Advanced Document Similarity With Apache Lucene

Alessandro Benedetti, Software Engineer, Sease Ltd.

Who I am

Alessandro Benedetti

- Search Consultant
- R&D Software Engineer
- Master in Computer Science
- Apache Lucene/Solr Enthusiast
- Semantic, NLP, Machine Learning Technologies passionate
- Beach Volleyball Player & Snowboarder

Sease Ltd

Search **Se**rvices

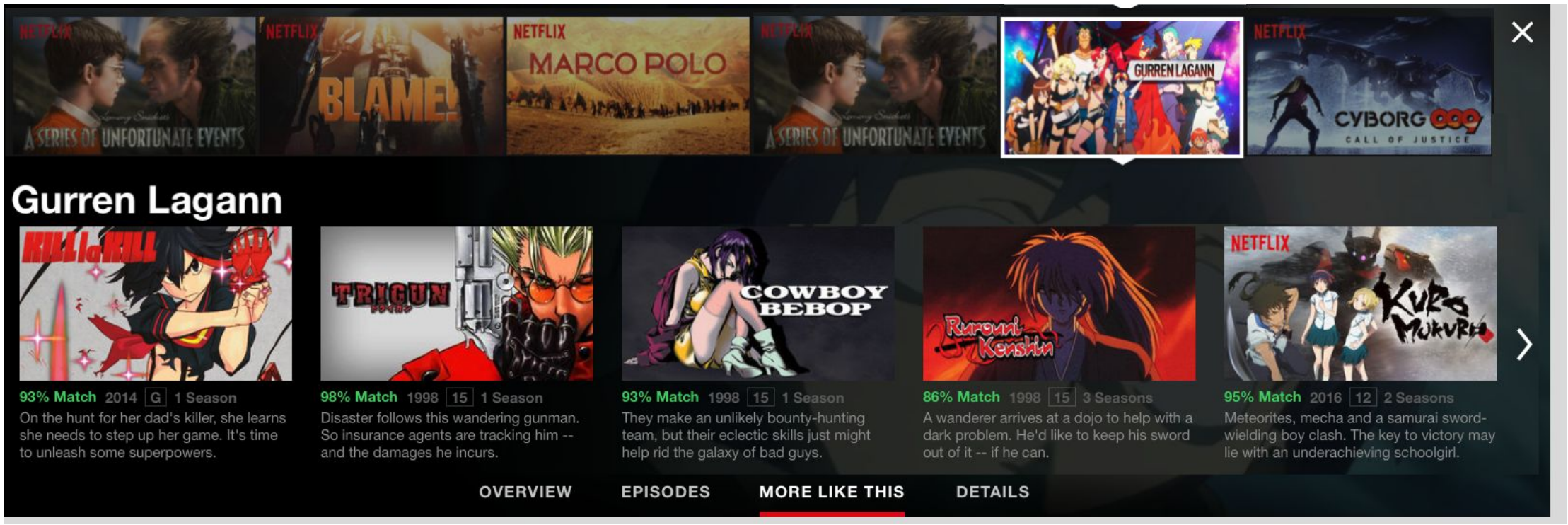
- Open Source Enthusiasts
- Apache Lucene/Solr experts
- Community Contributors
- Active Researchers
- Hot Trends : Learning To Rank, Document Similarity,
  Measuring Search Quality, Relevancy Tuning

Agenda

- Document Similarity
- Apache Lucene More Like This
- Term Scorer
- BM25
- Interesting Terms Retrieval
- Query Building
- DEMO
- Future Work
- JIRA References

SeaSe

Real World Use Cases - Streaming Services



Real World Use Cases - Hotels

Alessandro, we found properties like APA Hotel Iidabashi Ekimae that other travellers liked

APA Hotel Tokyo Kudanshita ★★★

A 3-minute walk from Kudanshita Subway Station and 1 km from Tokyo Dome, APA Hotel Tokyo Kudanshita features a Japanese restaurant and free Wi-Fi.

5 people are looking at this moment

Score from **1,277** reviews

Good 7.5 /10

Total price from:

£650

Book now

FLEXSTAY INN Iidabashi ★★★

Conveniently located a 8-minute walk from JR Iidabashi Train Station and Iidabashi Subway Station, Flexstay Inn Iidabashi offers self-catering accommodation with free WiFi access throughout.

