Black Boxes are Harmful

Sauparna Palchowdhury
NIST

http://kak.tx0.org

SEPTEMBER 2016

IR Experiment

- Test-collection
- Search system
- index — retrieve — evaluate loop

Why Evaluate?

- Measure effectiveness
- Establish baseline
- Render experiment reproducible

A Failed Experiment

Point of Failure

- Test-collection
- Retrieval system

Point of Failure

Test-collection

- Broken document corpus; checksum mismatch
- Wrong document-query-qrel triplet

Point of Failure

Configuration Pitfalls

- A counterintuitive interface
- One parameter, many meanings
- Switches are not mutually exclusive

Point of Failure

Word-of-mouth-heuristics

QUERY/MODEL	BM25	PL2	LM-Dirichlet	TF_IDF
TITLE	0.3–0.5	4—7	750—1000	0.3–0.5
DESCRIPTION	0.6—0.8	1—2	1500—2000	0.6—0.8

Point of Failure

A Bug

BM25 (A)	$$\frac{tf}{k_1\left(1-b+b\cdot\frac{dl}{avdl}\right)+tf}\cdot\log\left(\frac{N-n+0.5}{n+0.5}\right)\cdot\frac{\left(k_3+1\right)qtf}{k_3+qtf}$$
BM25 (B)	$$\frac{\left(k_1+1\right)tf}{k_1\left(1-b+b\cdot\frac{dl}{avdl}\right)+2\cdot tf}\cdot\log\left(\frac{N-n+0.5}{n+0.5}\right)\cdot\frac{\left(k_3+1\right)qtf}{k_3+qtf}$$

Point of Failure

Naming

Generics	Terrier	Lucene
Porter	PorterStemmer	PorterStemFilter
	WeakPorterStemmer	
Snowball	EnglishSnowballStemmer	SnowballStemFilter
S-Stemmer	SStemmer	**EnglishMinimal**StemFilter
Krovets		**K**StemFilter

Point of Failure

Naming

Terrier	Lucene
Tf	DefaultSimilarity
TF_IDF	BM25Similarity
LemurTF_IDF	TFIDFSimilarity
BM25	
DFRBM25	

Point of Failure

The Parser

- Tags/parts to include/exclude
- Stop-word removal
- Stemmer
- Curate the vocabulary

Recheck everything; to what length and end?

Alternative; Lucene

- LTR; mod of Lucene 5.4.0
- Not another blackbox
- Augment documentation

A Single Point of Reference

- TFxIDF Repository
- TXT; system with ‘correct’ implementations
- TRECBOX; facility to repeat experiments
- Evaluation table

TFxIDF Repository

SMART'S TERM-WEIGHTING TRIPLE NOTATION												
$tf(f_{ik})$				$df(N,n_k)$				$g(G,D_i)$				
	b	1	Binary weight	x	**n**	1	Multiplier of 1, disregards the collections frequency	x	**n**	1	1, disregards length normalization factor
t	**n**	f_{ik}	raw term frequency	f		$\log\left(\frac{N}{n_k}\right)$	inverse collection frequency		**c**	$\sqrt{\sum_{k=1}^{t} w_{ik}^2}$	cosine normalization
	a	$0.5+0.5\cdot\frac{f_{ik}}{\max(f_{ik})}$	augmented normalized term frequency (normalized to be in [0.5, 1])		**t**	$\log\left(\frac{N+1}{n_k}\right)$	inverse collection frequency		**u**	$1-s+s\cdot\frac{u_i}{avgu}$	pivoted unique normalization
	l	$1+\log(f_{ik})$	log	p		$\log\left(\frac{N-n_k}{n_k}\right)$	probabilistic inverse collection frequency		**b**	$1-s+s\cdot\frac{b_i}{avgb}$	pivoted byte size normalization
	L	$\frac{1+\log(f_{ik})}{1+\log(avg(f_{ik}))}$	average term frequency based normalization								
	d	$1+\log(1+\log(f_{ik}))$	double logarithm								

