

evaluating information retrieval systems

jussi karlgren, gavagai and kth, stockholm

january 2019

jussi karlgren

- ▶ real scale real time distributional semantics

jussi karlgren

- ▶ real scale real time distributional semantics
- ▶ stylistics and genre in text and conversation

jussi karlgren

- ▶ real scale real time distributional semantics
- ▶ stylistics and genre in text and conversation
- ▶ since 2010 at the text analysis company Gavagai

jussi karlgren

- ▶ real scale real time distributional semantics
- ▶ stylistics and genre in text and conversation
- ▶ since 2010 at the text analysis company Gavagai
- ▶ and as adjoint professor of language technology at KTH

jussi karlgren

- ▶ real scale real time distributional semantics
- ▶ stylistics and genre in text and conversation
- ▶ since 2010 at the text analysis company Gavagai
- ▶ and as adjoint professor of language technology at KTH
- ▶ easy to find on the internet

gavagai

- ▶ research-based lexical learning back end
lexicon.gavagai.se

gavagai

- ▶ research-based lexical learning back end
lexicon.gavagai.se
- ▶ media monitoring in >40 languages
monitor.gavagai.se

gavagai

- ▶ research-based lexical learning back end
lexicon.gavagai.se
- ▶ media monitoring in >40 languages
monitor.gavagai.se
- ▶ highly interactive clustering and theme
extraction from open answers to questionnaires
etc
explorer.gavagai.se

continuous evaluation

- ▶ crucial vehicle for successful technology development

continuous evaluation

- ▶ crucial vehicle for successful technology development
- ▶ testing comes first, not last!

continuous evaluation

- ▶ crucial vehicle for successful technology development
- ▶ testing comes first, not last!
- ▶ testing continuously developed

continuous evaluation

- ▶ crucial vehicle for successful technology development
- ▶ testing comes first, not last!
- ▶ testing continuously developed
- ▶ benchmarking

continuous evaluation

- ▶ crucial vehicle for successful technology development
- ▶ testing comes first, not last!
- ▶ testing continuously developed
- ▶ benchmarking
- ▶ validation → testing must be tested!

reliability vs validity

improve systems, but how?

challenges: am i afraid to miss good stuff? or am i
flooded with junk?

improving systems

- ▶ increase relevance?

improving systems

- ▶ increase relevance?
- ▶ increase precision?

improving systems

- ▶ increase relevance?
- ▶ increase precision?
- ▶ increase coverage?

relevance

relevance

in everyday language:

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

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, generalising from everything contextual

this is not a clear-cut issue!

improving systems

- ▶ use stoplists?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?
- ▶ query expansion?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?
- ▶ query expansion?
- ▶ user modelling?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?
- ▶ query expansion?
- ▶ user modelling?
- ▶ geolocation?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?
- ▶ query expansion?
- ▶ user modelling?
- ▶ geolocation?
- ▶ plan inference

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?
- ▶ query expansion?
- ▶ user modelling?
- ▶ geolocation?
- ▶ plan inference
- ▶ domain specificity?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?
- ▶ query expansion?
- ▶ user modelling?
- ▶ geolocation?
- ▶ plan inference
- ▶ domain specificity?
- ▶ genre and style analysis?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?
- ▶ query expansion?
- ▶ user modelling?
- ▶ geolocation?
- ▶ plan inference
- ▶ domain specificity?
- ▶ genre and style analysis?
- ▶ measures of reliability and authority?

improving systems

- ▶ use stoplists?
- ▶ lemmatise?
- ▶ use thesauri?
- ▶ query expansion?
- ▶ user modelling?
- ▶ geolocation?
- ▶ plan inference
- ▶ domain specificity?
- ▶ genre and style analysis?
- ▶ measures of reliability and authority?
- ▶ social factors?

information access is about making users happy.

who is the user?

- ▶ happiness thru three-way optimisation:

- ▶ happiness thru three-way optimisation:
- ▶ price-quality-timeliness

- ▶ happiness thru three-way optimisation:
- ▶ price-quality-timeliness
- ▶ (we'll mostly discuss quality here)

improvement pipeline? how does it help in the end?

what is quality in information access?

