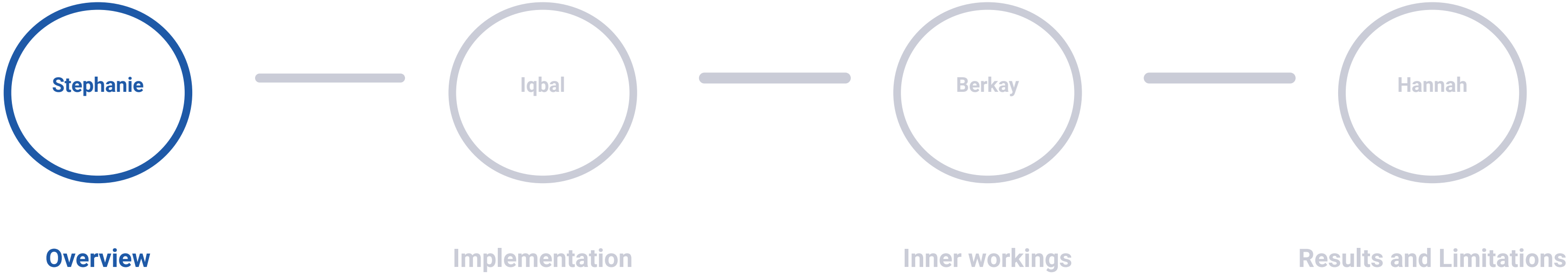
Search Engine Design with BM25 and DESM

ECS735P Information Retrieval - Group 18

Stephanie Nicole Garibay Lim
Berkay Dur
Hannah Melkemaryam Claus
Iqbal Singh

Presentation Structure



Stephanie

Overview

Iqbal

Implementation

Berkay

Inner workings

Hannah

Results and Limitations

Overview

- Importance of identification of the most relevant and accurate news information
  - Incapability to do so results to
    - Inaccurate or incomplete news information
    - Negative consequences such as misinformation and misinterpretation

- An IR system is capable to:
  - Accurate
  - Reliable
  - Timely

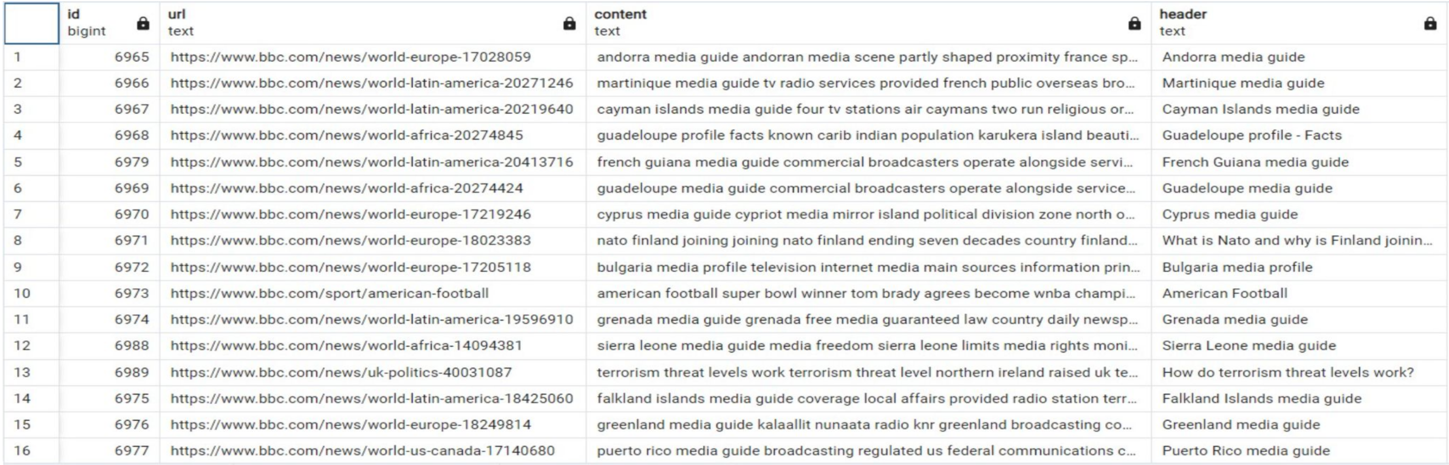
- Project scope:
  - Investigate and evaluate the design of news article search engine with 2 different models
  - Overview of the future design implementation of the system.

(Saracevic, 2010)

Problem statement

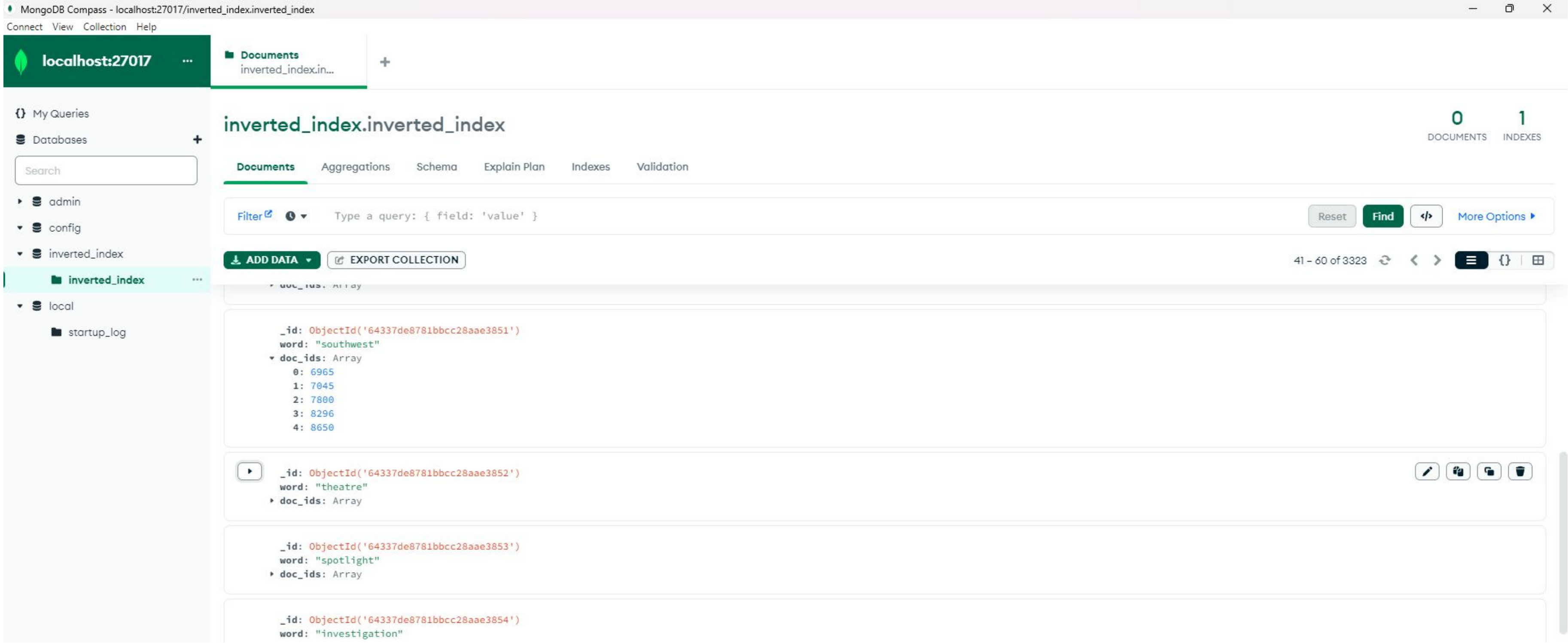
Dataset

- BBC News articles, scraped using Python.
- Contains all the recent news articles within the 20-day window, from the day it was scraped.
- Total data used for this project is **8,333**



	id bigint	url text	content text	header text
1	6965	https://www.bbc.com/news/world-europe-17028059	andorra media guide andorran media scene partly shaped proximity france sp...	Andorra media guide
2	6966	https://www.bbc.com/news/world-latin-america-20271246	martinique media guide tv radio services provided french public overseas bro...	Martinique media guide
3	6967	https://www.bbc.com/news/world-latin-america-20219640	cayman islands media guide four tv stations air caymans two run religious or...	Cayman Islands media guide
4	6968	https://www.bbc.com/news/world-africa-20274845	guadeloupe profile facts known carib indian population karukera island beauti...	Guadeloupe profile - Facts
5	6979	https://www.bbc.com/news/world-latin-america-20413716	french guiana media guide commercial broadcasters operate alongside servi...	French Guiana media guide
6	6969	https://www.bbc.com/news/world-africa-20274424	guadeloupe media guide commercial broadcasters operate alongside service...	Guadeloupe media guide
7	6970	https://www.bbc.com/news/world-europe-17219246	cyprus media guide cypriot media mirror island political division zone north o...	Cyprus media guide
8	6971	https://www.bbc.com/news/world-europe-18023383	nato finland joining joining nato finland ending seven decades country finland...	What is Nato and why is Finland joinin...
9	6972	https://www.bbc.com/news/world-europe-17205118	bulgaria media profile television internet media main sources information prin...	Bulgaria media profile
10	6973	https://www.bbc.com/sport/american-football	american football super bowl winner tom brady agrees become wnba champi...	American Football
11	6974	https://www.bbc.com/news/world-latin-america-19596910	grenada media guide grenada free media guaranteed law country daily newsp...	Grenada media guide
12	6988	https://www.bbc.com/news/world-africa-14094381	sierra leone media guide media freedom sierra leone limits media rights moni...	Sierra Leone media guide
13	6989	https://www.bbc.com/news/uk-politics-40031087	terrorism threat levels work terrorism threat level northern ireland raised uk te...	How do terrorism threat levels work?
14	6975	https://www.bbc.com/news/world-latin-america-18425060	falkland islands media guide coverage local affairs provided radio station terr...	Falkland Islands media guide
15	6976	https://www.bbc.com/news/world-europe-18249814	greenland media guide kalaallit nunaata radio knr greenland broadcasting co...	Greenland media guide
16	6977	https://www.bbc.com/news/world-us-canada-17140680	puerto rico media guide broadcasting regulated us federal communications c...	Puerto Rico media guide

Inverted Index



```
 ▸ doc_ids: Array

   _id: ObjectId('64337de8781bbcc28aae3851')
   word: "southwest"
 ▾ doc_ids: Array
     0: 6965
     1: 7045
     2: 7800
     3: 8296
     4: 8650

   _id: ObjectId('64337de8781bbcc28aae3852')
   word: "theatre"
 ▸ doc_ids: Array

   _id: ObjectId('64337de8781bbcc28aae3853')
   word: "spotlight"
 ▸ doc_ids: Array

   _id: ObjectId('64337de8781bbcc28aae3854')
   word: "investigation"
```

Problem statement

Dataset

Evaluation

Retrieval Models

Evaluation

- Limitations with dataset chosen:
  - Unlabelled data
- Manual scoring
  - Query and result relevance based on:
    - News category
    - News article content

Problem statement

Retrieval Models

- BM25
- Dual Embedding Space Model (DESM)

Retrieval Models

- **BM25**
  - It ranks the relevance of the document by weighing the similarity of the query terms in the document
    - Counting repetition of query terms in the document
  - Assumption:
    - Query terms are more useful for document ranking
  - Uses only original query terms and any additional query will be linked to the original query via relevance

- **DESM**
  - Two embeddings
    - Query words, Q,
    - Document words, D.
  - Ranking function is simply the mean cosine similarity of Q and D bar
  - Takes into account distributional semantics, which incorporates the relationship between words

$$DESM(Q,D) = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{\mathbf{q}_i^T \overline{\mathbf{D}}}{\|\mathbf{q}_i\| \|\overline{\mathbf{D}}\|}$$

Where:

$$\overline{\mathbf{D}} = \frac{1}{|D|}\sum_{\mathbf{d}_j \in D} \frac{\mathbf{d}_j}{\|\mathbf{d}_j\|}$$

(Nalisnick *et al.*, 2016)