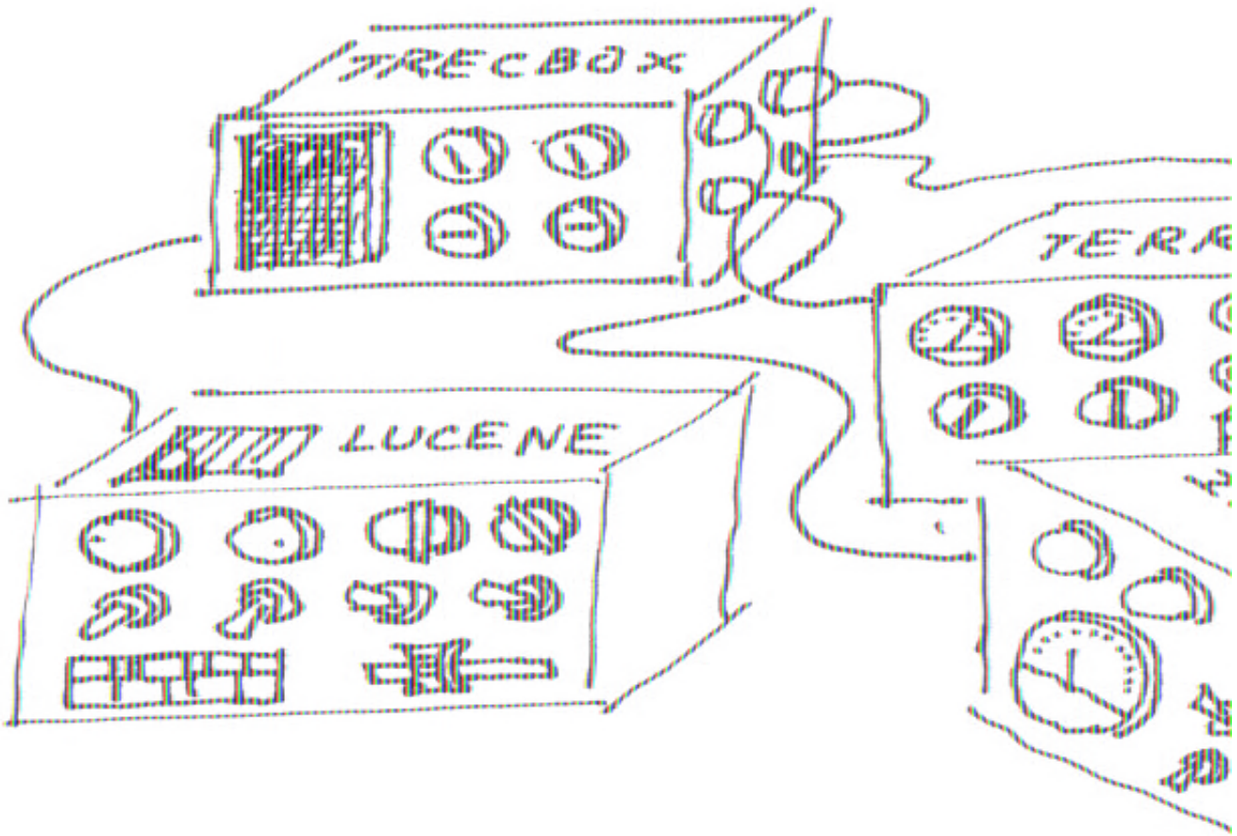
TFxIDF Repository

w	Scaling	TF	DF	QTF	Correction factor	Parameters
$BM0$		1				
$BM1$	s_3	1	$w^{(1)}$	$\frac{qtf}{k_3+qtf}$	$k_2 \cdot nq \cdot \frac{avdl-dl}{avdl+dl}$	
$BM15$	$s_1 s_3$	$\frac{tf}{k_1+tf}$	$w^{(1)}$	$\frac{qtf}{k_3+qtf}$	$k_2 \cdot nq \cdot \frac{avdl-dl}{avdl+dl}$	$s_i = \max(k_i,1)$ or 1 if $k_2 = 0$
$BM11$	$s_1 s_3$	$\frac{tf}{k_1 \cdot \frac{dl}{avdl}+tf}$	$w^{(1)}$	$\frac{qtf}{k_3+qtf}$	$k_2 \cdot nq \cdot \frac{avdl-dl}{avdl+dl}$	$s_i = \max(k_i,1)$ or 1 if $k_2 = 0$
$BM25$	$s_1 s_3$	$\frac{tf^c}{K+tf^c}$	$w^{(1)}$	$\frac{qtf}{k_3+qtf}$	$k_2 \cdot nq \cdot \frac{avdl-dl}{avdl+dl}$	$s_i = k_i+1$, $c = 1+mK$, $m \geq 0$ $K = k_1\left((1-b)+b \cdot \frac{dl}{avdl}\right)$
$BM25(k_1,k_2,k_3,b)$ The general form as a function of k_1, k_2, k_3, b and $m=0$.	$w = (k_1+1)\cdot(k_3+1)\cdot \frac{tf}{k_1\left((1-b)+b\cdot\frac{dl}{avdl}\right)+tf} \cdot \log\left(\frac{N-n+0.5}{n+0.5}\right) \cdot \frac{qtf}{k_3+qtf} + k_2 \cdot nq \cdot \frac{avdl-dl}{avdl+dl}$					
