Modeling search engine performance measurement

Mark Sanderson

marksanderson.org

@IR_oldie

RMIT UNIVERSITY

Mounia Lalmas
@mounialalmas

This is a special edition as @IR_oldie and I are giving the keynotes. We started zillions of year ago our PhD together at the University of Glasgow under the supervision of Keith van Rijsbergen. Thank you @ictir2020. I feel honored 🙏

RMIT UNIVERSITY


HELP!

http://ir.dcs.gla.ac.uk/sigirphotos/sigir94/

Test collection

Documents 

Topics *“How do you get a Maryland fishing license?”*

QRELS 

RMIT UNIVERSITY

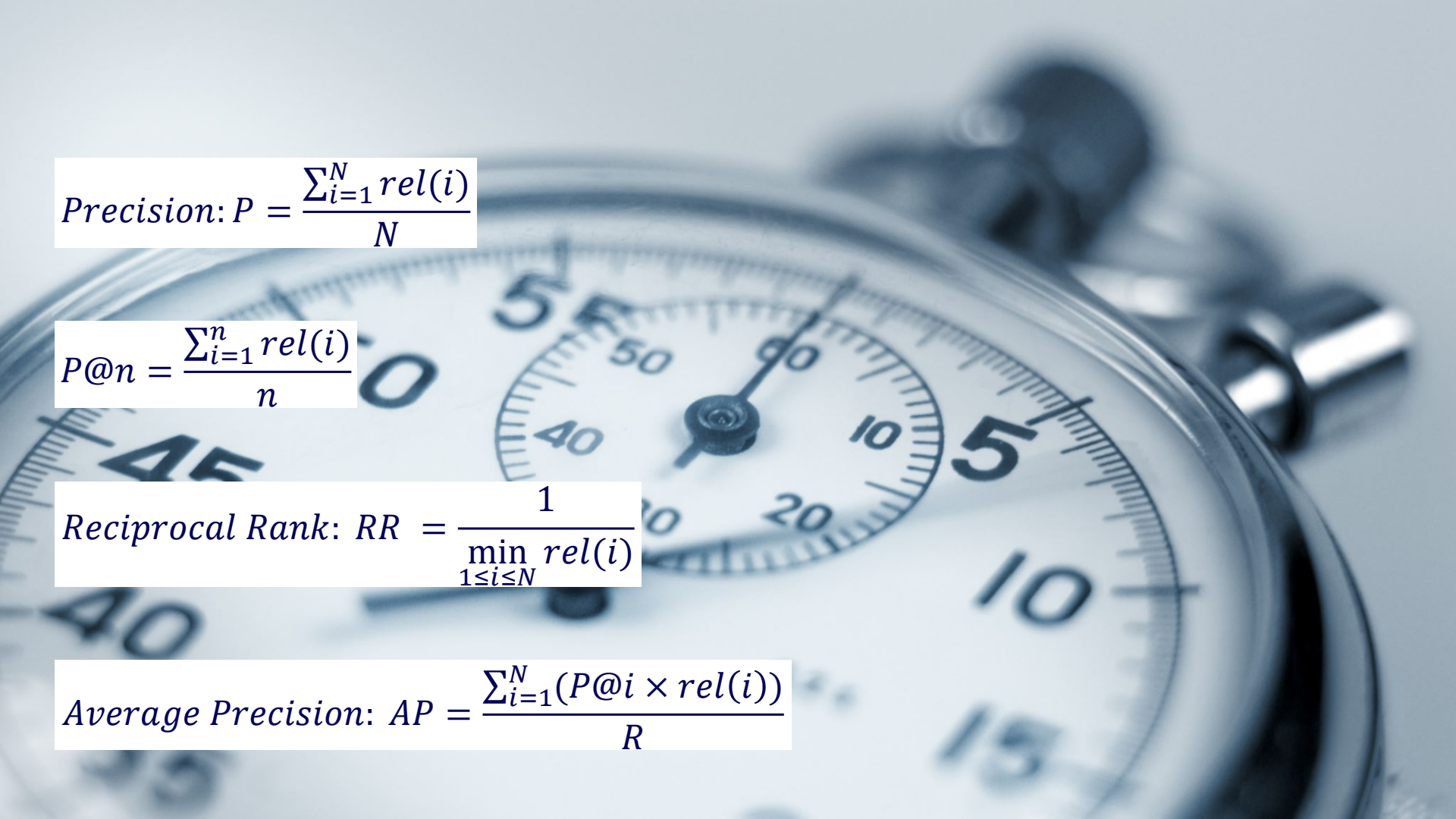
Test collection

https://www.designbuild-network.com/projects/national_stadium/

New method

https://www.designbuild-network.com/projects/national_stadium/

		NDCG@10	NDCG@20	P@10	P@20	MRR	MAP@100
GOV2	SDM$^{(\dagger)}$	0.4769	0.4751	0.5694	0.5469	0.7763	0.1802
	SDM+ILP (EmbYA)	0.4983$^{\dagger **}$	0.4951$^{\ddagger **}$	[illegible]	0.5779‡	0.7900**	0.1876$^{\ddagger \diamond **}$
	SDM+ILP (BM25)	0.5293‡	0.5171‡	[illegible]	[illegible]	[illegible]	**0.2009**$^{\ddagger \diamond *}$
	QSDM$^{(\diamond)}$	0.5127‡	0.5022‡	[illegible]	[illegible]	0.8174†	0.1919‡
	QSDM+ILP (EmbYA)	0.5197‡	0.5126‡	[illegible]	[illegible]	0.8258†	0.1891$^{\ddagger *}$
	QSDM+ILP (BM25)	0.5412$^{\ddagger \diamond}$	0.5245$^{\ddagger \diamond \diamond}$	[illegible]	[illegible]	0.8338†	0.2007$^{\ddagger \diamond \diamond *}$
	QSDM+EE$^{(*)}$	0.5339$^{\ddagger \diamond}$	0.5213$^{\ddagger \diamond \diamond}$	[illegible]	[illegible]	**0.8416**‡	0.1948‡
	QSDM+EE+ILP (EmbYA)	0.5329$^{\ddagger \diamond}$	0.5208$^{\ddagger \diamond}$	[illegible]	[illegible]	0.8044*	0.1947‡
	QSDM+EE+ILP (BM25)	**0.5442**$^{\ddagger \diamond \diamond}$	**0.5311**$^{\ddagger \diamond \diamond}$	**0.66**[illegible]	[illegible]	0.8407‡	0.1996$^{\ddagger \diamond \diamond **}$
CW09B	SDM$^{(\dagger)}$	0.2542	0.2462	[illegible]	[illegible]	0.5010	0.1053
	SDM+ILP (EmbYA)	0.2818$^{\ddagger *}$	0.2665$^{\dagger *}$	[illegible]	[illegible]	0.5579†	0.1092*
	SDM+ILP (BM25)	0.3115$^{\ddagger \diamond \diamond}$	0.2955$^{\ddagger \diamond \diamond}$	[illegible]	[illegible]	0.5902$^{\ddagger \diamond \diamond}$	0.1209$^{\ddagger \diamond \diamond *}$
	QSDM$^{(\diamond)}$	0.2735	0.2639†	[illegible].3938†	0.3467	[illegible]224	0.1094
	QSDM+ILP (EmbYA)	0.2853†	0.2691†	[illegible].3923	0.3485	[illegible]66†	0.1109
	QSDM+ILP (BM25)	0.3107$^{\ddagger \diamond \diamond}$	0.2959$^{\ddagger \diamond \diamond}$	[illegible].4333$^{\ddagger \diamond \diamond *}$	0.3774$^{\ddagger \diamond \diamond}$	[illegible]02$^{\ddagger \diamond \diamond}$	0.1190$^{\ddagger \diamond \diamond}$
	QSDM+EE$^{(*)}$	0.2985$^{\ddagger \diamond}$	0.2819‡+	0.4056‡	0.3610‡	0.57[illegible]9$^{\ddagger \diamond \diamond}$	0.1148$^{\ddagger \diamond}$
	QSDM+EE+ILP (EmbYA)	0.3042$^{\ddagger \diamond \diamond}$	0.2864$^{\ddagger \diamond \diamond}$	0.4174‡	0.3679$^{\ddagger \diamond}$	0.58[illegible]$^{\ddagger \diamond \diamond}$	0.1169$^{\ddagger \diamond}$
	QSDM+EE+ILP (BM25)	**0.3194**$^{\ddagger \diamond \diamond *}$	**0.3015**$^{\ddagger \diamond \diamond **}$	**0.4338**$^{\ddagger \diamond **}$	**0.3826**$^{\ddagger \diamond \diamond **}$	**0.6138**[illegible]$^{\diamond}$	**0.1210**$^{\ddagger \diamond \diamond **}$



