M
~~G~~OOGLE
^

MOOC Document Search Engine

Problem

Quality learning material is hard & time consuming to find

- Top ranked sites give terse explanations, more for experts (eg: Wikipedia)
- Blogs contain a large noise to content ratio
- May share bad practices
- Might miss the crux of the concept

Solution

Search engine, with only quality learning documents

- Use course material from top ranking universities
- Stanford, Princeton, etc
- Extremely popular, like Stanford’s Intro to Machine Learning
- Don’t need the whole course, links you to documents relevant to search

Data Used

What we have:

- Course material from about 80 coursera courses
- Contains mp4, subtitles, powerpoint, pdf, txt, code

What we index:

- ✓ text, subtitles, pdfs, powerpoint (no ocr)
- ✗ code (hard to match)
- ✗ video & image (requires ocr)
- ✗ html (contains unrelated info)

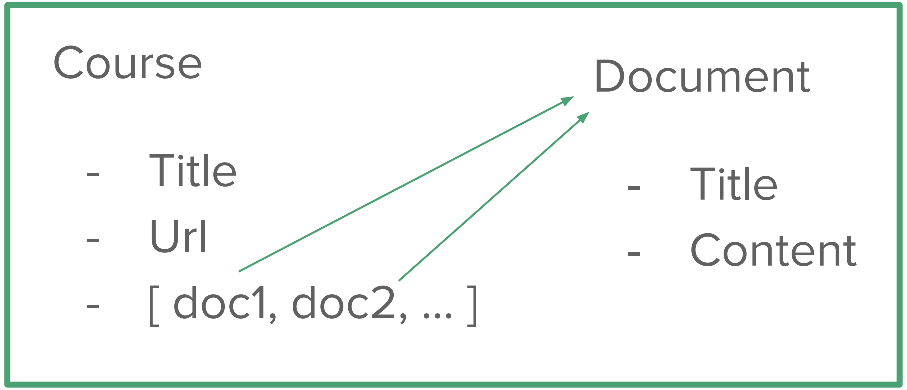
Structure & Indexing

Courses contain many documents

We index:

- Course Title
- Document Title
- Document Content

(BM25 is used for indexing)



Course

- Title
- Url
- [doc1, doc2, ...]