Presentation Structure

Stephanie

Overview

Iqbal

Implementation

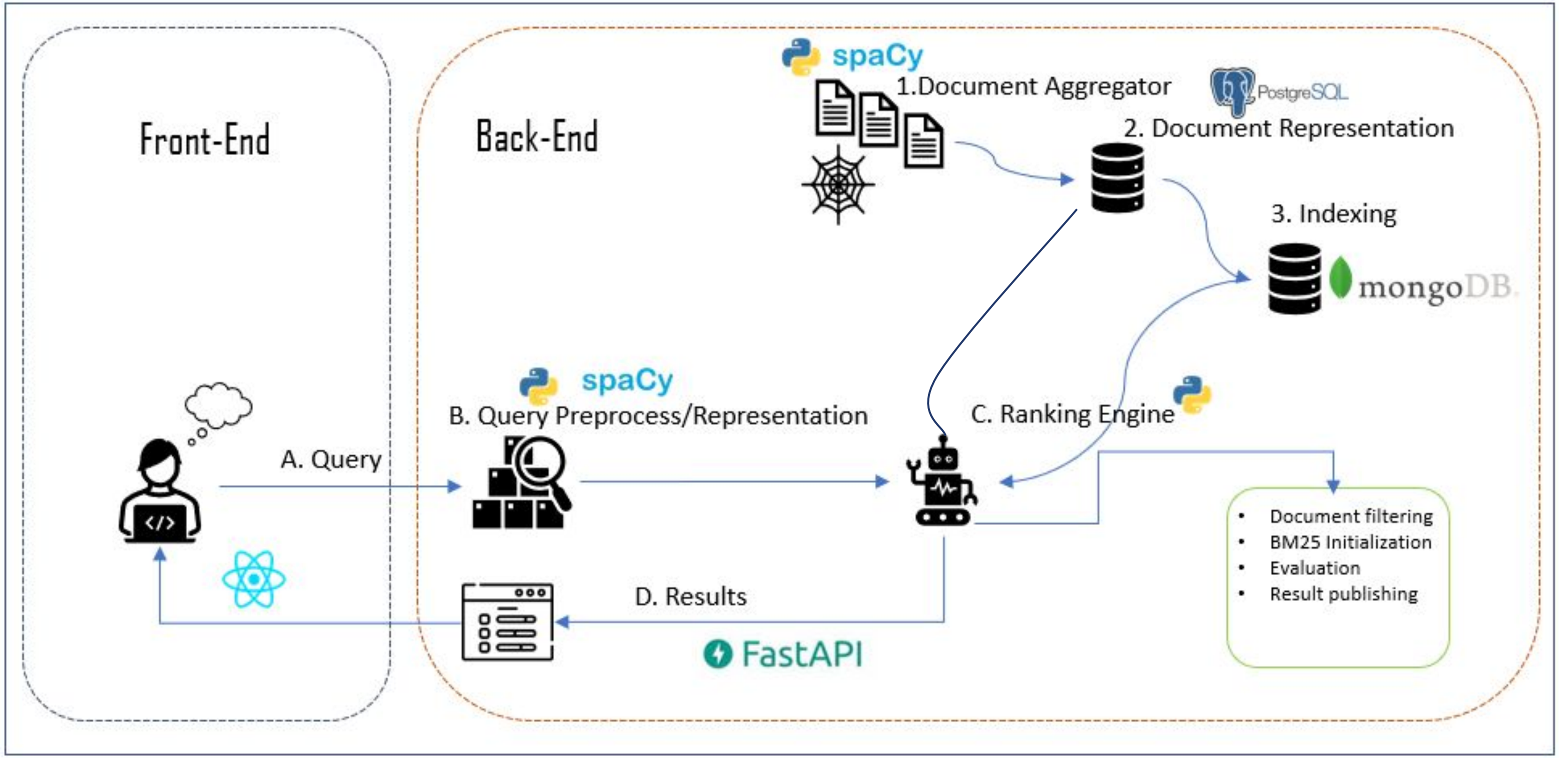
Berkay

Inner workings

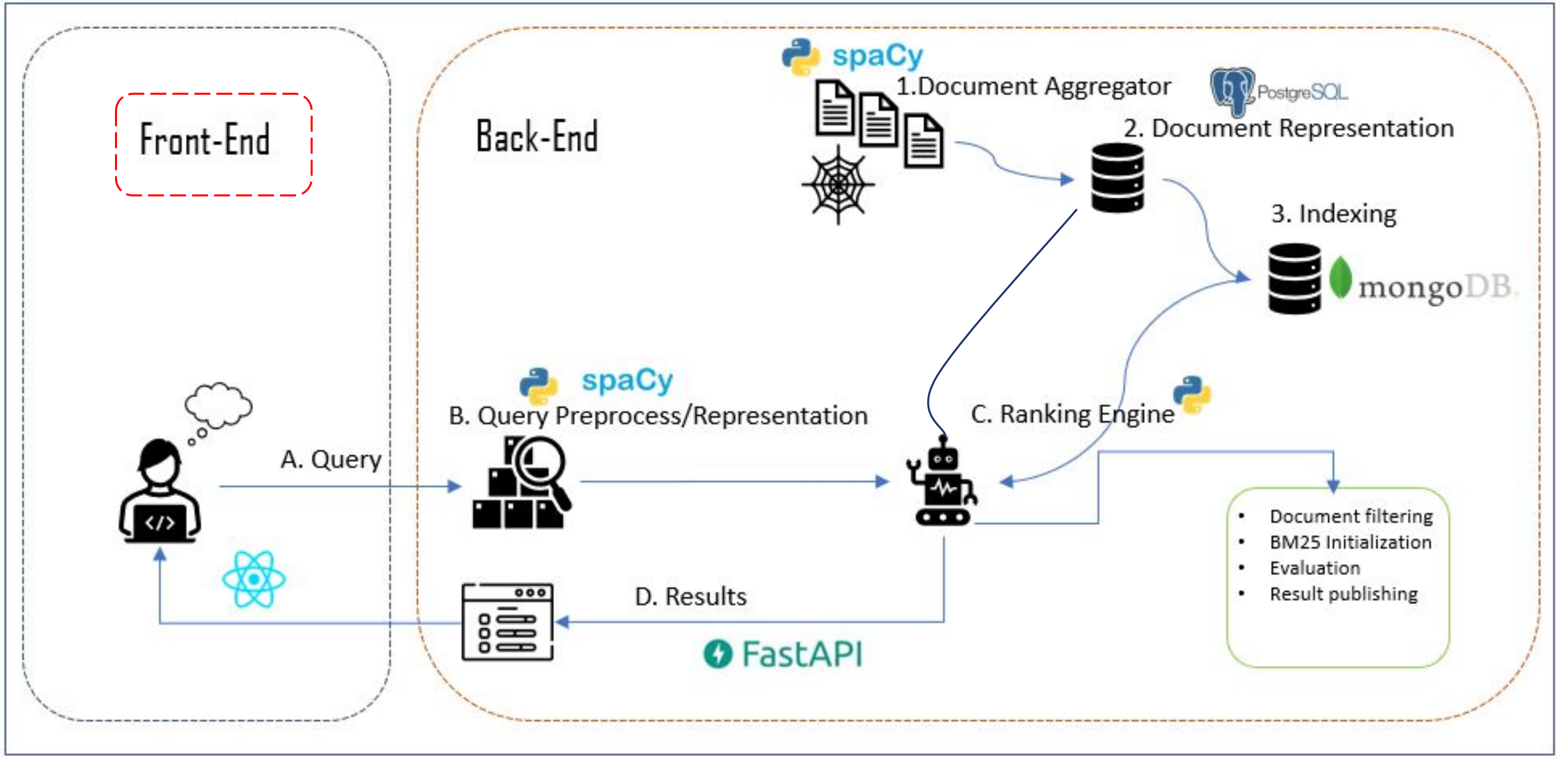
Hannah

Results and Limitations

Implementation

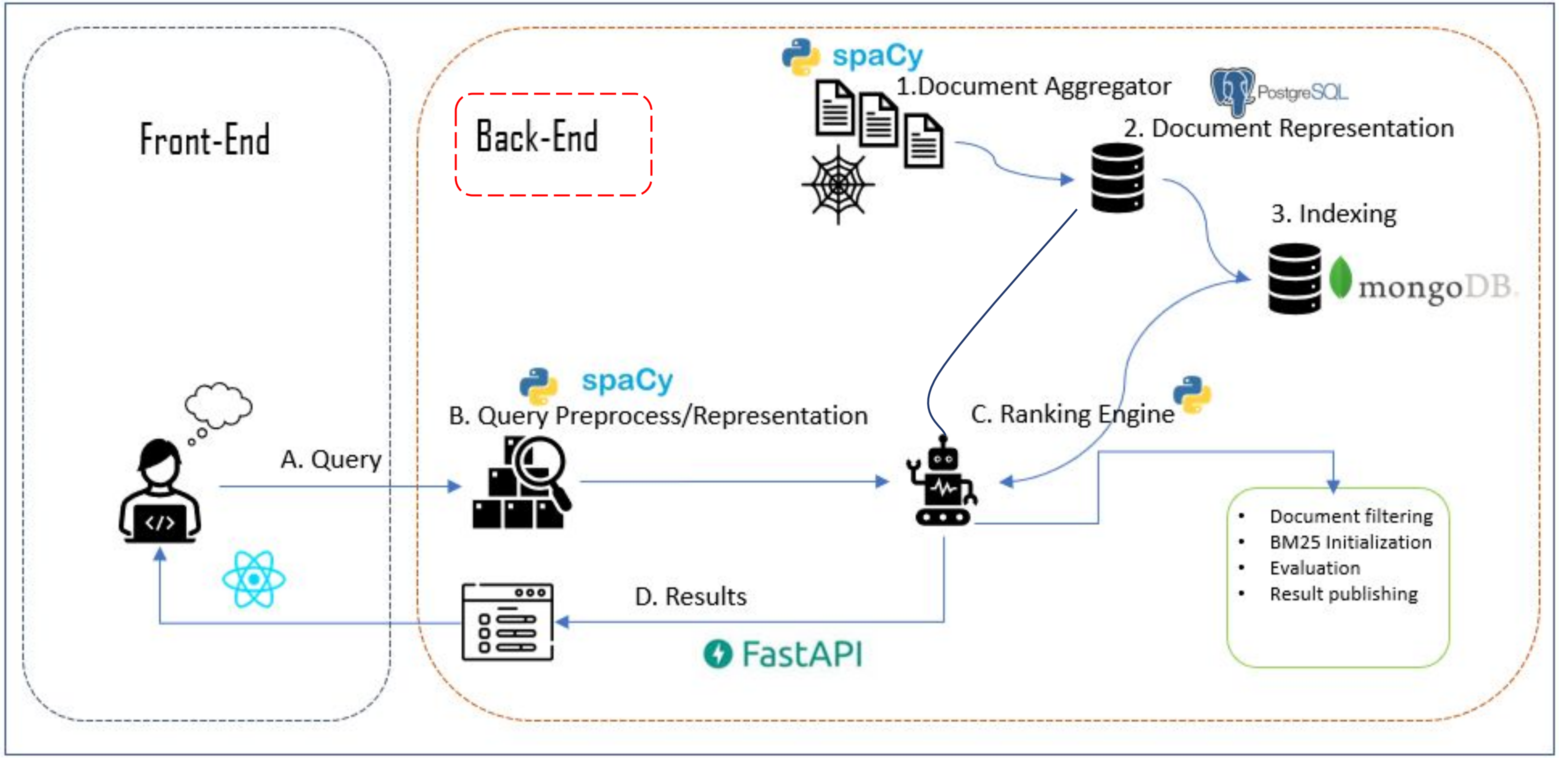


Implementation

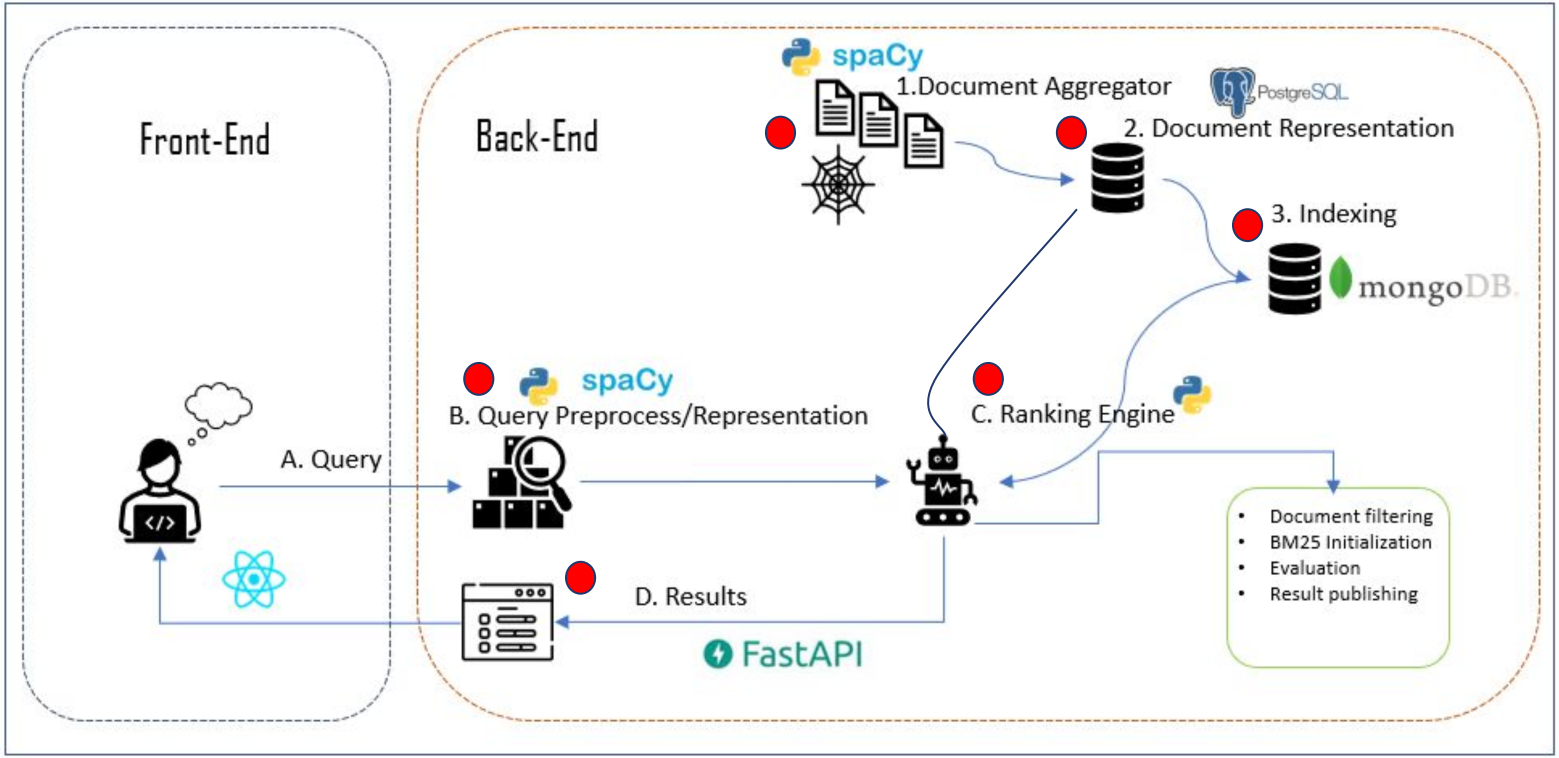


- Document filtering
- BM25 Initialization
- Evaluation
- Result publishing

Implementation

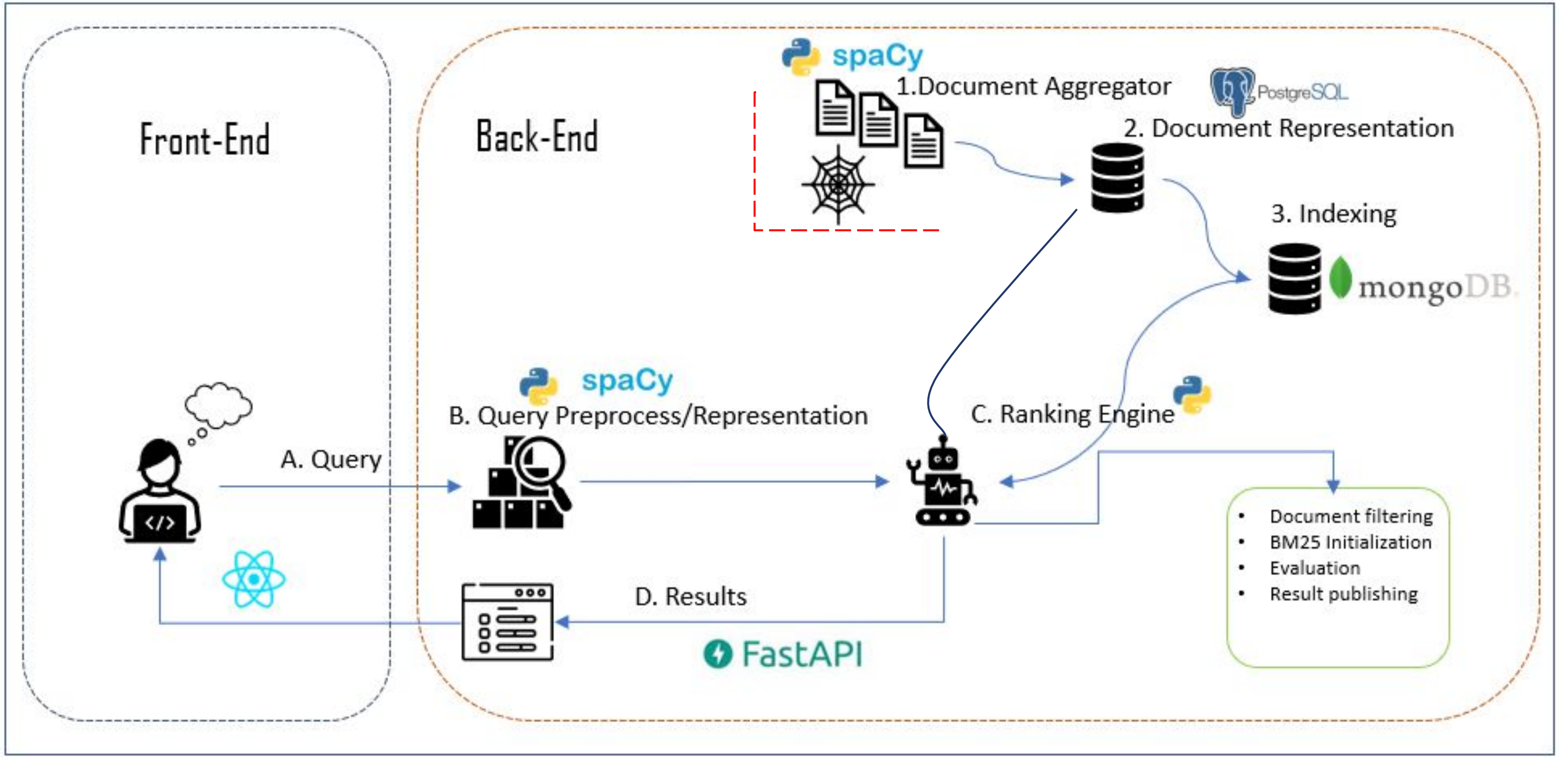


Implementation

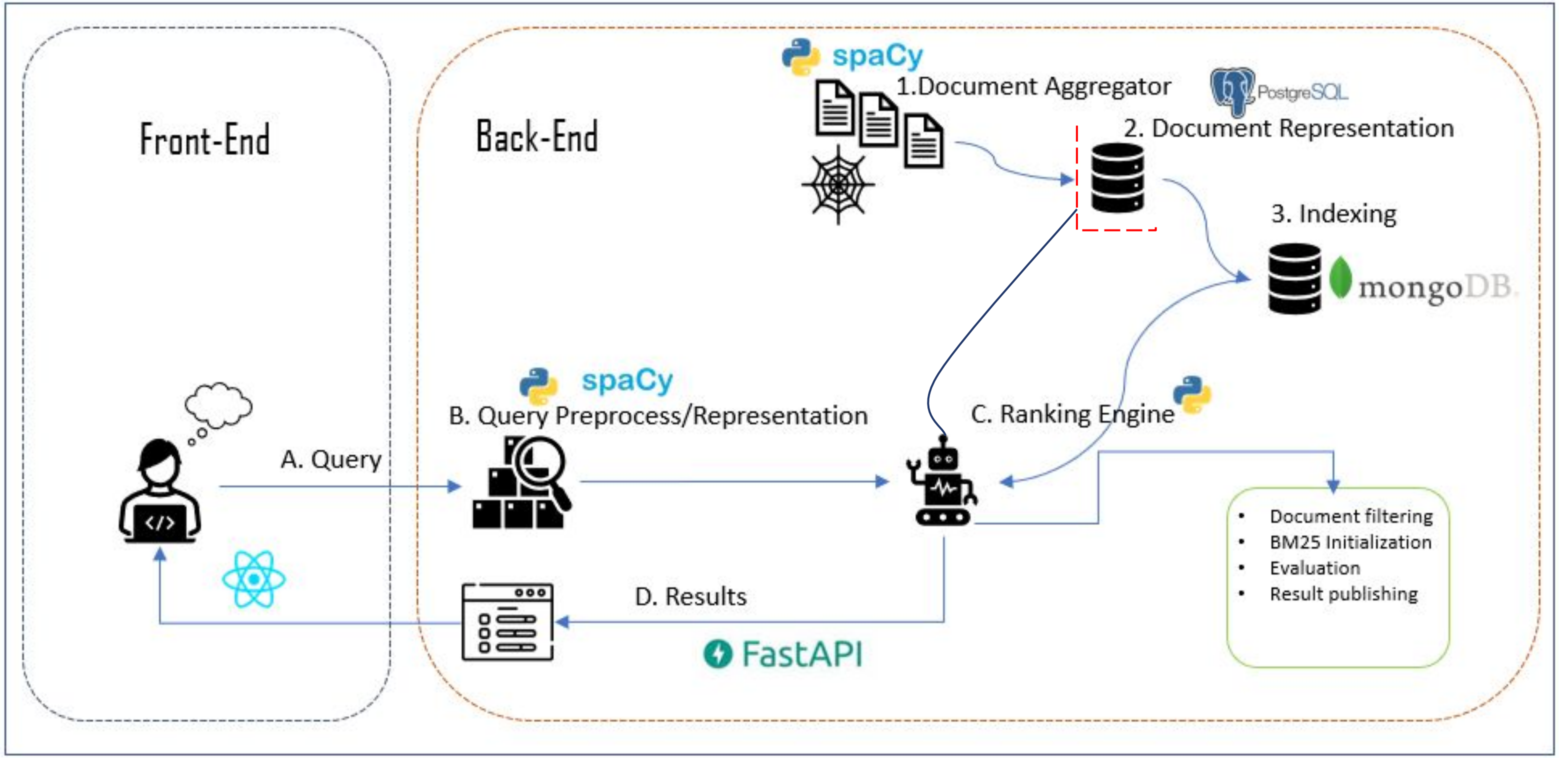


- Document filtering
- BM25 Initialization
- Evaluation
- Result publishing

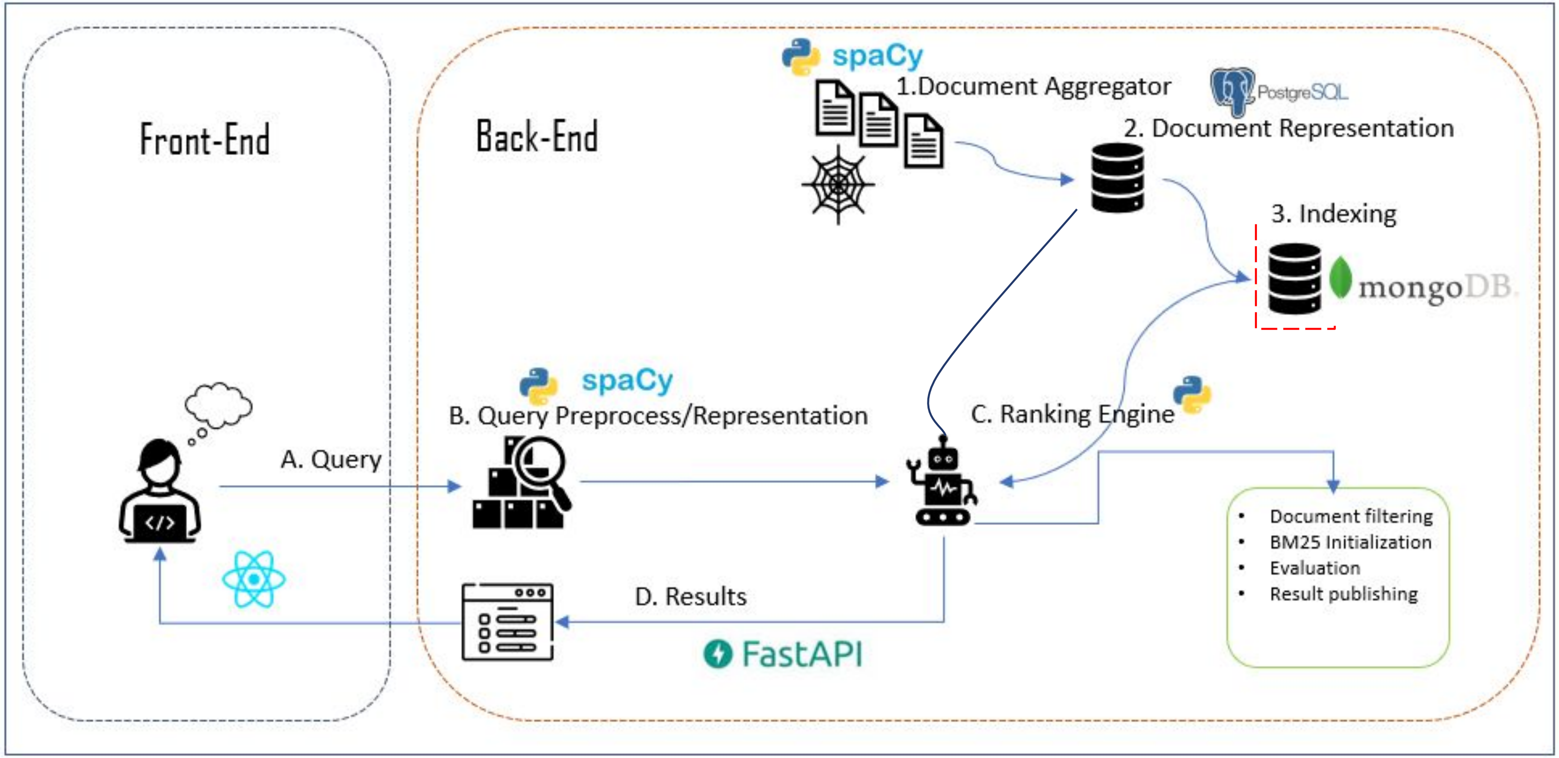
Implementation



Implementation

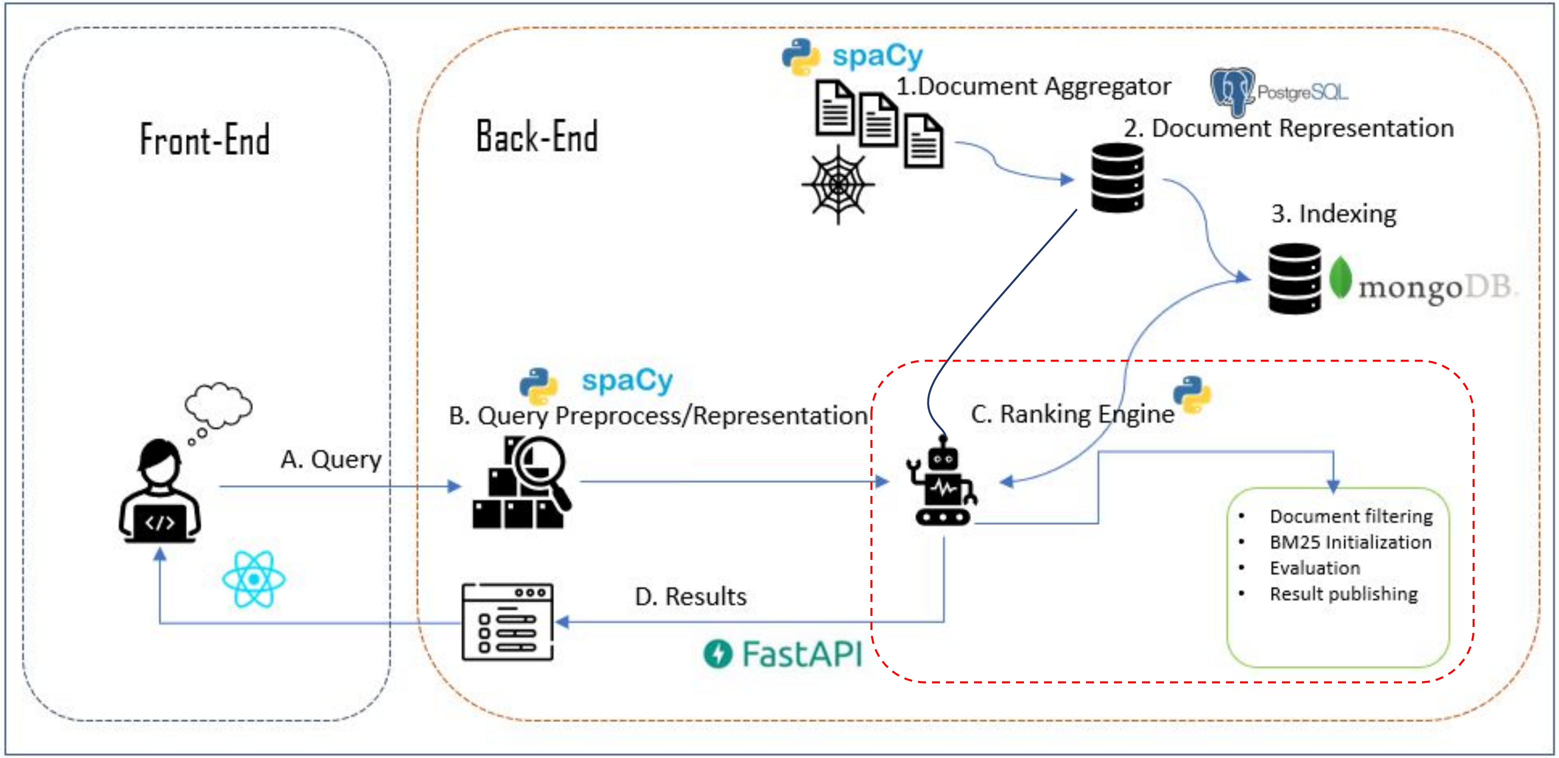


Implementation

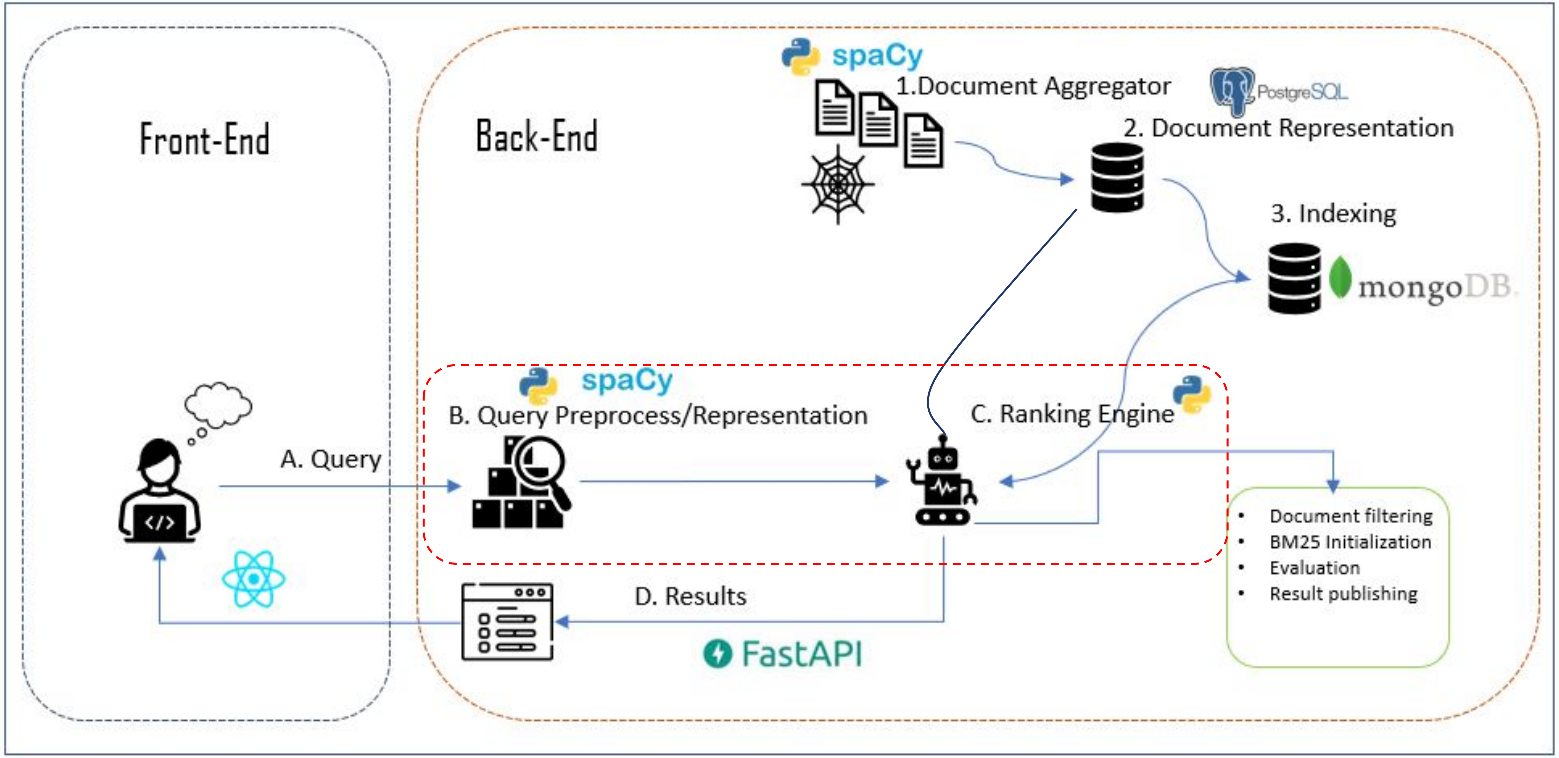


- Document filtering
- BM25 Initialization
- Evaluation
- Result publishing

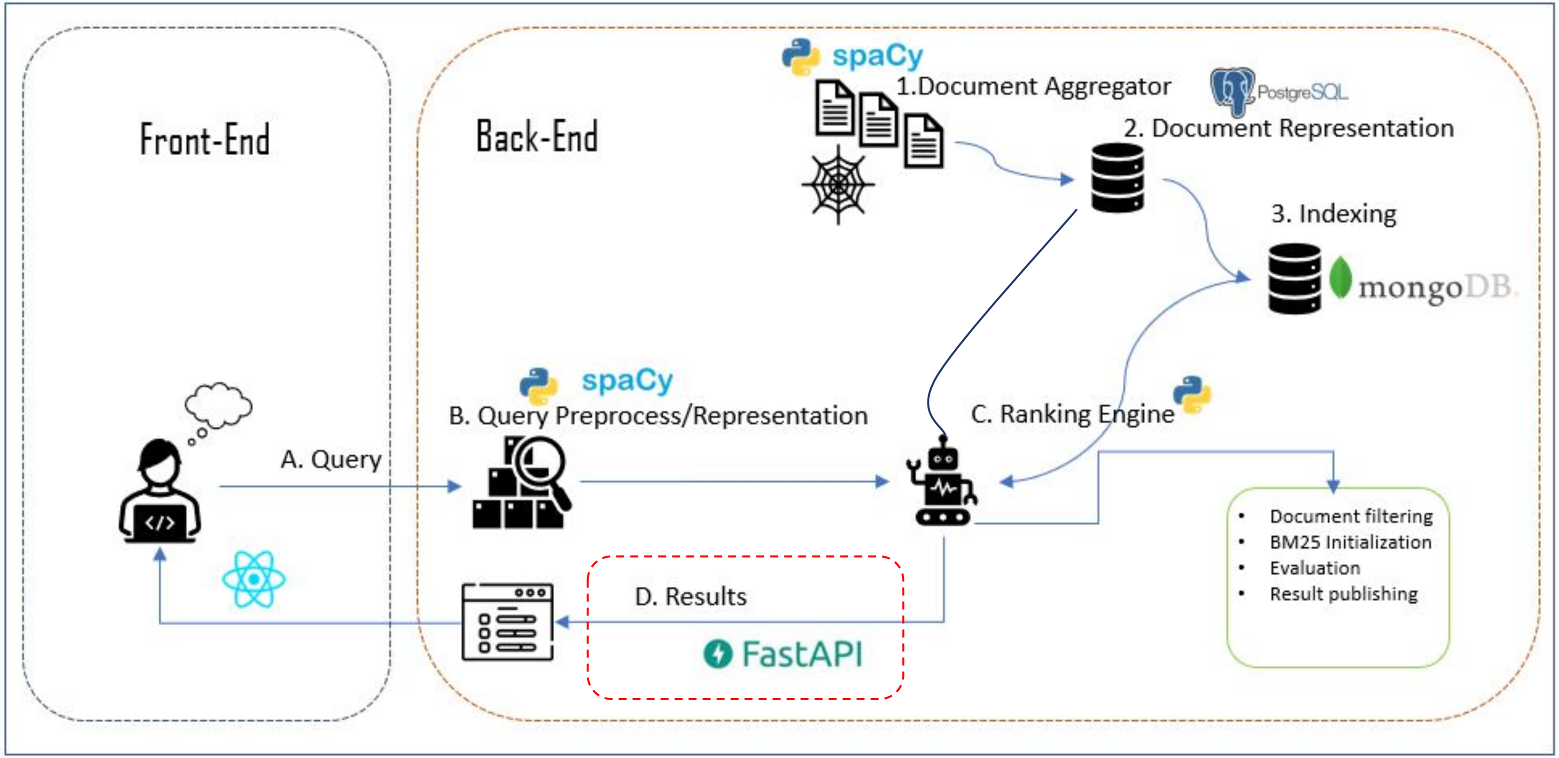
Implementation



Implementation

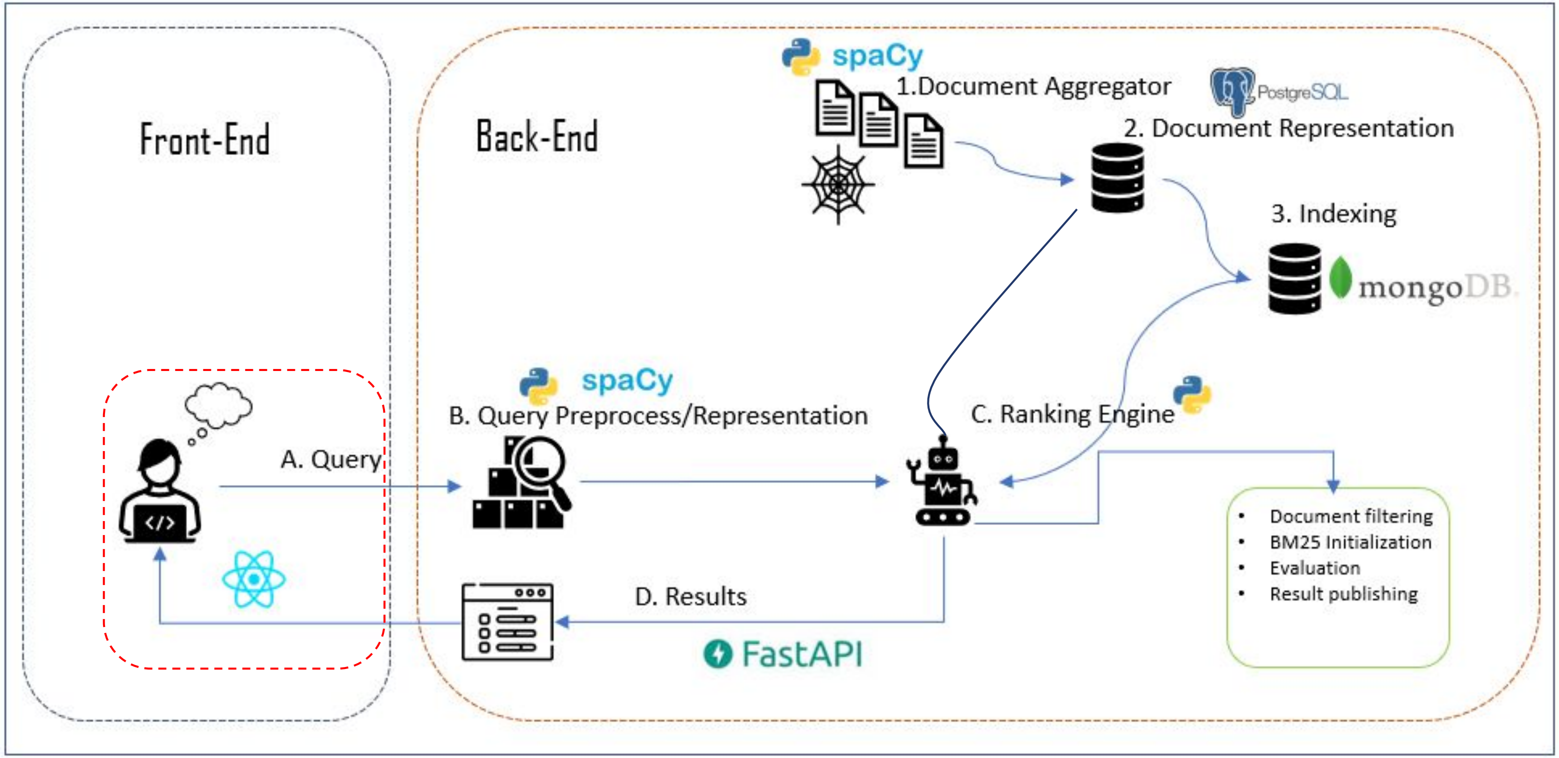


Implementation

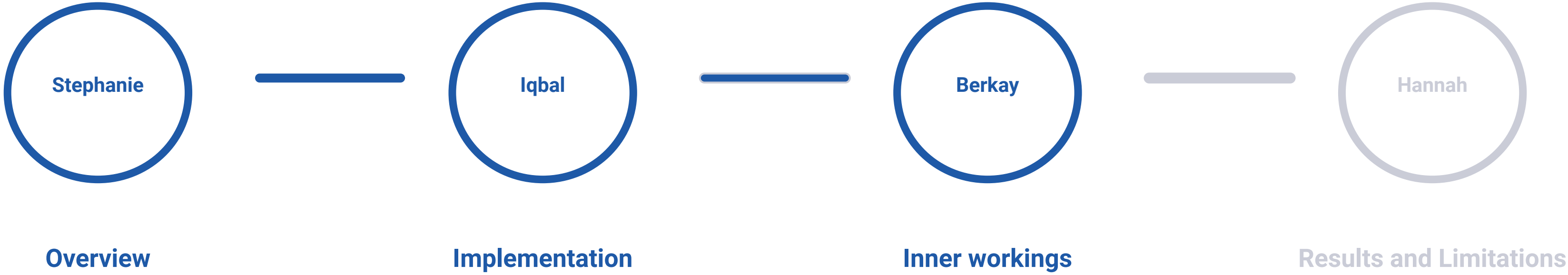


- Document filtering
- BM25 Initialization
- Evaluation
- Result publishing

Implementation



Presentation Structure



Stephanie

Overview

Iqbal

Implementation

