evaluating information retrieval systems

kth

jussi karlgren

february 2014

jussi karlgren

gavagai & kth

language technology applied to information retrieval

text styles and variation in text use

interactive information retrieval

large scale text analysis

Gavagai

continuous evaluation is the most important vehicle for successful technology development information access is about making the user happy

but who is our user here?

three-way optimisation:

price-quality-timeliness

what is quality in an information system?

usefulness and effectiveness for task

appealing presentation

authority and trustworthiness and sourceability

relevance and truthfulness

reusability and cost

happiness, trust, and satisfaction!

we'll focus on relevance

examples of design questions

use stoplists?

lemmatisation?

thesauri?

query expansion?

user modelling?

plan inference?

positional modelling?

genre analysis?

the target concept of relevance

in everyday language:

a function of task, collection characteristics, user preferences and background, situation, tool, temporal constraints, and untold other factors

in information retrieval research:

a (binary) relation between query and document, disregarding everything contextual

	relevant	non-relevant
retrieved	true positives	false positives
not retrieved	false negatives	true negatives

$$\text{accuracy} = (tp+tn)/(tp+tn+fp+fn)$$

$$\text{precision} = tp/(tp+fp)$$

$$\text{recall} = tp/(tp+fn)$$ (täckning)

5 min exercise

retrieve and assess relevance of top ten

compare two queries and two search engines use gold standards / ground truth

lock down the notion of relevance

