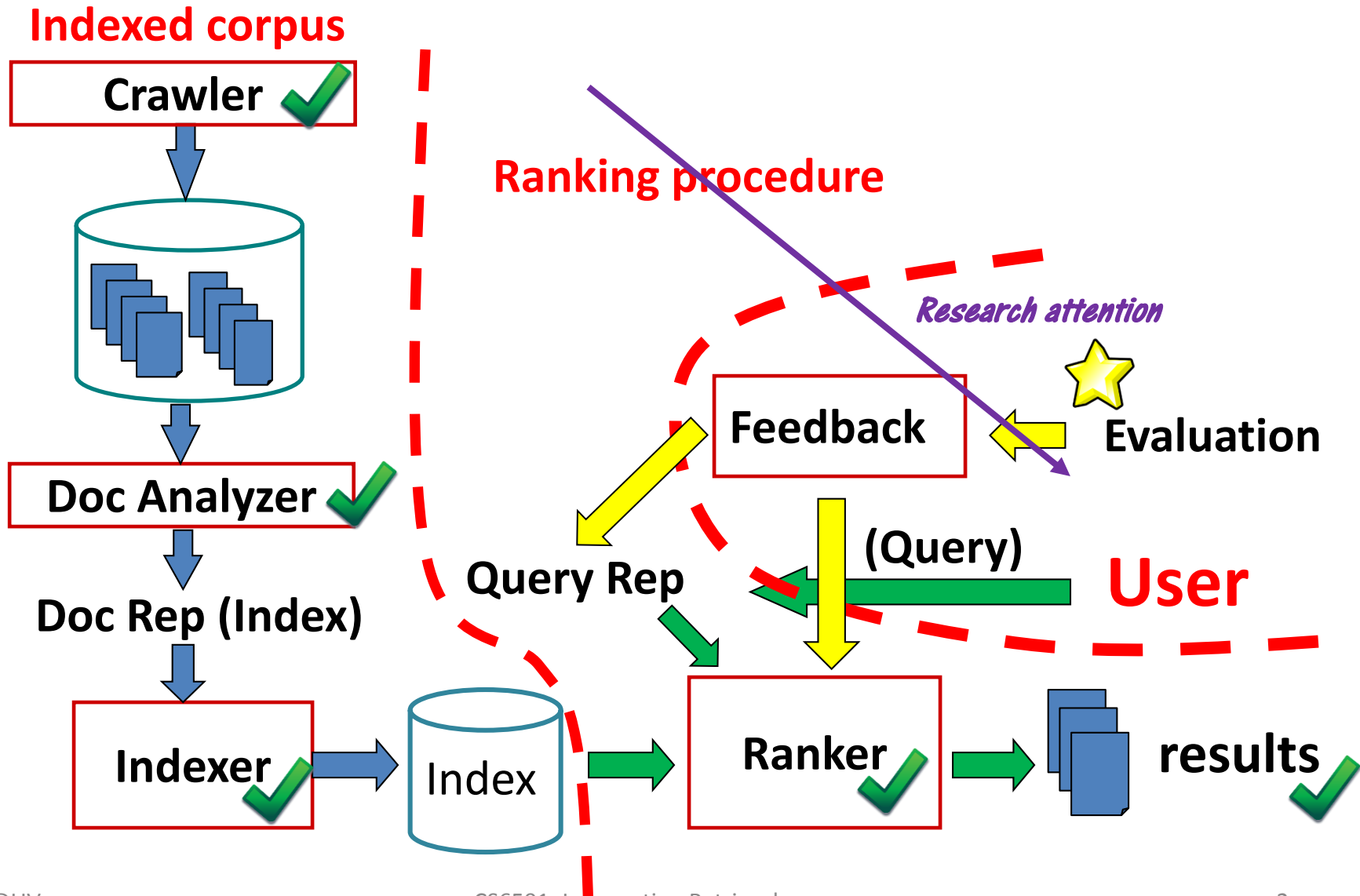


Retrieval Evaluation

Hongning Wang

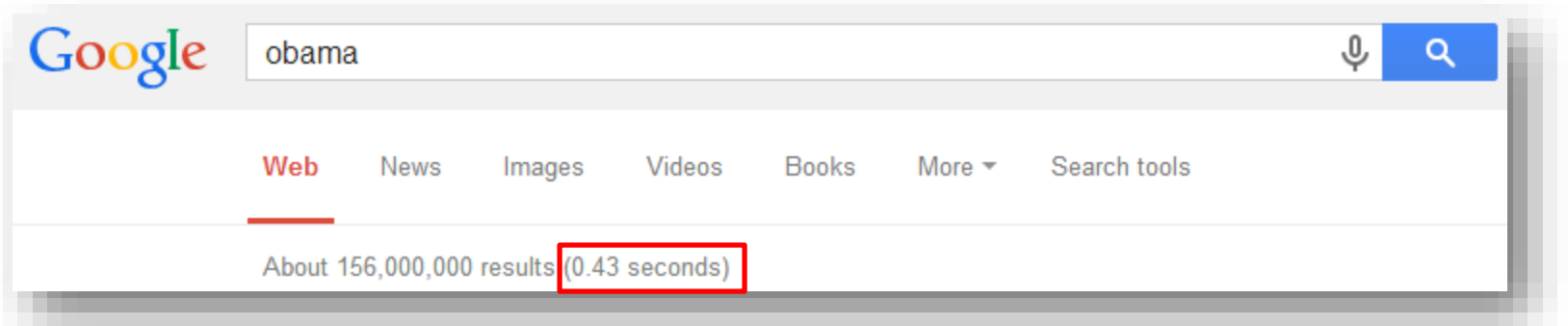
CS@UVa

What we have learned so far