$$Precision: P = \frac{\sum_{i=1}^{N} rel(i)}{N}$$

$$P@n = \frac{\sum_{i=1}^{n} rel(i)}{n}$$

$$Reciprocal\ Rank:\ RR = \frac{1}{\min\limits_{1 \le i \le N} rel(i)}$$

$$Average\ Precision:\ AP = \frac{\sum_{i=1}^{N} (P@i \times rel(i))}{R}$$

author:scholer author:moffat aut

scholar.google.com.au/scholar?hl=en&as_sdt=0%2C5&q=author%3Ascholer+author%3Amoffat+aut...

author:scholer author:moffat author:bailey author:thomas

Scholar 13 results (0.03 sec) YEAR

User variability and IR system evaluation [PDF] unimelb.edu.au

P Bailey, A Moffat, F Scholer, P Thomas - Proceedings of the 38th ..., 2015 - dl.acm.org

Test collection design eliminates sources of user variability to make statistical comparisons among information retrieval (IR) systems more affordable. Does this choice unnecessarily limit generalizability of the outcomes to real usage scenarios? We explore two aspects of ...

Cited by 47 Related articles All 7 versions

Incorporating user expectations and behavior into the measurement of search effectiveness [PDF] unimelb.edu.au

A Moffat, P Bailey, F Scholer, P Thomas - ACM Transactions on ..., 2017 - dl.acm.org

Information retrieval systems aim to help users satisfy information needs. We argue that the goal of the person using the system, and the pattern of behavior that they exhibit as they proceed to attain that goal, should be incorporated into the methods and techniques used to ...

Cited by 27 Related articles All 3 versions

UQV100: A test collection with query variability [PDF] unimelb.edu.au

P Bailey, A Moffat, F Scholer, P Thomas - Proceedings of the 39th ..., 2016 - dl.acm.org

We describe the UQV100 test collection, designed to incorporate variability from users. Information need? backstories? were written for 100 topics (or sub-topics) from the TREC 2013 and 2014 Web Tracks. Crowd workers were asked to read the backstories, and ...


Documents
Topics
QRELS
“How do you get a Maryland fishing license?”
https://www.designbuild-network.com/projects/national_stadium/

Urbano, J., Marrero, M., & Martín, D. (2013, July). On the measurement of test collection reliability. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 393-402)

Documents

Topics

QRELS

"How do you get a Maryland fishing license?"

https://abandonedplaygrounds.com/2019/04/13/gilroy-stadium-an-abandoned-sports-arena-of-gary/

A theory is a contemplative and rational type of abstract or generalizing thinking about a phenomenon, or the results of such thinking.

RMIT UNIVERSITY

Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

Overview

Other Evaluations

Publications

Information for Active Participants

Frequently Asked Questions

Tracks

Data

Past TREC Results

Contact Information

TREC-8 Adhoc Results

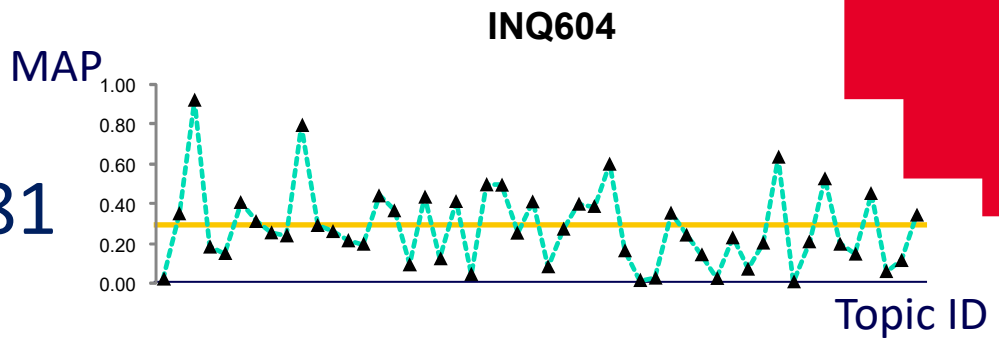
Publications home | NIST Special Publication 500-246 home page | Help

NIST

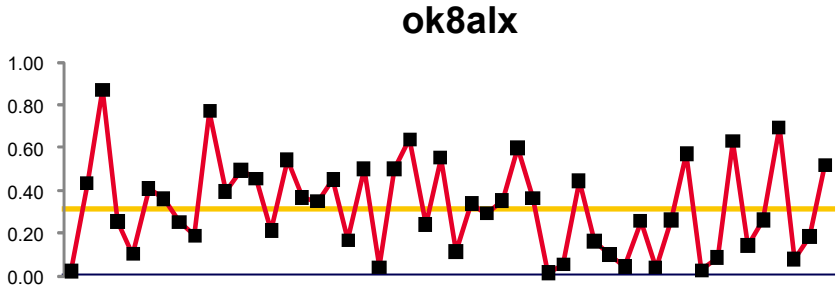
- 1
- 8manexT3D1N0
- AntHoc1
- CL99SD
- CL99SDopt1
- CL99SDopt2
- CL99XT
- CL99XTopt
- Drn8Nbn
- Drn8NbnR
- Drn8TFbn
- Drn8TFidf
- Flab8as
- Flab8at
- Flab8atd2
- Flab8atdn
- Flab8ax
- GE8ATD3
- GE8ATDN1
- GE8ATDN2
- GE8MTD2
- INQ601
- INQ602
- INQ603
- INQ604
- Mer8Adtd1
- Mer8Adtd2
- Mer8Adtd4
- Mer8Adtnd3
- MITSLStd
- MITSLStdn
- READWARE
- READWARE2

Noise

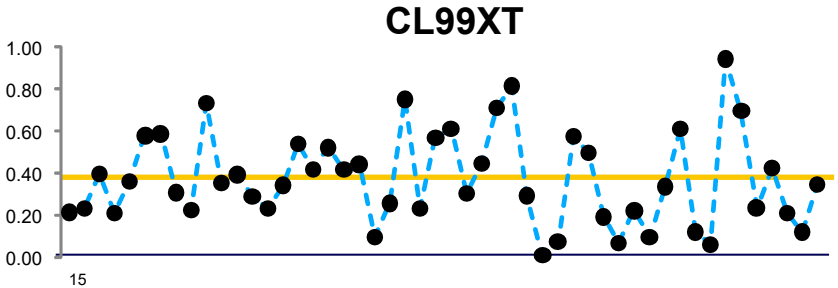
0.281



0.324



0.373



RMIT UNIVERSITY

Manage noise?