create test collections

define shared tasks

locking down the notion of relevance

TREC, US, 1992 -

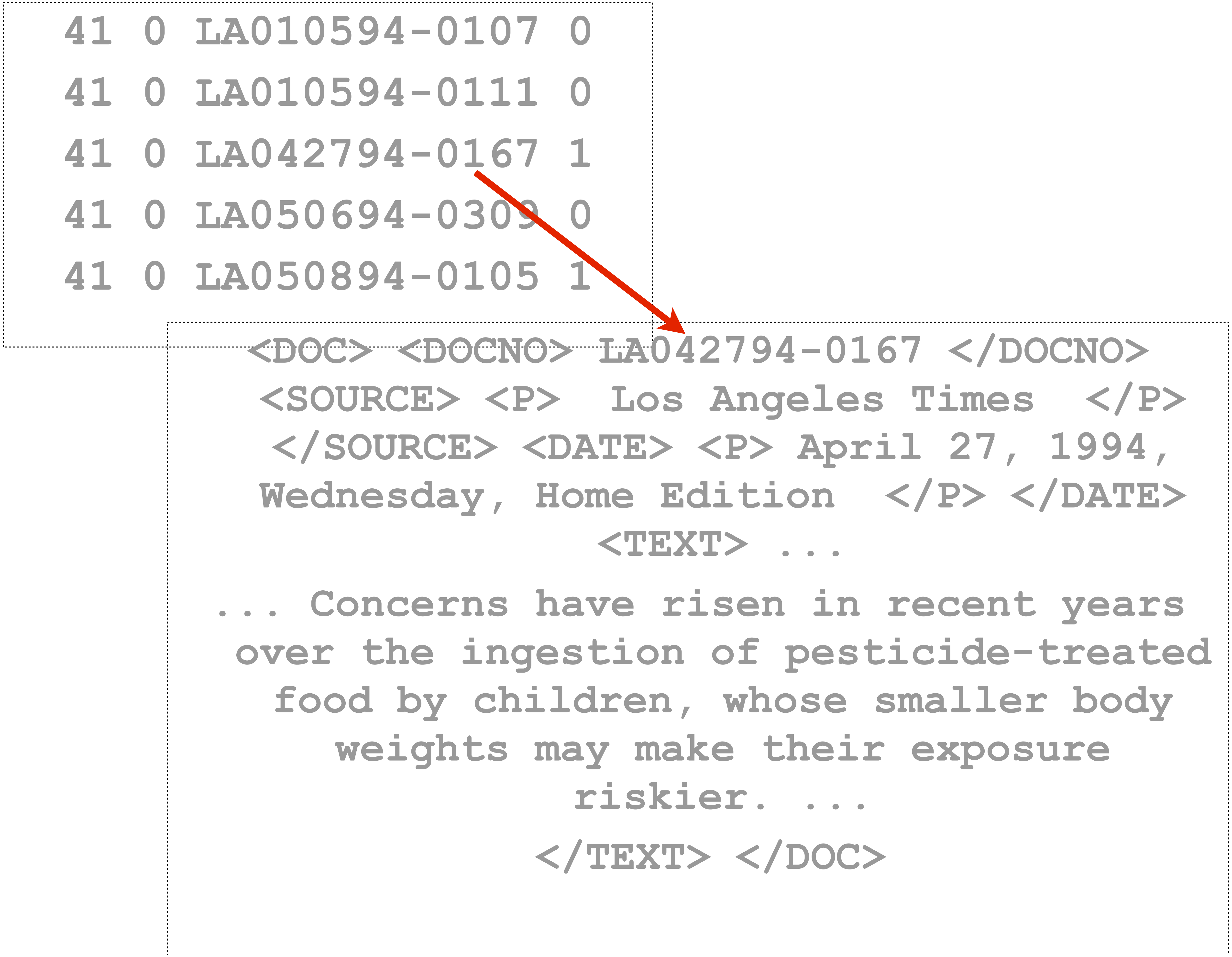
CLEF, EU, 1999 -

NTCIR, Japan, 1999 -

FIRE, India, 2008 -

plus many similar in ML, NLP etc

```
<top>
<num> C041 </num>
<EN-title> Pesticides in Baby Food </EN-title>
<EN-desc> Find reports on pesticides in baby food. </EN-desc>
<EN-narr> Relevant documents give information on the discovery of pesticides in baby food. They report on different brands, supermarkets, and companies selling baby food which contains pesticides. They also discuss measures against the contamination of baby food by pesticides. </EN-narr>
</top>
```



```
41 0 LA010594-0107 0
41 0 LA010594-0111 0
41 0 LA042794-0167 1
41 0 LA050694-0309 0
41 0 LA050894-0105 1
```

```
<DOC> <DOCNO> LA042794-0167 </DOCNO>
<SOURCE> <P>  Los Angeles Times  </P>
</SOURCE> <DATE> <P> April 27, 1994,
Wednesday, Home Edition  </P> </DATE>
<TEXT> ...
... Concerns have risen in recent years
over the ingestion of pesticide-treated
food by children, whose smaller body
weights may make their exposure
riskier. ...
</TEXT> </DOC>
```

imposing power on any observable variable creates bias!

risky!

risk 1: blocks creativity - what happened with e.g. context?

risk 2: overtraining (partial remedy: crossvalidation)

risk 3: variation across queries greater than variation across systems (partial remedy: more queries in test set)

example problem: sentiment polarity

And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been a expansion pack.

I loved The Spy Who Came In From The Cold but the movie is a bit dated in a way the book never will be.

