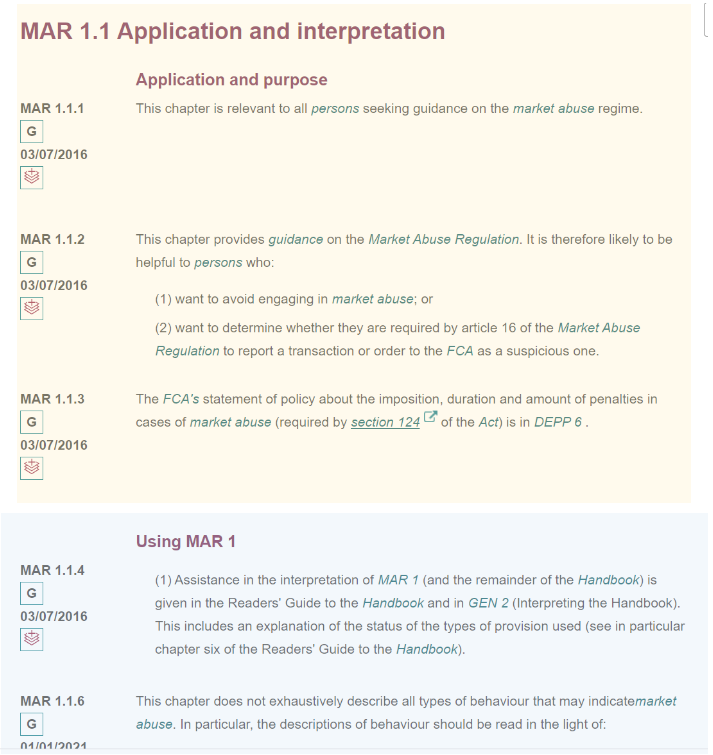
Article 1

The Annex to Regulation (EC) No 1126/2008 is amended as follows:

(a) International Accounting Standard (IAS) 1 *Presentation of Financial Statements* is amended as set out in the Annex to this Regulation;

(b) IAS 8 *Accounting Policies, Changes in Accounting Estimates and Errors* is amended as set out in the Annex to this Regulation.

FCA Handbook(UK)



MAR 1.1 Application and interpretation

Application and purpose

MAR 1.1.1
G
03/07/2016

This chapter is relevant to all *persons* seeking guidance on the *market abuse* regime.

MAR 1.1.2
G
03/07/2016

This chapter provides *guidance* on the *Market Abuse Regulation*. It is therefore likely to be helpful to *persons* who:

(1) want to avoid engaging in *market abuse*; or

(2) want to determine whether they are required by article 16 of the *Market Abuse Regulation* to report a transaction or order to the *FCA* as a suspicious one.

MAR 1.1.3
G
03/07/2016

The *FCA's* statement of policy about the imposition, duration and amount of penalties in cases of *market abuse* (required by section 124 of the *Act*) is in *DEPP 6* .

Using MAR 1

MAR 1.1.4
G
03/07/2016

(1) Assistance in the interpretation of *MAR 1* (and the remainder of the *Handbook*) is given in the Readers' Guide to the *Handbook* and in *GEN 2* (Interpreting the Handbook). This includes an explanation of the status of the types of provision used (see in particular chapter six of the Readers' Guide to the *Handbook*).

MAR 1.1.6
G
01/01/2021

This chapter does not exhaustively describe all types of behaviour that may indicate*market abuse*. In particular, the descriptions of behaviour should be read in the light of:

Tools

Tools utilised for the search engine architecture proposed.

Python was used to implement the search engine components. The libraries that were utilised to scrape and prepare the data, implement the search engine and run experiments with results are listed below.

Python libraryName	Description of use in proposed implementation
Beautiful Soup	It was used to scrape HTML from the web containing the relevant documents.
Requests	It was used to apply HTTP request and get URLs desired for the dataset.
Pickle	This library was used to move data.
Pandas	It was used to read the csv file and transform it to a data frame.
Pyterrier	It was used to index the data, apply the retrieval models, tune the parameters (field weights) for the BM25F model and run experiments.

The main library used to implement each step of the architecture proposed was PyTerrier.

PyTerrier is a platform used for information retrieval experiments in Python. It uses Java-based Terrier information retrieval platform to support indexing and retrieval operations.