- ▶ usefulness (and effectiveness) for task

what is quality in information access?

- ▶ usefulness (and effectiveness) for task
- ▶ appealing presentation, response time, and intuitive interaction

what is quality in information access?

- ▶ usefulness (and effectiveness) for task
- ▶ appealing presentation, response time, and intuitive interaction
- ▶ authority, trustworthiness, sourceability, and reliability

what is quality in information access?

- ▶ usefulness (and effectiveness) for task
- ▶ appealing presentation, response time, and intuitive interaction
- ▶ authority, trustworthiness, sourceability, and reliability
- ▶ relevance and truthfulness

what is quality in information access?

- ▶ usefulness (and effectiveness) for task
- ▶ appealing presentation, response time, and intuitive interaction
- ▶ authority, trustworthiness, sourceability, and reliability
- ▶ relevance and truthfulness
- ▶ reusability and cost

what is quality in information access?

- ▶ usefulness (and effectiveness) for task
- ▶ appealing presentation, response time, and intuitive interaction
- ▶ authority, trustworthiness, sourceability, and reliability
- ▶ relevance and truthfulness
- ▶ reusability and cost
- ▶ timeliness in content

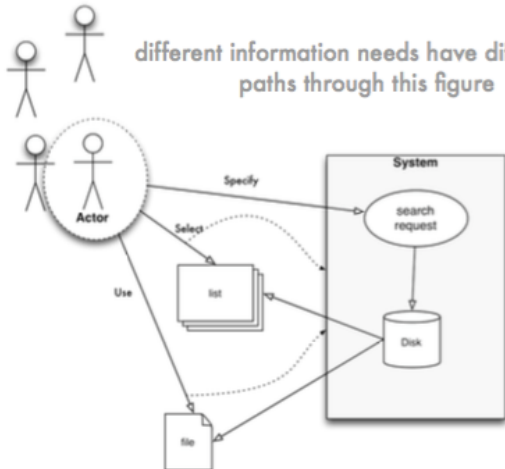
what is quality in information access?

- ▶ usefulness (and effectiveness) for task
- ▶ appealing presentation, response time, and intuitive interaction
- ▶ authority, trustworthiness, sourceability, and reliability
- ▶ relevance and truthfulness
- ▶ reusability and cost
- ▶ timeliness in content

how to measure nice things?

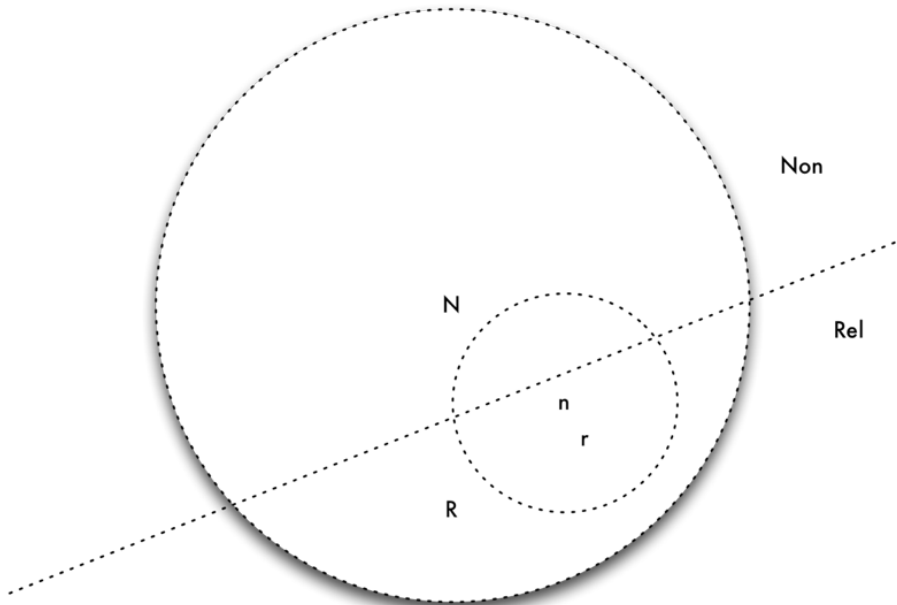
measuring depends on the envisioned situation of use

different information needs have different
paths through this figure



we will focus on relevance

but remember this decision!