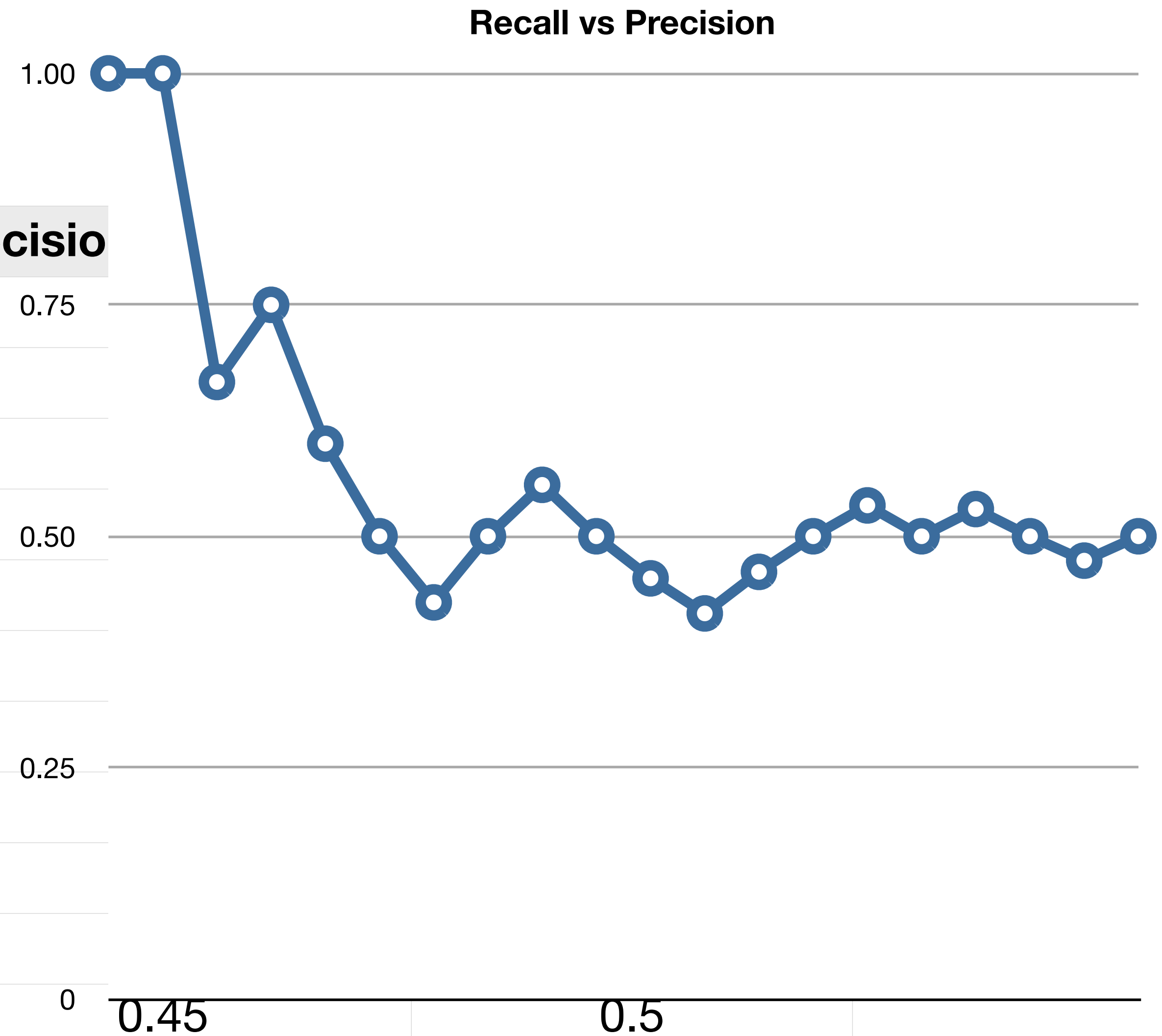
Meat is more environmentally friendly than seafood.

I am unsure about the feasibility of this knitting pattern.