$BM25(k_1,0,k_3,b)$ The form, rearranged, after six years of trial-and-error from TREC3 to TREC8 (1995-2000)	$w = \frac{(k_1+1)\cdot tf}{k_1\left((1-b)+b\cdot\frac{dl}{avdl}\right)+tf} \cdot \log\left(\frac{N-n+0.5}{n+0.5}\right) \cdot \frac{(k_3+1)\cdot qtf}{k_3+qtf}$					

TFxIDF Repository

BMxx CONSTANTS											
	s_1	s_2	s_3	k_1	k_2	k_3	b	k_4	k_5	k_6	m
TREC 1											
TREC 2	$s_i = \max(k_i, 1)$ or 1 if $k_2 = 0$			0.0–0.3	∞						
TREC 3	$s_i = k_i + 1$			2	0	8, ∞	0.75				0
TREC 4	$s_i = k_i + 1$			1–2	0	8	0.6–0.75				0
TREC 5	$s_i = k_i + 1$			1–2	0	8, 1000	0.6–0.75				0
TREC 6	$s_i = k_i + 1$			1.2	0	0–1000	0.75	-0.7 or 0	0–4	4–∞	0
TREC 7	$s_i = k_i + 1$			1.2, 2	0	0–1000	0.75, 0.8	-0.7 or 0	0–4	4–∞	0
TREC 8	$s_i = k_i + 1$			1.2	0	7 or 1000	0.75				0

TRECBOX



TRECBOX

Settings.txt

```
EVAL /Users/rup/ir/trec_eval.9.0
LUCENE /Users/rup/ir/LTR
TERRIER /Users/rup/ir/TTR
LEMUR /Users/rup/ir/indri
EXP /Users/rup/ir/sub-collections
```