5 people are looking at this moment

Score from **970** reviews

Good 7.5 /10

Total price from:

£512

Book now

APA Hotel Hanzomon Hirakawacho ★★★

Open from June 2014, APA Hotel Hanzomon Hirakawacho is conveniently located within a 6-minute walk from 3 subway stations, including Hanzomon and Kojimachi stations. It offers free WiFi in all areas.

2 people are looking at this moment

Score from **2,513** reviews

Very good 8 /10

Total price from:

£758

Book now

APA Hotel Asakusabashi-Ekikita ★★★

Opening in June 2015, this hotel is conveniently located a 4-minute walk away from Asakusabashi subway Station on the Asakusa Line and JR Sobu Line.

5 people are looking at this moment

Score from **1,813** reviews

Very good 8.1 /10

Total price from:

£637

Book now

» Hotels in Tokyo

Document Similarity

Problem : find similar documents to a seed one

Solution(s) :

- Collaborative approach (users interactions)
- **Content Based**
- Hybrid

Similar ?

- Documents accessed in association to the input one by users close to you
- **Terms distributions**
- All of above

SeaSe

Apache Lucene

Apache Lucene™ is a high-performance, full-featured text search engine library written entirely in Java.

It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

Apache Lucene is an open source project available for free download.

Apache Lucene

- Search Library (java)
- Structured Documents
- Inverted Index
- Similarity Metrics (TF-IDF, BM25)
- Fast Search
- Support for advanced queries
- Relevancy tuning

```
Doc0
  { "id":"c",
    "title":"video game history"
},

Doc1
  { "id":"a",
    "title":"game video review game"
},

Doc2
  { "id":"b",
    "title":"game store"
},
```

SeaSe

Inverted Index

Doc0
```
{ "id":"c",
  "title":"video game history"
},
```

Doc1
```
{ "id":"a",
  "title":"game video review game"
},
```

Doc2
```
{ "id":"b",
  "title":"game store"
},
```

Indexing

Field	id		
Ordinal	Term	Document Frequency	Posting List
0	a	1	1 : 1 : [1] : [0-1]
1	b	1	2 : 1 : [1] : [0-1]
2	c	1	0 : 1 : [1] : [0-1]

Field	title		
Ordinal	Term	Document Frequency	Posting List
0	game	3	0 : 1 : [2] : [6-10], 1 : 2 : [1, 4] : [0-4, 18-22], 2 : 1 : [1] : [0-4]
1	history	1	0 : 1 : [3] : [11-18]
2	review	1	1 : 1 : [3] : [11-17]
3	store	1	2 : 1 : [2] : [5-10]
4	video	2	0 : 1 : [1] : [0-5], 1 : 1 : [2] : [5-10],

SeaSe

More Like This

Pros

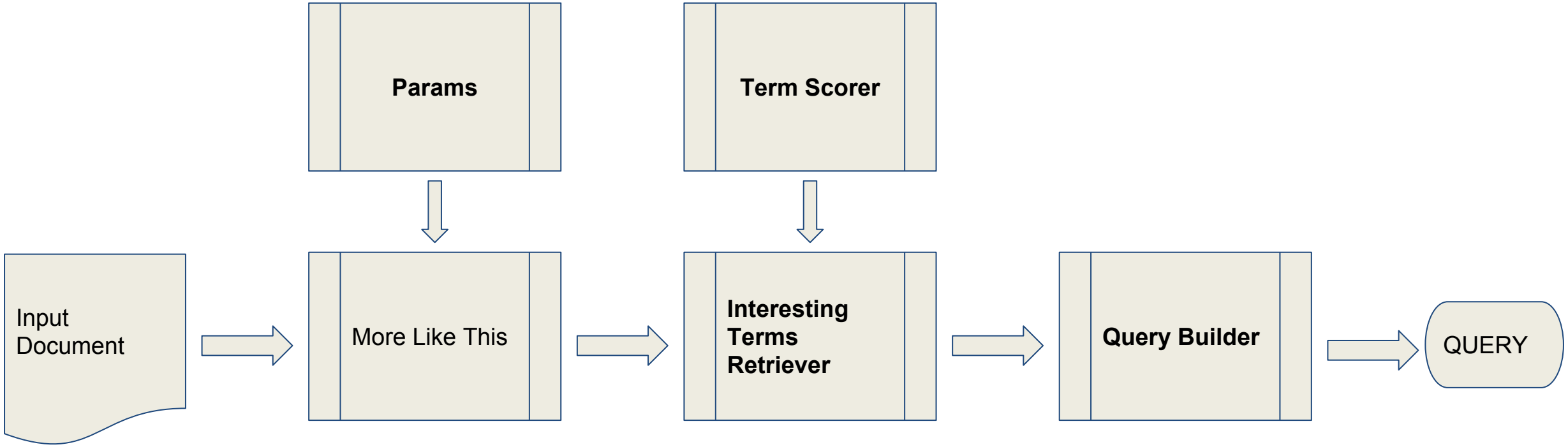
- Apache Lucene Module
- Advanced Params
- Input :
  - structured document
  - just text
- Build an advanced query
- Leverage the Inverted Index
  (and additional data structures)