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

	relevant	non-relevant
delivered	true positives (good)	false positives
not delivered	false negatives	true negatives

	relevant	non-relevant
delivered	true positives (good)	false positives
not delivered	false negatives	true negatives (uninteresting)

	relevant	non-relevant
delivered	true positives (good)	false positives (errors)
not delivered	false negatives	true negatives (uninteresting)

	relevant	non-relevant
delivered	true positives (good)	false positives (errors)
not delivered	false negatives (misses)	true negatives (uninteresting)

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

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

$$= (\text{good} + \text{uninteresting})/\text{all docs}$$

$$= \text{correct labels}/\text{all docs}$$

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

$$\begin{aligned}\text{precision} &= \text{tp}/(\text{tp}+\text{fp}) \\ &= \text{good}/\text{delivered}\end{aligned}$$

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

(täckning)

= good/relevant

= good/(good + miss)

5 min exercise

retrieve and assess relevance of top ten
compare two queries and two search engines

was this easy?

the practice of evaluation

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>

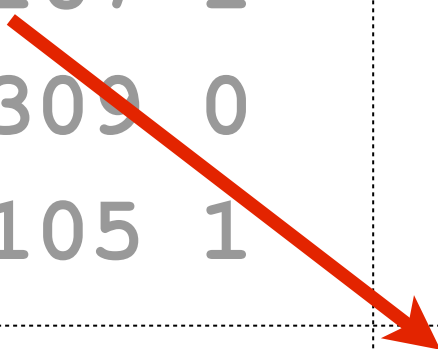
<EN-title> Pesticides in Baby Food </EN-title>

<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!

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!

And the sound quality - my God!

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

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.

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.

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.

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.

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.

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.

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.

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.

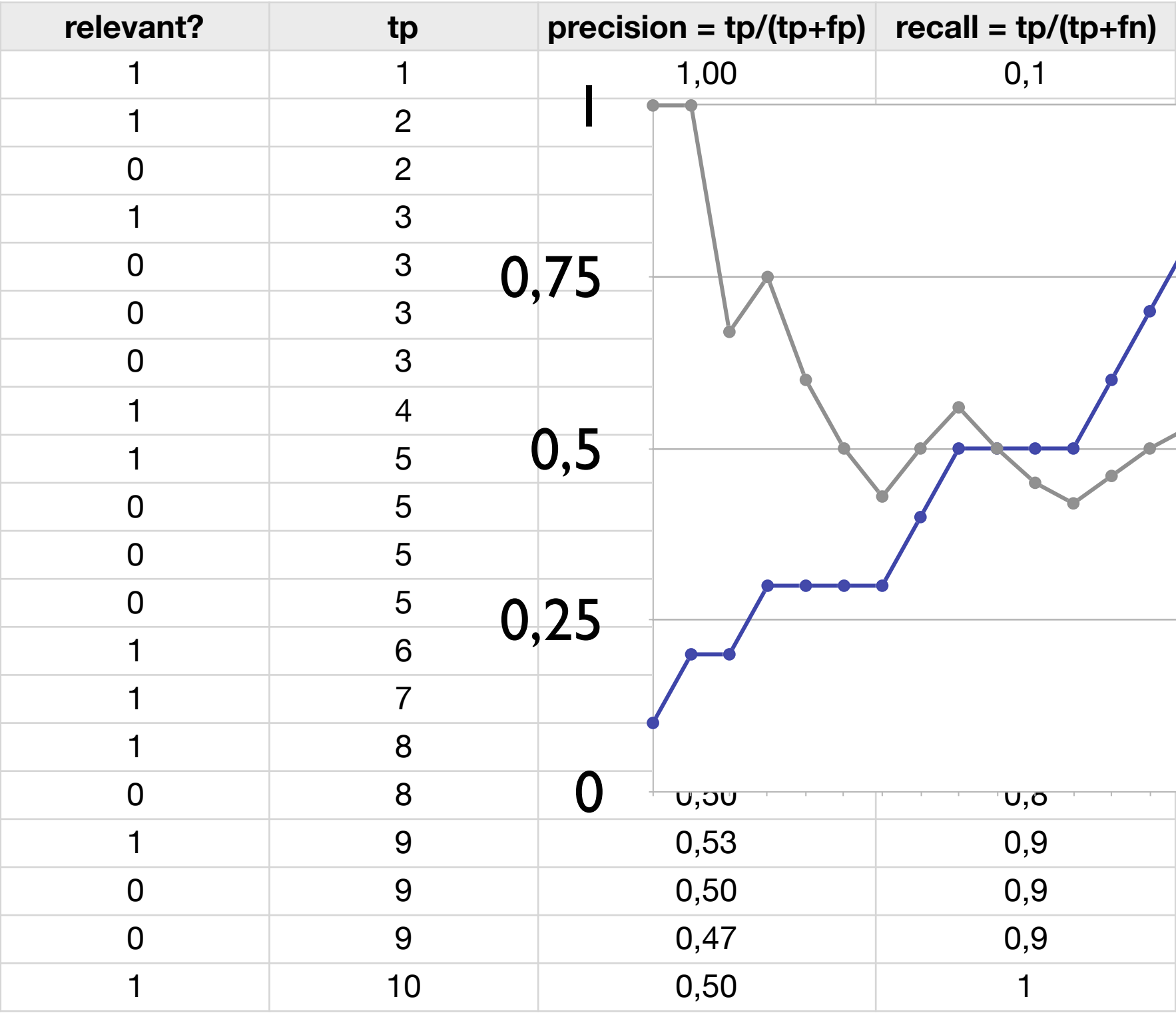
This is true.