Experiment.txt

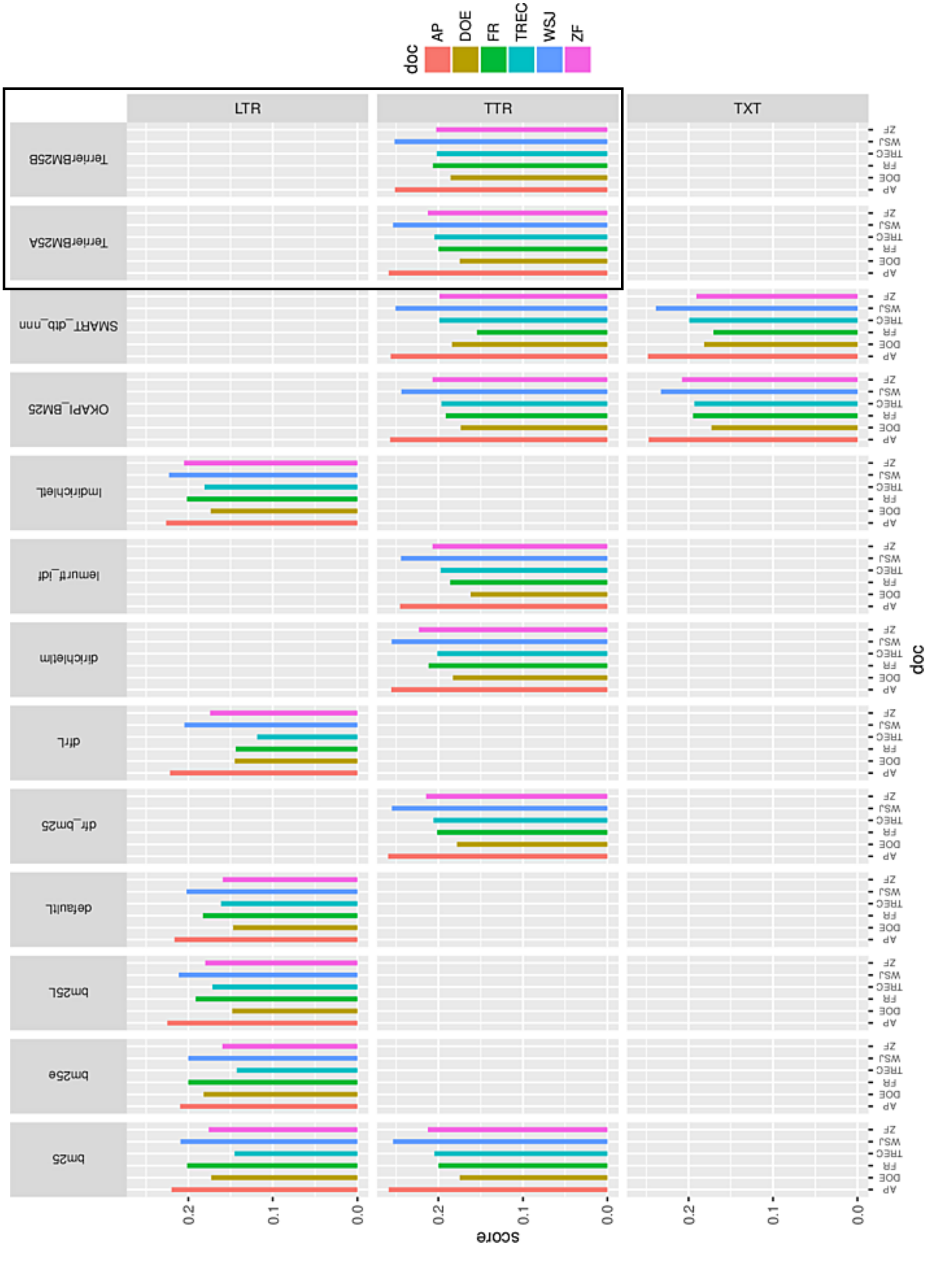
```
TESTCOL AP       AP     1-450:T:1-200.AP.196   1-200.AP.196.qrel
TESTCOL DOE      DOE    1-450:T:1-200.DOE.80   1-200.DOE.80.qrel
TESTCOL FR       FR     1-450:T:1-200.FR.111   1-200.FR.111.qrel
TESTCOL TREC     cd12   1-450:T:1-200.TREC.200 1-200.TREC.200.qrel
TESTCOL WSJ      WSJ    1-450:T:1-200.WSJ.200  1-200.WSJ.200.qrel
TESTCOL ZF       ZF     1-450:T:1-200.ZF.122   1-200.ZF.122.qrel

MODEL bm25 dirichletlm lemurtf_idf dfr_bm25
MODEL SMART_dtb_nnn OKAPI_BM25 TerrierBM25A TerrierBM25B
STEM x porter
STOP smart571
QEXP x
SYS terrier
```