Cons

- Massive single class
- Low cohesion
- Low readability
- Minimum test coverage
- Difficult to extend
  (and improve)

SeaSe

More Like This - Break Up



More Like This Params

Responsibility : define a set of parameters (and defaults) that affect the various components of the More Like This module

- Regulate MLT behavior
- Groups parameters specific to each component
- Javadoc documentation
- Default values
- Useful container for various parameters to be passed

Term Scorer

Responsibility : assign a score to a term that measure how distinctive is the term for the document in input

- Field Name
- Field Stats (Document Count)
- Term Stats (Document Frequency)
- Term Frequency

- **TF-IDF** -> tf * (log (numDocs / docFreq + 1) + 1)
- **BM25**

SeaSe

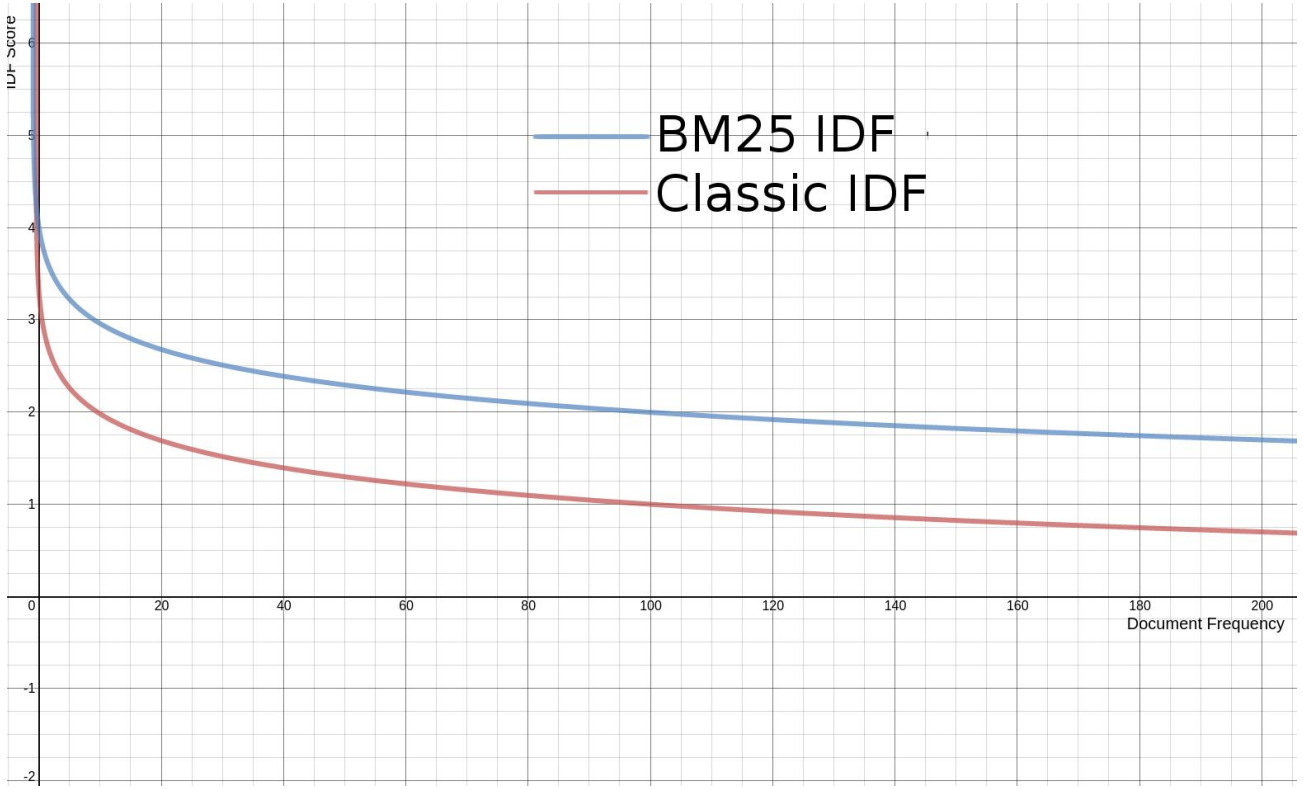
BM25 Term Scorer

- Origin from Probabilistic Information Retrieval
- Default Similarity from Lucene 6.0 [1]
- 25th iteration in improving TF-IDF
- TF
- IDF
- Document Length

[1] LUCENE-6789

BM25 Term Scorer - Inverse Document Frequency

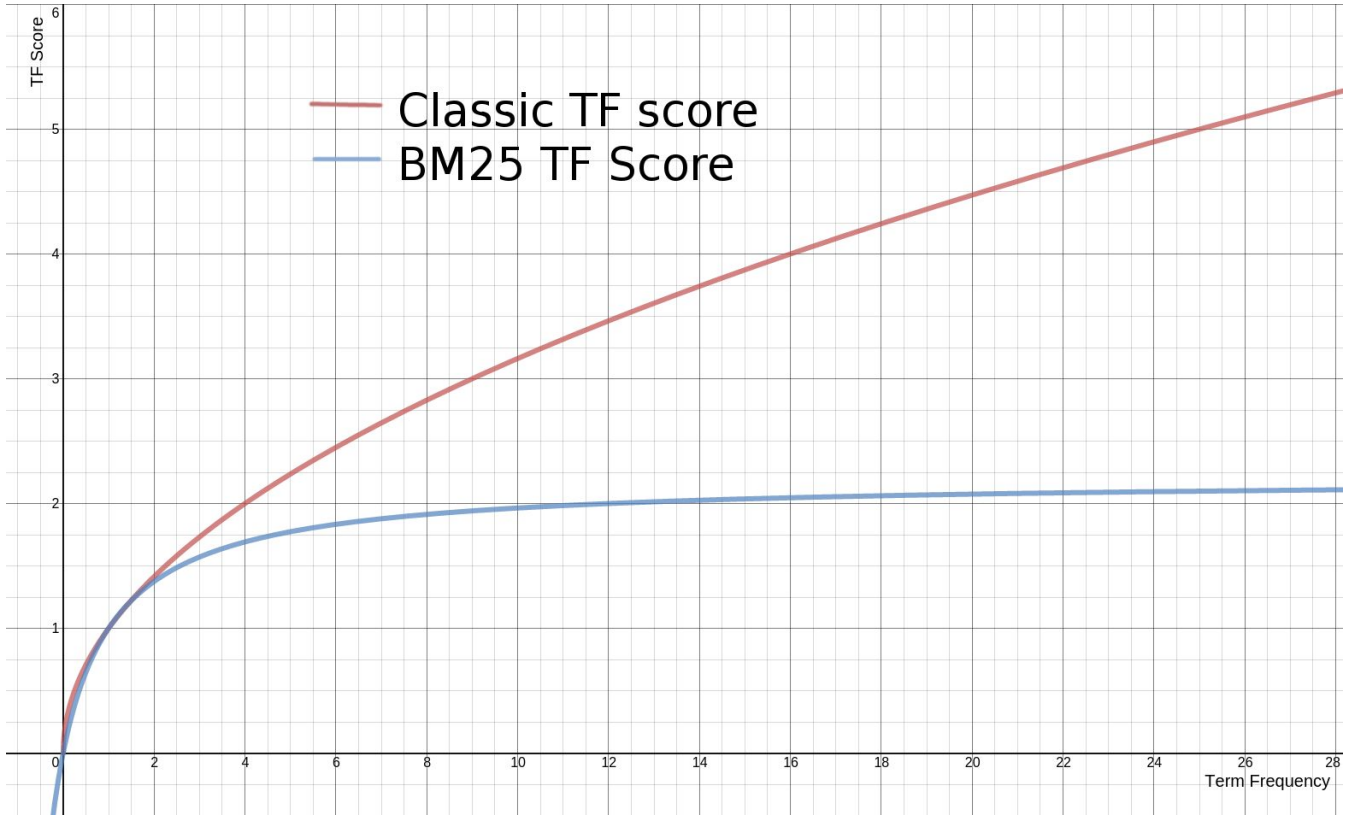
IDF Score
has very similar
behavior



BM25 Term Scorer - Term Frequency

TF Score approaches asymptotically (k+1)

