TF-IDF Representation

Dot Product, Norm, and Cosine Review?

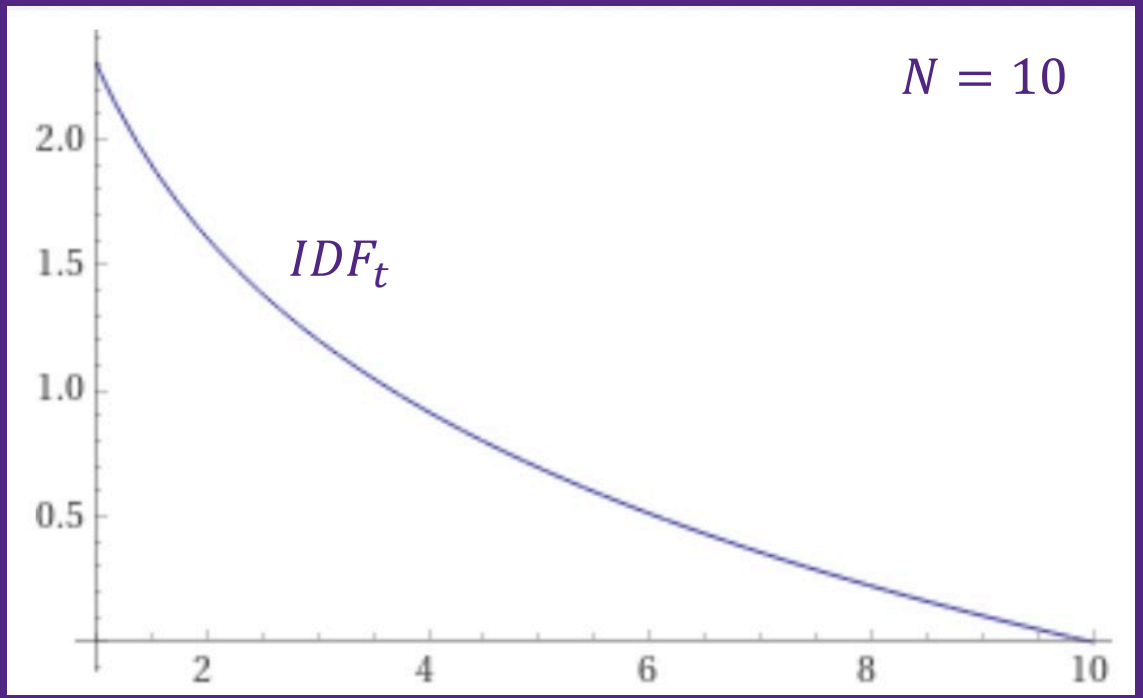
TF-IDF

- Term Frequency – Inverse Document Frequency
- Different vector representation for documents
- Replaces BoW counts to reflect term “importance” **relative to the corpus**.
  - Words that are *less* widespread in the corpus get *more* weight.
- *Many* variants

Inverse Document Frequency

- For term t,
  - Let N be number of documents
  - Let N_t be number of documents containing term t

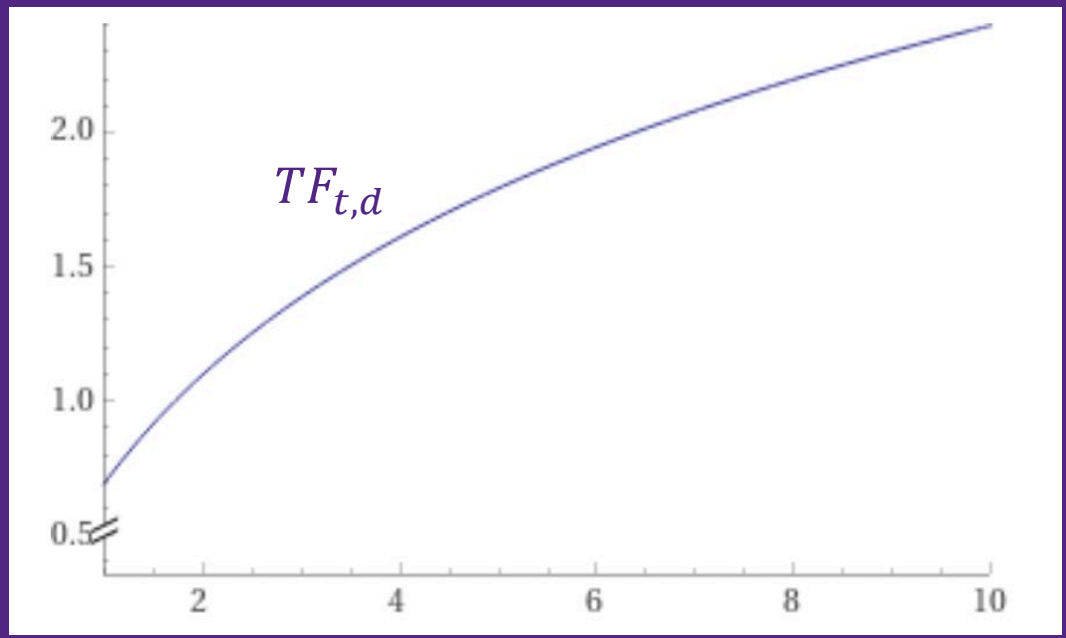
- $IDF_t = log\left(\frac{N}{N_t}\right)$