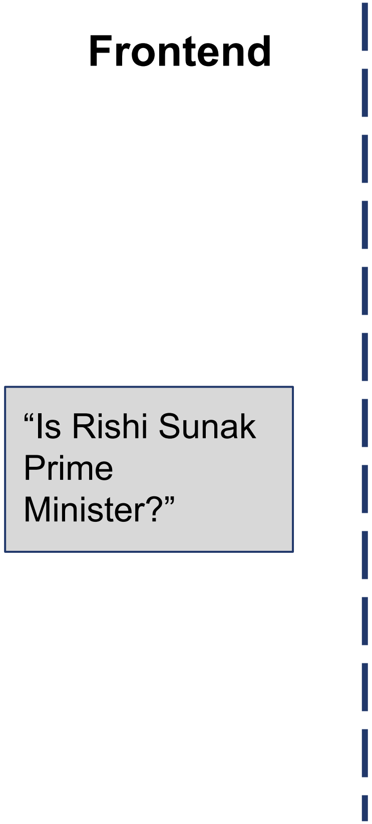
Berkay

Inner workings

Hannah

Results and Limitations

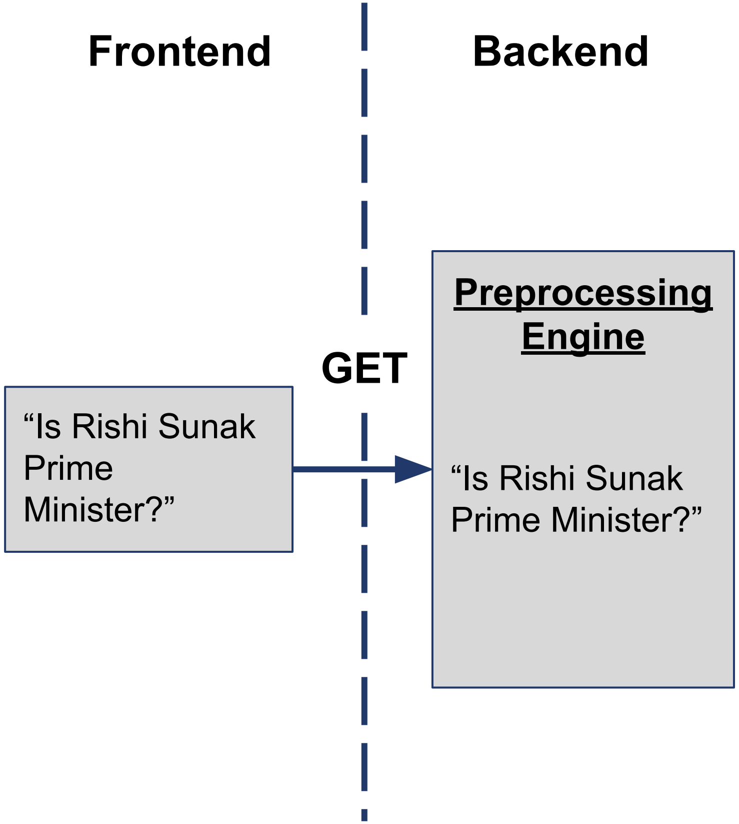
Inner Workings



Frontend

“Is Rishi Sunak Prime Minister?”

Inner Workings



Frontend

Backend

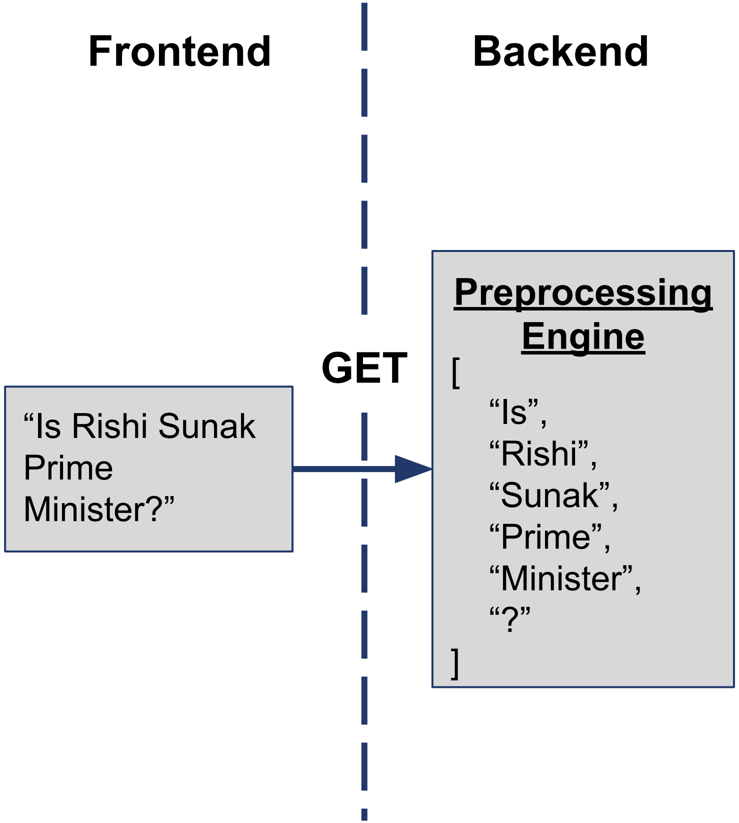
“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

“Is Rishi Sunak Prime Minister?”

Inner Workings



Frontend

Backend

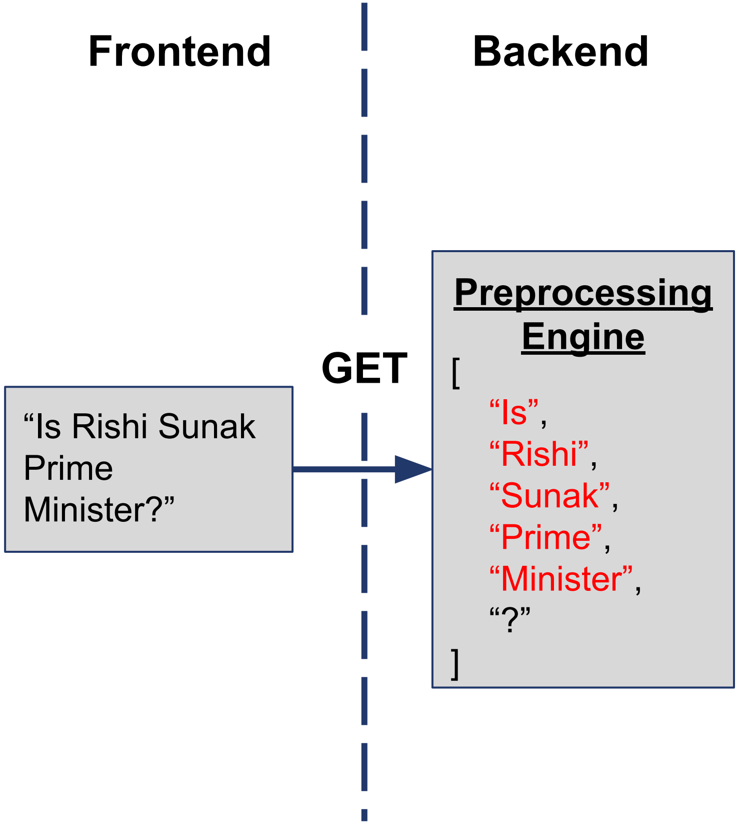
GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

```
[
    “Is”,
    “Rishi”,
    “Sunak”,
    “Prime”,
    “Minister”,
    “?”
]
```

Inner Workings



Frontend

Backend

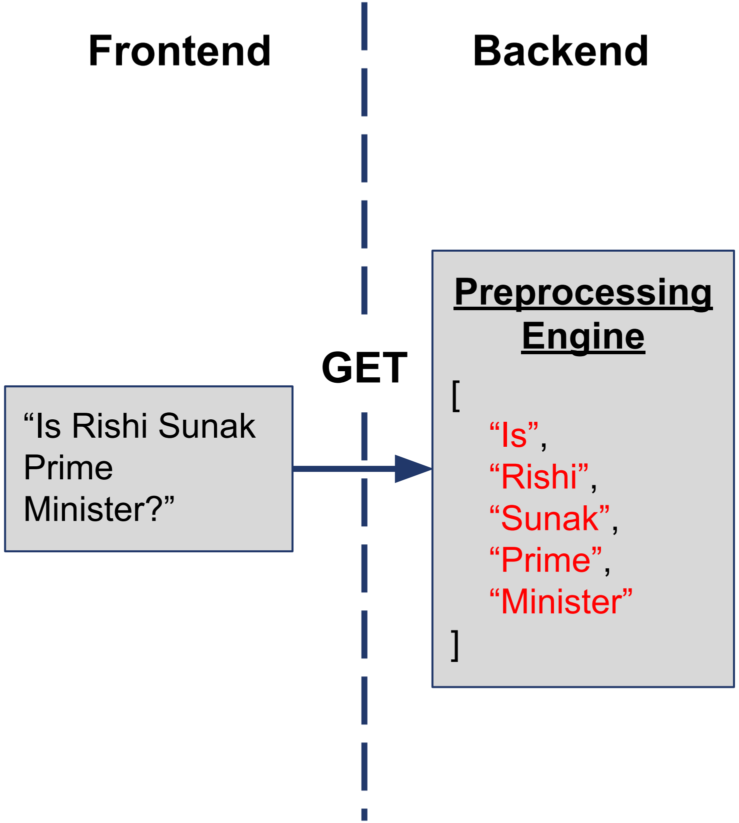
GET

"Is Rishi Sunak Prime Minister?"

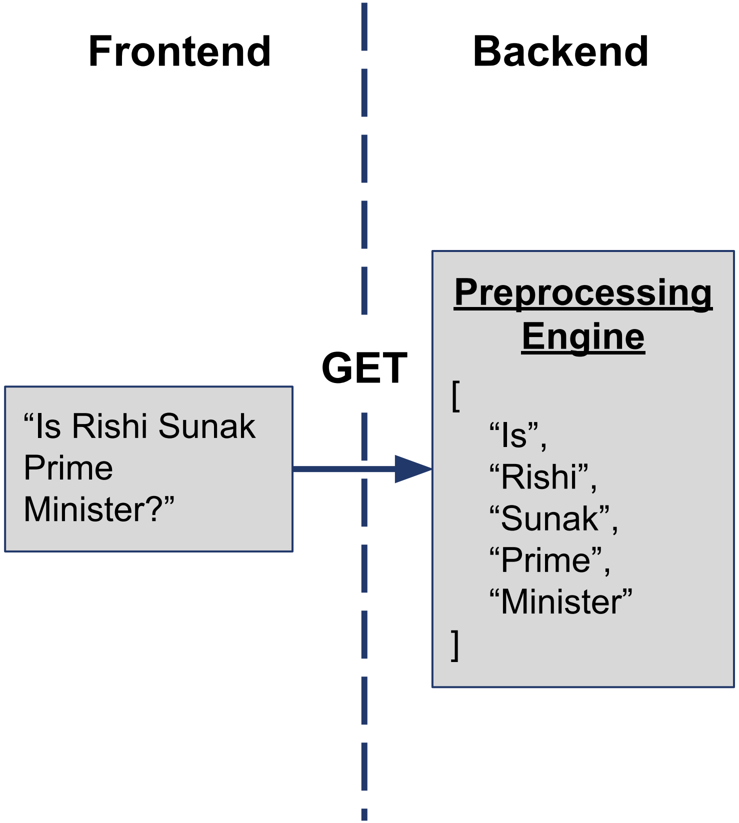
Preprocessing Engine

```
[
    "Is",
    "Rishi",
    "Sunak",
    "Prime",
    "Minister",
    "?"
]
```

Inner Workings



Inner Workings



Frontend

Backend

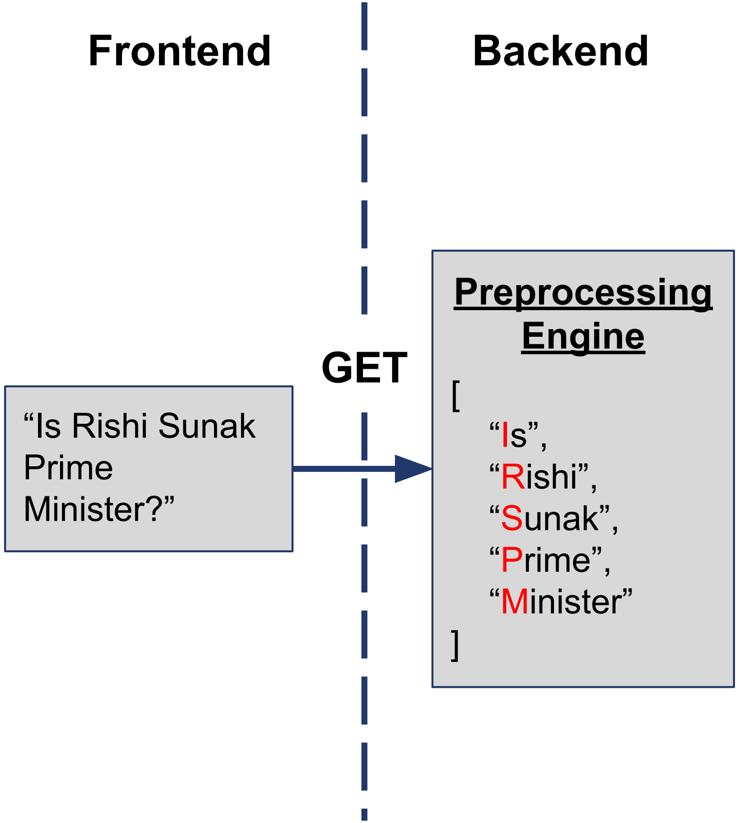
“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

```
[
    “Is”,
    “Rishi”,
    “Sunak”,
    “Prime”,
    “Minister”
]
```

Inner Workings



Frontend

Backend

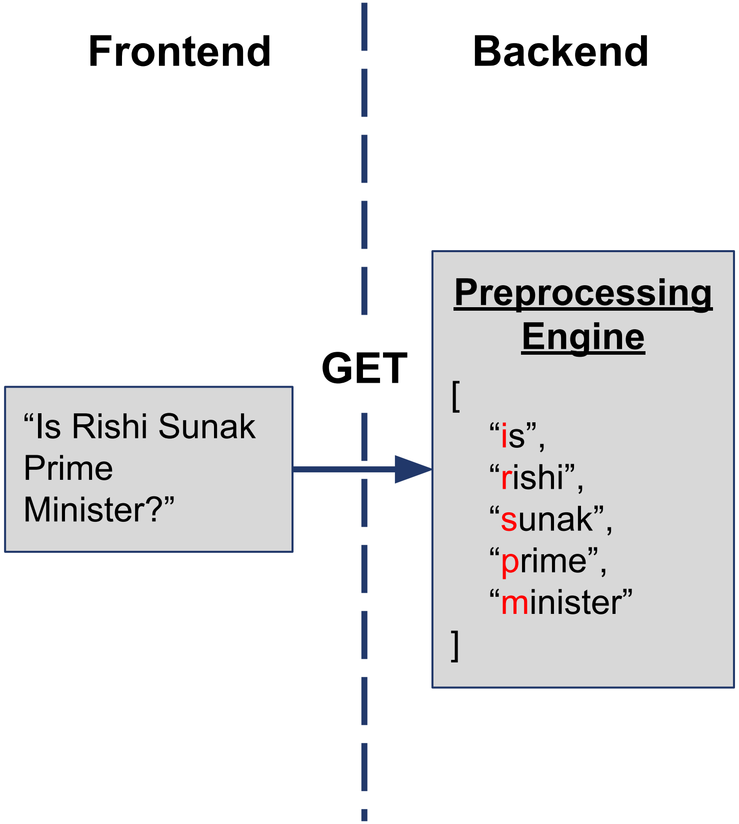
"Is Rishi Sunak Prime Minister?"