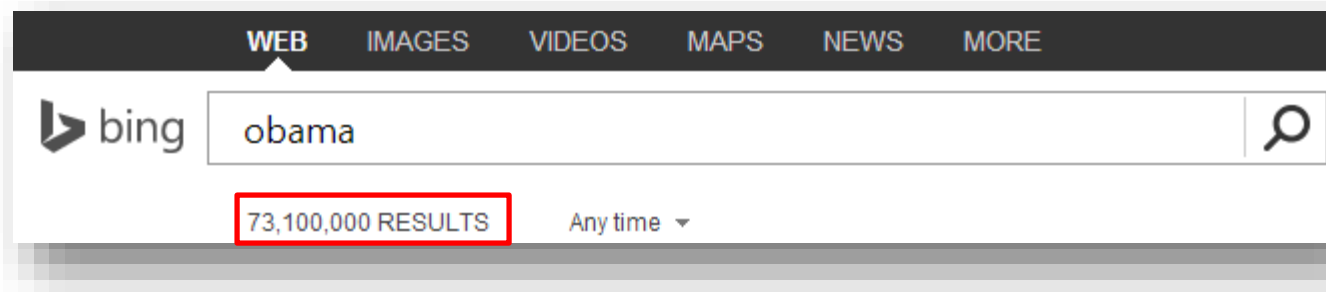


Which search engine do you prefer: Bing or Google?

- What are your judging criteria?
 - How fast does it response to your query?