Term Frequency

- Higher frequency implies higher importance
  - Empirically, a diminishing return is helpful
- Let $f_{t,d}$ be the frequency of term *t* in document *d*

- $TF_{t,d} = \log(1 + f_{t,d})$



TF-IDF

- $TFIDF_{t,d} = log(1 + f_{t,d}) \times log\left(\frac{N}{N_t}\right)$

The TF-IDF "Vector model" Dense Representation

DocID	first	hurlyburly	in	thunder	witch	witchcraft	witches	witching	…
1	0.931	0	0.931	0.388	0.931	0.931	0	0	…
4	0	0	0	0	0	0	1.376	0	…
5	0	0	0	0.388	0	0.587	0	0	…
8	0	0	0	0	0	0	0	1.349	…
9	0	0.868	0	0	0	0	0	0	…
22	1.175	0.868	0.931	0.615	1.364	0	1.376	0	…
37	0.587	0	0.587	0.388	0.587	0.587	0	0	…
…	…	…	…	…	…	…	…	…	…

The TF-IDF "Vector model" Sparse Representation

DocID	Words
1	first:0.931, in:0.931, thunder:0.388, witch:0.931, witchcraft:0.931
4	witches:1.376
5	thunder:0.388, witchcraft:0.587
8	witching:1.349
9	hurlyburly:0.868
22	first:1.175, hurlyburly:0.868, in:0.931, thunder:0.615, witch:1.364, witches:1.376
37	first:0.587, in:0.587, thunder:0.388, witch:0.587, witchcraft:0.587
…	…

Similarity – TFIDF Cosine measure

DocID	Words	Similarity to {baseball:0.13, season:0.13, opener:1.24}
1	baseball:0.44, season:0.13, opener:1.24	0.972
2	baseball:0.44, season:0.33	0.141
6	season:0.13	0.101
7	baseball:0.44	0.101
10	baseball:0.44, season:0.25	0.138
35	baseball:0.44, season:0.20	0.134
…	…	

Generalized Similarity Measures

BM-25

General Similarity Measures

- "Tune" the ideas of bag-of-words, TFIDF, dot product/cosine to perform well for information retrieval

- Queries and documents are not treated the same. (Similarity from query-to-document not same as from document-to-query.)

- Default in Lucene

BM25

- Lucene default similarity function, related to TF-IDF
- $BM25_{d,q} = \sum_t q_t \frac{(k+1) f_{t,d}}{f_{t,d} + k\left((1-b)+b\frac{|d|}{avg\ doc\ length}\right)} IDF_t$

- The numbers k and b are "tuning parameters"
- q_t is frequency of term *t* in the query
- $|d|$ is document length; $avg\ doc\ length$ is average document length

BM25

- Lucene defaults: $k = 1.2, b = 0.75$

- $$BM25_{d,q} = \sum_t q_t \frac{2.2 f_{t,d}}{f_{t,d} + 0.3 + 0.9 \frac{|d|}{avg\ doc\ length}} IDF_t$$

- q_t is frequency of term *t* in the query
- $|d|$ is document length; $avg\ doc\ length$ is average document length
- Think:
  - What happens as IDF goes up?
  - What happens as $f_{t,d}$ goes up?
  - What happens as $|d|$ goes up?

Summary – Document Representation and Retrieval

- Term search and Boolean Queries
- Similarity-based Search
  - Dot product
  - Cosine
- Representations
  - Bag of Words (counts)
  - TFIDF
- Generalized similarity measures
  - BM25

Linear Algebra

- If you need to refresh the basics of linear algebra:
  - Multiplication of vectors and matrices
  - Transpose
  - Matrix Inverse

- http://www.cs.cmu.edu/~zkolter/course/15-884/linalg-review.pdf

- I'll review these with examples.