GET

Preprocessing Engine

```
[
    "Is",
    "Rishi",
    "Sunak",
    "Prime",
    "Minister"
]
```

Inner Workings



Frontend

Backend

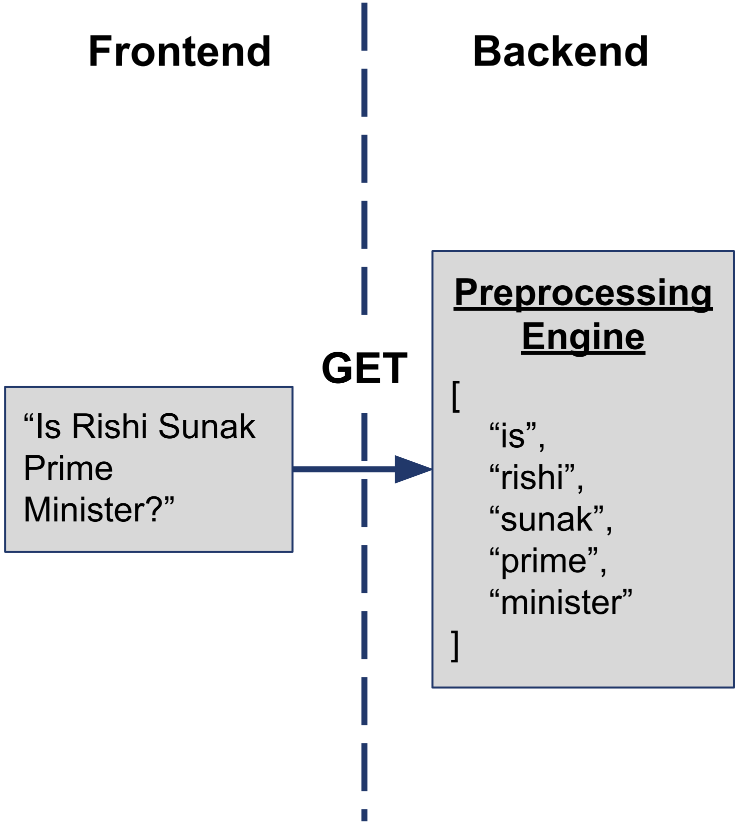
“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

```
[
    “is”,
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Inner Workings



Frontend

Backend

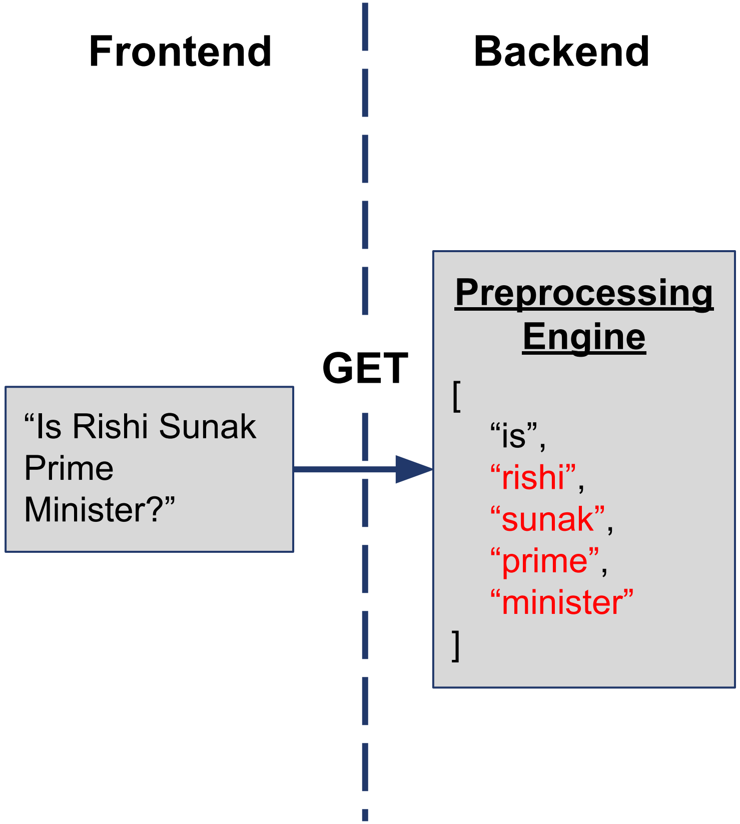
“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

```
[
    “is”,
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Inner Workings



Frontend

Backend

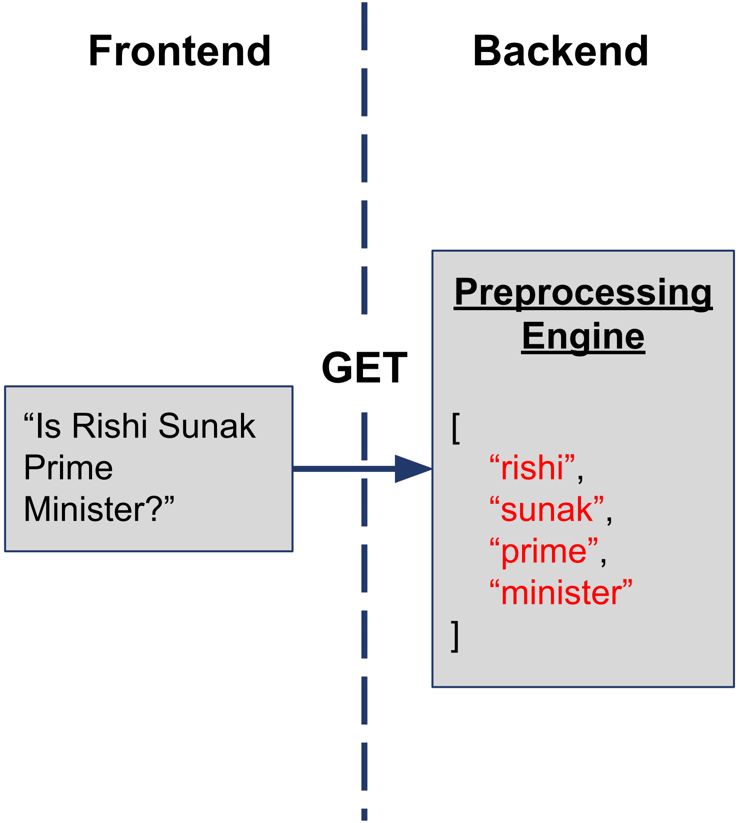
“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

```
[
    “is”,
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Inner Workings



Frontend

Backend

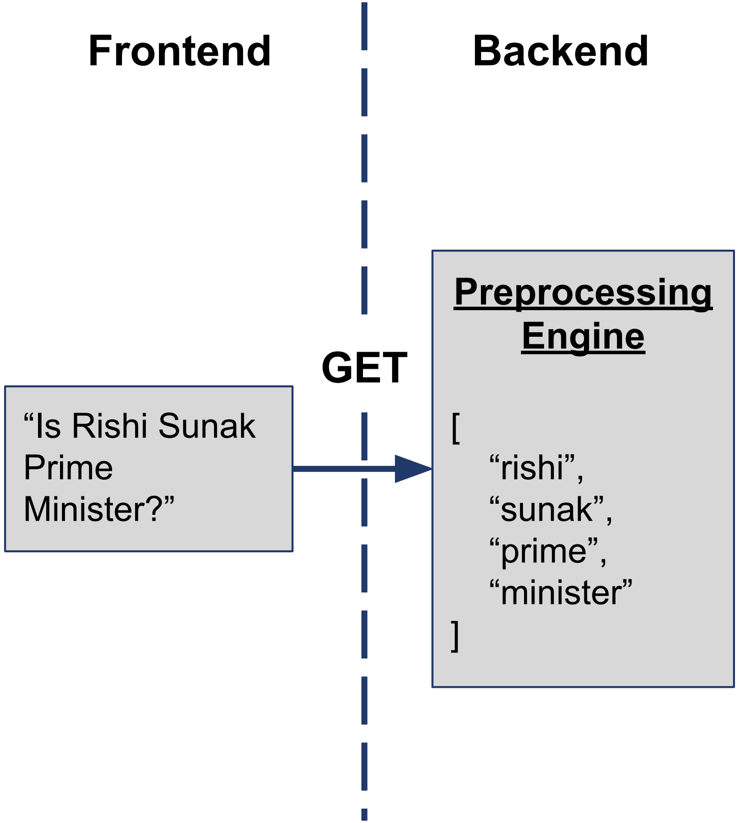
“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Inner Workings


Frontend

Backend

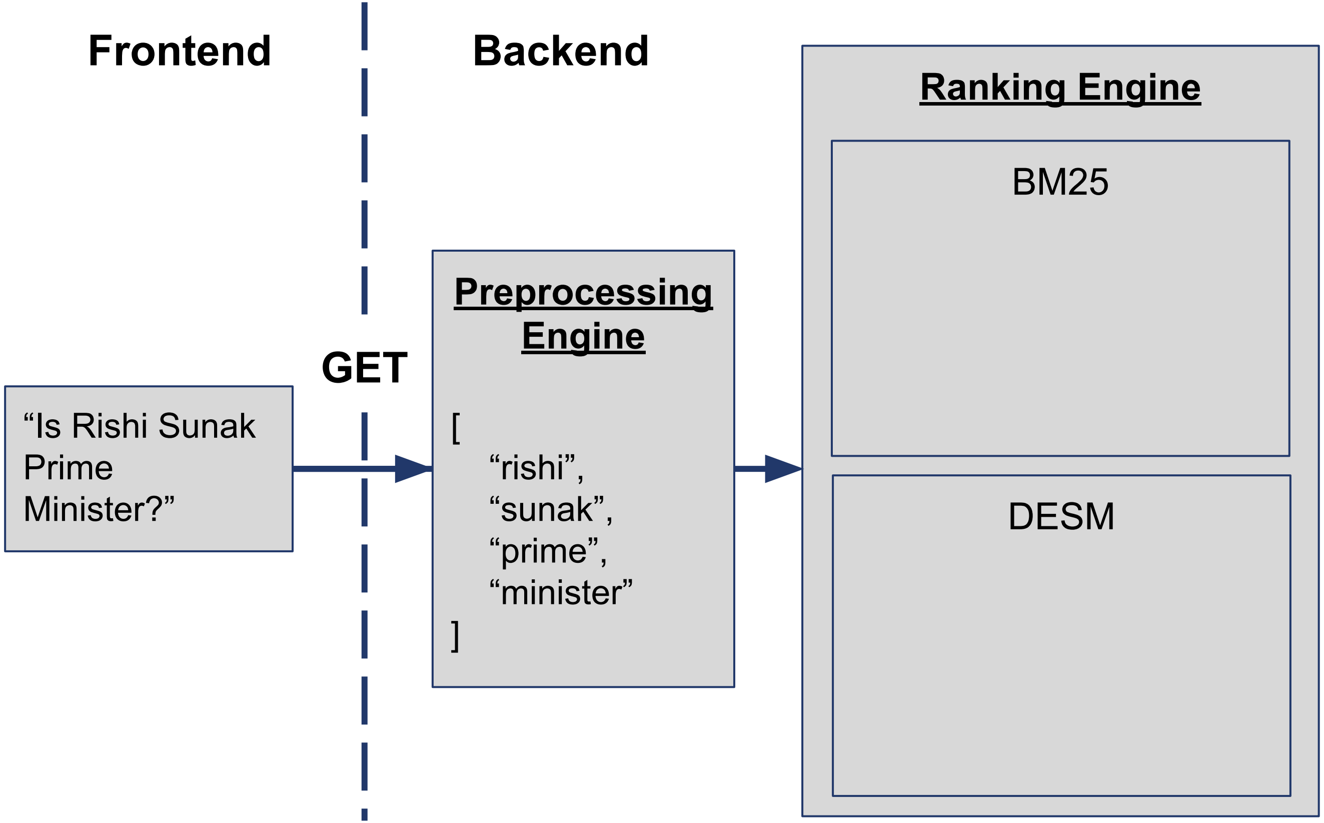
GET

“Is Rishi Sunak Prime Minister?”

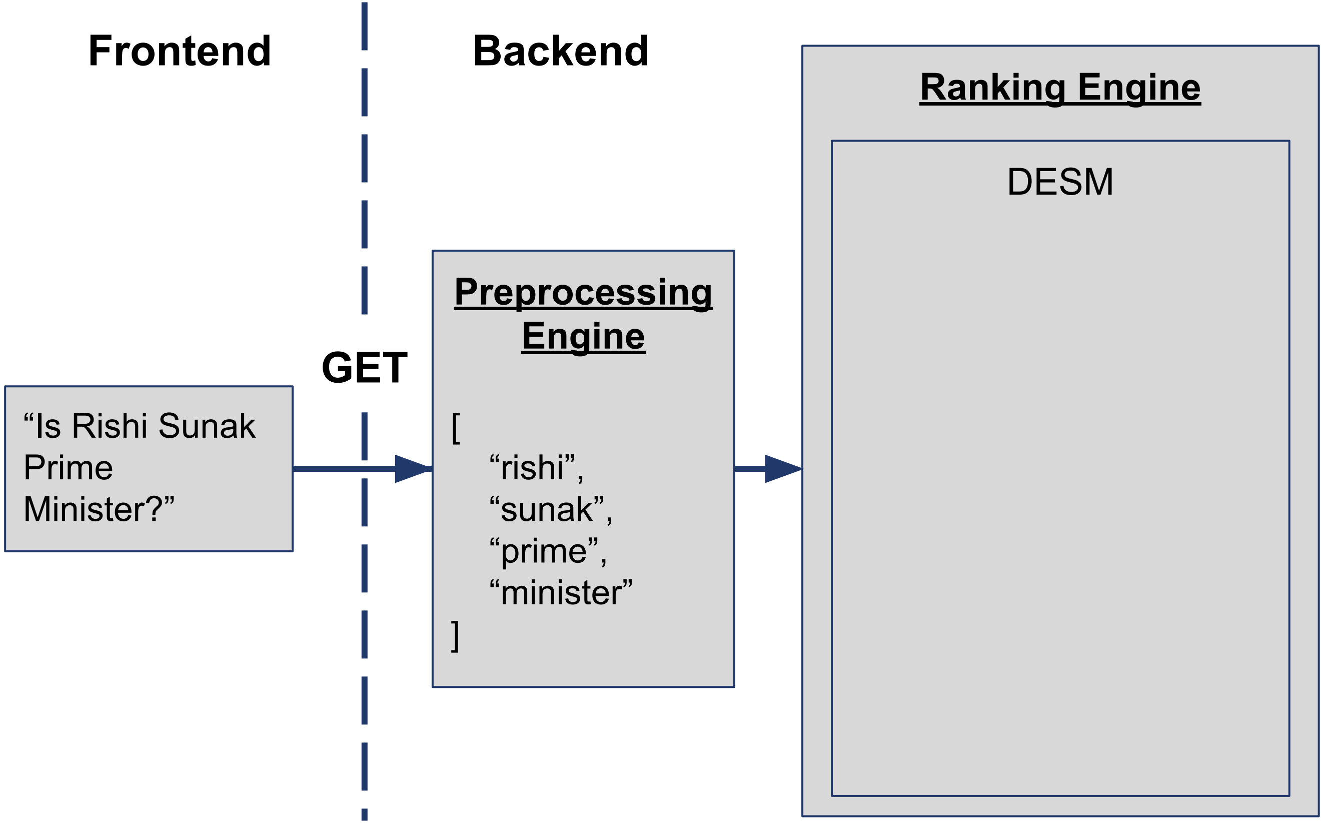
Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Inner Workings



Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

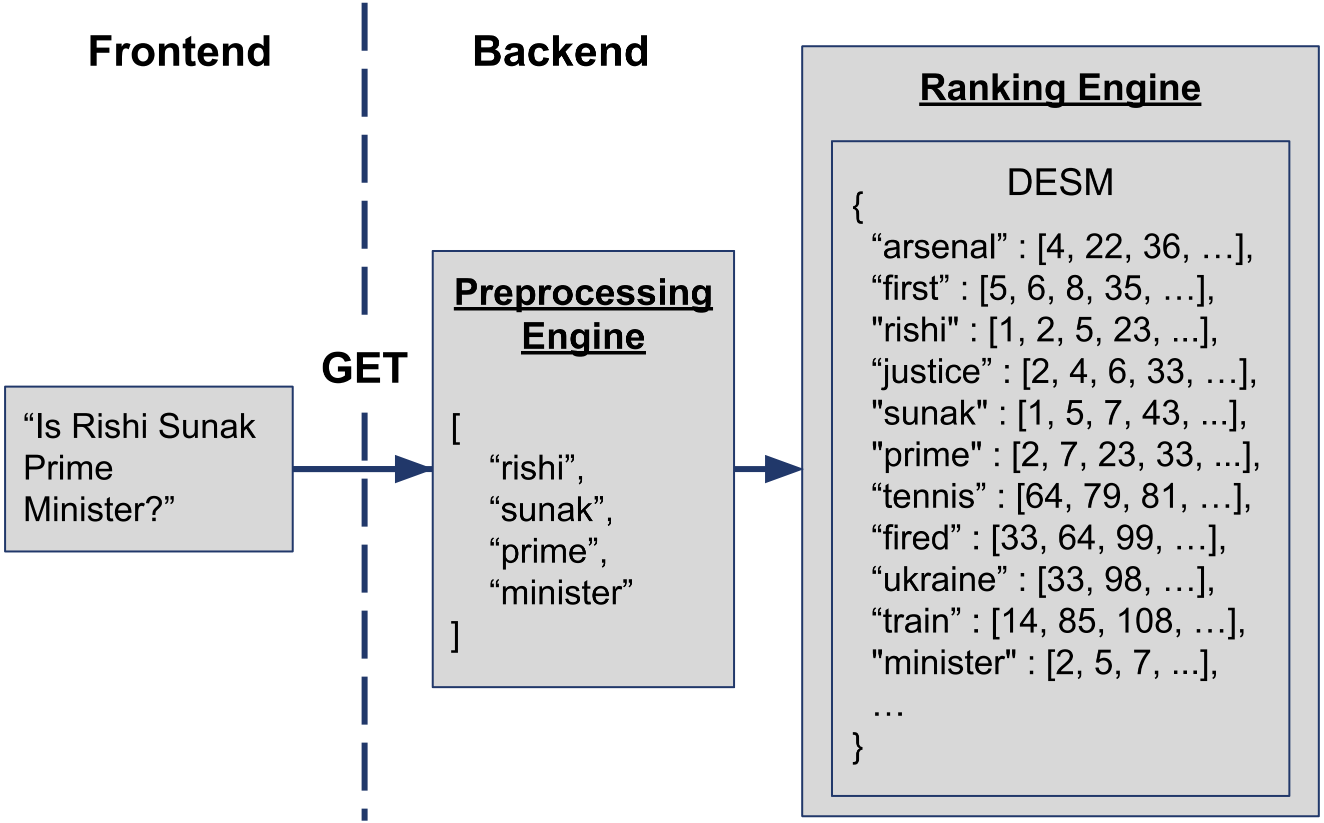
Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

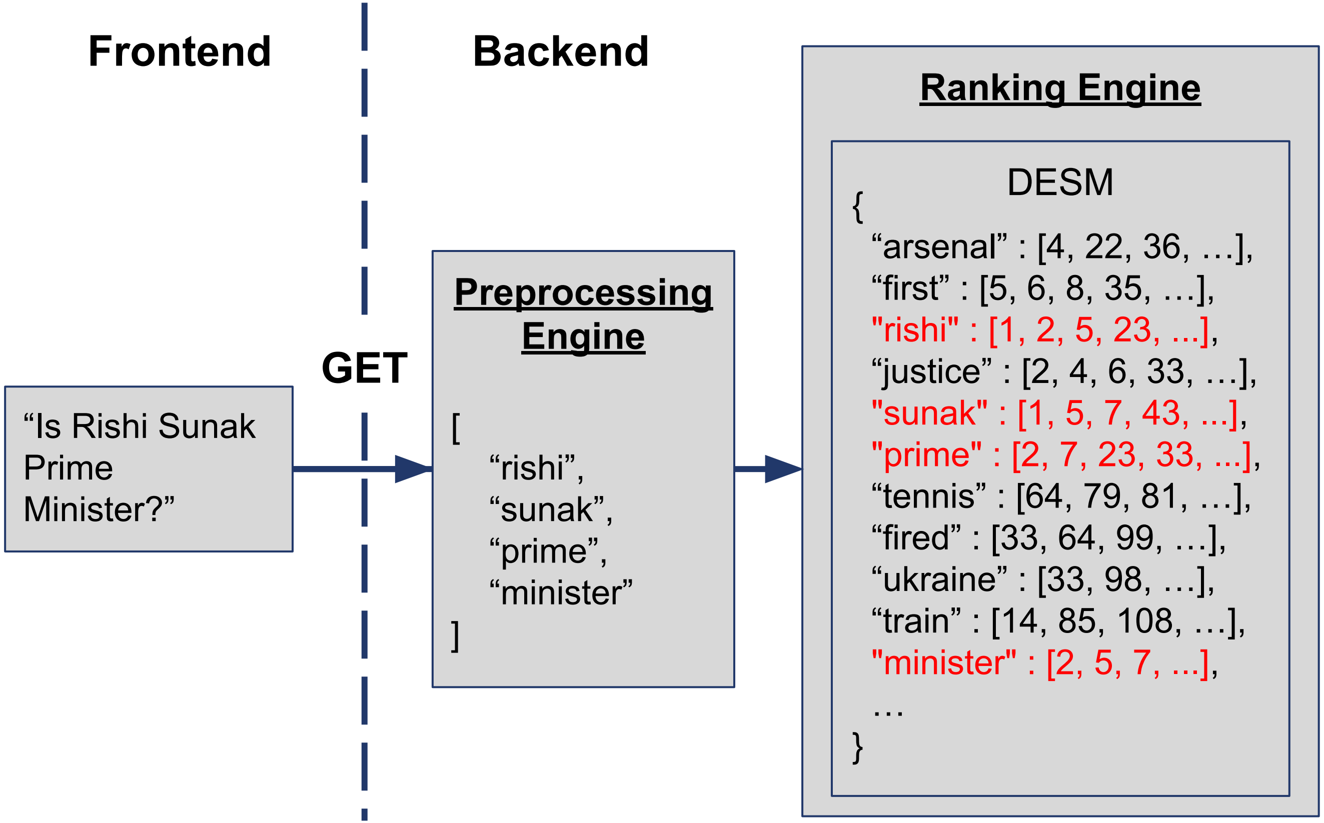
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
{
  “arsenal” : [4, 22, 36, …],
  “first” : [5, 6, 8, 35, …],
  "rishi" : [1, 2, 5, 23, ...],
  “justice” : [2, 4, 6, 33, …],
  "sunak" : [1, 5, 7, 43, ...],
  "prime" : [2, 7, 23, 33, ...],
  “tennis” : [64, 79, 81, …],
  “fired” : [33, 64, 99, …],
  “ukraine” : [33, 98, …],
  “train” : [14, 85, 108, …],
  "minister" : [2, 5, 7, ...],
  …
}
```

Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

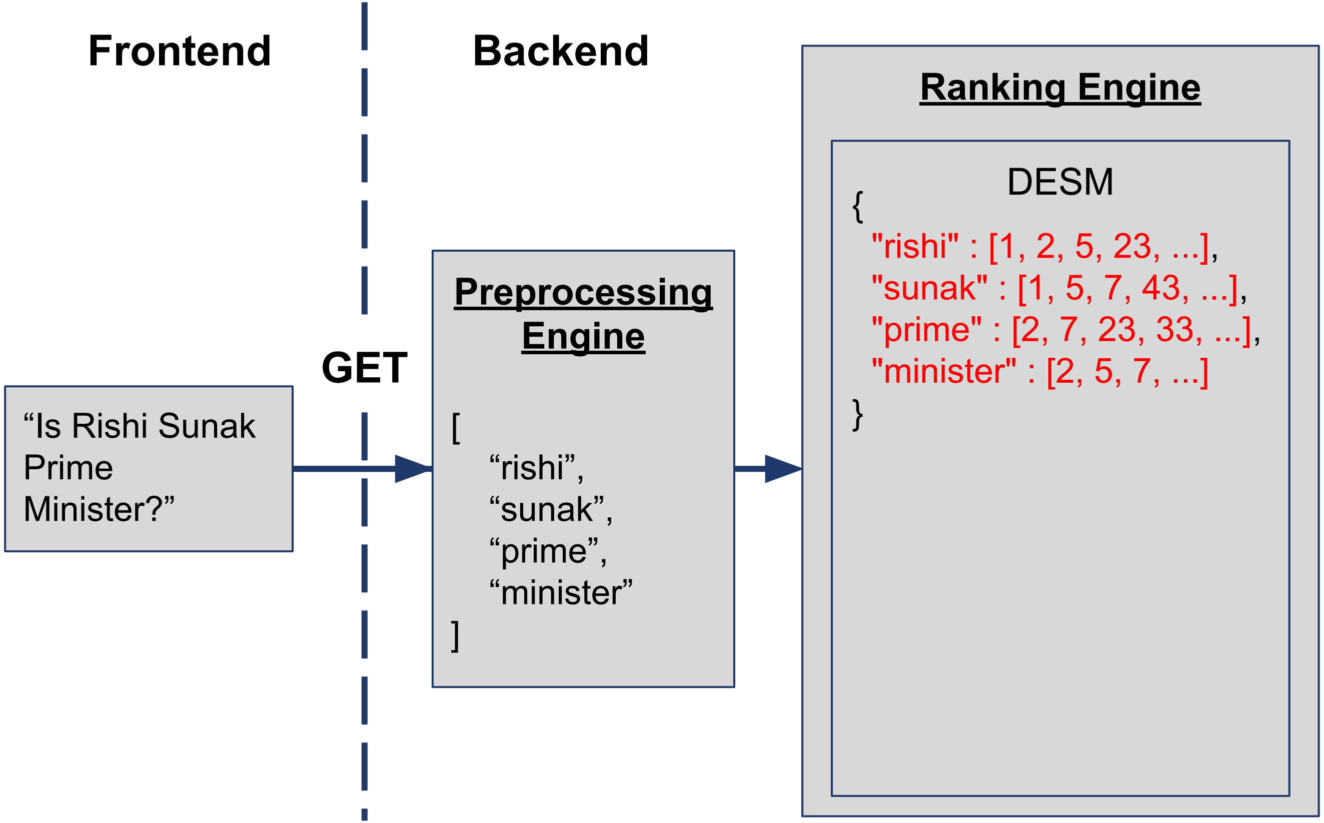
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
{
  “arsenal” : [4, 22, 36, …],
  “first” : [5, 6, 8, 35, …],
  "rishi" : [1, 2, 5, 23, ...],
  “justice” : [2, 4, 6, 33, …],
  "sunak" : [1, 5, 7, 43, ...],
  "prime" : [2, 7, 23, 33, ...],
  “tennis” : [64, 79, 81, …],
  “fired” : [33, 64, 99, …],
  “ukraine” : [33, 98, …],
  “train” : [14, 85, 108, …],
  "minister" : [2, 5, 7, ...],
  …
}
```

Inner Workings



Frontend

Backend

“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

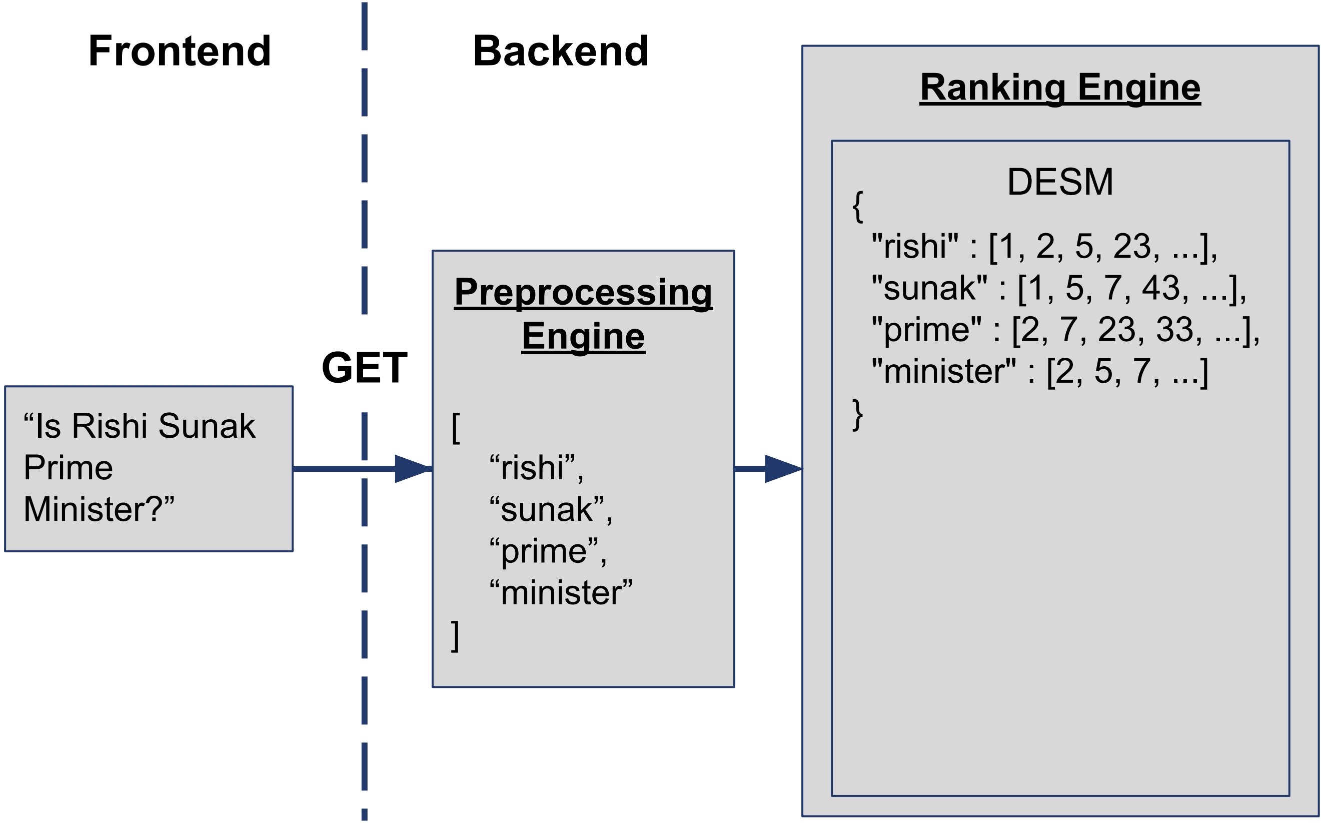
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
{
  "rishi" : [1, 2, 5, 23, ...],
  "sunak" : [1, 5, 7, 43, ...],
  "prime" : [2, 7, 23, 33, ...],
  "minister" : [2, 5, 7, ...]
}
```

Inner Workings



"Is Rishi Sunak Prime Minister?"