ANOVA: Data = Model + Error

Model: Linear mixture of factors

Sum of Squares (SS): Variance from mean

Effect size: Normalised measure of variance (η, ω)

RMIT UNIVERSITY

First go

Tague-Sutcliffe and Blustein, 1995

Factors	DF	Sum of Squares	Mean Square	F Value
Systems	41	15.42	0.38	34.44**
Topics	49	46.25	0.94	86.46**
Error	2009	21.93	0.01	
Total	2099	83.60		

**Probability of F < .0001.

RMIT UNIVERSITY

Topic*System interaction?

Tague-Sutcliffe and Blustein, 1995

Factors	DF	Sum of Squares	Mean Square	F Value
Systems	41	15.42	0.38	34.44**
Topics	49	46.25	0.94	86.46**
Error	2009	21.93	0.01	
Total	2099	83.60		

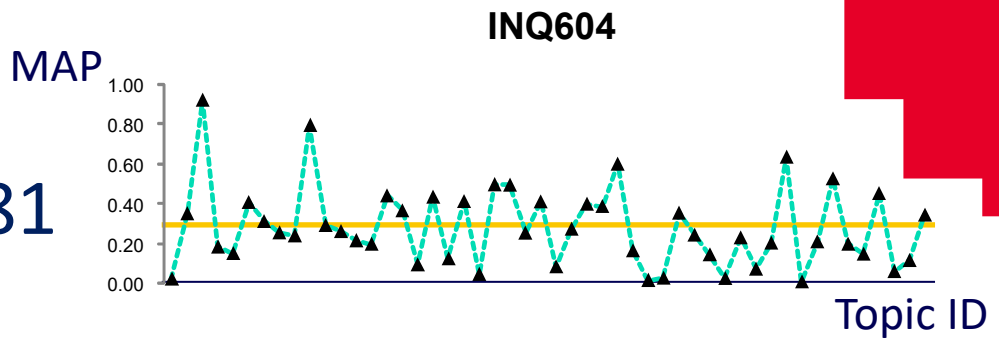
**Probability of F < .0001.

RMIT UNIVERSITY

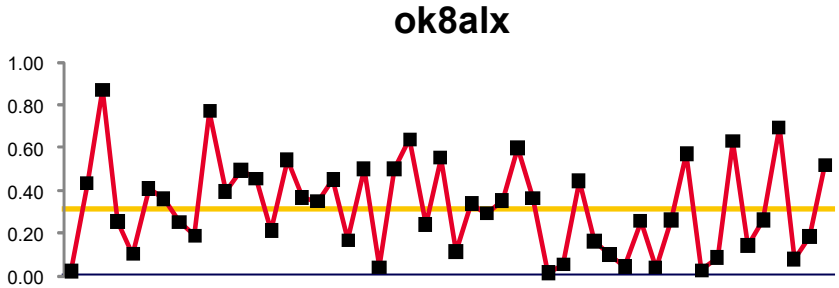
Interactions?

