Information Retrieval Competition

Technion - Israel Institute of Technology

Methodology

1. Testing a few algorithms with default parameters : TF-IDF, BM-25
2. Get some statistics from : Queries, Index, Vocabulary
3. Discussion on how to exploit the meta data
4. Discussion on wether include Pseudo-Relevance Feedback
5. Infer all of the above for designing the algorithm and the parameters selection

1. The algorithm

BM-25

2. Statistics

Document Statistics

Metric	Value / Range	Interpretation / Use
Sample size	20,000	Used to estimate distributions
Total number of documents (N)	528,155	Required for IDF computation
Average document length est. (words)	457	Not very short documents
Document length min–max est.	0 – 56,654	Highly variable corpus
25th–50th–75th percentiles est.	155 – 314 – 604	Central mass of document sizes
Standard deviation est.	699.6	Indicates strong heterogeneity

2. Statistics

Table 1: Top 30 Terms (without stopwords) by Collection Frequency and Document Frequency

Rank	Term (CF)	CF	DF	Term (DF)	CF	DF
1	said	1038774	266638	said	1038774	266638
2	year	529423	208723	ft	242202	211865
3	mr	476692	119772	year	529423	208723
4	new	465557	198857	new	465557	198857
5	company	452521	145376	time	307947	162947
6	government	408431	146836	years	295176	153284
7	cent	387247	100136	government	408431	146836
8	market	363877	114741	company	452521	145376
9	state	343715	133089	report	262633	137040
10	pound	337836	86268	use	324174	135928
11	use	324174	135928	state	343715	133089
12	people	320602	130980	people	320602	130980
13	time	307947	162947	group	262928	124101
14	years	295176	153284	make	198957	122971
15	country	274720	115755	edition	125675	122869
16	group	262928	124101	month	202423	121667
17	report	262633	137040	text	127333	120697
18	say	259705	107339	mr	476692	119772
19	party	253645	78727	end	181043	119488
20	service	249802	95417	issue	247465	117613
21	issue	247465	117613	home	190943	116635
22	ft	242202	211865	bfn	116384	116358
23	bank	241230	65504	country	274720	115755
24	plan	240176	112637	market	363877	114741
25	dollar	218719	63999	long	187912	113100
26	trade	218247	85535	plan	240176	112637
27	president	217474	98806	say	259705	107339
28	price	216283	77391	national	214476	106887
29	national	214476	106887	week	176467	105600
30	public	212361	99574	international	174191	102493

2. Statistics

Query Statistics (249 Queries)

Metric	Value / Range	Interpretation / Use
Total number of queries	249	Used in training/evaluation
Average number of terms per query	2.64	Very short queries; favors PRF
Minimum – Maximum terms	1 – 4	Low expressiveness
Median number of terms	3	Most queries are concise
Top terms	disease, treatment, crime, etc.	Societal or health-related focus
Relevant Document Ratio est. (per query)		
Mean	0.171	On average, 17.1% are relevant
Standard deviation	0.192	High variance across queries
Min – Max	0.006 – 0.893	Very extreme min/max
25th – 50th – 75th percentile	0.024 – 0.121 – 0.221	Most have low to moderate coverage

3. The meta data

What are our Meta-Data ?

XML Tag	Description
<DOCNO>	Unique document identifier (e.g., FT942-6276)
<PROFILE>	Internal code or identifier (often unused)
<DATE>	Document date (e.g., 940531 → May 31, 1994)
<HEADLINE>	Full title or summary of the content
<BYLINE>	Author of the document
<XX>	Category separator (used as grouping tag)
<CO>	**Companies** mentioned (e.g., Dresdner Bank)
<CN>	**Countries** concerned (e.g., Germany, EC)
<IN>	**Industries**: sector of activity (e.g., Foreign Banking...)
<TP>	**Types**: editorial nature (e.g., Comment & Analysis, Market shares)
<PUB>	Name of the publication (e.g., The Financial Times)
<PAGE>	Page location (e.g., London Page III)
<TEXT>	Main content of the document

What/How to use ?

Field	Usage	What / How
<TEXT>	Main indexed field	The main content.
<HEADLINE>	Score boosting	We can boost in scoring. Can be useful for improving short query relevance.
<CO>	Fielded search	Allows targeted search on companies.
<CN>	Filtering / Reranking	Country-based filtering or reranking.
<IN>	Vertical search	Industry classification. Useful for filtering or reranking by domain.
<DATE>	Temporal filtering	Can be used to boost recent content, when filtering or reranking.

4. Pseudo Relevance Feedback

RM3

5. The final algorithm

BM25 + RM3

Parameters range for grid search and selection :

Component	Parameter	Values
BM25	k_1	$\{1.0,\ 1.2,\ 1.4\}$
	b	$\{0.5,\ 0.75,\ 1.0\}$
RM3 (Pseudo Relevance Feedback)	$fbDocs$	$\{5,\ 10,\ 20\}$
	$fbTerms$	$\{10,\ 20,\ 30\}$
	$fbOrigWeight$	$\{0.5,\ 0.8\}$

Parameter selection :

Best MAP reached = 0.2334

(Estimated on the first 50 queries and their relevance feedbacks)

Parameter	Value
k_1	1.0
b	0.5
$fbDocs$	5
$fbTerms$	30
$fbOrigWeight$	0.8