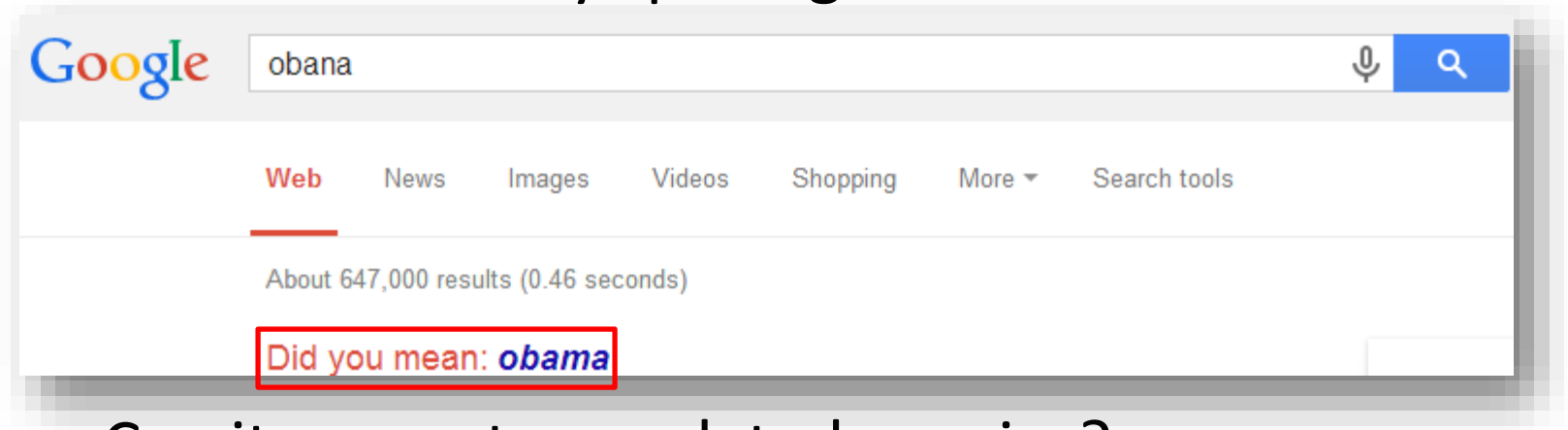


- How many documents can it return?



Which search engine do you prefer: Bing or Google?

- What are your judging criteria?
 - Can it correct my spelling errors?





- Can it suggest me related queries?








Retrieval evaluation

- Aforementioned evaluation criteria are all good, but not essential
 - Goal of any IR system
 - Satisfying users' information need
 - Core quality measure criterion
 - *“how well a system meets the information needs of its users.” – wiki*
 - Unfortunately vague and hard to execute

Bing v.s. Google?

BING  **bing**
— VS —
IT ON  **Google**

ROUNDS COMPLETED     

MAKE YOUR SELECTION FOR: who is the 34th president of US

[I choose the results on the left](#) ☐ [I can't decide, it's a draw](#) ☐ [I choose the results on the right](#) ☐

United States 34th president

Dwight D. Eisenhower

Data from Wikipedia

[Who was the 34th President of the United States - The Q&A ...](#)
wiki.answers.com › ... › Mortgages › Home Equity and Refinancing
Who is the 34th President of Mexico? Pedro Lascuráin was the 34th president of Mexico. Was Dwight Eisenhower the 34th president? Ike was the 34th if you count ...

[Dwight D. Eisenhower - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Dwight_D._Eisenhower
Dwight David "Ike" Eisenhower was the 34th President of the United States from 1953 until 1961. He was a five-star general in the United States Army during ...
Early life and education · Personal life · Early military career · World War II


[List of Presidents of the United States - Wikipedia, the ...](#)
en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States
List of Presidents of the United States, sortable by previous experience; List of Vice Presidents of the United States; Presidential portrait (United States)
List of presidents · Living former presidents

[Who was the 34th president of the US? - Yahoo](#)
answers.yahoo.com › Politics & Government › Government
Resolved · 29 posts · 28 total answers · Published Aug 18, 2006
Aug 18, 2006 · President Dwight David Eisenhower Served 1953-1957, 1957-1961 Dwight David Eisenhower was born October 14, 1890, in Denison, Texas, as David Dwight Eisenhower.

[Dwight D. Eisenhower | The White House](#)
www.whitehouse.gov/about/presidents/dwighteisenhower
The Presidential biographies on WhiteHouse.gov are from "The Presidents of the United

Dwight D. Eisenhower

The 34th President of the United States



Feedback

[Dwight D. Eisenhower - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Dwight_D._Eisenhower Wikipedia
Dwight David "Ike" Eisenhower (pronounced /ˈaɪzənhaʊər/, EYE-zan-how-ar; October 14, 1890 – March 28, 1969) was the 34th President of the United States ...
[Eisenhower administration](#) · [Military career of Dwight D.](#) · [John Eisenhower](#)

[List of Presidents of the United States - Wikipedia, the free ...](#)
en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States Wikipedia
For lists of U.S. Presidents based on other criteria such as age or home state, see
"Dwight D. Eisenhower – Democratic Party – 34th President – American ...
[Names](#) · [Whig](#) · [Federalist Party](#) · [Presidents Who Died In Office](#)

[American President: Dwight David Eisenhower - Miller Center](#)
millercenter.org/president/eisenhower Miller Center of Public Affairs
Expert University of Virginia biography of President Dwight D. Eisenhower, including facts on being a ... Term: 34th President of the United States (1953–1961).