Replicates

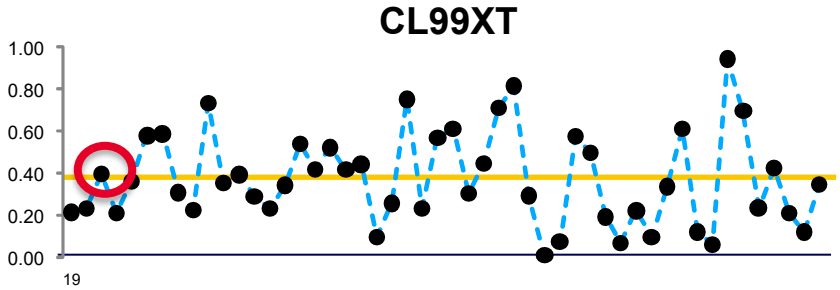
0.281



0.324

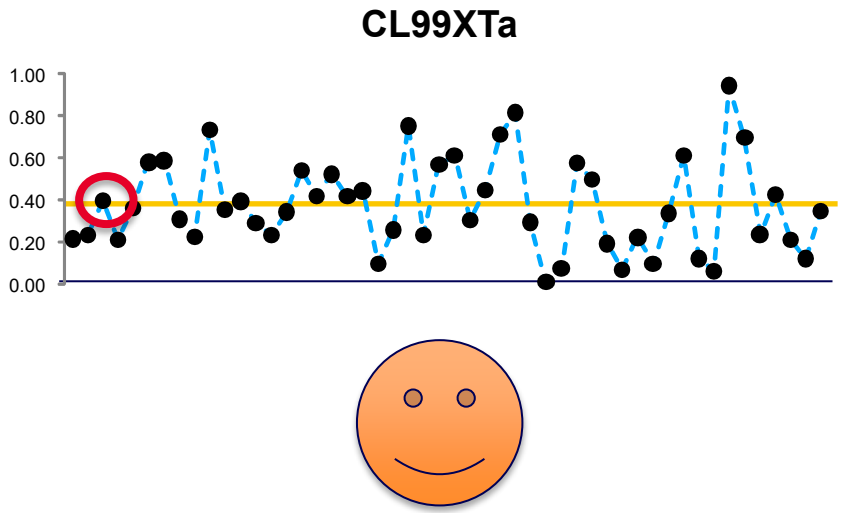


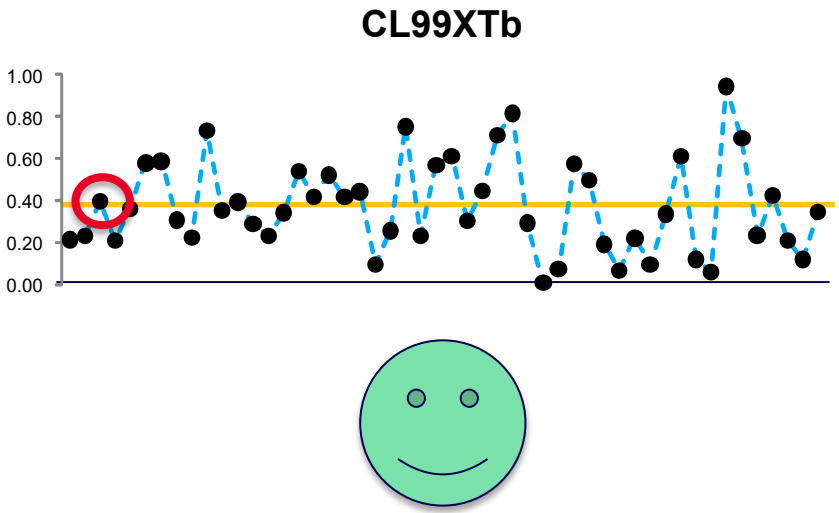
0.373



RMIT UNIVERSITY

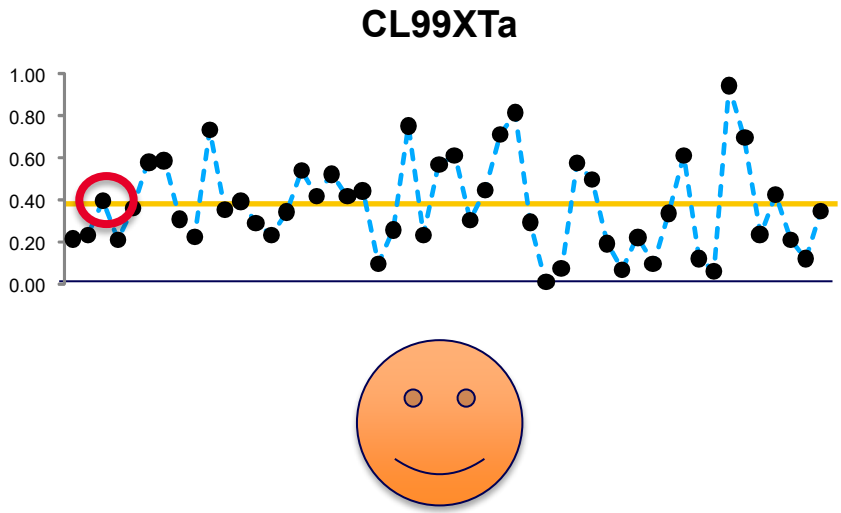
Replicates – QRELS

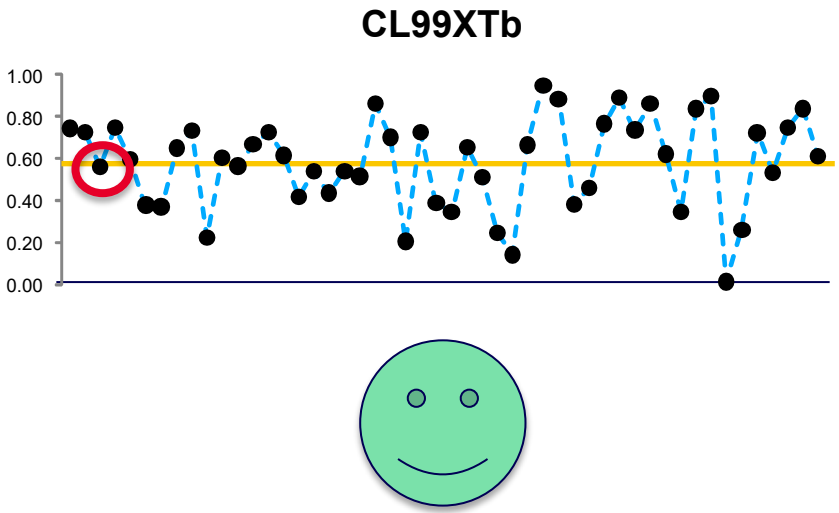




RMIT UNIVERSITY

Replicates – QRELS





RMIT UNIVERSITY

System*Topic Interaction

Bodoff and Li, 2007

Replicates:

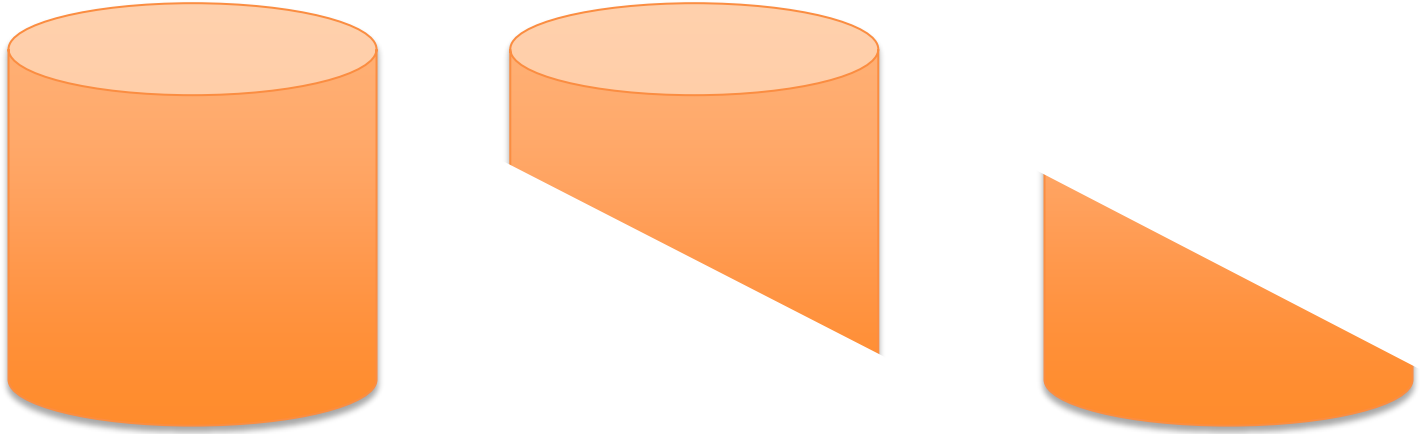
QRELS

Factors	MSE	Variance	% of total variance
System	0.7485	.00751	19.13%
Topic	1.1274	.01596	40.67%
Assessor main effect	0	0	0%
System*Topic	.0269	.01258	32.04%