Preprocessing Engine

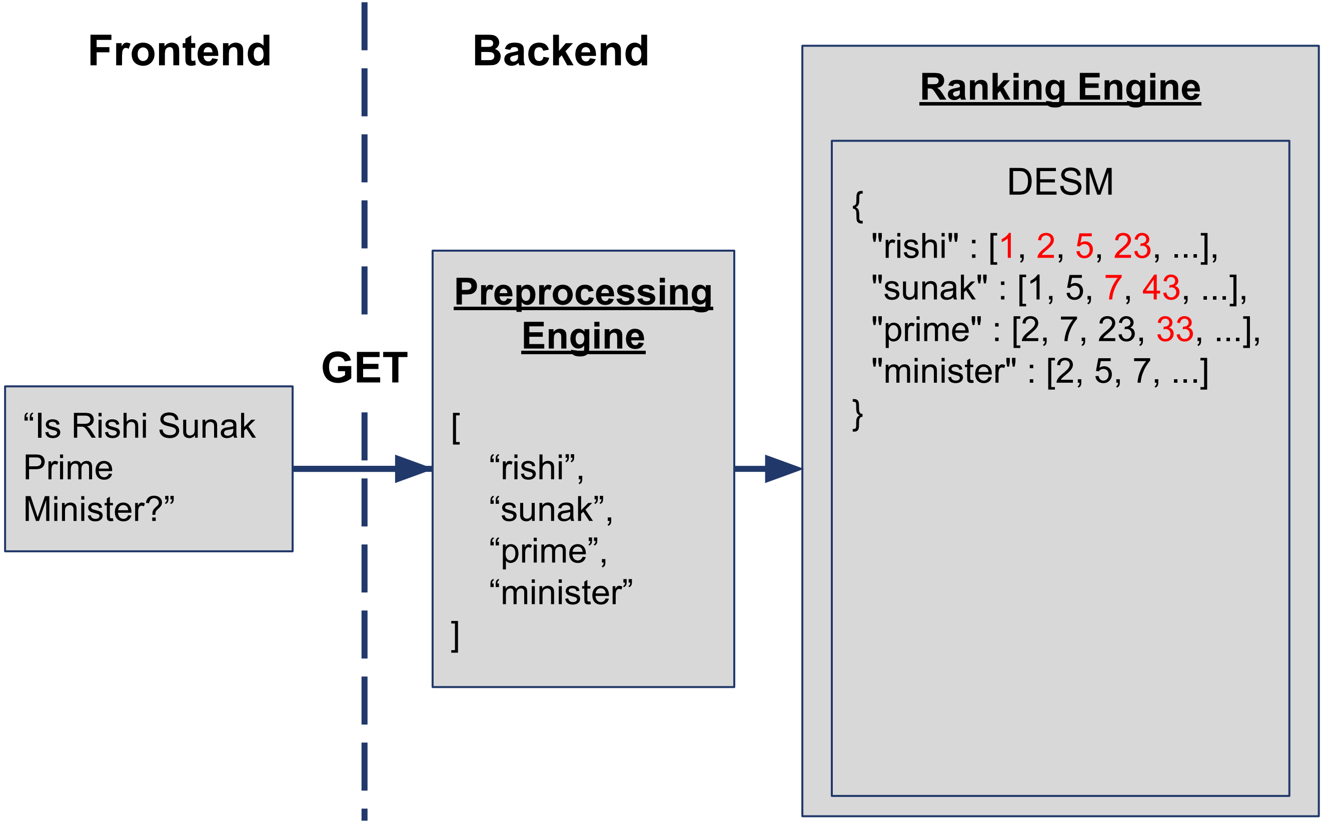
```
[
    "rishi",
    "sunak",
    "prime",
    "minister"
]
```

Ranking Engine

DESM

```
{
  "rishi" : [1, 2, 5, 23, ...],
  "sunak" : [1, 5, 7, 43, ...],
  "prime" : [2, 7, 23, 33, ...],
  "minister" : [2, 5, 7, ...]
}
```

Inner Workings



Frontend

Backend

GET

"Is Rishi Sunak Prime Minister?"

Preprocessing Engine

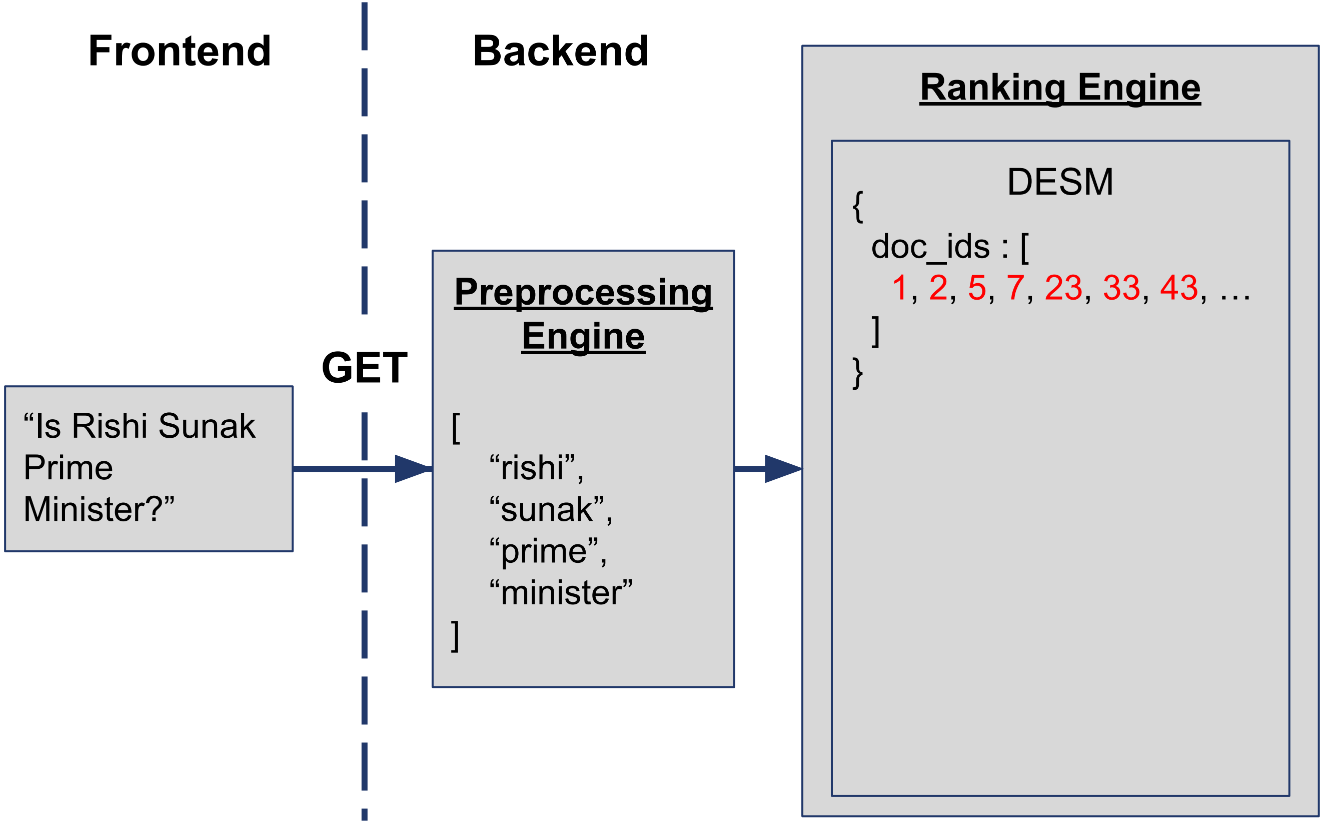
```
[
    "rishi",
    "sunak",
    "prime",
    "minister"
]
```

Ranking Engine

DESM

```
{
  "rishi" : [1, 2, 5, 23, ...],
  "sunak" : [1, 5, 7, 43, ...],
  "prime" : [2, 7, 23, 33, ...],
  "minister" : [2, 5, 7, ...]
}
```

Inner Workings



Frontend

Backend

GET

"Is Rishi Sunak Prime Minister?"

Preprocessing Engine

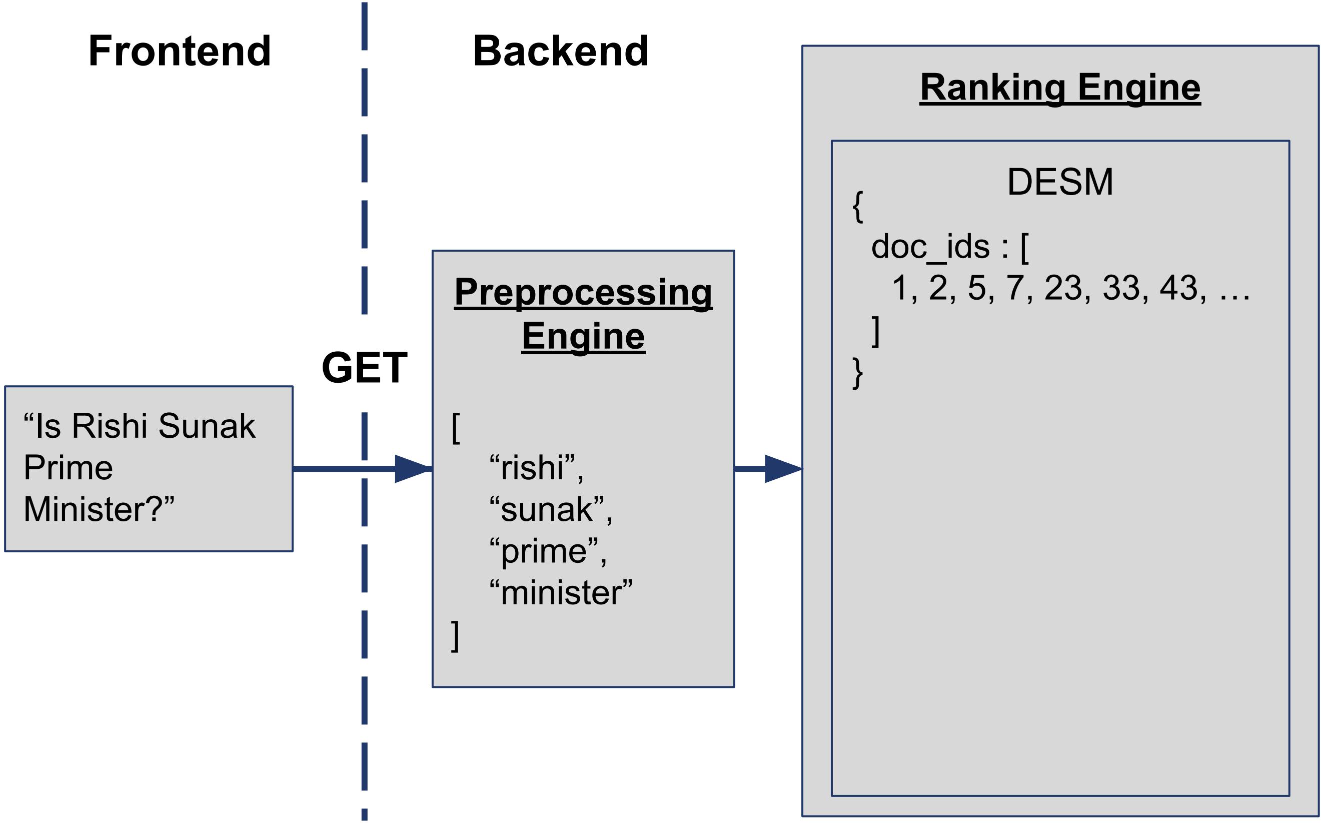
```
[
    "rishi",
    "sunak",
    "prime",
    "minister"
]
```

Ranking Engine

DESM

```
{
  doc_ids : [
    1, 2, 5, 7, 23, 33, 43, …
  ]
}
```

Inner Workings



Frontend

Backend

“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

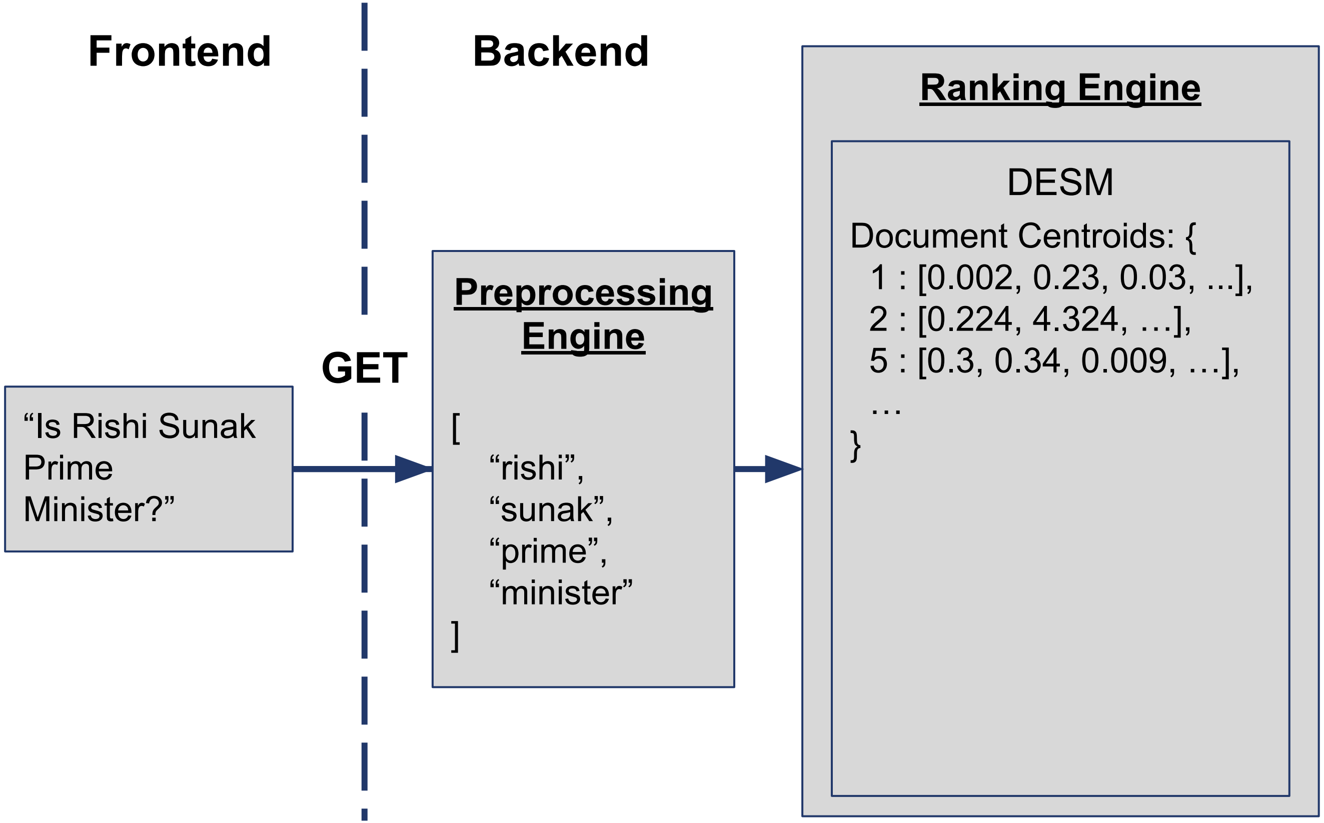
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
{
  doc_ids : [
    1, 2, 5, 7, 23, 33, 43, …
  ]
}
```

Inner Workings



Frontend

Backend

“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

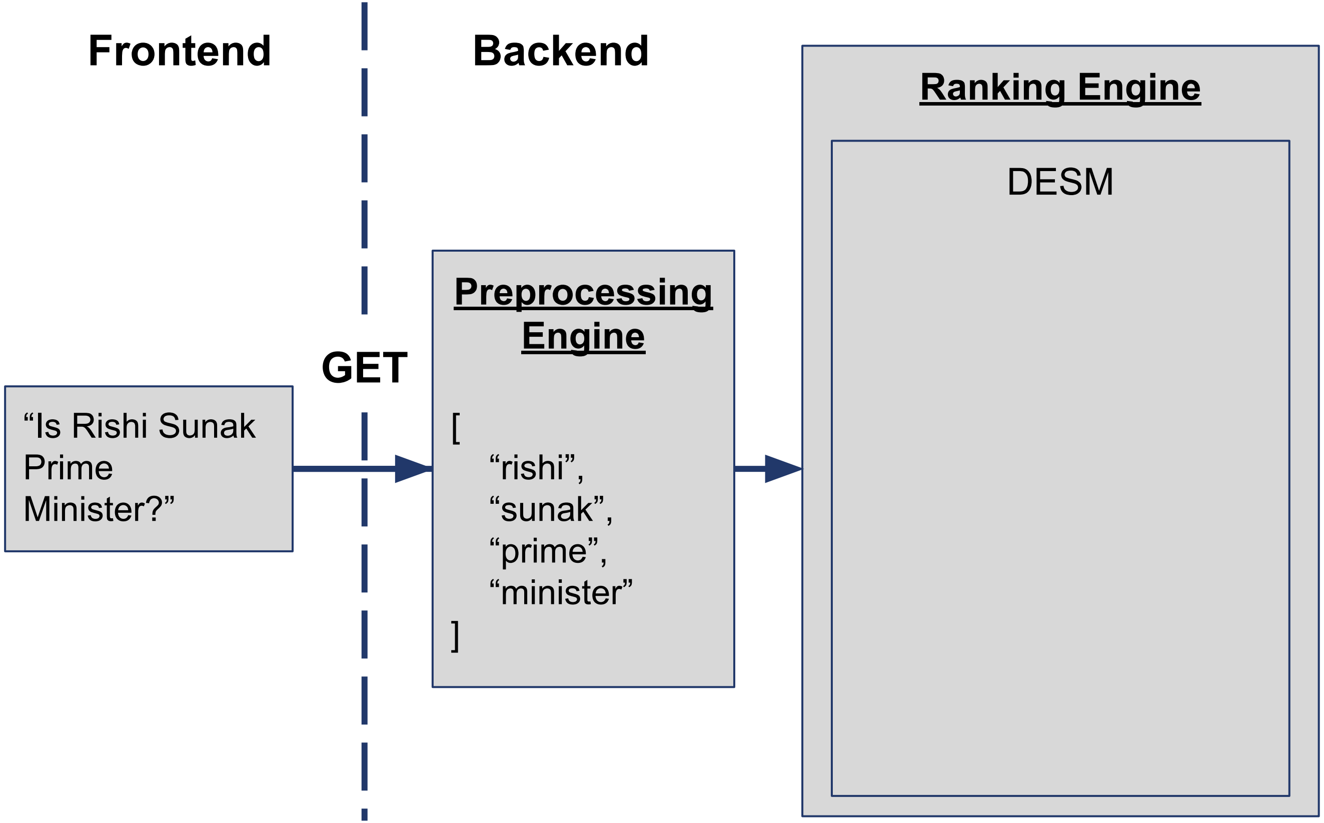
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
Document Centroids: {
  1 : [0.002, 0.23, 0.03, ...],
  2 : [0.224, 4.324, …],
  5 : [0.3, 0.34, 0.009, …],
  …
}
```

Inner Workings



Frontend

Backend

"Is Rishi Sunak Prime Minister?"