but let's assume we can swing it

first: we will now focus on ranked retrieval

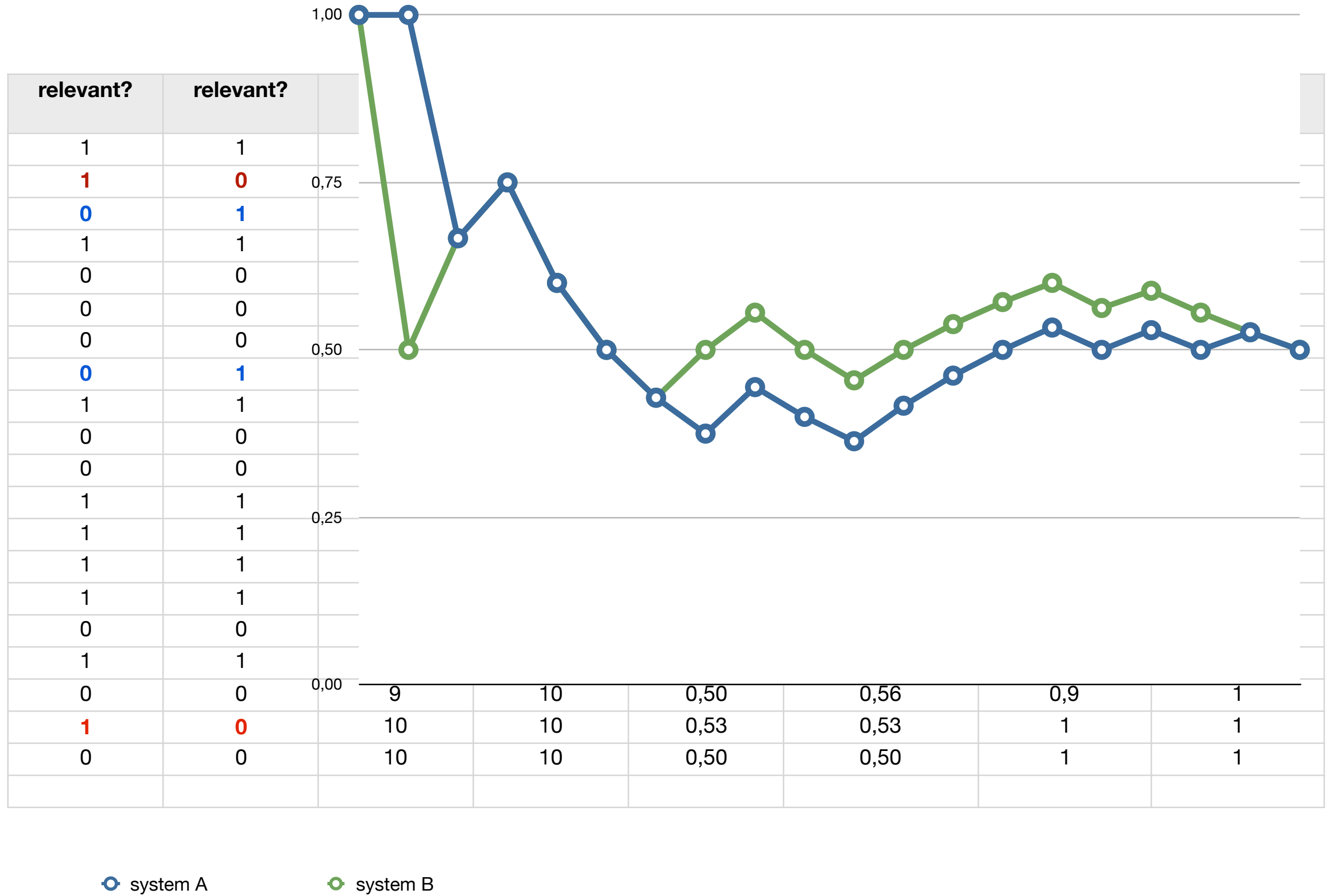
precision and recall combined

relevant?	tp	precision = $tp/(tp+fp)$	recall = $tp/(tp+fn)$
1	1	1,00	0,1
1	2	1,00	0,2
0	2	0,67	0,2
1	3	0,75	0,3
0	3	0,60	0,3
0	3	0,50	0,3
0	3	0,43	0,3
1	4	0,50	0,4
1	5	0,56	0,5
0	5	0,50	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?	tp	tp	precision	precision	recall	recall
1	1	1	1	1,00	1,00	0,1	0,1
1	0	2	1	1,00	0,50	0,2	0,1
0	1	2	2	0,67	0,67	0,2	0,2
1	1	3	3	0,75	0,75	0,3	0,3
0	0	3	3	0,60	0,60	0,3	0,3
0	0	3	3	0,50	0,50	0,3	0,3
0	0	3	3	0,43	0,43	0,3	0,3
0	1	3	4	0,38	0,50	0,3	0,4
1	1	4	5	0,44	0,56	0,4	0,5
0	0	4	5	0,40	0,50	0,4	0,5
0	0	4	5	0,36	0,45	0,4	0,5
1	1	5	6	0,42	0,50	0,5	0,6
1	1	6	7	0,46	0,54	0,6	0,7
1	1	7	8	0,50	0,57	0,7	0,8
1	1	8	9	0,53	0,60	0,8	0,9
0	0	8	9	0,50	0,56	0,8	0,9
1	1	9	10	0,53	0,59	0,9	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

Precision



a curve is fine, but a scalar would be simpler

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

(8.8)

map

relevant?	relevant?	map	map
1	1	1,000	1,000
1	0	1,000	0,500
0	1	0,567	0,567
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	
		0,666	0,673

average precision at the rank of each retrieved document

(8.8)

11-pt interpolated precision

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

(8.7)