[Dwight D. Eisenhower | The White House](#)
www.whitehouse.gov › [About the White House](#) White House

Quantify the IR quality measure

- Information need
 - *“an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need” – wiki*
 - Reflected by user query
 - Categorization of information need
 - Navigational
 - Informational
 - Transactional

Quantify the IR quality measure

- Satisfaction
 - *“the opinion of the user about a specific computer application, which they use” – wiki*
 - Reflected by
 - Increased result clicks
 - Repeated/increased visits
 - Result relevance

Classical IR evaluation



- Cranfield experiments
 - Pioneer work and foundation in IR evaluation
 - Basic hypothesis
 - Retrieved documents' relevance is a good proxy of a system's utility in satisfying users' information need
 - Procedure
 - 1,398 abstracts of aerodynamics journal articles
 - 225 queries
 - Exhaustive relevance judgments of all (query, document) pairs
 - Compare different indexing system over such collection

Classical IR evaluation

- Three key elements for IR evaluation
 1. A document collection
 2. A test suite of information needs, expressible as queries
 3. A set of relevance judgments, e.g., binary assessment of either *relevant* or *nonrelevant* for each query-document pair

Search relevance

- Users' information needs are translated into queries
- Relevance is judged with respect to the information need, **not** the query
 - E.g., Information need: “When should I renew my Virginia driver’s license?”
Query: “Virginia driver’s license renewal”
Judgment: whether a document contains the right answer, e.g., every 8 years; rather than if it literally contains those four words

Text REtrieval Conference (TREC)

- Large-scale evaluation of text retrieval methodologies
 - Since 1992, hosted by NIST
 - Standard benchmark for IR studies
 - A wide variety of evaluation collections
 - Web track
 - Question answering track
 - Cross-language track
 - Microblog track
 - And more...

Public benchmarks

TABLE 4.3 Common Test Corpora

<i>Collection</i>	<i>NDocs</i>	<i>NQrys</i>	<i>Size (MB)</i>	<i>Term/Doc</i>	<i>Q-D RelAss</i>
ADI	82	35			
ATT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Table from Manning Stanford CS276, Lecture 8

Evaluation metric

- To answer the questions
 - Is Google better than Bing?
 - Which smoothing method is most effective?
 - Is BM25 better than language models?
 - Shall we perform stemming or stopword removal?
- We need a quantifiable metric, by which we can compare different IR systems
 - As unranked retrieval sets
 - As ranked retrieval results

Evaluation of unranked retrieval sets

- In a Boolean retrieval system
 - Precision: fraction of retrieved documents that are relevant, i.e., $p(\text{relevant} | \text{retrieved})$
 - Recall: fraction of relevant documents that are retrieved, i.e., $p(\text{retrieved} | \text{relevant})$

	relevant	nonrelevant
retrieved	true positive (TP)	false positive (FP)
not retrieved	false negative (FN)	true negative (TN)

Precision:

$$P = \frac{TP}{TP + FP}$$

Recall: $R = \frac{TP}{TP + FN}$

Evaluation of unranked retrieval sets

- Precision and recall trade off against each other
 - Precision decreases as the number of retrieved documents increases (unless in perfect ranking), while recall keeps increasing
 - These two metrics emphasize different perspectives of an IR system
 - Precision: prefers systems retrieving fewer documents, but highly relevant
 - Recall: prefers systems retrieving more documents

Evaluation of unranked retrieval sets

- Summarizing precision and recall to a single value
 - In order to compare different systems
 - F-measure: weighted harmonic mean of precision and recall, α balances the trade-off

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} \quad \left(F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} \right)$$

- Why harmonic mean?