RMIT UNIVERSITY

Partition collections

Shards



RMIT UNIVERSITY

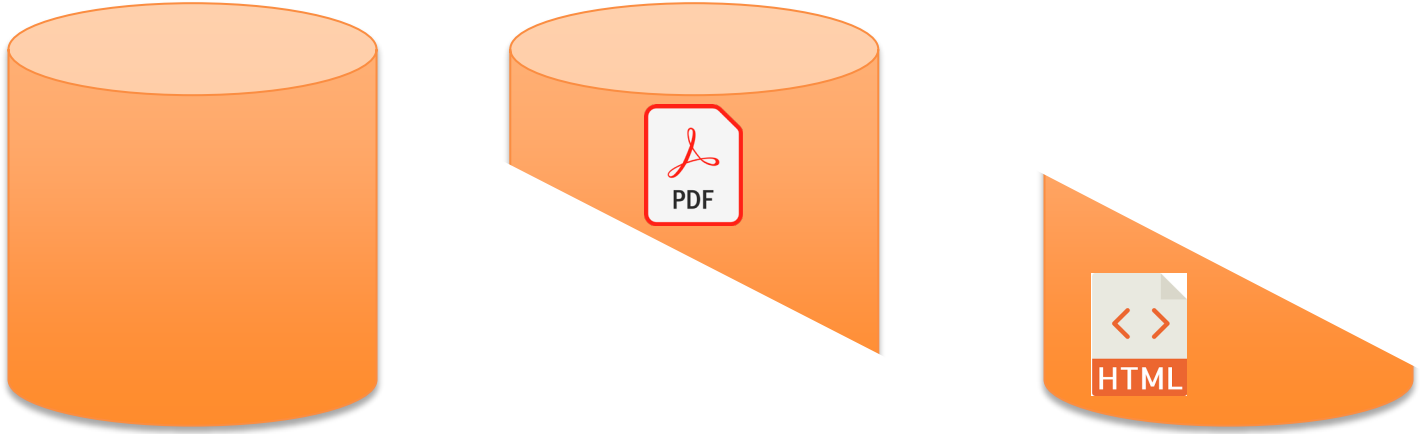
Collections

Type of documents

RMIT UNIVERSITY

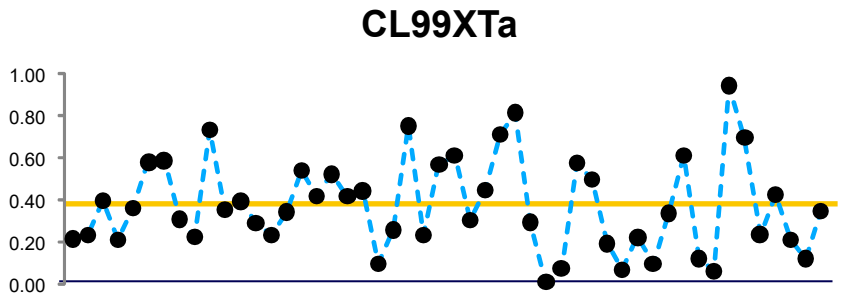
Study collections?

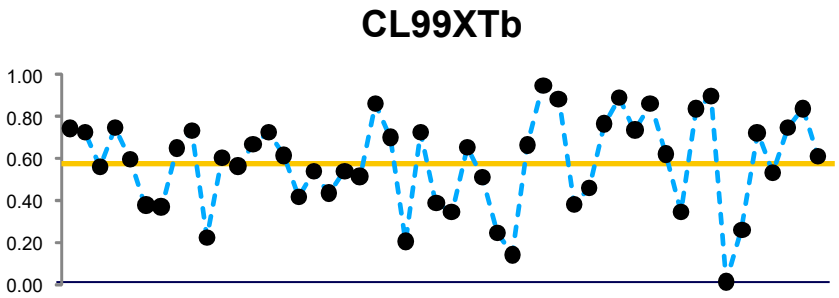
Partition them – shards

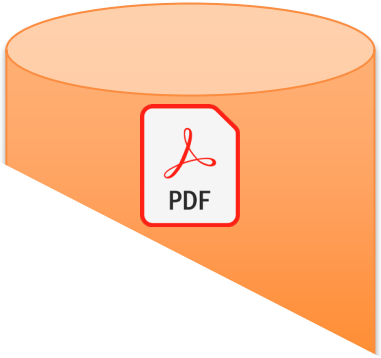


RMIT UNIVERSITY

Replicates – shard by type







RMIT UNIVERSITY

Type of documents?

Measure	T13 – GOV2_PDF_NOTPDF					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.67	0.42	0.01	0.18	0.37	0.24
R-prec	0.63	0.35	0.03	0.19	0.29	0.20
RBP	0.58	0.24	0.11	0.11	0.29	0.18
nDCG	0.64	0.60	<1e-2	0.43	0.45	0.17
ERR	0.42	0.12	0.14	0.07	0.10	0.13
Twist	0.60	0.41	0.03	0.25	0.30	0.20
P@5	0.48	0.18	0.08	0.08	0.24	0.14
P@10	0.58	0.24	0.10	0.11	0.30	0.17
P@20	0.65	0.30	0.14	0.12	0.31	0.21
P@50	0.73	0.32	0.28	0.15	0.32	0.30
P@100	0.75	0.29	0.38	0.16	0.25	0.37
nDCG@5	0.44	0.17	0.03	0.08	0.18	0.20
nDCG@10	0.53	0.23	0.04	0.11	0.25	0.21
nDCG@20	0.57	0.31	0.03	0.14	0.30	0.21
nDCG@50	0.62	0.40	0.01	0.22	0.38	0.26
nDCG@100	0.61	0.43	–	0.27	0.38	0.28

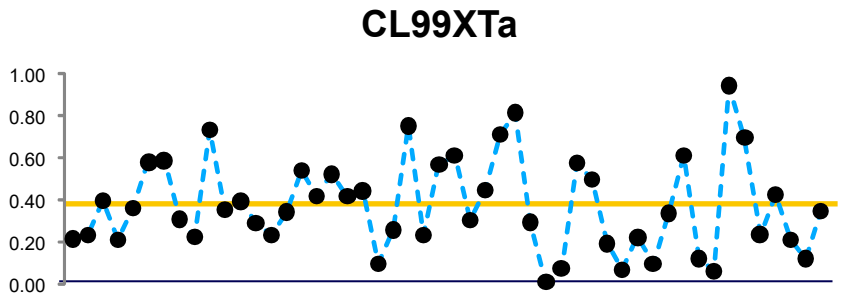
Type of documents?

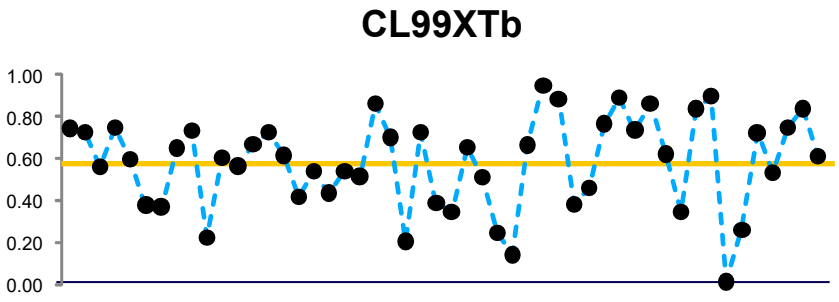
Measure	T14 – GOV2_PDF_NOTPDF					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.76	0.25	<1e-2	0.04	0.39	0.51
R-prec	0.70	0.16	<1e-2	0.03	0.33	0.49
RBP	0.71	0.10	0.05	0.02	0.29	0.49
nDCG	0.75	0.35	0.03	0.05	0.51	0.34
ERR	0.58	0.03	0.05	0.01	0.17	0.40
Twist	0.69	0.20	–	0.02	0.33	0.46
P@5	0.60	0.06	0.02	0.01	0.19	0.40
P@10	0.69	0.08	0.04	0.01	0.26	0.45
P@20	0.75	0.11	0.07	0.01	0.25	0.51
P@50	0.82	0.14	0.22	0.02	0.27	0.55
P@100	0.82	0.14	0.34	0.03	0.24	0.56
nDCG@5	0.60	0.06	–	0.02	0.19	0.43
nDCG@10	0.68	0.09	–	0.02	0.26	0.45
nDCG@20	0.71	0.12	–	0.02	0.25	0.44
nDCG@50	0.73	0.15	0.03	0.04	0.30	0.44
nDCG@100	0.69	0.19	0.10	0.05	0.34	0.36

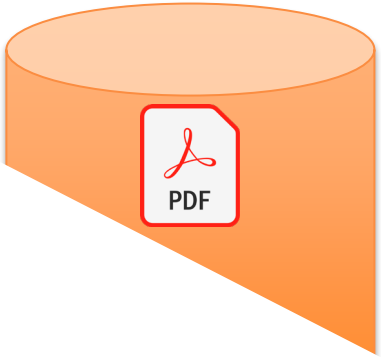
Type of documents?