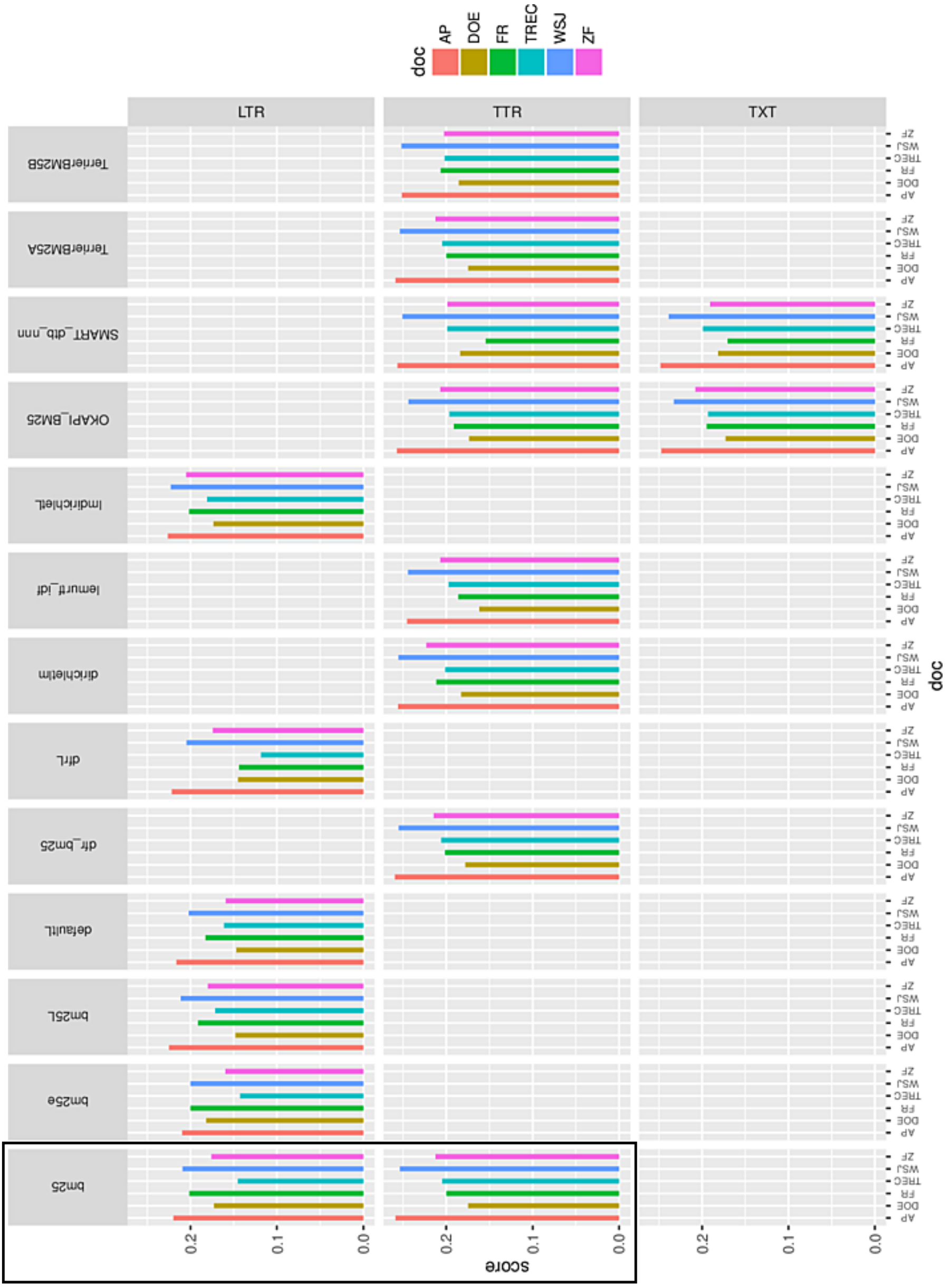
Evaluation Table

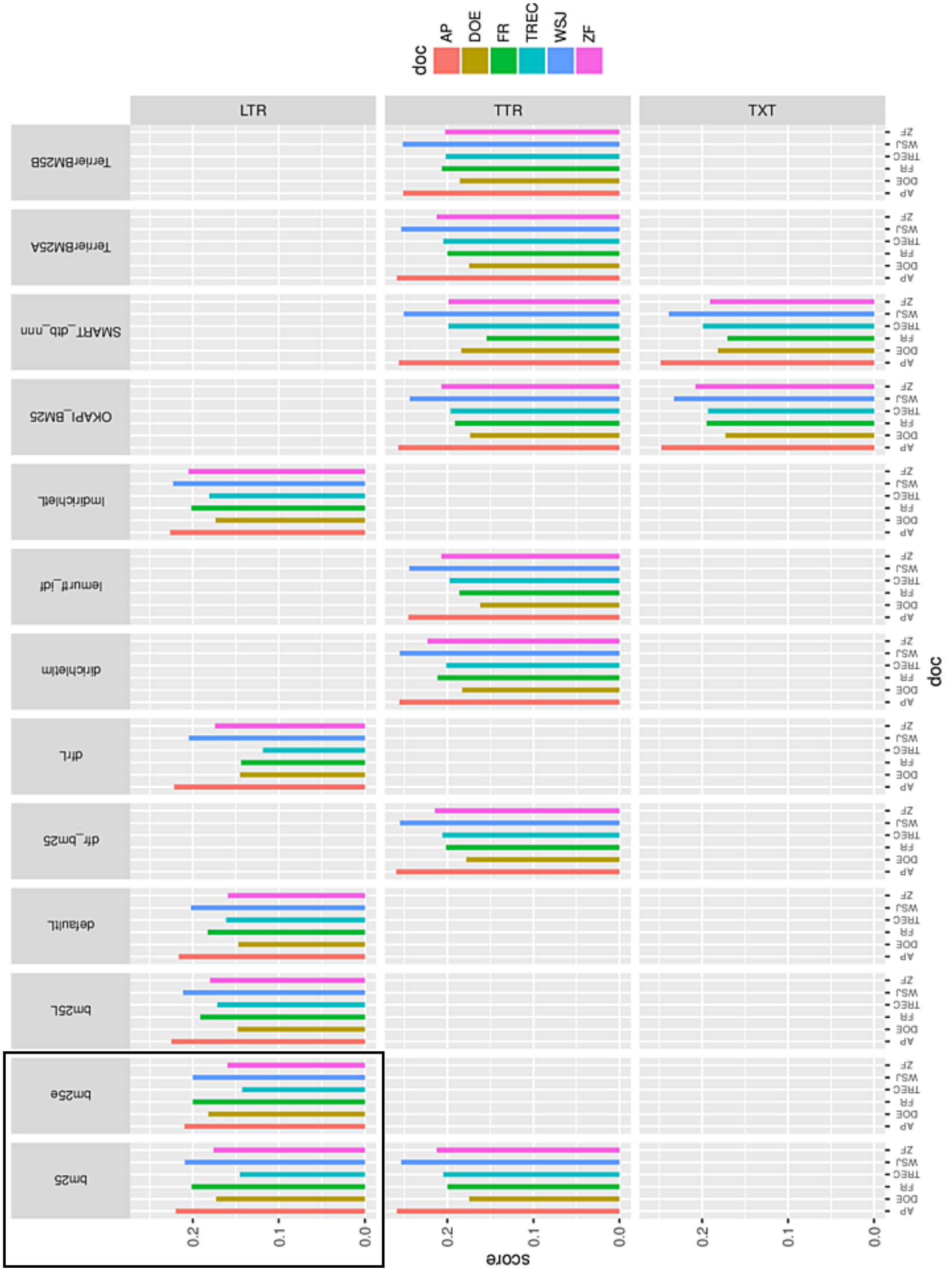
- System x Model x Doc

- Sanity-check

bm25
bm25e
bm25L
defaultL
dfr_bm25
dfrL
dirichletlm
lemurtf_idf
lmdirichletL
OKAPI_BM25
SMART_dtb_nnn
TerrierBM25A
TerrierBM25B
LTR
TTR
TXT
score
0.0
0.1
0.2
doc
AP
DOE
FR
TREC
WSJ
ZF


doc
AP
DOE
FR
TREC
WSJ
ZF
LTR
TTR
TXT
TerrierBM25B
TerrierBM25A
SMART_dtb_nnn
OKAPI_BM25
lmdirichlelL
lemurtf_idf
dirichletlm
dfrL
dfr_bm25
defaultL
bm25L
bm25e
bm25
score
0.0
0.1
0.2