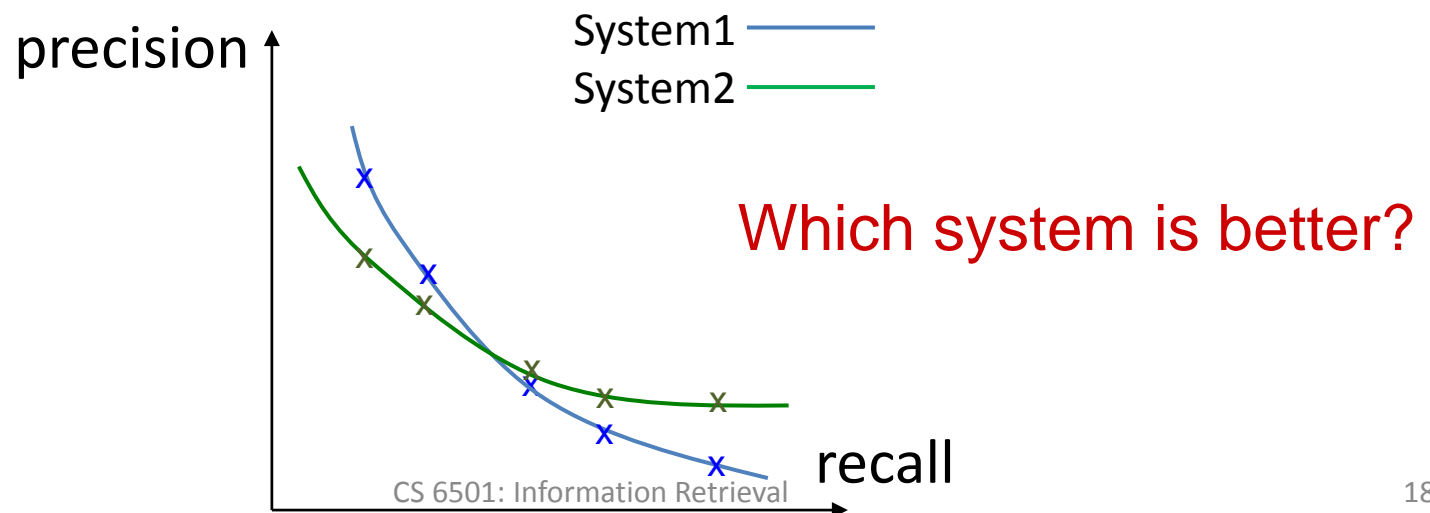
- System1: P:0.53, R:0.36
- System2: P:0.01, R:0.99