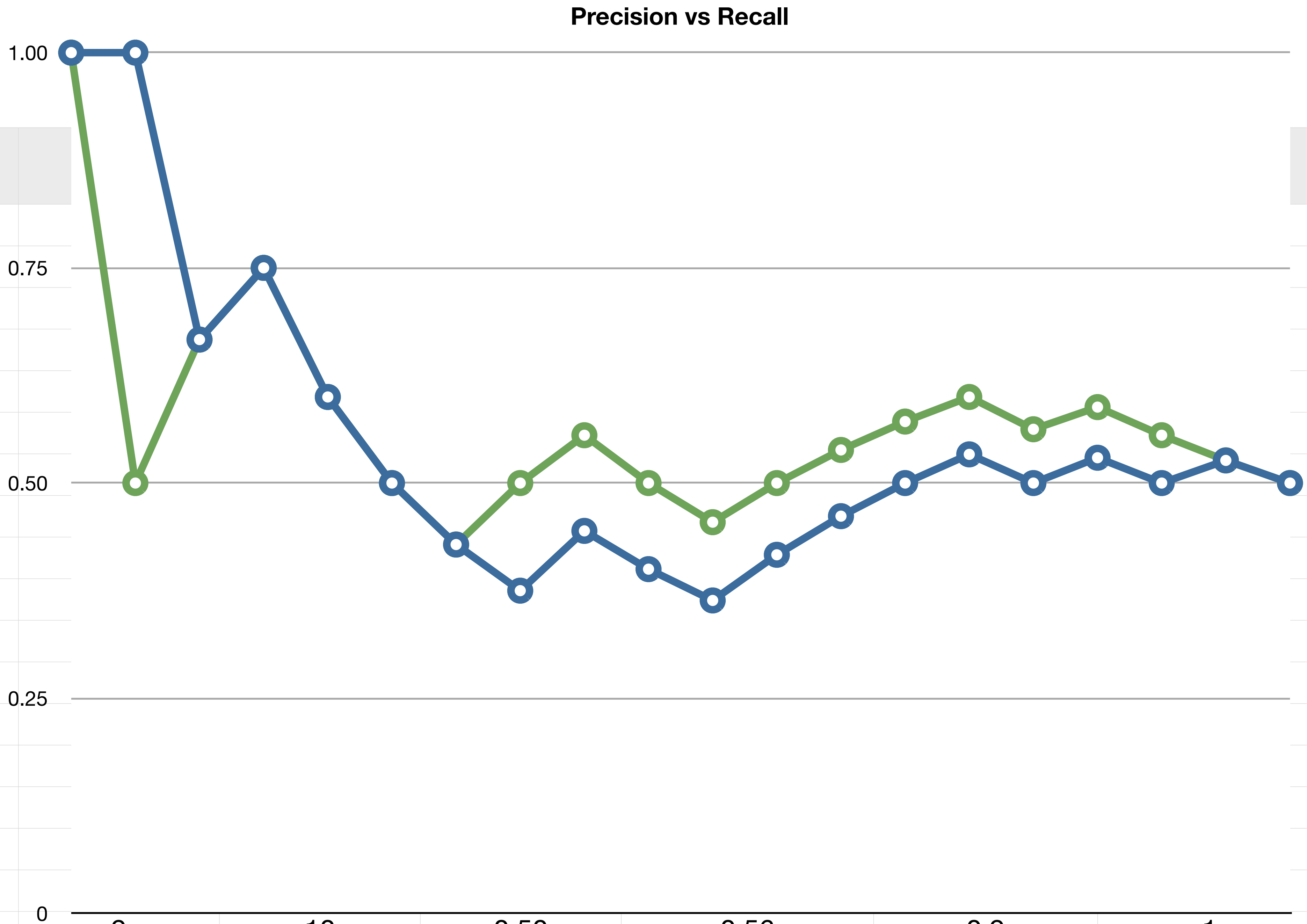
I love the Samsung B2710 but I would not recommend it to my colleagues.

I don't know if I should call her up – I liked her when I met her last weekend.

This is true.



relevant?	tp	precisio	
1	1		
1	2		
0	2		
1	3		
0	3		
0	3		
0	3		
1	4		
1	5		
0	5		
0	5	0.45	0.5
0	5	0.42	0.5
1	6	0.46	0.6
1	7	0.50	0.7
1	8	0.53	0.8
0	8	0.50	0.8
1	9	0.53	0.9
0	9	0.50	0.9
0	9	0.47	0.9
1	10	0.50	1



relevant?	relevant?						
1	1						
1	0						
0	1						
1	1						
0	0						
0	0						
0	0						
0	1						
1	1						
0	0						
0	0						
1	1						
1	1						
1	1						
1	1						
0	0						
1	1						
0	0	9	10	0.50	0.56	0.9	1
1	0	10	10	0.53	0.53	1	1
0	0	10	10	0.50	0.50	1	1

F-score

harmonic mean of precision and recall

$F_1 = 2PR / (P + R)$

you will not be able to avoid the F-score

(8.5; 8.6)