Measure	T15 – GOV2_PDF_NOTPDF					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.75	0.13	<1e-2	0.01	0.49	0.46
R-prec	0.70	0.09	0.01	0.01	0.40	0.42
RBP	0.69	0.06	0.09	0.02	0.29	0.49
nDCG	0.78	0.17	0.03	0.04	0.58	0.44
ERR	0.54	0.03	0.14	0.01	0.15	0.43
Twist	0.68	0.10	0.04	0.02	0.42	0.40
P@5	0.56	0.03	0.05	0.02	0.19	0.38
P@10	0.67	0.05	0.10	0.02	0.23	0.48
P@20	0.76	0.06	0.20	0.03	0.33	0.56
P@50	0.82	0.09	0.44	0.03	0.31	0.57
P@100	0.86	0.10	0.61	0.02	0.32	0.64
nDCG@5	0.53	0.04	0.01	0.02	0.20	0.34
nDCG@10	0.59	0.05	<1e-2	0.02	0.24	0.34
nDCG@20	0.62	0.06	0.01	0.03	0.32	0.33
nDCG@50	0.66	0.11	0.06	0.04	0.42	0.37
nDCG@100	0.71	0.16	0.14	0.05	0.49	0.37

Replicates – shard by type

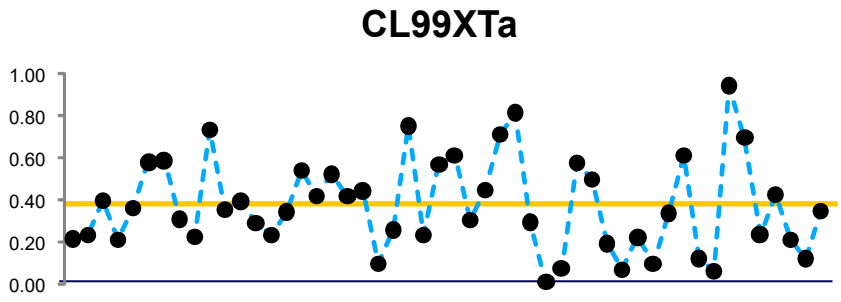


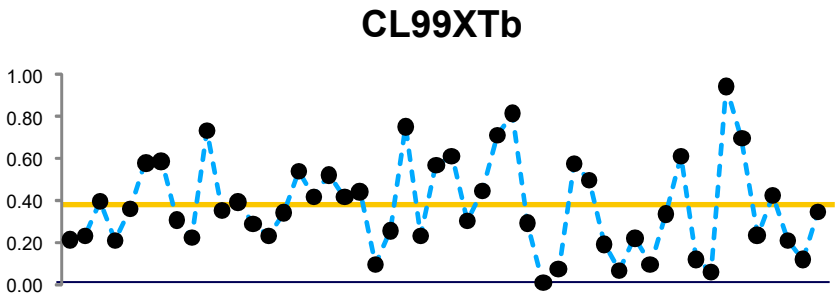


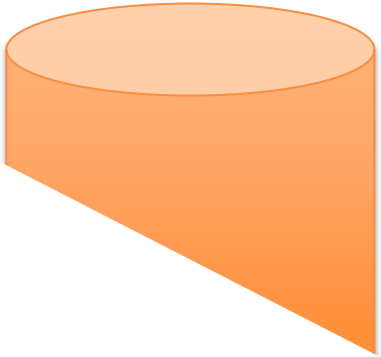


RMIT UNIVERSITY

Replicates – shard randomly

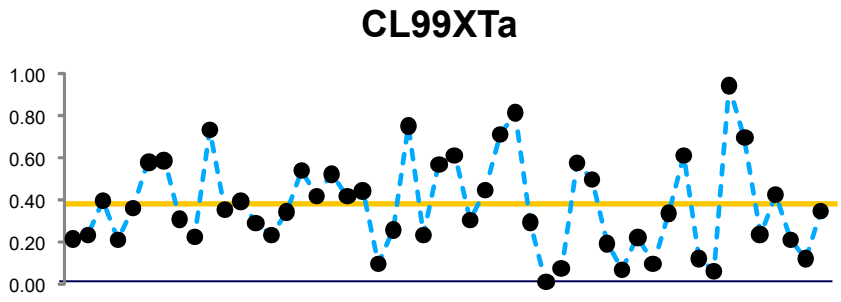


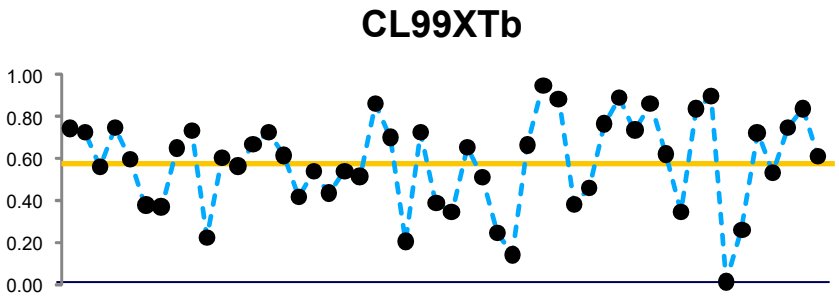


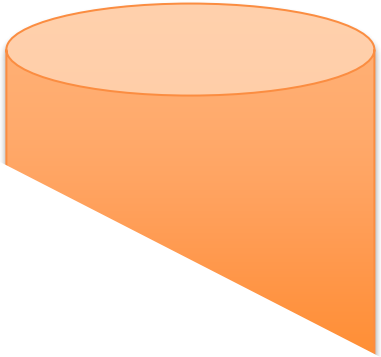