11-pt interpolated precision

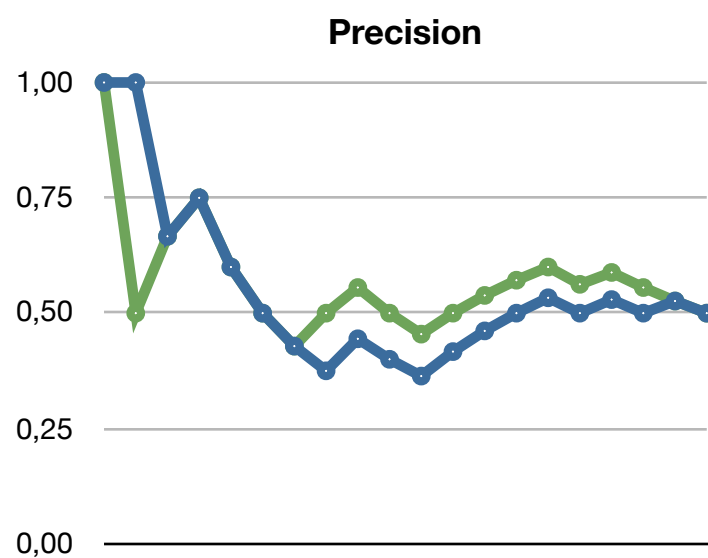
1. precision at recall level r is the highest precision for every recall level $\geq r$
2. compute this for $r = 0.0, 0.1 \dots 0.9, 1.0$
3. equivalent of smoothing recall-precision curve