References: [pyterrier.readthedocs.io.], [terrier.org.],

Indexing Architecture

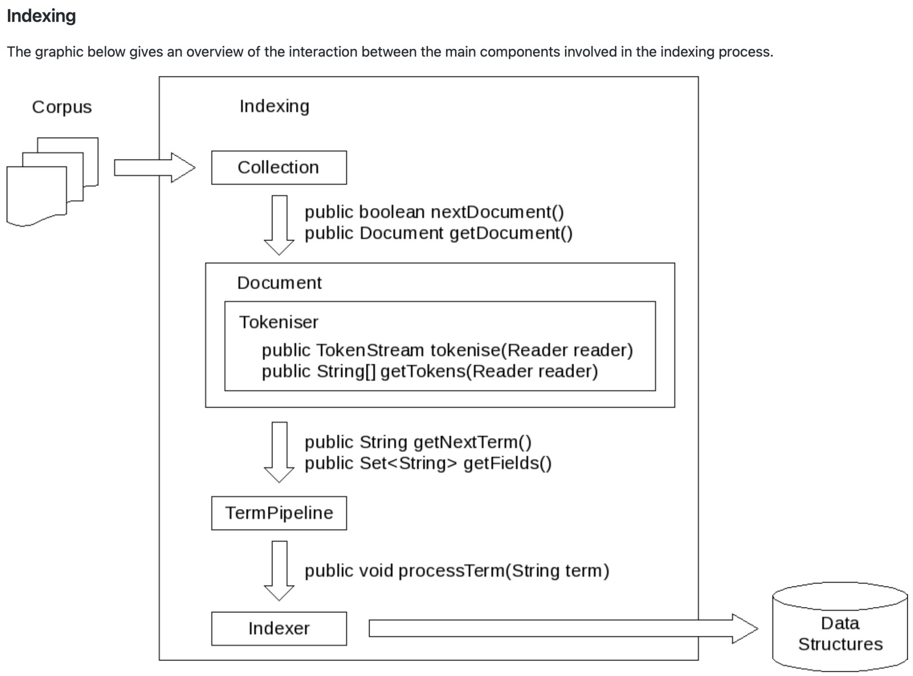
Method used for the search engine architecture proposed.

Indexing

1. A corpus is represented in the form of a collection object. Documents are provided with an instance of a Tokeniser class that breaks pieces of text into single indexing tokens.

2. The indexer manages the indexing process. It iterates over the documents of the collection and sends each term found through a TermPipeline component, which where stemming and stop word removal takes place (default PorterStemmer).

3. Once the terms have been processed through the TermPipeline, they are aggregated and data structures are created by their corresponding DocumentBuilders.



Indexing

The graphic below gives an overview of the interaction between the main components involved in the indexing process.

Corpus

Indexing

Collection

public boolean nextDocument()
public Document getDocument()

Document

Tokeniser

public TokenStream tokenise(Reader reader)
public String[] getTokens(Reader reader)

public String getNextTerm()
public Set<String> getFields()

TermPipeline

public void processTerm(String term)

Indexer

Data Structures

Reference: [GitHub. (2022). *terrier-org/terrier-core*]

Model Configurations

How the models are configured in PyTerrier

Retrieval Models

- **BM25**: PyTerrier uses a java class to implement the Okapi BM25 weighting model.

$$w_i^{BM25}(tf) = \frac{tf}{k_1\left((1-b) + b\frac{dl}{avdl}\right) + tf} \cdot w_i^{RSJ}$$

- **BM25F**: PyTerrier uses a java subclass of PerFieldNormWeightingModel setup to implement BM25F as described by [Zaragoza TREC-2004], which is the same set up described in the search engine proposal.

$$\widetilde{tf}_i = \sum_{s=1}^{S} v_s \frac{tf_{si}}{B_s}$$

$$B_s = \left((1-b_s) + b_s\frac{sl_s}{avsl_s}\right), \qquad 0 \le b \le 1$$

$$w_i^{BM25F} = \frac{\widetilde{tf}_i}{k_1 + \widetilde{tf}_i} w_i^{RSJ}$$

- The parameters k1 and b are the default for both models and were kept consistent for the experiments.

References: [pyterrier.readthedocs.io.], [Robertson, S. and Zaragoza, H., 20]