H	A
0.429	0.445
0.019	0.500

Equal weight between precision and recall

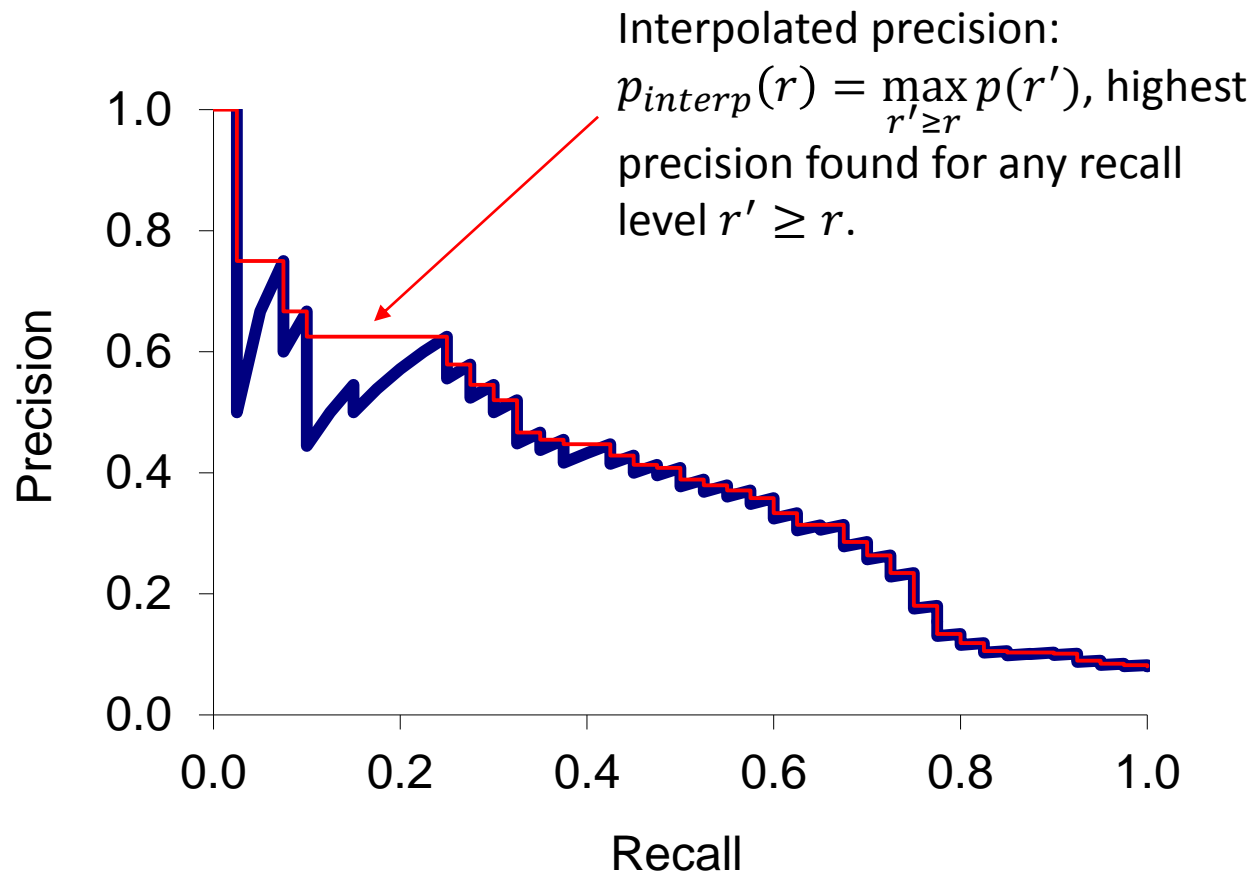
Evaluation of ranked retrieval results

- Ranked results are the core feature of an IR system
 - Precision, recall and F-measure are set-based measures, that cannot assess the ranking quality
 - Solution: evaluate precision at every recall point