GET

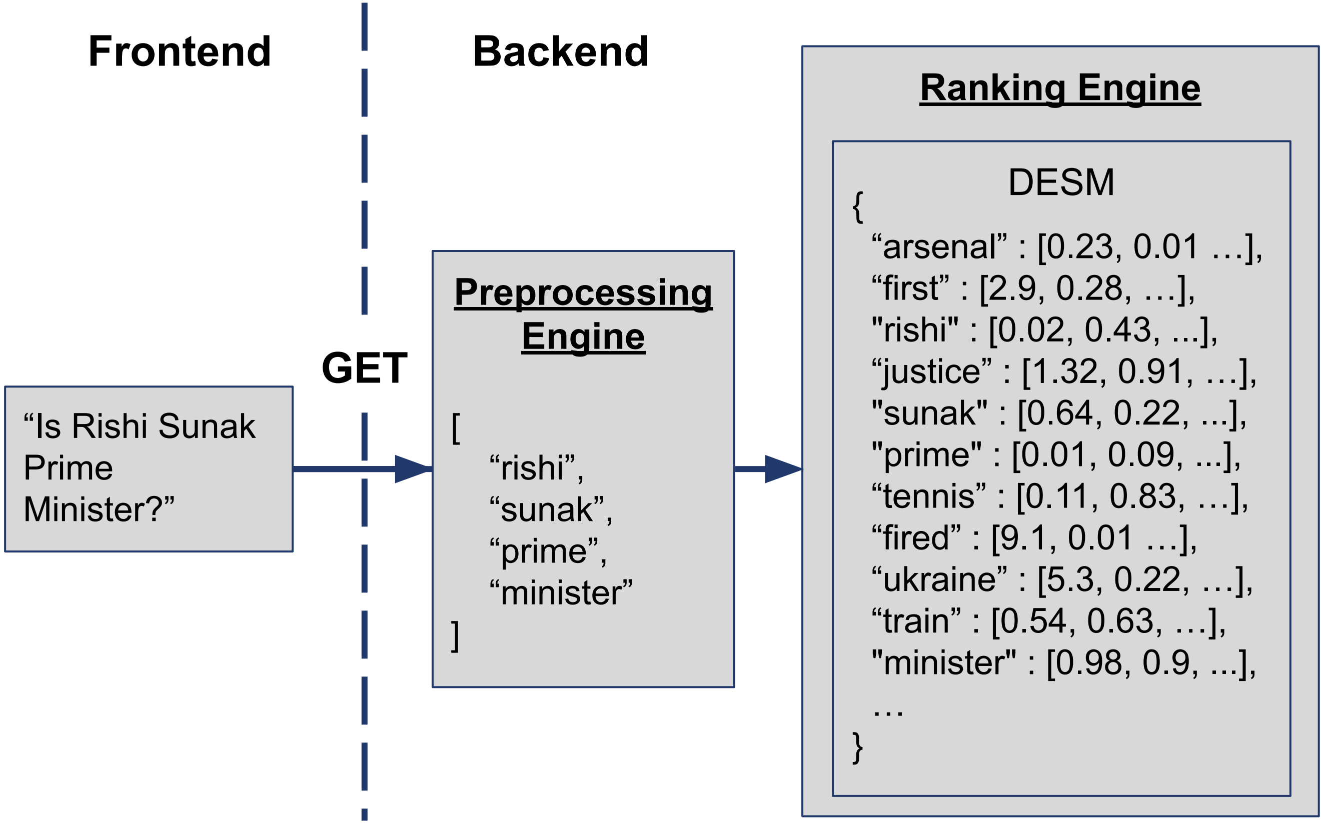
Preprocessing Engine

[
 "rishi",
 "sunak",
 "prime",
 "minister"
]

Ranking Engine

DESM

Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

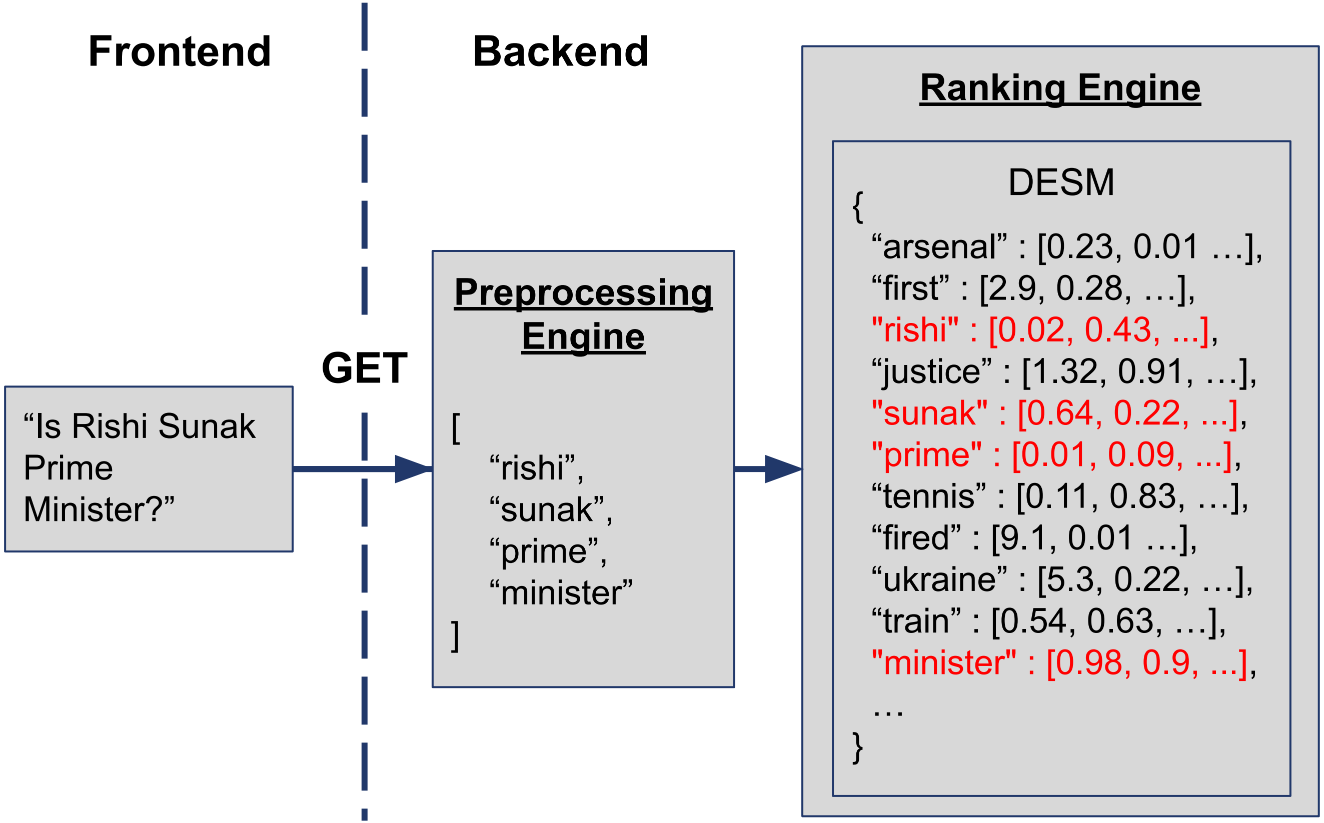
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
{
  “arsenal” : [0.23, 0.01 …],
  “first” : [2.9, 0.28, …],
  "rishi" : [0.02, 0.43, ...],
  “justice” : [1.32, 0.91, …],
  "sunak" : [0.64, 0.22, ...],
  "prime" : [0.01, 0.09, ...],
  “tennis” : [0.11, 0.83, …],
  “fired” : [9.1, 0.01 …],
  “ukraine” : [5.3, 0.22, …],
  “train” : [0.54, 0.63, …],
  "minister" : [0.98, 0.9, ...],
  …
}
```

Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

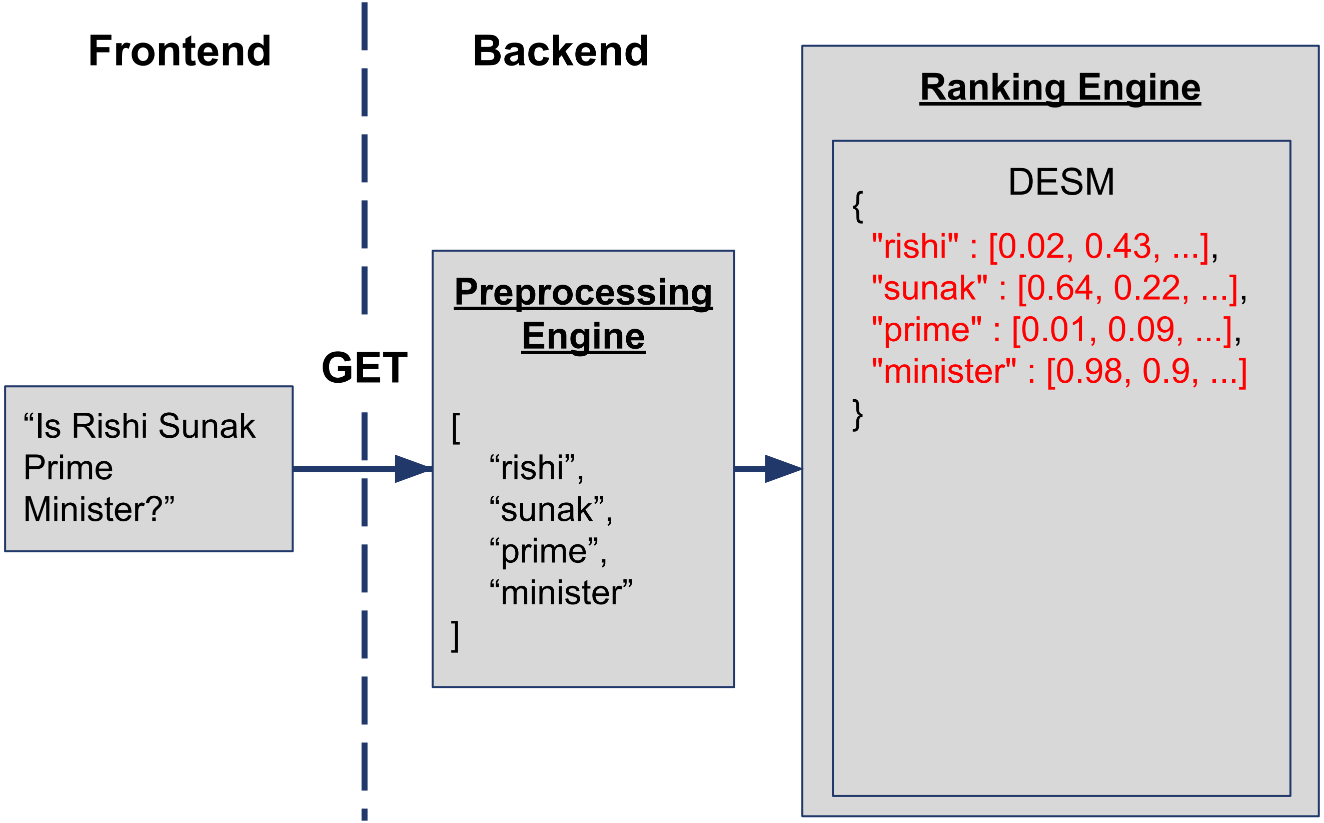
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
{
  “arsenal” : [0.23, 0.01 …],
  “first” : [2.9, 0.28, …],
  "rishi" : [0.02, 0.43, ...],
  “justice” : [1.32, 0.91, …],
  "sunak" : [0.64, 0.22, ...],
  "prime" : [0.01, 0.09, ...],
  “tennis” : [0.11, 0.83, …],
  “fired” : [9.1, 0.01 …],
  “ukraine” : [5.3, 0.22, …],
  “train” : [0.54, 0.63, …],
  "minister" : [0.98, 0.9, ...],
  …
}
```

Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

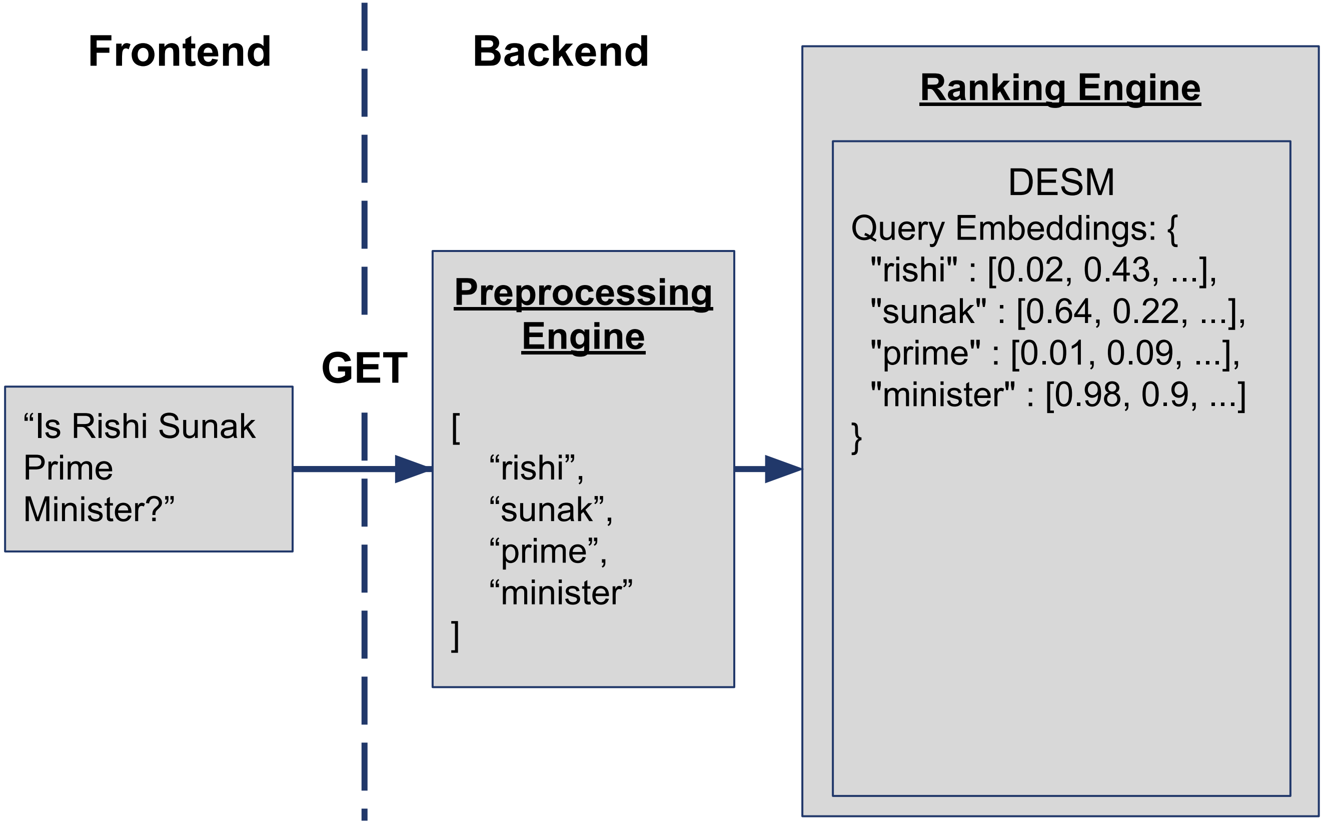
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
{
  "rishi" : [0.02, 0.43, ...],
  "sunak" : [0.64, 0.22, ...],
  "prime" : [0.01, 0.09, ...],
  "minister" : [0.98, 0.9, ...]
}
```

Inner Workings



Frontend

Backend

GET

"Is Rishi Sunak Prime Minister?"

Preprocessing Engine

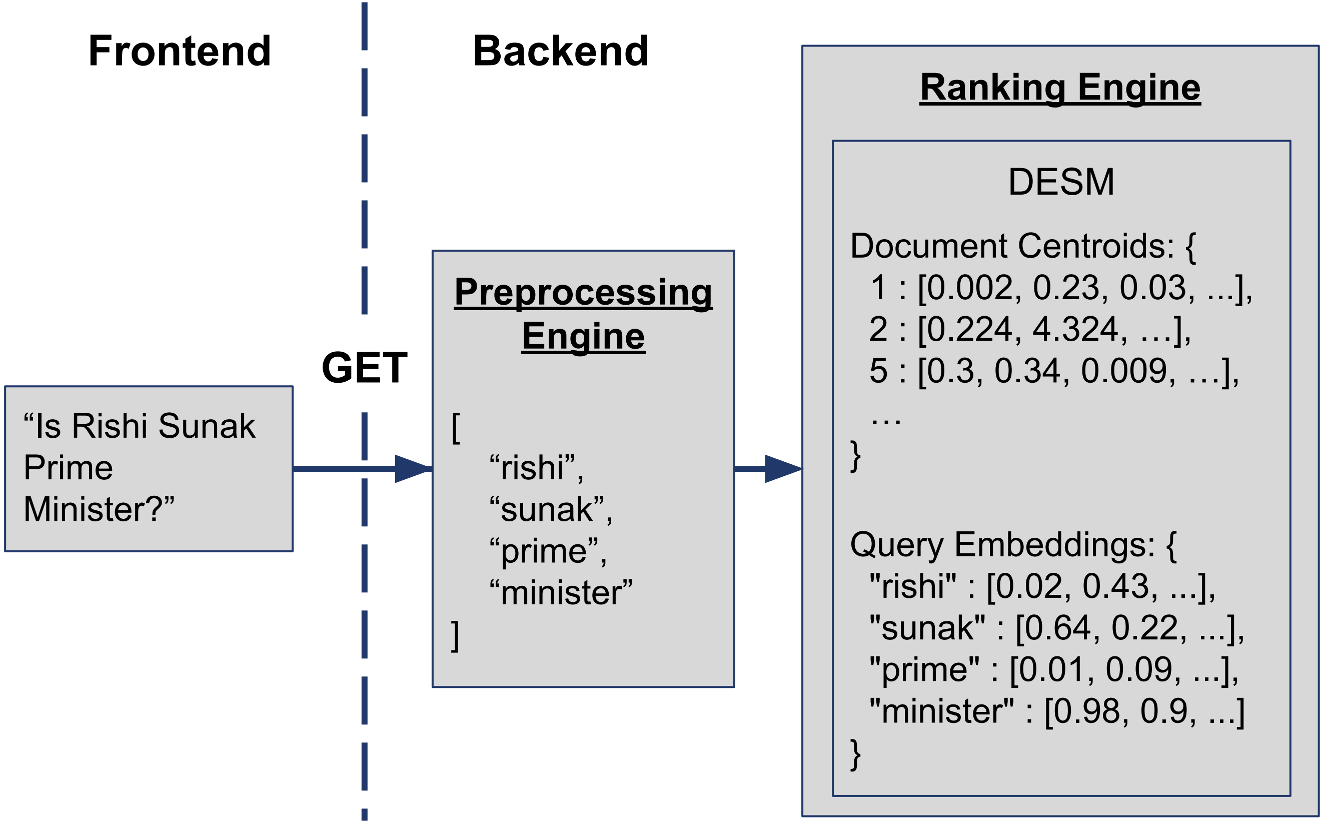
```
[
    "rishi",
    "sunak",
    "prime",
    "minister"
]
```

Ranking Engine

DESM

```
Query Embeddings: {
  "rishi" : [0.02, 0.43, ...],
  "sunak" : [0.64, 0.22, ...],
  "prime" : [0.01, 0.09, ...],
  "minister" : [0.98, 0.9, ...]
}
```

Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

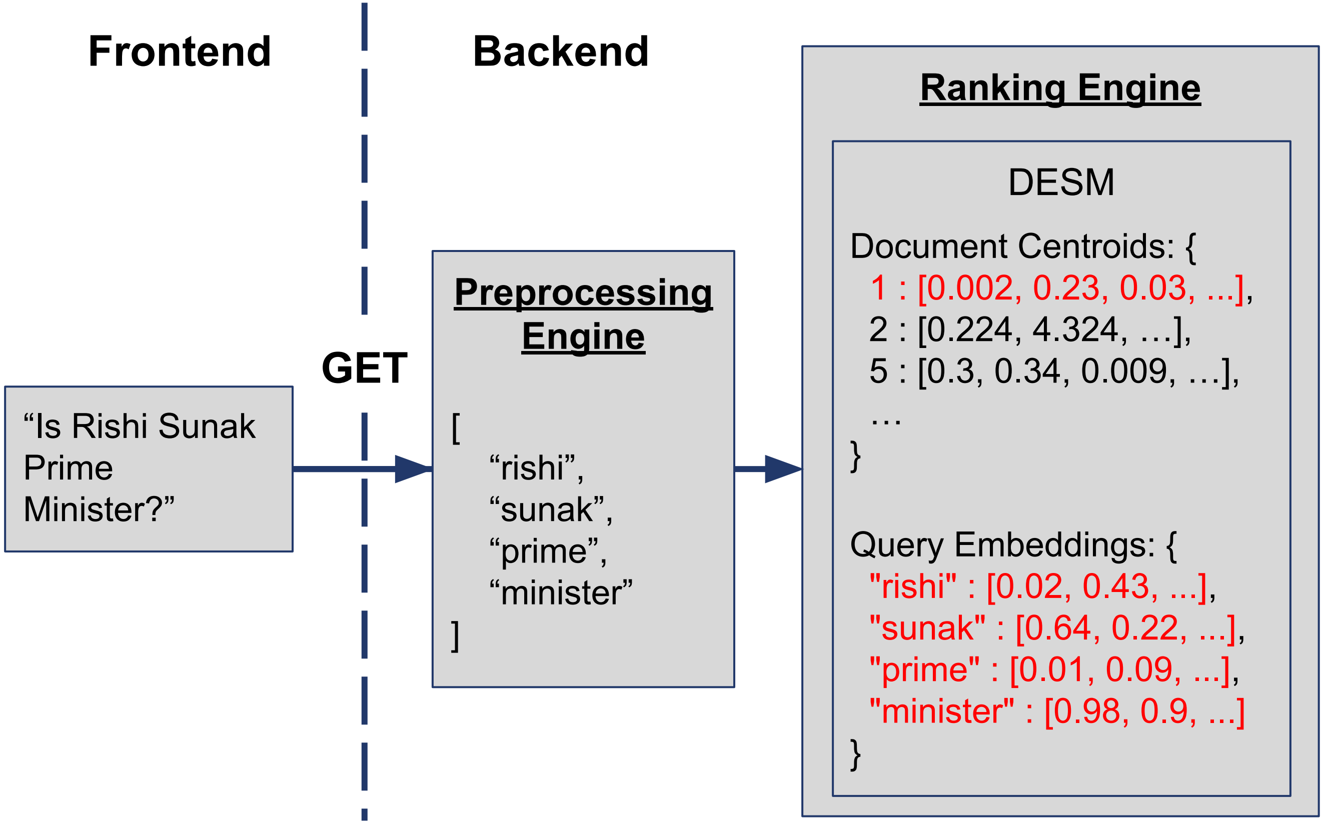
Ranking Engine

DESM

```
Document Centroids: {
  1 : [0.002, 0.23, 0.03, ...],
  2 : [0.224, 4.324, …],
  5 : [0.3, 0.34, 0.009, …],
  …
}

Query Embeddings: {
  "rishi" : [0.02, 0.43, ...],
  "sunak" : [0.64, 0.22, ...],
  "prime" : [0.01, 0.09, ...],
  "minister" : [0.98, 0.9, ...]
}
```

Inner Workings



“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

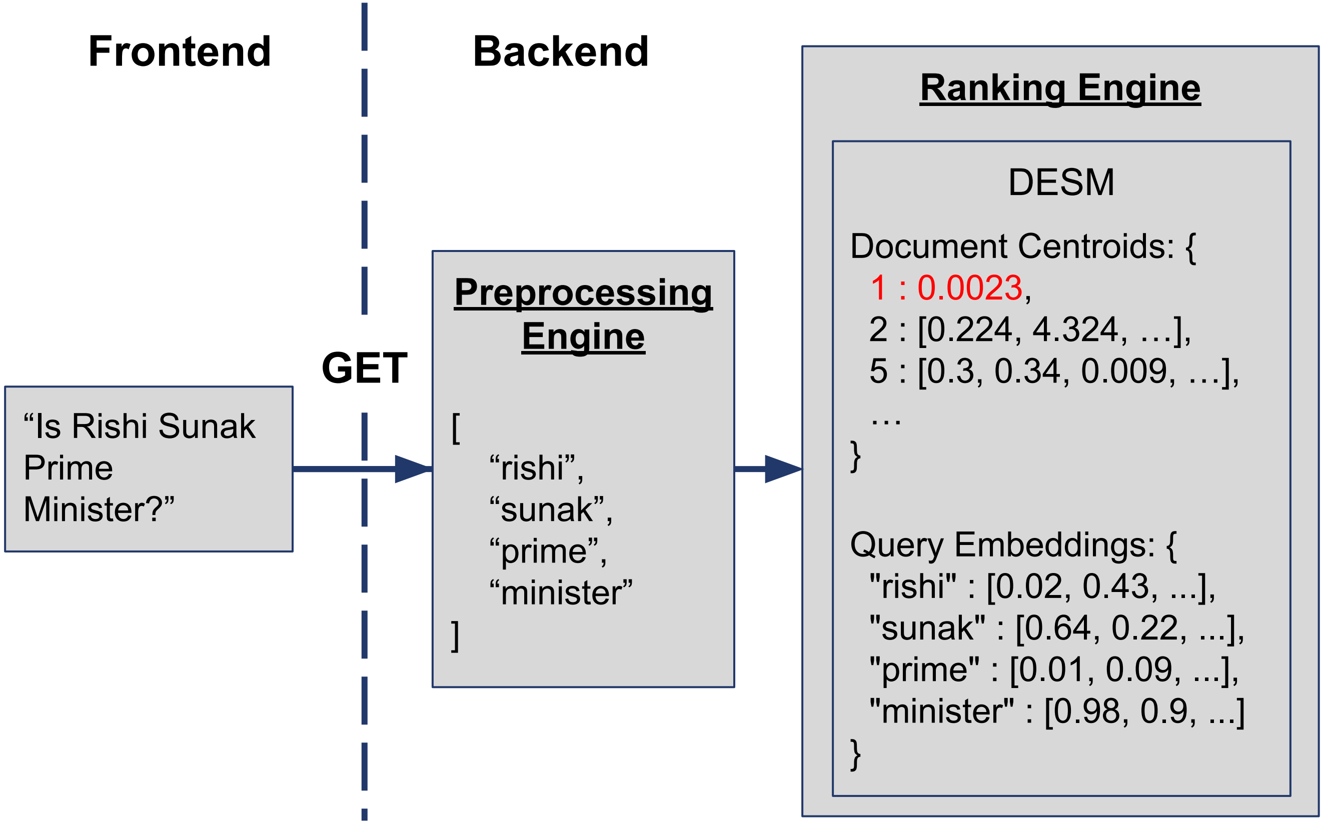
Ranking Engine

DESM

```
Document Centroids: {
  1 : [0.002, 0.23, 0.03, ...],
  2 : [0.224, 4.324, …],
  5 : [0.3, 0.34, 0.009, …],
  …
}