(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

11-pt interpolated precision

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

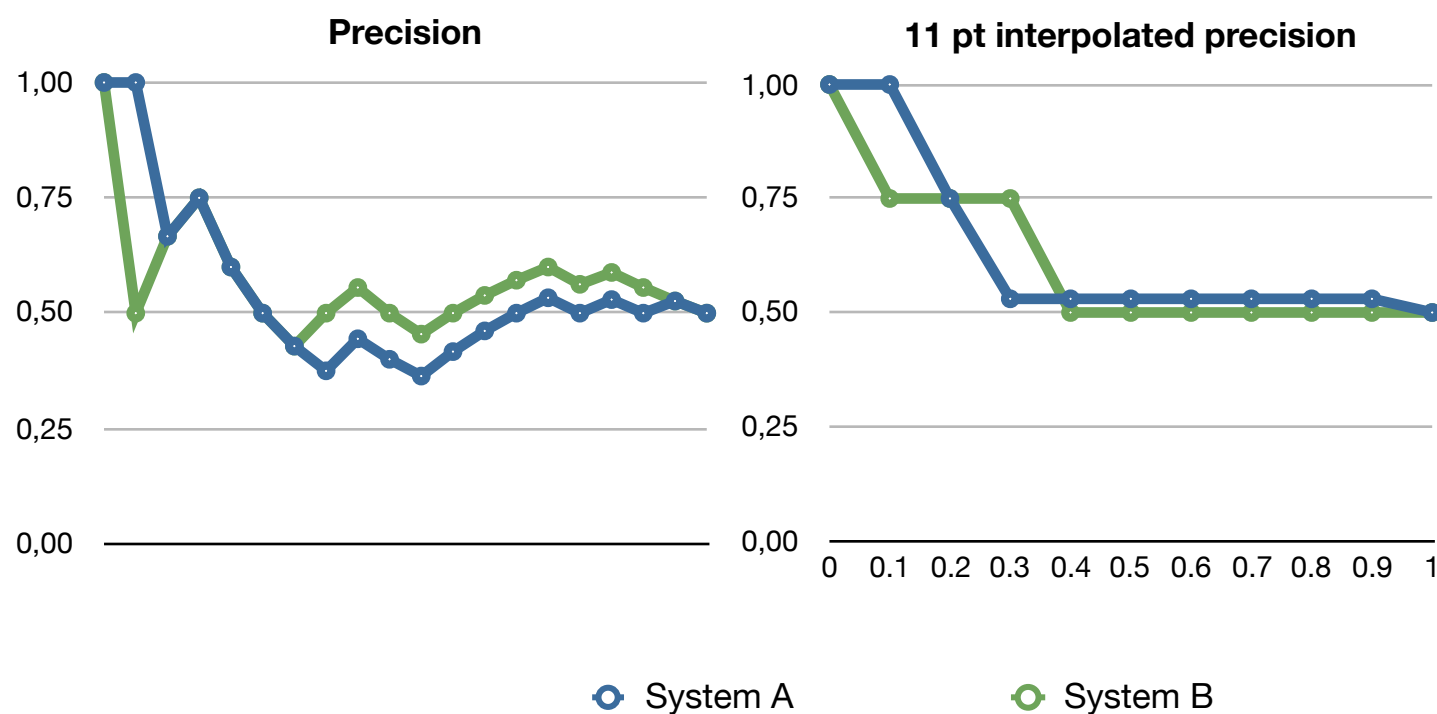


(8.7)

	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

11-pt interpolated precision