map

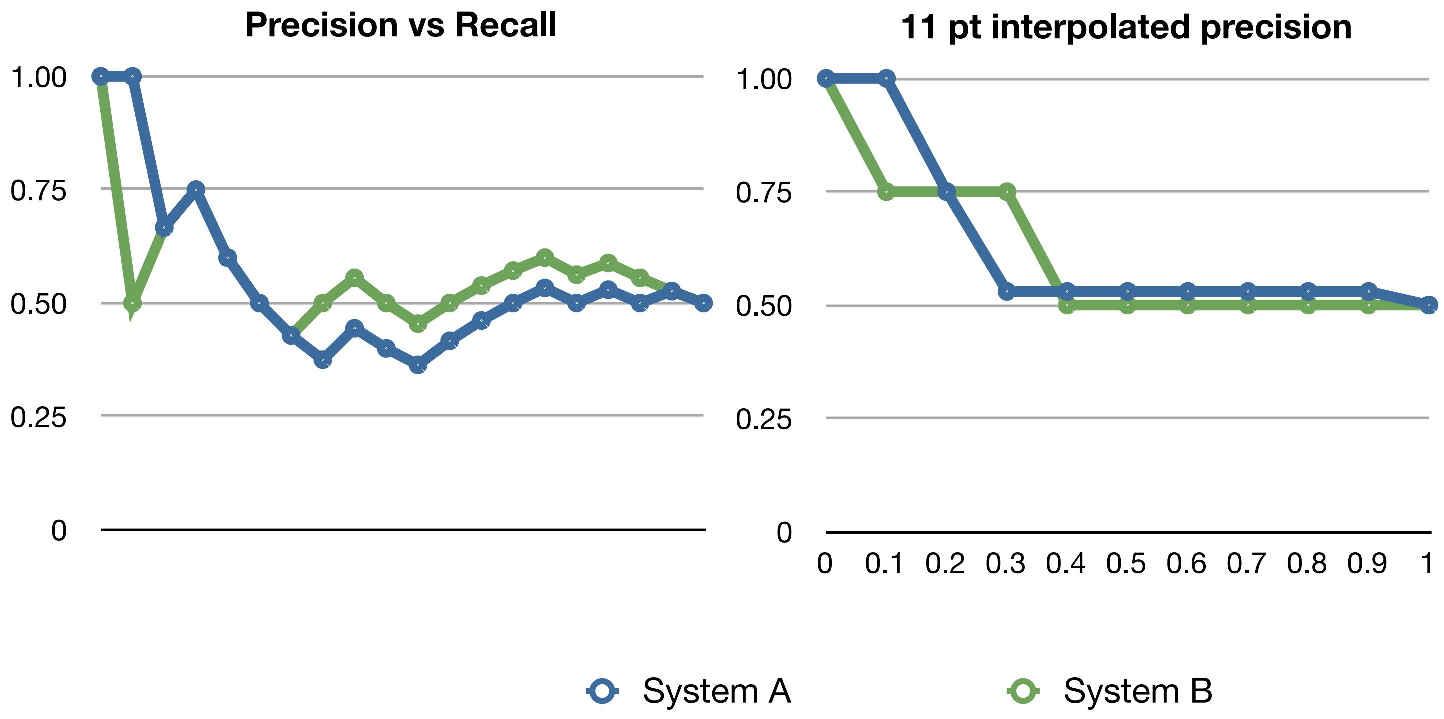
average precision at the rank of each retrieved document

relevant?	relevant?	precision	precision
1	1	1.000	1.000
1	0	1.000	0.500
0	1	0.667	
1	1		0.750
0	0	0.600	0.600
0	0		
0	0		
0	1		
1	1		0.556
0	0	0.400	0.500
0	0		
1	1		
1	1	0.462	0.538
1	1	0.500	0.571
1	1	0.533	0.600
0	0	0.500	0.563
1	1		
0	0	0.500	0.556
1	0		
0	0	0.500	
	MAP:	**0.666**	**0.673**

(8.8)

11-pt interpolated precision

1. precision at recall level r is the highest precision for every recall level ≥ r
2. compute this for r = 0.0, 0.1 ... 0.9, 1.0
3. equivalent of smoothing recall-precision curve