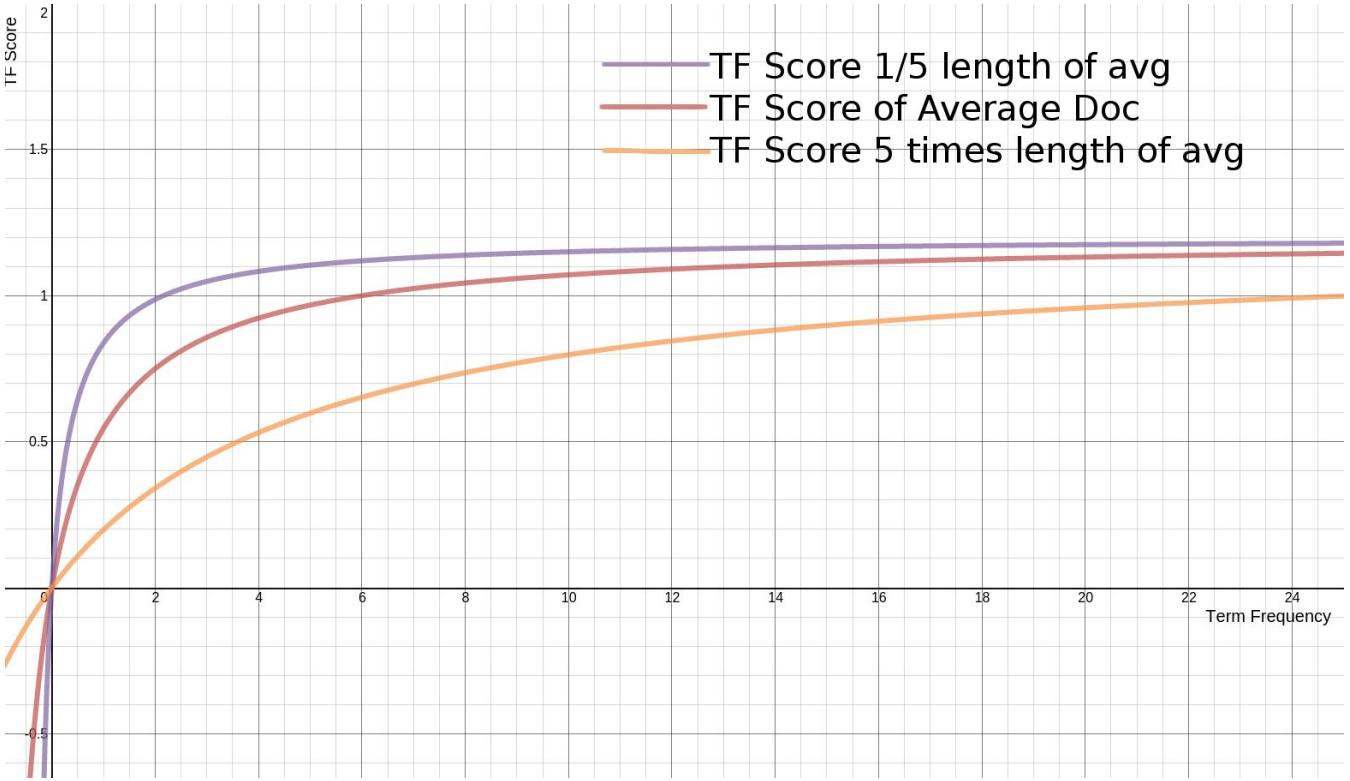
k=1.2 in this example



BM25 Term Scorer - Document Length

Document Length / Avg Document Length

affects how fast we saturate TF score



Interesting Term Retriever

Responsibility : retrieve from the document a queue of weighted interesting terms

- Analyze content / Term Vector
- Skip Tokens
- Score Tokens
- Build Queue of Top Scored terms