RMIT UNIVERSITY

Replicates – shard randomly





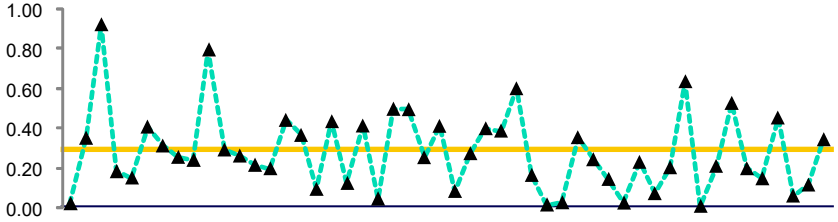


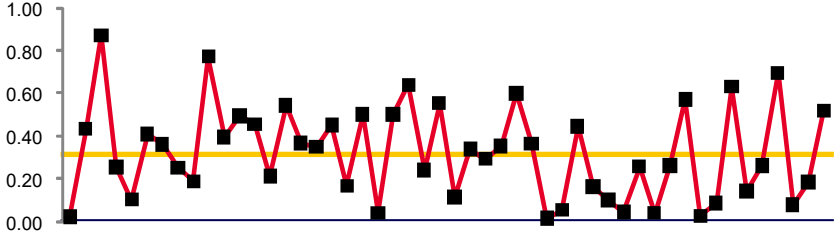
RMIT UNIVERSITY

Topics*Shard – random split

Measure	T09 – WT10g_RANDOM_10					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.42	0.12	0.03	<1e-2	0.30	0.61
R-prec	0.30	0.06	0.02	–	0.19	0.48
RBP	0.71	0.16	0.02	–	0.48	0.56
nDCG	0.67	0.20	0.02	–	0.45	0.71
ERR	0.50	0.07	0.02	<1e-2	0.23	0.62
Twist	0.37	0.08	0.02	–	0.21	0.52
P@5	0.59	0.11	0.01	–	0.37	0.47
P@10	0.74	0.16	0.02	–	0.52	0.51
P@20	0.84	0.21	0.03	–	0.64	0.59
P@50	0.90	0.26	0.04	–	0.73	0.68
P@100	0.93	0.29	0.05	–	0.78	0.72
nDCG@5	0.36	0.09	0.02	0.01	0.24	0.54
nDCG@10	0.40	0.12	0.02	<1e-2	0.29	0.56
nDCG@20	0.50	0.15	0.02	–	0.35	0.60
nDCG@50	0.61	0.18	0.02	–	0.42	0.67
nDCG@100	0.66	0.21	0.02	–	0.45	0.70

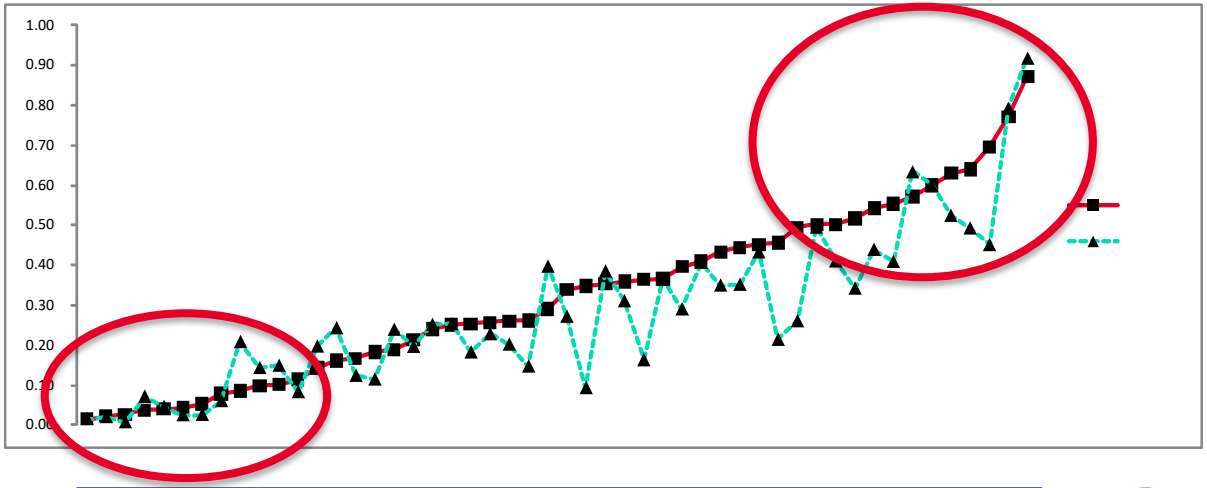
IR evaluation is noisy

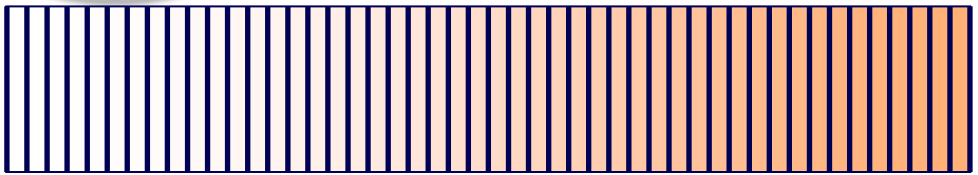




RMIT UNIVERSITY

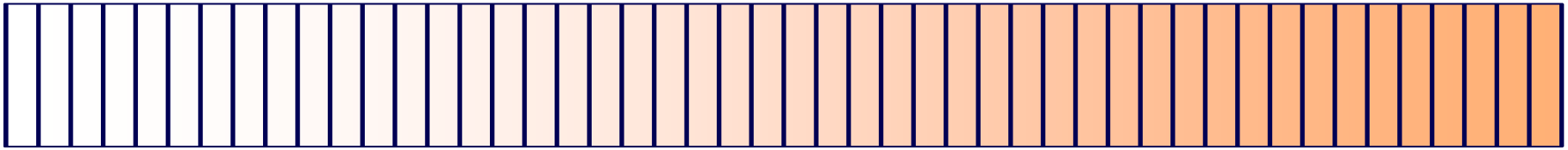
Hard vs Easy Topics?