doc
AP
DOE
FR
TREC
WSJ
ZF
LTR
TTR
TXT
TerrierBM25B
TerrierBM25A
SMART_dtb_nnn
OKAPI_BM25
lmdirichlelL
lemurtf_idf
dirichletlm
dfrL
dfr_bm25
defaultL
bm25L
bm25e
bm25
score
0.0
0.1
0.2


doc
AP
DOE
FR
TREC
WSJ
ZF
LTR
TTR
TXT
TerrierBM25B
TerrierBM25A
SMART_dtb_nnn
OKAPI_BM25
lmdirichlelL
lemurtf_idf
dirichletlm
dfrL
dfr_bm25
defaultL
bm25L
bm25e
bm25
score
0.0
0.1
0.2

bm25
bm25e
bm25L
defaultL
dfr_bm25
dfrL
dirichletlm
lemurtf_idf
lmdirichletL
OKAPI_BM25
SMART_dtb_nnn
TerrierBM25A
TerrierBM25B
LTR
TTR
TXT
score
0.0
0.1
0.2
AP
DOE
FR
TREC
WSJ
ZF
doc

Lucene's Similarity-Score Computation

‘Conceptual’ formula

$$score(q,d) = coord\text{-}factor(q,d) * query\text{-}boost(q) * \frac{V(q) * V(d)}{|V(q)|} * doc\text{-}len\text{-}norm(d) * doc\text{-}boost(d)$$

‘Practical’ scoring formula

$$score(q,d) = coord(q,d) * queryNorm(q) * \sum_{t\ in\ q} (tf(t\ in\ d) * idf(t)^2 * t.getBoost() * norm(t,d))$$

Generalized

$$score(Q,D) = f_c(Q,D) \cdot f_q(Q) \cdot \sum_{T_k \in Q \cap D} \left(tf(T_k) \cdot df(T_k) \cdot f_b(T_k) \cdot f_n(T_k, D) \right)$$

	w_i	w_j
BM25(k1,0,k3,b)	$\frac{(k_1+1)f_{ik}}{k_1\left((1-b)+b\cdot\frac{dl_i}{avdl}\right)+f_{ik}}\cdot\log\left(\frac{N-n_k+0.5}{n_k+0.5}\right)$	$\frac{(k_3+1)f_{jk}}{k_3+f_{jk}}$
	T * I	Q
dtb.nnn	$\frac{1+\log\left(1+\log\left(f_{ik}\right)\right)\cdot\log\left(\frac{N+1}{n_k}\right)}{1-s+s\cdot\frac{b_i}{avgb}}$	f_{jk}
	T * I / L	Q

$$score(D_i,D_j)=\sum_{T_k\in D_i\cap D_j} w_i\cdot w_j$$

$score(Q,D)$	=	$f_c(Q,D)$	•	$f_q(Q)$	•	$\sum_{T_k \in Q \cap D}$	(	$tf(T_k)$	•	$df(T_k)$	•	$f_b(T_k)$	•	$f_n(T_k,D)$	)
BM25(k1,0,k3,b)	=	1	.	1	.	$\sum_{T_k \in Q \cap D}$	(	T	.	I	.	Q	.	1	)
dtb.nnn	=	1	.	1	.	$\sum_{T_k \in Q \cap D}$	(	T	.	I	.	Q	.	L	)

	Description	Function names in code
	Coordination factor	coord()
	Query normalization factor	queryNorm()
	Term-frequency transformation	tf()
	Document-frequency transformation	idf()
	Query boost	in computeWeight()
	Document length normalization	lengthNorm()

In Conclusion

- Test-collection statistics
- Design documentation
- Consistent naming, well-defined notation
- Evaluation table
- Sharable experimental artifacts
- Implementations traceable to a source

Thank you.

Resources

- Experimental Methods for Information Retrieval (Donald Metzler and Oren Kurland, SIGIR 2012)

http://iew3.technion.ac.il/~kurland/sigir12-tutorial.pdf

- TFxIDF Repository (and other notes/tools)

http://kak.tx0.org/On-TFIDF