Params Used

- Analyzer
- Max Num Token Parsed
- Min Term Frequency
- Min/Max Document Frequency
- Max Query Terms
- Query Time Field Boost

SeaSe

More Like This Query Builder

Field1 : Term1	**Field2** : Term2	**Field1** : Term3	**Field1** : Term4	**Field3** : Term5
3.0	**4.0**	**4.5**	**4.8**	**7.5**

Q = *Field1*:Term1^3.0 *Field2*:Term2^4.0 *Field1*:Term3^4.5 *Field1*:Term4^4.8 *Field3*:Term5^7.5

Params Used

- Term Boost Enabled

SeaSe

More Like This Boost

Field Boost

- *field1*^5.0 *field2*^2.0 *field3*^1.5
- Affect Term Scorer
- Affect the interesting terms retrieved

N.B. a highly boosted field can dominate the interesting terms retrieval

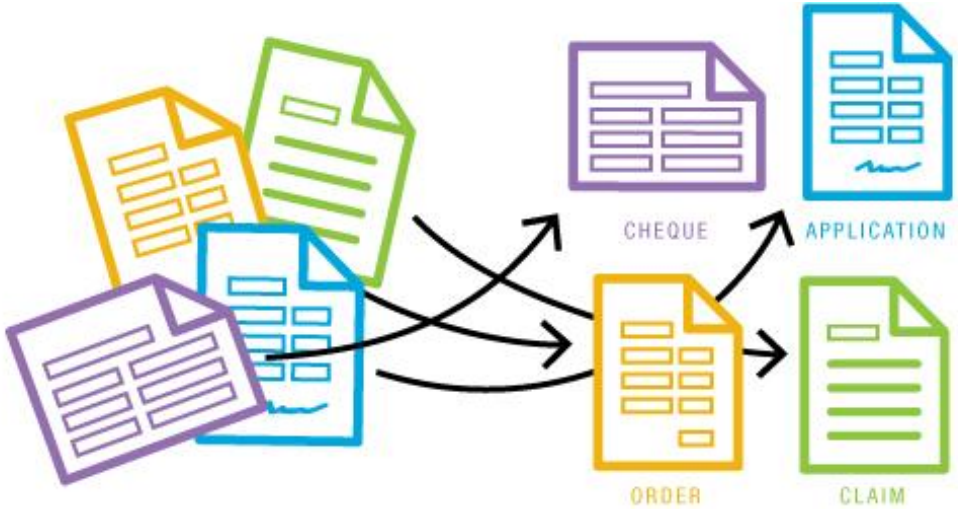
Term Boost

- on/off
- Affect each term weight in the MLT query
- It is the term score
  (it depends of the Term Scorer implementation chosen)

SeaSe

More Like This Usage - Lucene Classification

- Given a document D to classify
- K Nearest Neighbours Classifier
- Find Top K similar documents to D (MLT)
- Classes are extracted
- Class Frequency + Class ranking -> Class



SeaSe

More Like This Usage - Apache Solr

- **More Like This query parser**
  (can be concatenated with other queries)
- **More Like This search component**
  (can be assigned to a Request Handler)
- **More Like This handler**
  (handler with specific request parameters)

More Like This Demo - Movie Data Set

This data consists of the following fields:

- **id** - unique identifier for the movie
- **name** - Name of the movie
- **directed_by** - The person(s) who directed the making of the film
- **initial_release_date** - The earliest official initial film screening date in any country
- **genre** - The genre(s) that the movie belongs to

More Like This Demo - Tuned

- Enable/Disable Term Boost
- Min Term Frequency
- Min Document Frequency
- Field Boost
- Ad Hoc fields (ngram analysis)

SeaSe

Future Work

- Query Builder just use Terms and Term Score
- Term Positions ?
- Phrase Queries Boost
  (for terms close in position)
- Sentence boundaries
- Field centric vs Document centric
  (should high boosted fields kick out
  relevant terms from low boosted fields)

SeaSe

Future Work - More Like These

- Multiple documents in input
- Interesting terms across documents
- Useful for Content Based recommender engines

SeaSe

JIRA References

- LUCENE-7498 - Introducing BM25 Term Scorer
- LUCENE-7802 - Architectural Refactor

Questions ?

SeaSe

Arigato！

ありがとう！

SeaSe