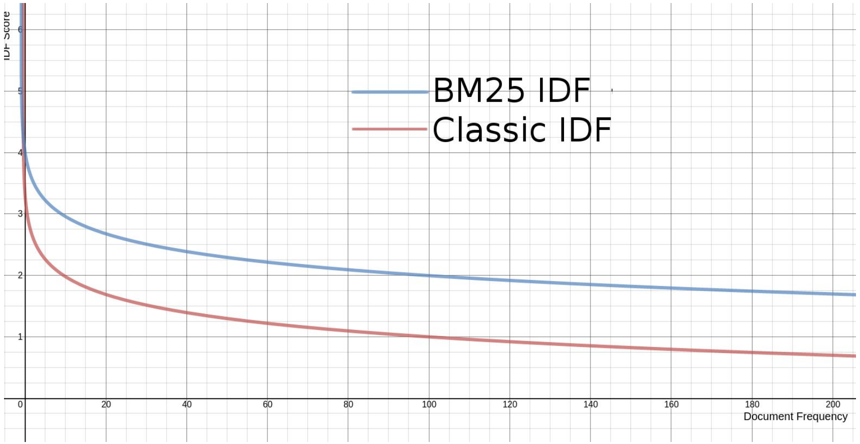
Document

- Title
- Content

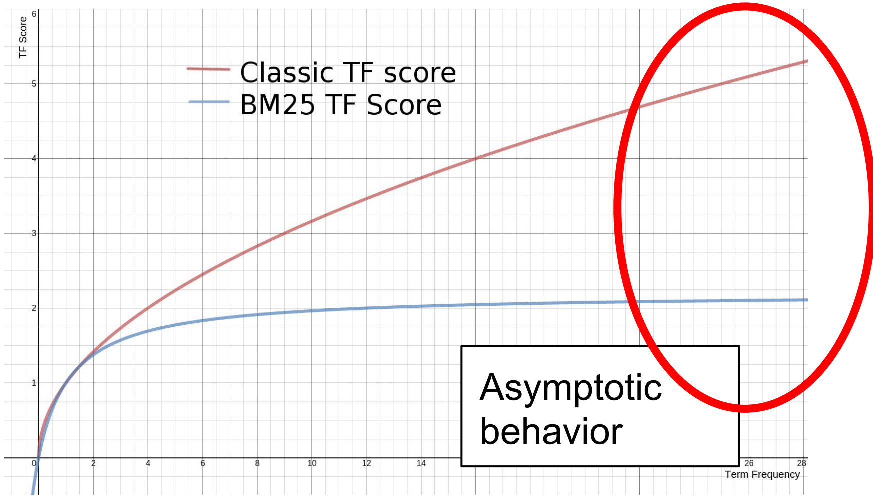
BM25 vs TD-IDF

Almost exactly the same, uses a structure like TF x IDF

IDF Score to Document Count



TF Score to Term Frequency



Asymptotic behavior

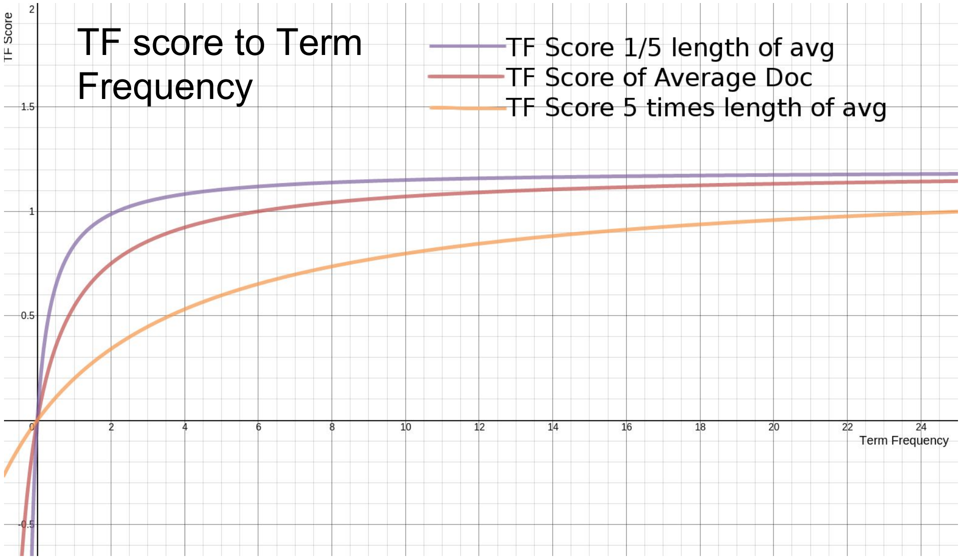
Super high frequency score is similar to high frequency

Adjusting b

- BM25 considers document length
- Longer documents need higher term frequency to reach the same score
- 0 <= b <= 1

Lowering b

- lower focus on document length, long & short documents are similar
- b = 0.25, for long supporting docs & short subtitles



Processing Text

Generating tokens for fields & queries

- Remove stopwords
- Ignorecase
- Minimum length of 2
- Append length 2 shingles (token sized bigrams)

Eg: “Why MPEG or JPEG” => “Why MPEG”, “MPEG or”, “or JPEG”

Computing Relevance

- Get score for all fields (course title, doc title, doc content)
- For each course, sum score for course title & documents
- Return highest scoring courses

Boosting

- Course title: 2 (whole course is related)
- Document title: 1.5 (whole document is related)
- Document content: 1

Live Demo

“Prisoner’s Dilemma”, “CNN last layer”, “Time Dilation”

Good matches:

- Computing related (bias in dataset)

Not always in the right order:

- Eg: “neural networks”, scores a networks module quite highly
- Boosting shingles might help