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



(8.7)

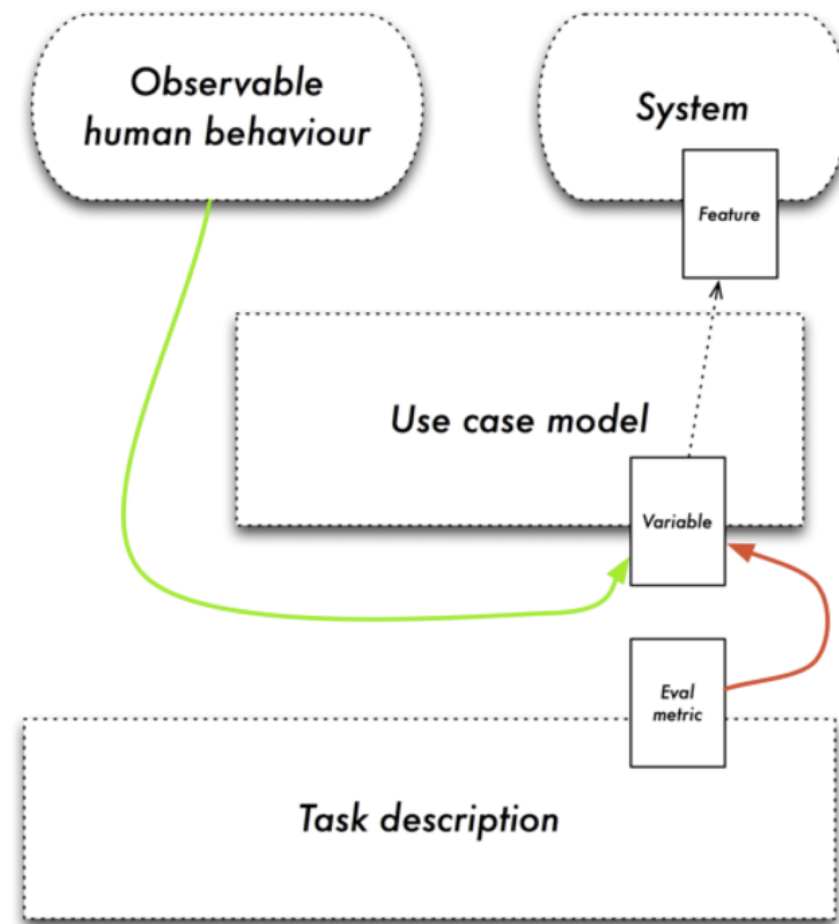
	11pt	11pt
0	1	0,5
0,1	1	0,5
0,2	0,75	0,5
0,3	0,53	0,5
0,4	0,53	0,5
0,5	0,53	0,5
0,6	0,53	0,5
0,7	0,53	0,5
0,8	0,53	0,5
0,9	0,53	0,5
1	0,5	0,5

back to usefulness for task

modelling usage:

1.87 wds / q

use case as a modelling framework



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

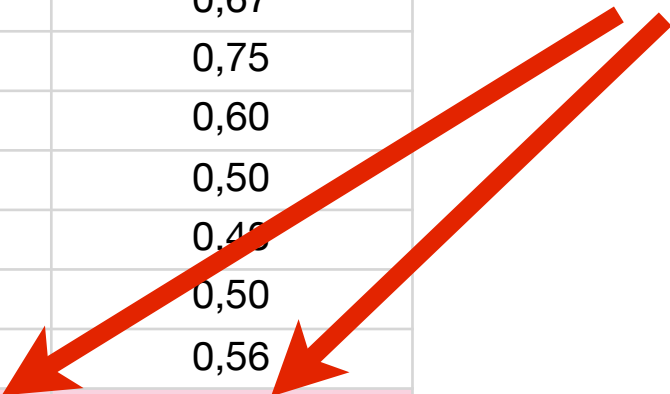
p@N

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

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 $\leq p$

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

relevant?	relevant?	CG	CG
3	3	3	3
2	0	5	3
0	0	5	3
0	2	5	5
1	2	6	7
2	1	8	8
3	1	11	9
1	3	12	12
0	0	12	12

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

relevant?	relevant?	CG	CG
3	3	3	3
2	0	5	3
0	0	5	3
0	2	5	5
1	2	6	7
2	1	8	8
3	1	11	9
1	3	12	12
0	0	12	12

non-binary relevance!

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

relevant?	relevant?	CG	CG
3	3	3	3
2	0	5	3
0	0	5	3
0	2	5	5
1	2	6	7
2	1	8	8
3	1	11	9
1	3	12	12
0	0	12	12

$$CG_p = \sum_{i=1}^p rel_i$$

non-binary relevance!

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

DCG: discounted cumulative gain at rank p

penalise relevant documents if delivered late

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

DCG: discounted cumulative gain at rank p

penalise relevant documents if delivered late

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

DCG: discounted cumulative gain at rank p

penalise relevant documents if delivered late

p	relevant?	relevant?	CG	CG	DCG	DCG
1	3	3	3	3	3,00	3,00
2	2	0	5	3	5,00	3,00
3	0	0	5	3	5,00	3,00
4	0	2	5	5	5,00	3,00
5	1	2	6	7	5,43	3,43
6	2	1	8	8	6,20	4,20
7	3	1	11	9	7,27	5,27
8	1	3	12	12	7,61	5,61
9	0	0	12	12	7,61	5,61

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

DCG: discounted cumulative gain at rank p

penalise relevant documents if delivered late

p	relevant?	relevant?	CG	CG	DCG	DCG
1	3	3	3	3	3,00	3,00
2	2	0	5	3	5,00	3,00
3	0	0	5	3	5,00	3,00
4	0	2	5	5	5,00	3,00
5	1	2	6	7	5,43	3,43
6	2	1	8	8	6,20	4,20
7	3	1	11	9	7,27	5,27
8	1	3	12	12	7,61	5,61
9	0	0	12	12	7,61	5,61

$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2(i)}$$

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

DCG: discounted cumulative gain at rank p

penalise relevant documents if delivered late

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

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

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

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

CG: simple cumulative gain at rank p

sum of relevance scores for all documents with rank $\leq p$

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

p	relevant?	relevant?	CG	CG	DCG	DCG	Ideal system	IDCG	nDCG	nDCG
1	3	3	3	3	3,00	3,00	3	3,00	1,00	1,00
2	2	0	5	3	5,00	3,00	3	6,00	0,83	0,50
3	0	0	5	3	5,00	3,00	2	7,26	0,69	0,41
4	0	2	5	5	5,00	3,00	2	8,26	0,61	0,36
5	1	2	6	7	5,43	3,43	1	8,69	0,62	0,39
6	2	1	8	8	6,20	4,20	1	9,08	0,68	0,46
7	3	1	11	9	7,27	5,27	0	9,08	0,80	0,58
8	1	3	12	12	7,61	5,61	0	9,08	0,84	0,62
9	0	0	12	12	7,61	5,61	0	9,08	0,84	0,62

take home

1. reliability vs validity

take home

1. reliability vs validity
2. relevance

take home

1. reliability vs validity
2. relevance
3. precision and recall

take home

1. reliability vs validity
2. relevance
3. precision and recall
4. some measures based on p & r

take home

1. reliability vs validity
2. relevance
3. precision and recall
4. some measures based on p & r
5. averages are risky