Query Embeddings: {
  "rishi" : [0.02, 0.43, ...],
  "sunak" : [0.64, 0.22, ...],
  "prime" : [0.01, 0.09, ...],
  "minister" : [0.98, 0.9, ...]
}
```

Inner Workings



Frontend

Backend

GET

"Is Rishi Sunak Prime Minister?"

Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

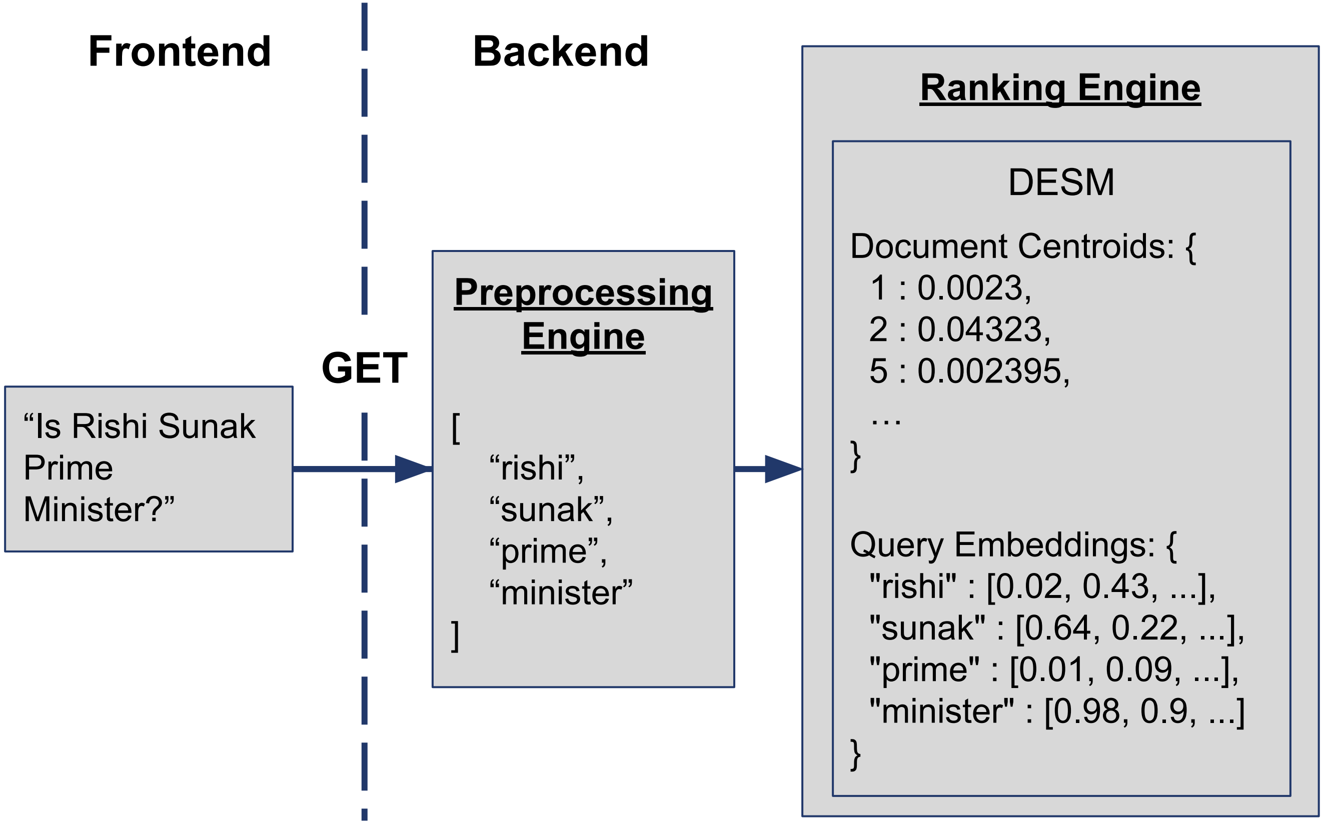
Ranking Engine

DESM

```
Document Centroids: {
  1 : 0.0023,
  2 : [0.224, 4.324, …],
  5 : [0.3, 0.34, 0.009, …],
  …
}

Query Embeddings: {
  "rishi" : [0.02, 0.43, ...],
  "sunak" : [0.64, 0.22, ...],
  "prime" : [0.01, 0.09, ...],
  "minister" : [0.98, 0.9, ...]
}
```

Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

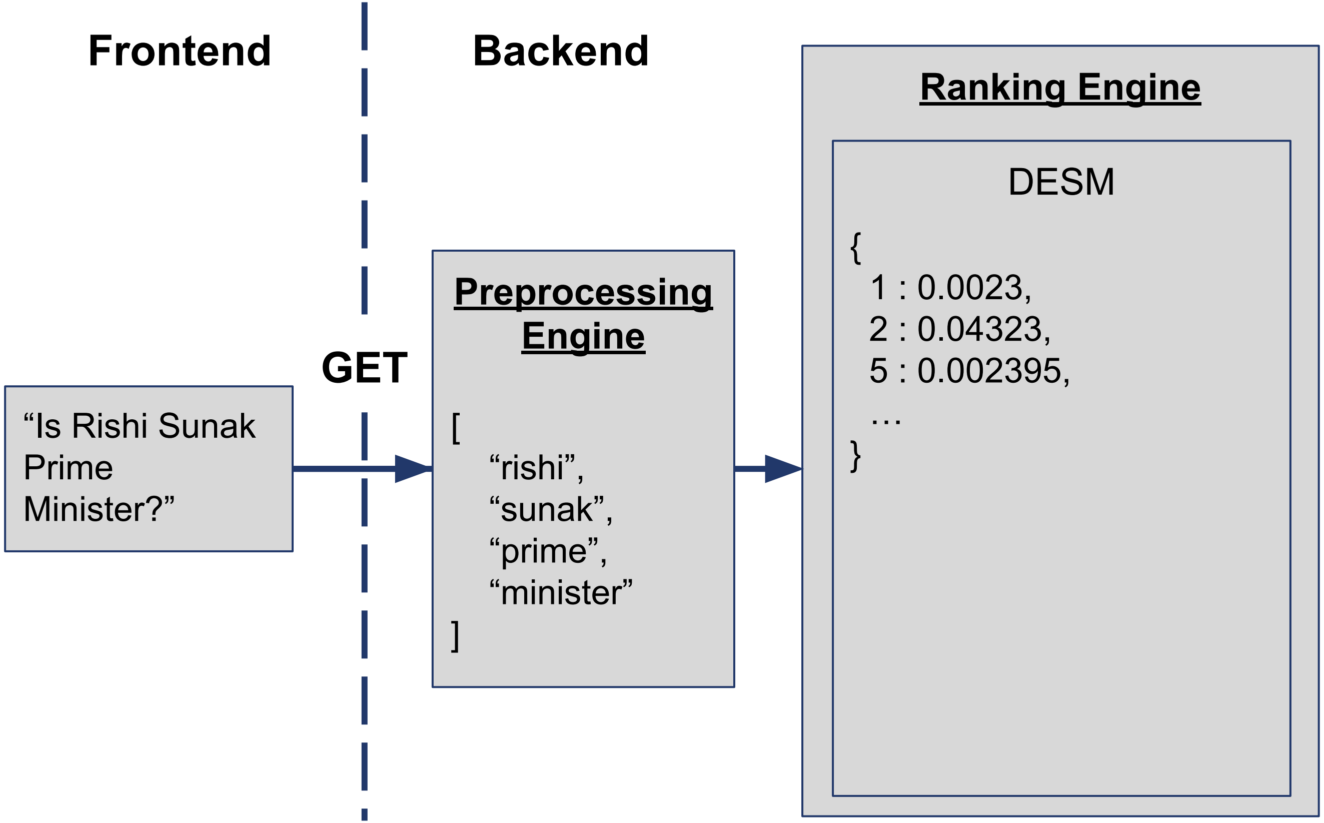
Ranking Engine

DESM

```
Document Centroids: {
  1 : 0.0023,
  2 : 0.04323,
  5 : 0.002395,
  …
}

Query Embeddings: {
  "rishi" : [0.02, 0.43, ...],
  "sunak" : [0.64, 0.22, ...],
  "prime" : [0.01, 0.09, ...],
  "minister" : [0.98, 0.9, ...]
}
```

Inner Workings



Frontend

Backend

GET

“Is Rishi Sunak Prime Minister?”

Preprocessing Engine

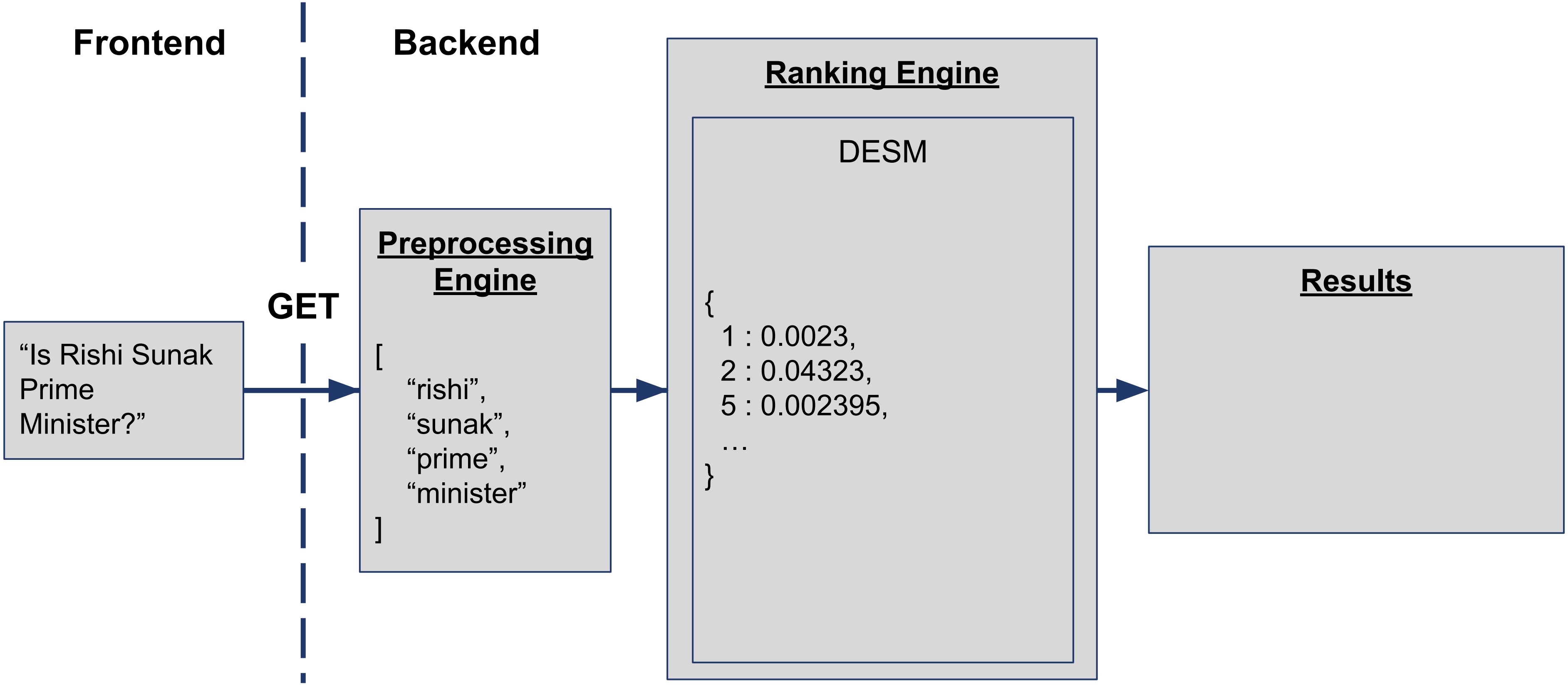
```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

```
{
  1 : 0.0023,
  2 : 0.04323,
  5 : 0.002395,
  …
}
```

Inner Workings



Frontend

Backend

“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

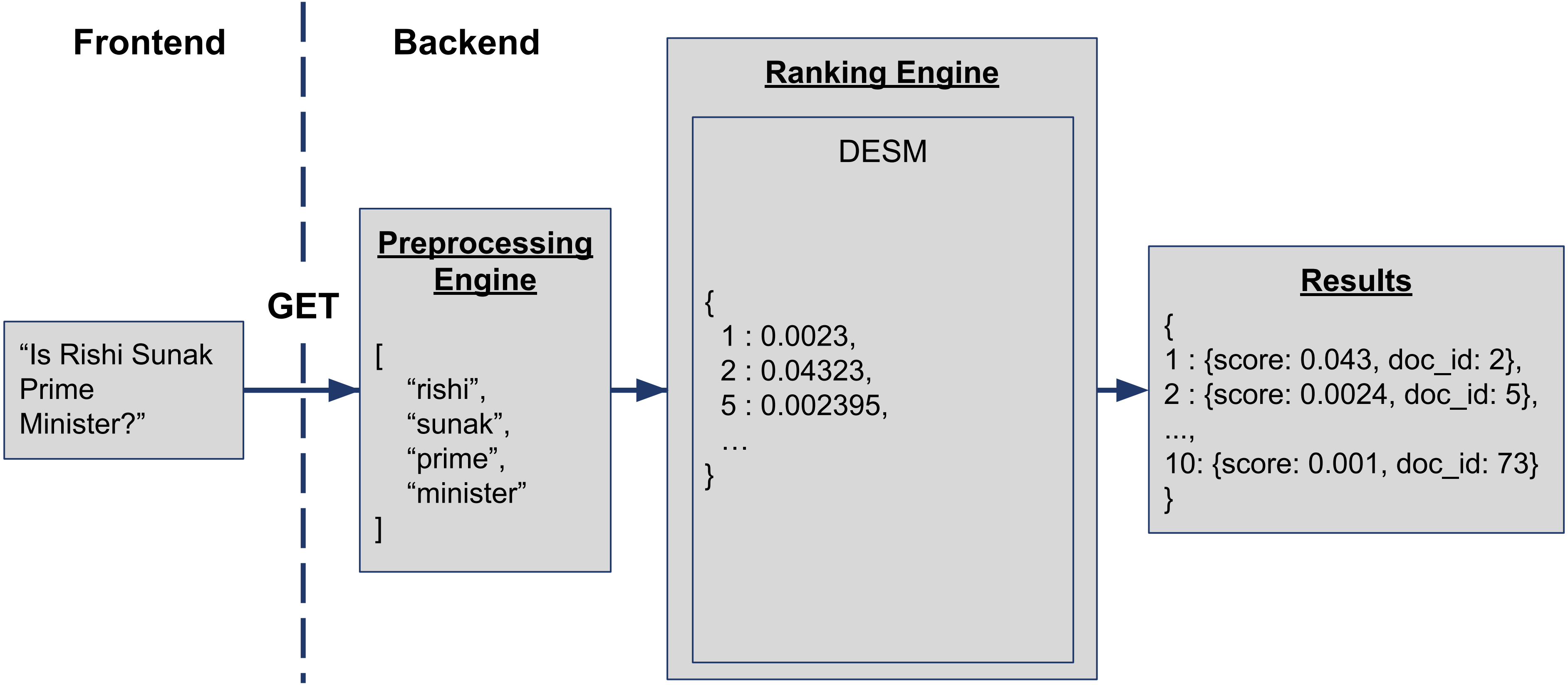
Ranking Engine

DESM

```
{
  1 : 0.0023,
  2 : 0.04323,
  5 : 0.002395,
  …
}
```

Results

Inner Workings



"Is Rishi Sunak Prime Minister?"

Preprocessing Engine

```
[
    "rishi",
    "sunak",
    "prime",
    "minister"
]
```

Ranking Engine

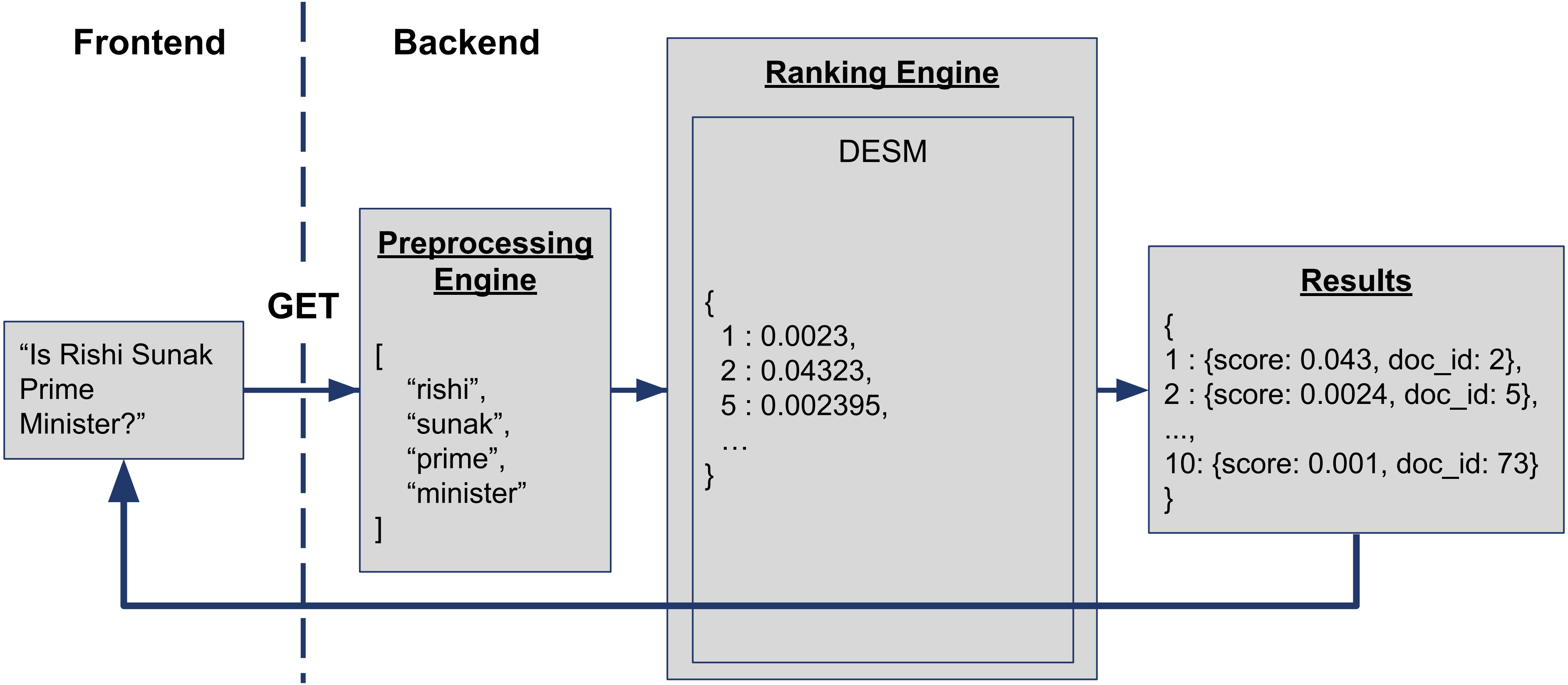
DESM

```
{
  1 : 0.0023,
  2 : 0.04323,
  5 : 0.002395,
  …
}
```

Results

```
{
1 : {score: 0.043, doc_id: 2},
2 : {score: 0.0024, doc_id: 5},
...,
10: {score: 0.001, doc_id: 73}
}
```

Inner Workings



Frontend

Backend

“Is Rishi Sunak Prime Minister?”

GET

Preprocessing Engine

```
[
    “rishi”,
    “sunak”,
    “prime”,
    “minister”
]
```

Ranking Engine

DESM

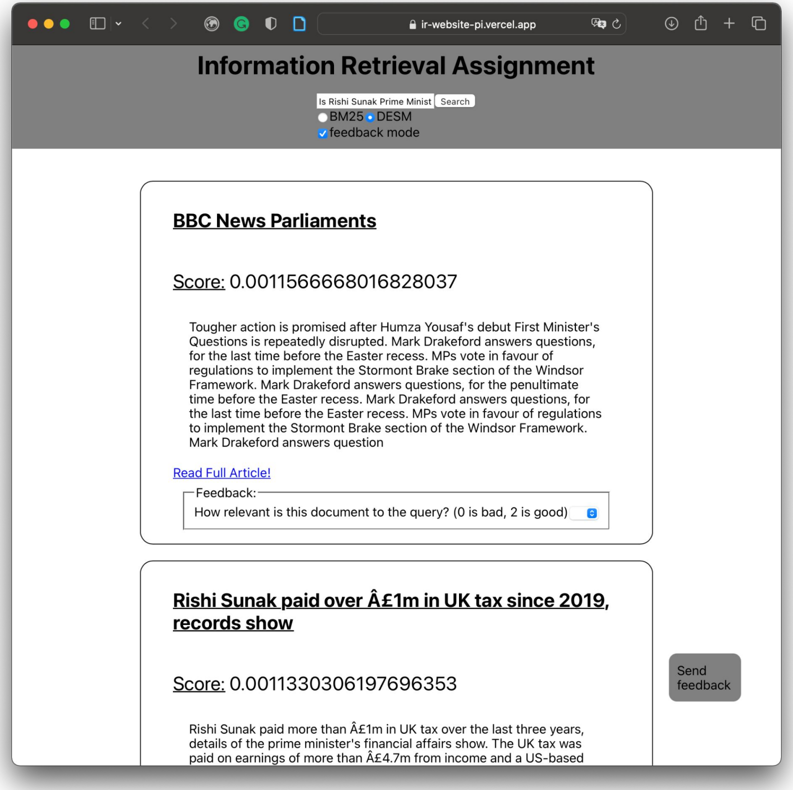
```
{
  1 : 0.0023,
  2 : 0.04323,
  5 : 0.002395,
  …
}
```

Results

```
{
1 : {score: 0.043, doc_id: 2},
2 : {score: 0.0024, doc_id: 5},
...,
10: {score: 0.001, doc_id: 73}
}
```

Queen Mary
University of London

Inner Workings



Information Retrieval Assignment

Is Rishi Sunak Prime Minist [Search]

○ BM25 ● DESM

☑ feedback mode

BBC News Parliaments

Score: 0.0011566668016828037

Tougher action is promised after Humza Yousaf's debut First Minister's Questions is repeatedly disrupted. Mark Drakeford answers questions, for the last time before the Easter recess. MPs vote in favour of regulations to implement the Stormont Brake section of the Windsor Framework. Mark Drakeford answers questions, for the penultimate time before the Easter recess. Mark Drakeford answers questions, for the last time before the Easter recess. MPs vote in favour of regulations to implement the Stormont Brake section of the Windsor Framework. Mark Drakeford answers question

Read Full Article!

Feedback:

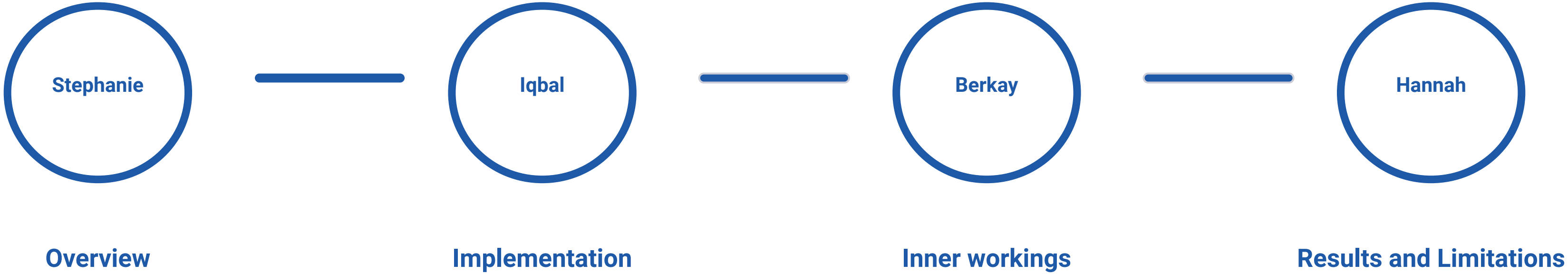
How relevant is this document to the query? (0 is bad, 2 is good)

Rishi Sunak paid over Â£1m in UK tax since 2019, records show

Score: 0.0011330306197696353

Rishi Sunak paid more than Â£1m in UK tax over the last three years, details of the prime minister's financial affairs show. The UK tax was paid on earnings of more than Â£4.7m from income and a US-based