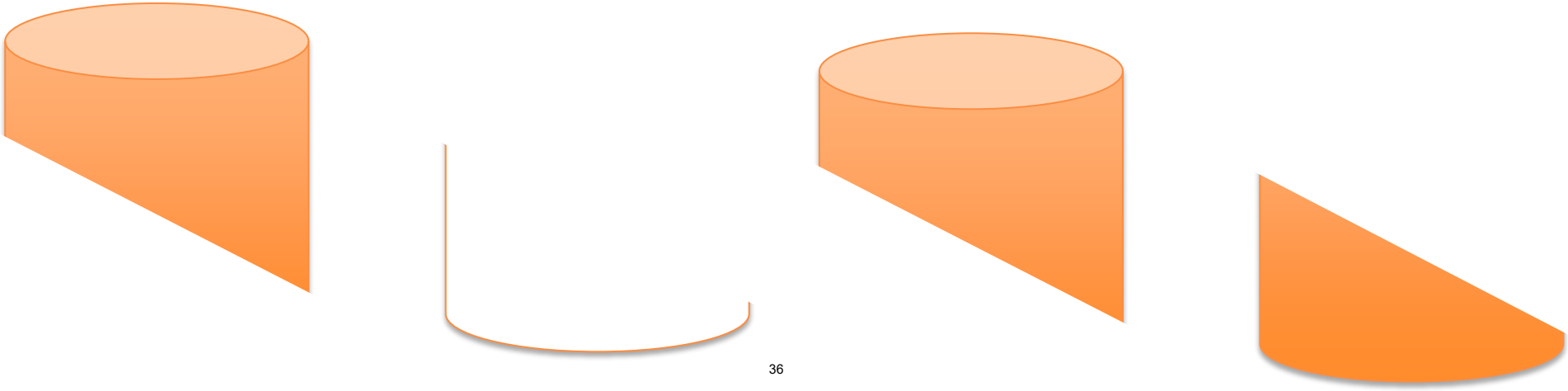




RMIT UNIVERSITY

Few vs Many QRELs



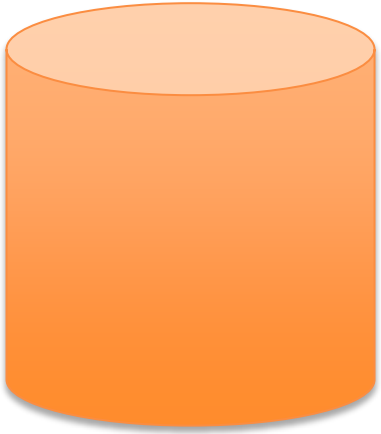


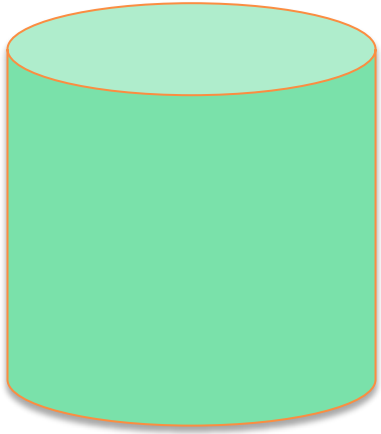
Topics*Shard – random split

Measure	T09 – WT10g_RANDOM_10					
	$\hat{\omega}^2_{\langle Tpc \rangle}$	$\hat{\omega}^2_{\langle Sys \rangle}$	$\hat{\omega}^2_{\langle Shr \rangle}$	$\hat{\omega}^2_{\langle Sys*Shr \rangle}$	$\hat{\omega}^2_{\langle Tpc*Sys \rangle}$	$\hat{\omega}^2_{\langle Tpc*Shr \rangle}$
AP	0.42	0.12	0.03	<1e-2	0.30	0.61
R-prec	0.30	0.06	0.02	–	0.19	0.48
RBP	0.71	0.16	0.02	–	0.48	0.56
nDCG	0.67	0.20	0.02	–	0.45	0.71
ERR	0.50	0.07	0.02	<1e-2	0.23	0.62
Twist	0.37	0.08	0.02	–	0.21	0.52
P@5	0.59	0.11	0.01	–	0.37	0.47
P@10	0.74	0.16	0.02	–	0.52	0.51
P@20	0.84	0.21	0.03	–	0.64	0.59
P@50	0.90	0.26	0.04	–	0.73	0.68
P@100	0.93	0.29	0.05	–	0.78	0.72
nDCG@5	0.36	0.09	0.02	0.01	0.24	0.54
nDCG@10	0.40	0.12	0.02	<1e-2	0.29	0.56
nDCG@20	0.50	0.15	0.02	–	0.35	0.60
nDCG@50	0.61	0.18	0.02	–	0.42	0.67
nDCG@100	0.66	0.21	0.02	–	0.45	0.70

TREC Robust track

Topics: hard on one collection, easy on another





RMIT UNIVERSITY

Topics? Queries!

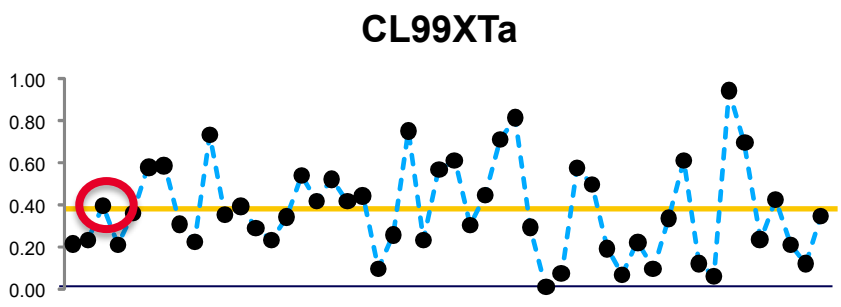
Moffat, Bailey, Scholer, Thomas (2017) – UQV100

Topic: "*How do you get a Maryland fishing license?*"

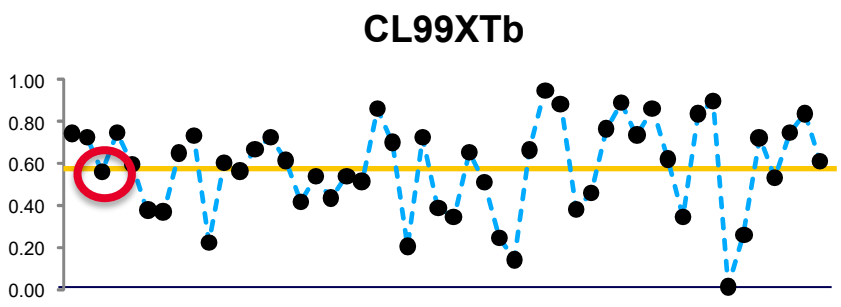
Queries: "*maryland fishing license*", "*how do i get a fishing license in maryland*", "*who can get a fishing license in maryland*", and "*is a fishing license needed in maryland*"

RMIT UNIVERSITY

Replicates – queries



"maryland fishing license"



"how do i get a fishing license in maryland"

RMIT UNIVERSITY

Queries vs Topics (2017)

Metric		η^2	SS	df	F
AP	query	0.53	152.26	4894	4.32
	system	0.23	39.24	149	36.55
	topic	0.15	23.26	178	15.90
NDCG	query	0.57	270.20	4894	5.17
	system	0.30	84.37	149	52.98
	topic	0.16	38.16	178	20.06
Q 1	query	0.56	142.03	4894	4.88
	system	0.20	28.54	149	32.20
	topic	0.18	24.18	178	22.84
RBP 0.85	query	0.51	328.77	4894	4.00
	system	0.21	84.09	149	33.62
	topic	0.13	47.53	178	15.90
INST	query	0.50	448.30	4894	3.90
	system	0.18	93.40	149	26.69
	topic	0.13	63.64	178	15.22

Queries vs Topics (2003)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
a) All Seven Runs					
Model	2449	694.5274	0.2836	95.24	0.0001
Error	12600	37.5185	0.0030		
Corrected total	15049	732.0459			
	R-Square	C.V.	Root MSE		AVP Mean
	0.9487	26.1391	0.0546		0.2088
Source	*DF*	Anova SS	Mean square	*F*-Value	Pr > *F*
TOPICID	49	418.8830	8.5486	2870.93	0.0001
QID (TOPICID)	2100	245.7583	0.1170	39.3	0.0001
RUNID	6	10.2069	1.7011	571.30	0.0001
TOPICID*RUNID	294	19.6792	0.0669	22.48	0.0001

RMIT UNIVERSITY

Topics

Topic*System

Queries

Shards

Topic*Shard

System*Shard

RMIT UNIVERSITY

A theory is a contemplative and rational type of abstract or generalizing thinking about a phenomenon, or the results of such thinking.

RMIT UNIVERSITY

System components

Ferro & Silvello (2016)

Source	SS	DF	MS	F	p
Topics′	820.99	49	16.75	3713.90	0.00
Stop list	9.89	4	2.47	548.06	0.00
Stemmer	4.16	4	1.04	230.76	0.00
Model	5.16	15	0.3443	76.32	0.00
Stop list*Stemmer	0.05	16	0.03	0.67	0.83
Stop list*Model	17.01	60	0.28	62.84	0.00
Stemmer*Model	0.07	60	0.001	0.26	1.00
Stop list*Stemmer*Model	0.09	240	0.00	0.08	1.00
Error	88.20	19551	0.005		
Total	945.63	19999			