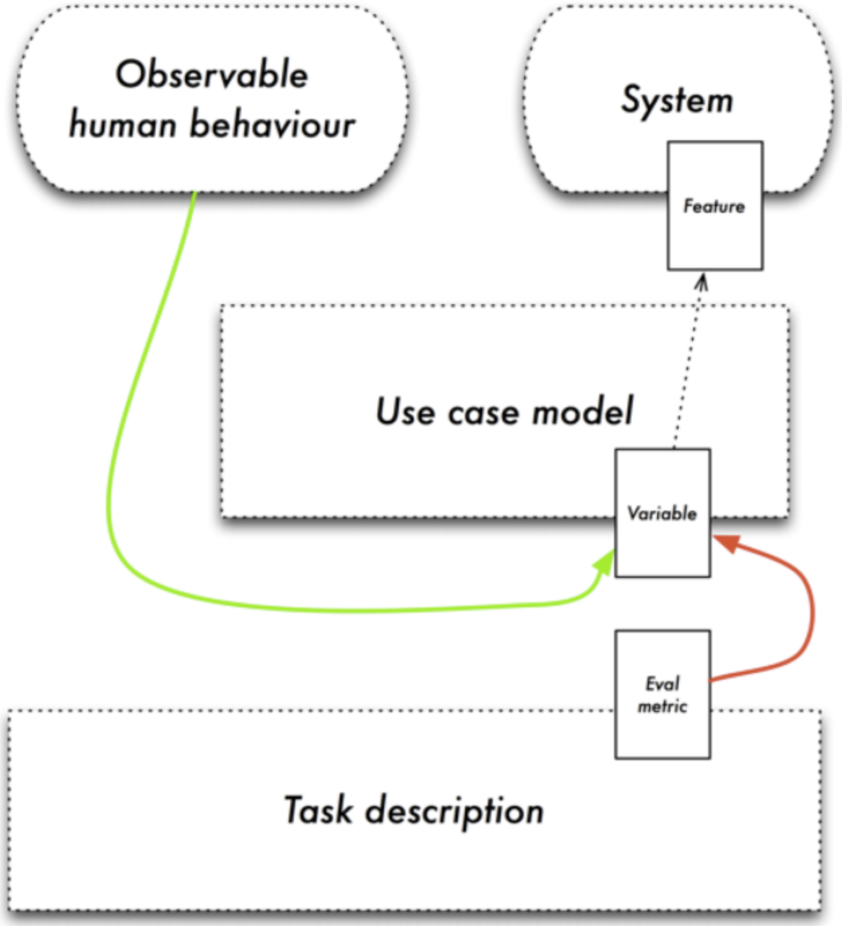
(8.7)

	11pt	11pt
0	1	0.5
0.1	1	0.5
0.2	0.75	0.5
0.3	0.53	0.5
0.4	0.53	0.5
0.5	0.53	0.5
0.6	0.53	0.5
0.7	0.53	0.5
0.8	0.53	0.5
0.9	0.53	0.5
1	0.5	0.5

modelling usage:

1.87 wds / q

use case as a modelling framework



(don't worry, we'll probably return to this next time)

p@N
assumes that N is a sensible number

relevant?	relevant?	tp	tp	precision	precision
1	1	1	1	1.00	1.00
1	0	2	1	1.00	0.50
0	1	2	2	0.67	0.67
1	1	3	3	0.75	0.75
0	0	3	3	0.60	0.60
0	0	3	3	0.50	0.50
0	0	3	3	0.43	0.43
0	1	3	4	0.38	0.50
1	1	4	5	0.44	0.56
0	0	4	5	**0.40**	**0.50**
0	0	4	5	0.36	0.45
1	1	5	6	0.42	0.50
1	1	6	7	0.46	0.54
1	1	7	8	0.50	0.57
1	1	8	9	0.53	0.60
0	0	8	9	0.50	0.56
1	1	9	10	0.53	0.59
0	0	9	10	0.50	0.56
1	0	10	10	0.53	0.53
0	0	10	10	**0.50**	**0.50**

P@10

cumulative gain measures

measure gain at rank p

introducing graded relevance values

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank ≤ p

DCG: discounted cumulative gain at rank p

penalise relevant documents if delivered late

nDCG: normalized discounted cumulative gain at rank p

compared to perfect system

relevant?	relevant?	CG	CG	DCG	DCG	Ideal system	IDCG	nDCG	nDCG
3	3	3	3	3.00	3.00	3	3.00	1.00	1.00
2	0	5	3	9.64	3.00	3	12.97	0.74	0.23
0	0	5	3	9.64	3.00	2	17.16	0.56	0.17
0	2	5	5	9.64	6.32	2	20.48	0.47	0.31
1	2	6	7	11.07	9.18	1	21.91	0.51	0.42
2	1	8	8	13.64	10.47	1	23.20	0.59	0.45
3	1	11	9	17.19	11.65	0	23.20	0.74	0.50
1	3	12	12	18.30	14.97	0	23.20	0.79	0.65
0	0	12	12	18.30	14.97	0	23.20	0.79	0.65

take home message

you should understand

evaluation and systematic testing

(the thing to do, whatever you do)

precision and recall

various measures based on p & r

perils of averages

crucial and central target notion of "relevance"

challenges to "relevance"