Presentation Structure



Stephanie

Overview

Iqbal

Implementation

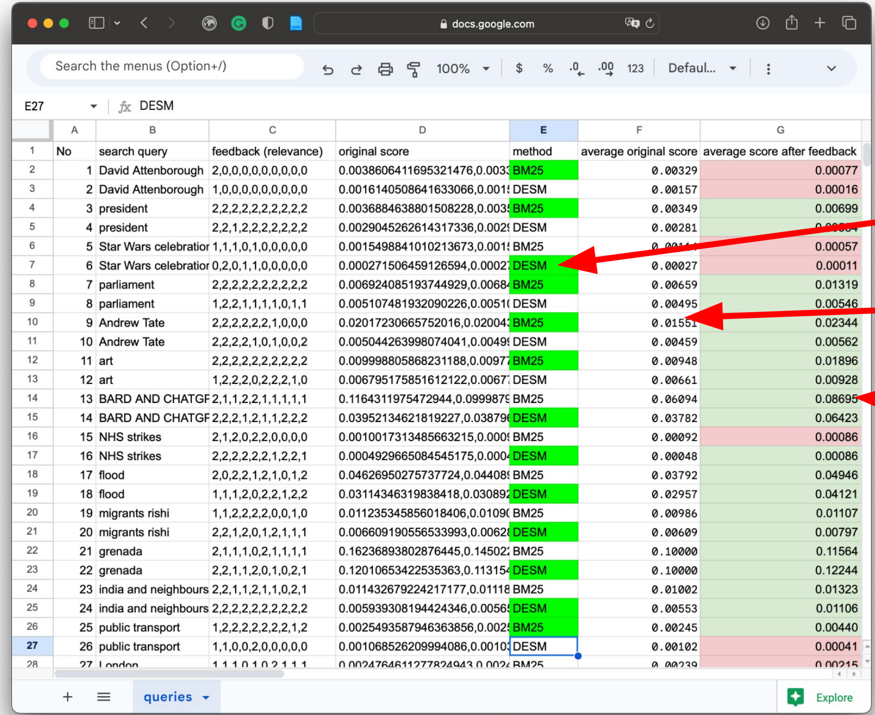
Berkay

Inner workings

Hannah

Results and Limitations

Results



	A	B	C	D	E	F	G
1	No	search query	feedback (relevance)	original score	method	average original score	average score after feedback
2	1	David Attenborough	2,0,0,0,0,0,0,0,0,0	0.0038606411695321476,0.0033	BM25	0.00329	0.00077
3	2	David Attenborough	1,0,0,0,0,0,0,0,0,0	0.0016140508641633066,0.0015	DESM	0.00157	0.00016
4	3	president	2,2,2,2,2,2,2,2,2,2	0.0036884638801508228,0.0035	BM25	0.00349	0.00699
5	4	president	2,2,1,2,2,2,2,2,2,2	0.0029045262614317336,0.0029	DESM	0.00281	0.00584
6	5	Star Wars celebratior	1,1,1,0,1,0,0,0,0,0	0.0015498841010213673,0.0015	BM25	0.00114	0.00057
7	6	Star Wars celebratior	0,2,0,1,1,0,0,0,0,0	0.000271506459126594,0.00027	DESM	0.00027	0.00011
8	7	parliament	2,2,2,2,2,2,2,2,2,2	0.006924085193744929,0.0068	BM25	0.00659	0.01319
9	8	parliament	1,2,2,1,1,1,1,0,1,1	0.005107481932090226,0.0051	DESM	0.00495	0.00546
10	9	Andrew Tate	2,2,2,2,2,2,1,0,0,0	0.02017230665752016,0.02004	BM25	0.01551	0.02344
11	10	Andrew Tate	2,2,2,2,1,0,1,0,0,2	0.005044263998074041,0.0049	DESM	0.00459	0.00562
12	11	art	2,2,2,2,2,2,2,2,2,2	0.009998805868231188,0.00977	BM25	0.00948	0.01896
13	12	art	1,2,2,2,0,2,2,2,1,0	0.006795175851612122,0.00677	DESM	0.00661	0.00928
14	13	BARD AND CHATGF	2,1,1,2,2,1,1,1,1,1	0.1164311975472944,0.0999879	BM25	0.06094	0.08695
15	14	BARD AND CHATGF	2,2,2,1,2,1,1,2,2,2	0.03952134621819227,0.038796	DESM	0.03782	0.06423
16	15	NHS strikes	2,1,2,0,2,2,0,0,0,0	0.0010017313485663215,0.0009	BM25	0.00092	0.00086
17	16	NHS strikes	2,2,2,2,2,2,1,2,2,1	0.0004929665084545175,0.0004	DESM	0.00048	0.00086
18	17	flood	2,0,2,2,1,2,1,0,1,2	0.04626950275737724,0.04408	BM25	0.03792	0.04946
19	18	flood	1,1,1,2,0,2,2,1,2,2	0.03114346319838418,0.030892	DESM	0.02957	0.04121
20	19	migrants rishi	1,1,2,2,2,2,0,0,1,0	0.011235345856018406,0.0109	BM25	0.00986	0.01107
21	20	migrants rishi	2,2,1,2,0,1,2,1,1,1	0.006609190556533993,0.0062	DESM	0.00609	0.00797
22	21	grenada	2,1,1,1,0,2,1,1,1,1	0.16236893802876445,0.14502	BM25	0.10000	0.11564
23	22	grenada	2,2,1,1,2,0,1,0,2,1	0.12010653422535363,0.11315	DESM	0.10000	0.12244
24	23	india and neighbours	2,2,1,1,2,1,1,0,2,1	0.011432679224217177,0.01118	BM25	0.01002	0.01323
25	24	india and neighbours	2,2,2,2,2,2,2,2,2,2	0.005939308194424346,0.00565	DESM	0.00553	0.01106
26	25	public transport	1,2,2,2,2,2,2,2,1,2	0.0025493587946363856,0.0025	BM25	0.00245	0.00440
27	26	public transport	1,1,0,0,2,0,0,0,0,0	0.001068526209994086,0.0010	DESM	0.00102	0.00041
28	27	London	1,1,1,0,1,0,2,1,1,1	0.0024764611277824943,0.0024	BM25	0.00239	0.00215

Legend:

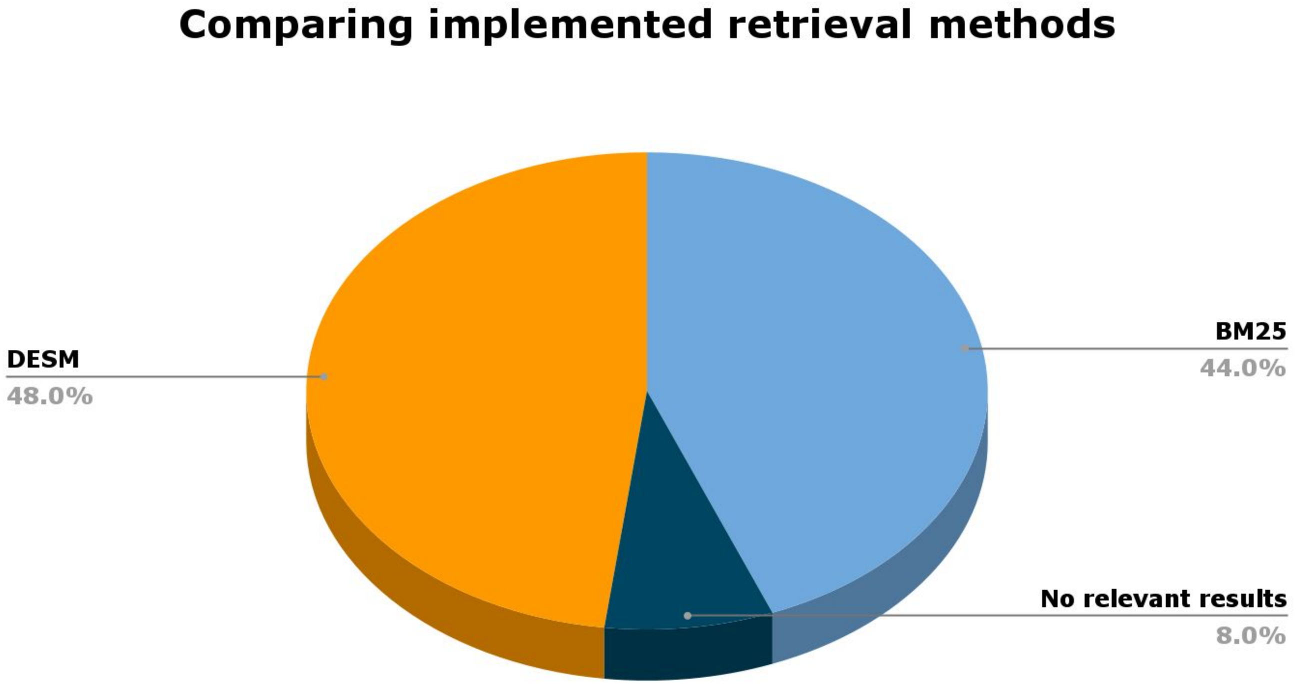
Retrieval method with the best search results for the particular query

Average score of the 10 shown articles

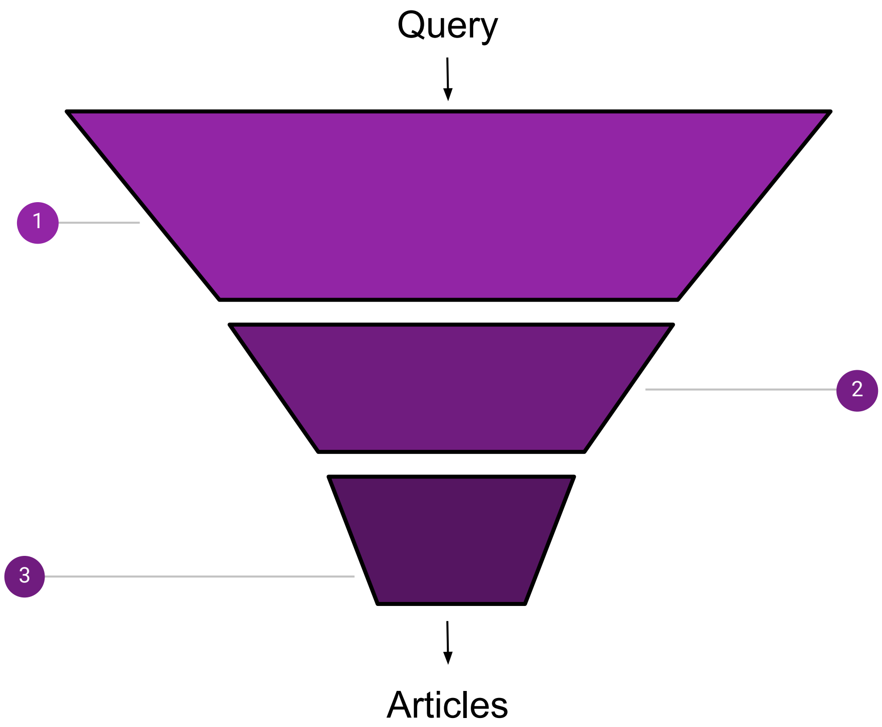
Average score of the 10 shown articles after the relevance feedback has been submitted (Evaluation of the performance)

Green: Improvement of the average score
Red: Decrease of the average score

Results


Comparing implemented retrieval methods
DESM
48.0%
BM25
44.0%
No relevant results
8.0%

Results



Preprocessing

Stopwords, punctuation and numbers are removed.

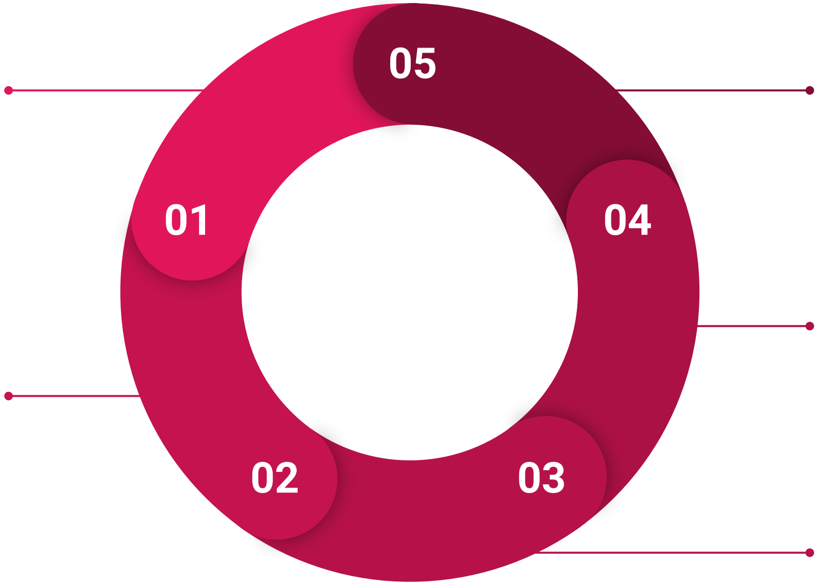
Inverted Index

Only get documents that contain at least one of the preprocessed and tokenised words of the query

Limited number of articles

Choose the 10 most relevant articles and show the results

Limitations



Indexing

Search engines cannot index all web pages due to lack of access or the great amount of data available

Accuracy

Search results cannot be fully accurate all the time, due to certain factors such as spam, clickbait, or ads

Language

Some language might not be as well represented as others, i.e. English, French, etc.The search engine might also prioritise results from specific regions

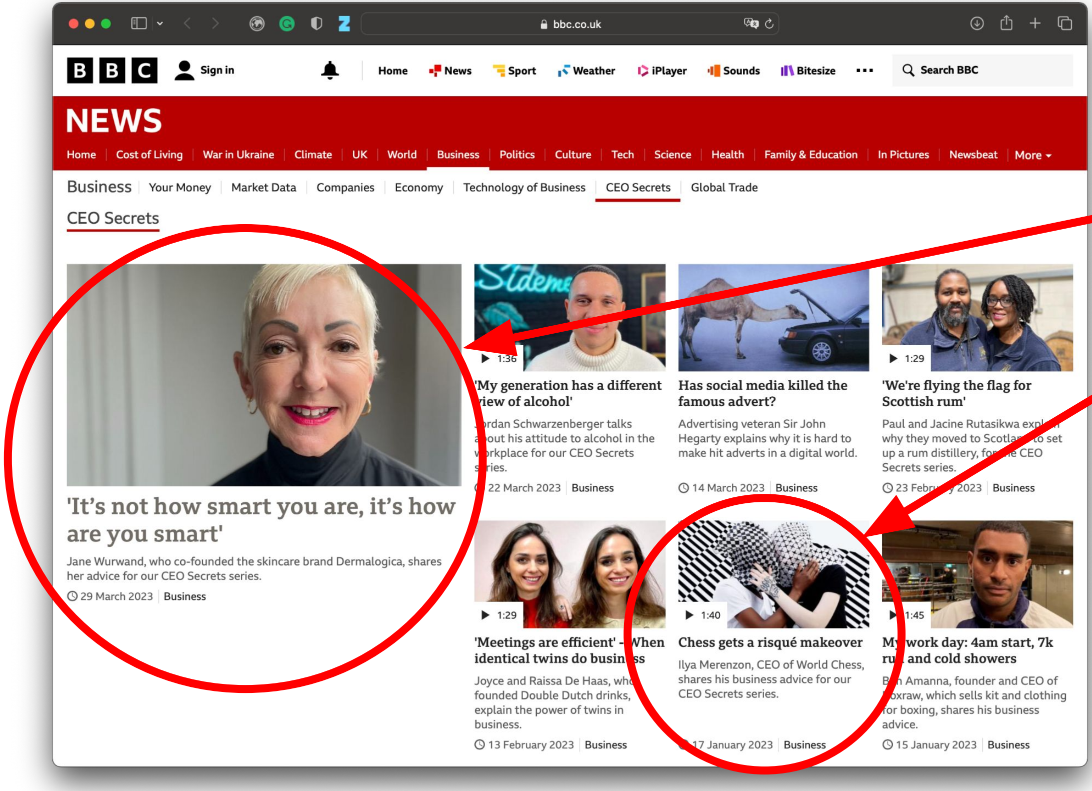
Bias

Search results can be influenced by personalisation, user data tracking or ads

Context

Search engines might not always understand the context of the query, i.e. apple as a fruit or the company

Limitations



bbc.co.uk

BBC Sign in

Home News Sport Weather iPlayer Sounds Bitesize ··· Search BBC

NEWS

Home | Cost of Living | War in Ukraine | Climate | UK | World | Business | Politics | Culture | Tech | Science | Health | Family & Education | In Pictures | Newsbeat | More

Business | Your Money | Market Data | Companies | Economy | Technology of Business | CEO Secrets | Global Trade

CEO Secrets

'It's not how smart you are, it's how are you smart'

Jane Wurwand, who co-founded the skincare brand Dermalogica, shares her advice for our CEO Secrets series.

29 March 2023 | Business

1:36

'My generation has a different iew of alcohol'

rdan Schwarzenberger talks out his attitude to alcohol in the orkplace for our CEO Secrets ries.

22 March 2023 | Business

Has social media killed the famous advert?

Advertising veteran Sir John Hegarty explains why it is hard to make hit adverts in a digital world.

14 March 2023 | Business

1:29

'We're flying the flag for Scottish rum'

Paul and Jacine Rutasikwa expl why they moved to Scotla to set up a rum distillery, fo e CEO Secrets series.

23 Febru y 2023 | Business

1:29

'Meetings are efficient' - When identical twins do busin ss

Joyce and Raissa De Haas, wh founded Double Dutch drinks, explain the power of twins in business.

13 February 2023 | Business

1:40

Chess gets a risqué makeover

Ilya Merenzon, CEO of World Chess, shares his business advice for our CEO Secrets series.

17 January 2023 | Business

1:45

My work day: 4am start, 7k ru and cold showers

B n Amanna, founder and CEO of oxraw, which sells kit and clothing for boxing, shares his business advice.

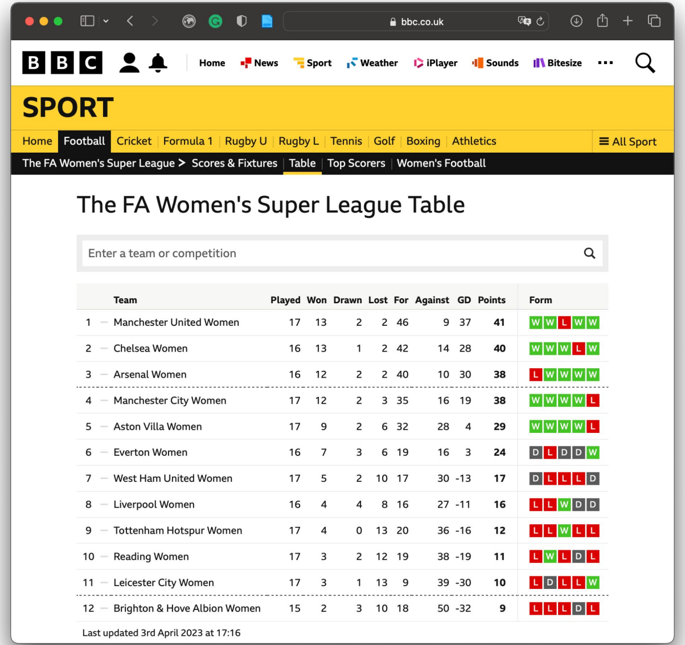
15 January 2023 | Business

Query: Chess

Article with first index: BBC News CEO Secrets

Intended article: Chess gets a risqué makeover

Limitations



The FA Women's Super League Table

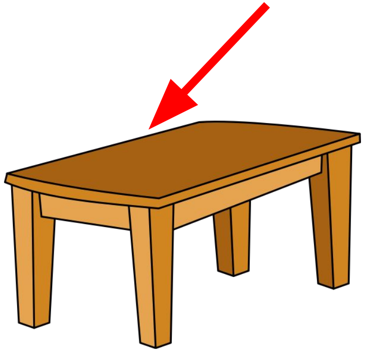
	Team	Played	Won	Drawn	Lost	For	Against	GD	Points	Form
1	Manchester United Women	17	13	2	2	46	9	37	**41**	W W L W W
2	Chelsea Women	16	13	1	2	42	14	28	**40**	W W W L W
3	Arsenal Women	16	12	2	2	40	10	30	**38**	L W W W W
4	Manchester City Women	17	12	2	3	35	16	19	**38**	W W W W L
5	Aston Villa Women	17	9	2	6	32	28	4	**29**	W W W W L
6	Everton Women	16	7	3	6	19	16	3	**24**	D L D D W
7	West Ham United Women	17	5	2	10	17	30	-13	**17**	D L L L D
8	Liverpool Women	16	4	4	8	16	27	-11	**16**	L L W D D
9	Tottenham Hotspur Women	17	4	0	13	20	36	-16	**12**	L L W L L
10	Reading Women	17	3	2	12	19	38	-19	**11**	L W L D L
11	Leicester City Women	17	3	1	13	9	39	-30	**10**	L D L L W
12	Brighton & Hove Albion Women	15	2	3	10	18	50	-32	**9**	L L L D L

Last updated 3rd April 2023 at 17:16

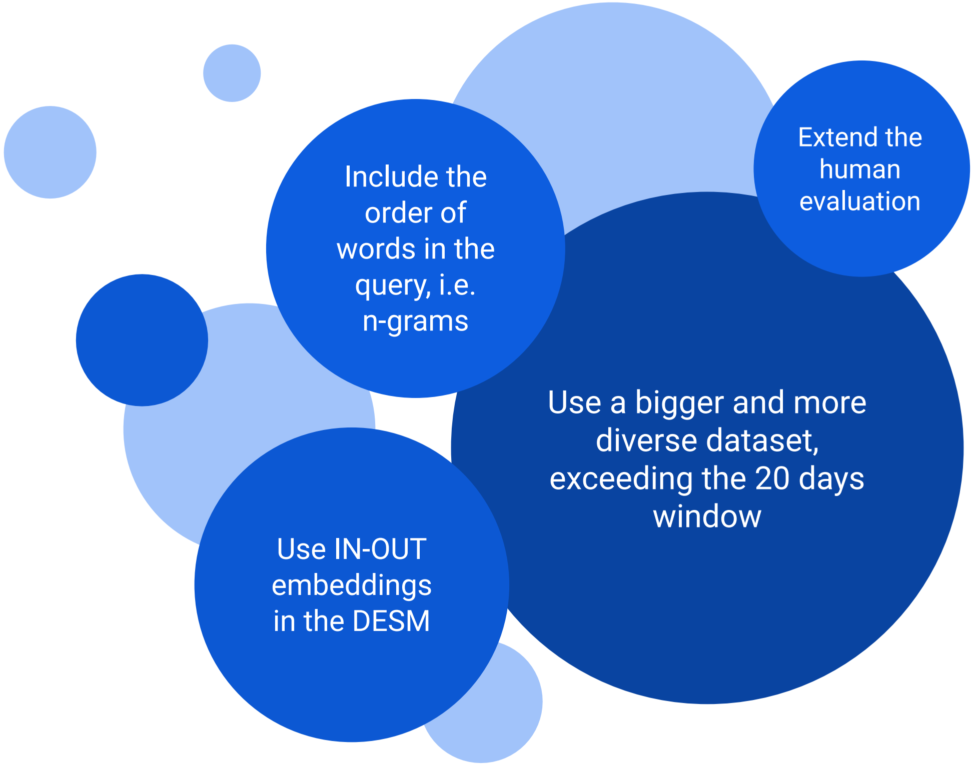
<u>Query: Table</u>

Results: BBC Sport News

Intended article: Something about tables (furniture)



Outlook



Include the order of words in the query, i.e. n-grams

Extend the human evaluation

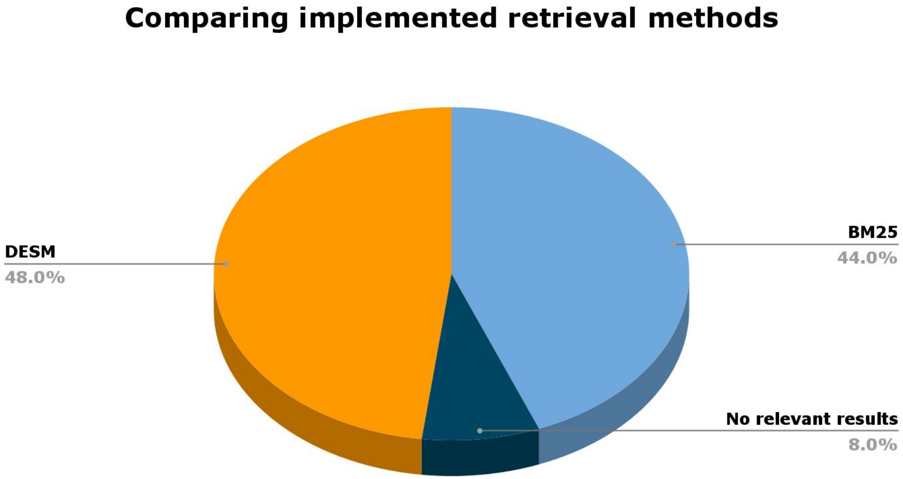
Use a bigger and more diverse dataset, exceeding the 20 days window

Use IN-OUT embeddings in the DESM

Conclusion

DESM performs better than BM25

- DESM is more capable of showcasing relevant news according to the given query.



DESM > BM25

Appendix

Presentation's Journal Paper:

Nalisnick, E. *et al.* (2016) "Improving document ranking with dual word embeddings," *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, pp. 83–84. Available at: https://doi.org/10.1145/2872518.2889361.

Notable citations:

- Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R. and Cheng, X. (2022). Semantic Models for the First-Stage Retrieval: A Comprehensive Review. ACM Transactions on Information Systems, [online] 40(4), pp.1–42. doi:https://doi.org/10.1145/3486250.
- Muhammad, I., Bollegala, D., Coenen, F., Gamble, C., Kearney, A. and Williamson, P. (2021). Document Ranking for Curated Document Databases Using BERT and Knowledge Graph Embeddings: Introducing GRAB-Rank. Big Data Analytics and Knowledge Discovery, pp.116–127. doi:https://doi.org/10.1007/978-3-030-86534-4_10.
- Chy, A.N., Ullah, M.Z. and Aono, M. (2019). Query Expansion for Microblog Retrieval Focusing on an Ensemble of Features. Journal of Information Processing, [online] 27, pp.61–76. doi:https://doi.org/10.2197/ipsjjip.27.61.
- Khattab, O., Hammoud, M. and Elsayed, T. (2020). Finding the Best of Both Worlds. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. doi:https://doi.org/10.1145/3397271.3401076.
- Järvelin, K. and Kekäläinen, J. (2017). IR evaluation methods for retrieving highly relevant documents. ACM SIGIR Forum, 51(2), pp.243–250. doi:https://doi.org/10.1145/3130348.3130374.
- Mitra, B., Nalisnick, E., Craswell, N., Caruana, R. and Redmond, M. (2016). *A Dual Embedding Space Model for Document Ranking*. [online] Available at: https://arxiv.org/pdf/1602.01137.pd
- Weng, L. (2017) *Learning word embedding*, *Lil'Log (Alt + H)*. Available at: https://lilianweng.github.io/posts/2017-10-15-word-embedding/ (Accessed: February 20, 2023).