Retrieval Architecture

Steps of how retrieval results are computed

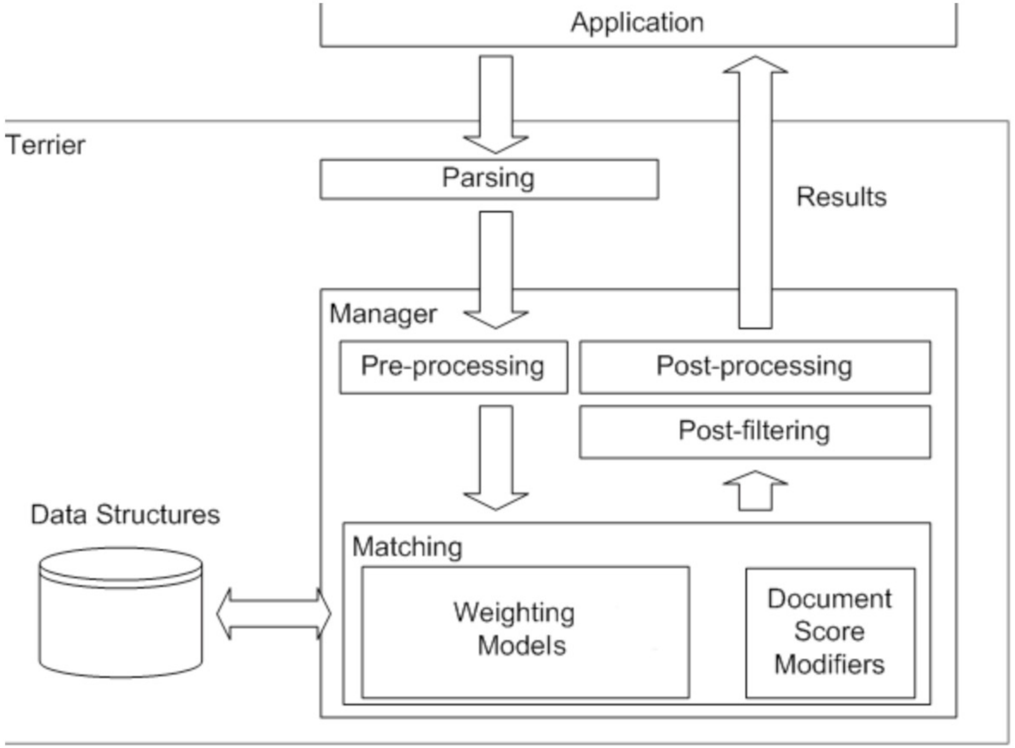
1. The “Application” in the diagram in our case refers to the Google Colab notebook which in the first step issues a query

2. The query is parsed and an instantiation of a query object takes place.

3. The query is then handled to the Manager component, which pre-processes the query by applying it to the configured TermPipeline.

4. The pre-processed query is handled to the Matching component which initializes the Weighting Model (BM25/BM25F) and Document Score Modifiers. Once all components have been instantiated the score computation with respect to the query will take place.

5. The PostProcessing and PostFiltering takes place and so the score document list is returned to the application.



Reference: [GitHub. (2022). *terrier-org/terrier-core*]

Queen Mary
University of London

DEMO

Results

Overall results using BM25 vs BM25F

Overall Results

Model	MAP	Recall@5	Recall@10	Recall@15	Recall@20	Recall@30	Recall@100	Recall@200	Recall@500	Recall@1000
BM25F	56.24%	22.37%	33.87%	41.14%	48.23%	58.09%	74.40%	82.65%	89.55%	93.28%
BM25	57.69%	21.55%	34.97%	42.63%	49.96%	59.55%	76.97%	84.29%	89.75%	93.12%

Recall@k (R@k): The fraction of relevant documents for a query that have been retrieved by rank k.

Parameters

Parameters k1 and b used were the default values for both models. The field weights for the BM25F model can be seen below:

Field	Weight
Text	0.5
Preamble	1.0

Queen Mary
University of London

Results

Results per query using BM25 vs BM25F

Query	Model	MAP	Recall@5	Recall@10	Recall@15	Recall@20	Recall@30	Recall@100	Recall@200	Recall@500	Recall@1000
Commodity derivative	BM25F	52.43%	8.77%	17.54%	21.05%	22.81%	29.82%	63.16%	87.72%	100.00%	100.00%
	BM25	53.24%	8.77%	15.79%	21.05%	21.05%	26.32%	71.93%	92.98%	100.00%	100.00%
Commodity pool operator	BM25F	62.95%	10.81%	18.92%	32.43%	40.54%	59.46%	81.08%	94.59%	100.00%	100.00%
	BM25	70.96%	10.81%	21.62%	35.14%	45.95%	62.16%	97.30%	100.00%	100.00%	100.00%
Derivatives clearing organizations	BM25F	80.99%	26.32%	52.63%	63.16%	68.42%	84.21%	100.00%	100.00%	100.00%	100.00%
	BM25	86.32%	26.32%	52.63%	63.16%	78.95%	89.47%	100.00%	100.00%	100.00%	100.00%
Escheatment	BM25F	100.00%	83.33%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	BM25	100.00%	83.33%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Liquidity risk	BM25F	13.84%	0.00%	0.00%	0.00%	0.00%	0.00%	1.89%	26.42%	76.42%	99.06%
	BM25	14.12%	0.00%	0.00%	0.00%	0.00%	0.00%	0.94%	21.70%	80.19%	99.06%
Major swap participant	BM25F	81.94%	16.13%	32.26%	48.39%	61.29%	74.19%	87.10%	93.55%	96.77%	100.00%
	BM25	82.70%	16.13%	32.26%	48.39%	58.06%	77.42%	90.32%	96.77%	96.77%	100.00%
National bank	BM25F	0.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.42%	20.28%
	BM25	0.42%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	18.40%
Physical commodity swaps	BM25F	57.99%	42.86%	42.86%	57.14%	85.71%	100.00%	100.00%	100.00%	100.00%	100.00%
	BM25	44.09%	28.57%	42.86%	57.14%	85.71%	100.00%	100.00%	100.00%	100.00%	100.00%
Proprietary trading	BM25F	32.44%	13.64%	22.73%	27.27%	36.36%	40.91%	81.82%	100.00%	100.00%	100.00%
	BM25	39.52%	18.18%	27.27%	31.82%	40.91%	45.45%	77.27%	100.00%	100.00%	100.00%
Swap data repositories	BM25F	56.14%	21.05%	42.11%	47.37%	47.37%	52.63%	89.47%	89.47%	100.00%	100.00%
	BM25	60.07%	21.05%	42.11%	47.37%	52.63%	57.89%	94.74%	100.00%	100.00%	100.00%
Swap execution facility	BM25F	66.34%	14.71%	23.53%	35.29%	47.06%	55.88%	88.24%	100.00%	100.00%	100.00%
	BM25	68.21%	14.71%	23.53%	38.24%	47.06%	55.88%	91.18%	100.00%	100.00%	100.00%
Whistleblower	BM25F	69.28%	30.77%	53.85%	61.54%	69.23%	100.00%	100.00%	100.00%	100.00%	100.00%
	BM25	72.62%	30.77%	61.54%	69.23%	69.23%	100.00%	100.00%	100.00%	100.00%	100.00%

Queen Mary
University of London

Conclusion

Final thoughts and further research for BM25F and regulation information retrieval

Conclusions

- BM25 is still the model to beat!
- BM25F showed potential for the more difficult queries where even BM25 had low performance.
  - The preamble is still an important section of regulation and the highly valuable content it contains continues to prove evasive for search engines to capture in relevance rankings.

Further research

- Field length normalization
  - Preambles had varying lengths that could affect the results. Further research into field length normalization
- Improved use of queries for retrieval
  - It would be beneficial for multi-word queries to rank documents that contain ‘word 1’ AND ‘word2’ higher
- N-grams
  - Using n-grams could better capture the proximity of terms in the text and queries to improve relevant document retrieval for domain-specific topics that are made up of general terms

Division of Work

Task	Team Member	Description	Percentage of work
Data Import and Structure	GD	Bring in and prepare data, both structured and unstructured	15%
Define Queries	GD	Define 15-20 queries and relevant document pairs	5%
Indexing of Data	DD, AT	Using team coding, index data to be scored-Including but not limited to, stop word removal, stemming, tf/idf calculations	5%
Implementation of BM25	DD, AT	Using team coding, implement BM25 model on the unstructured document data.	20%
Implementation of BM25F	GD, DD, AT	Using team coding, implement BM25F model on structured document data	20%
Optimization of Parameters	DD, AT	Optimize BM25F weight parameters and hyper parameters for both models (potentially using gradient descent)	15%
Results Evaluation	GD	Calculate precision, recall and F-scores to determine quality of retravel results	2.5%
Presentation Write-Up	GD, DD, AT	Write presentation of	15%
Record Presentation/Demo	GD, DD, AT	All team members will record and present search engine and results	2.5%

Sources

Na, S., Kang, I., & Lee, J. (2008). Improving Term Frequency Normalization for Multi-topical Documents and Application to Language Modeling Approaches. *ECIR*.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *SIGIR Forum, 51*, 176-184.

pyterrier.readthedocs.io. (n.d.). *Installing and Configuring — PyTerrier 0.8.1 documentation*. [online] Available at: https://pyterrier.readthedocs.io/en/latest/installation.html.

terrier.org. (n.d.). *Terrier IR Platform - Homepage*. [online] Available at: http://terrier.org [Accessed 14 Apr. 2022].

GitHub. (2022). *terrier-org/terrier-core*. [online] Available at: https://github.com/terrier-org/terrier-core/blob/5.x/doc/basicArchitecture.md [Accessed 14 Apr. 2022].

[Zaragoza TREC-2004] . H. Zaragoza, N. Craswell, M. Taylor, S. Saria, S. Robertson: Microsoft Cambridge at TREC 13: Web and Hard Tracks. In Proc. of TREC 2004

[Robertson, S. and Zaragoza, H., 2009.] The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4), pp.333-389.