Precision-Recall curve

- A sawtooth shape curve

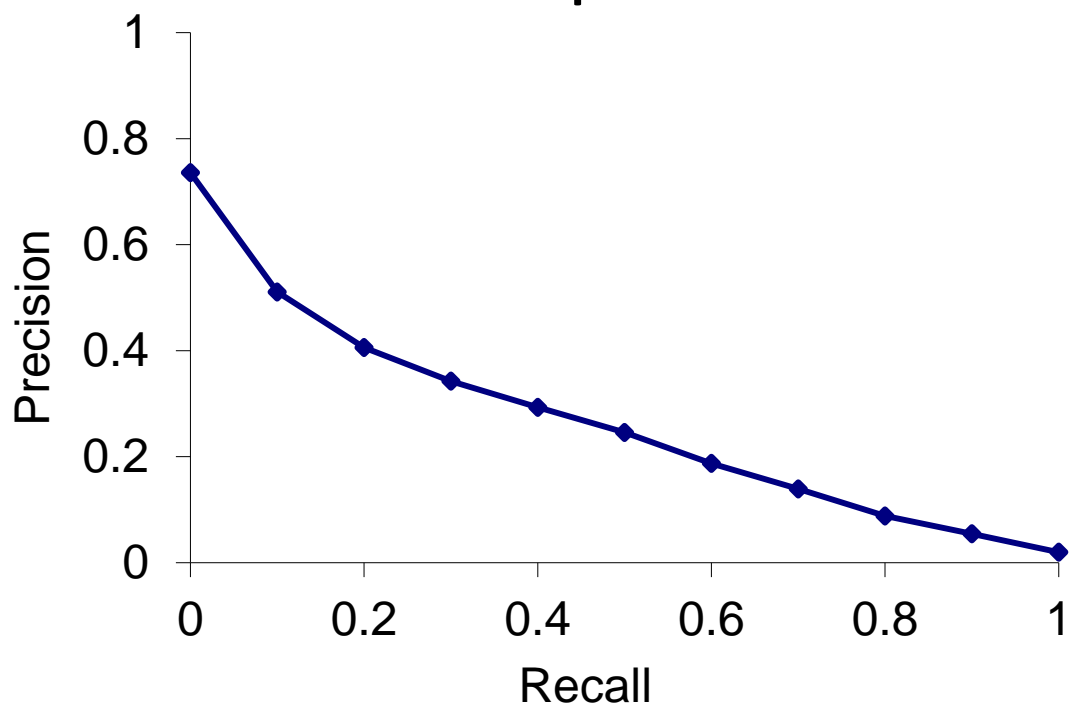


Evaluation of ranked retrieval results

- Summarize the ranking performance with a single number
 - Binary relevance
 - Eleven-point interpolated average precision
 - Precision@K ($P@K$)
 - Mean Average Precision (MAP)
 - Mean Reciprocal Rank (MRR)
 - Multiple grades of relevance
 - Normalized Discounted Cumulative Gain (NDCG)

Eleven-point interpolated average precision

- At the 11 recall levels $[0, 0.1, 0.2, \dots, 1.0]$, compute arithmetic mean of interpolated precision over all the queries



Precision@K

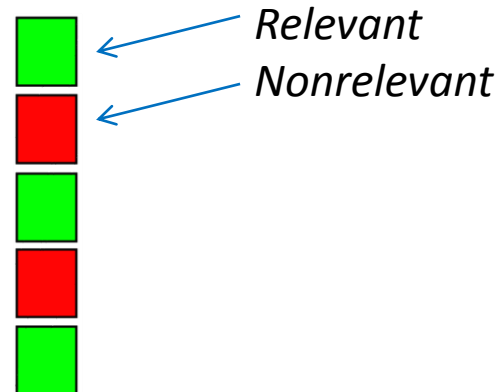
- Set a ranking position threshold K
- Ignores all documents ranked lower than K
- Compute precision in these top K retrieved documents

— E.g.,:

P@3 of 2/3

P@4 of 2/4

P@5 of 3/5



- In a similar fashion we have Recall@K

Mean Average Precision

- Consider rank position of each relevant doc
 - E.g., $K_1, K_2, \dots K_R$
- Compute $P@K$ for each $K_1, K_2, \dots K_R$
- Average precision = average of those $P@K$

– E.g.,



$$AvgPrec = \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) / 3$$

- MAP is mean of Average Precision across multiple queries/rankings

AvgPrec is about one query

 = the relevant documents

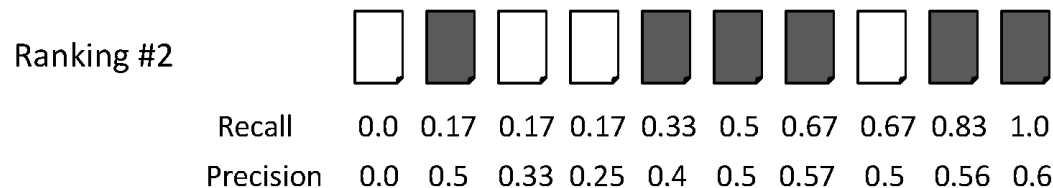
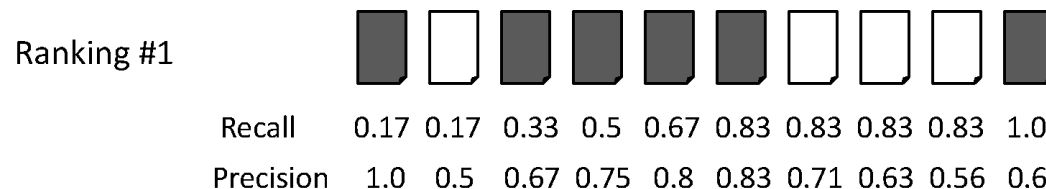


Figure from Manning Stanford CS276, Lecture 8

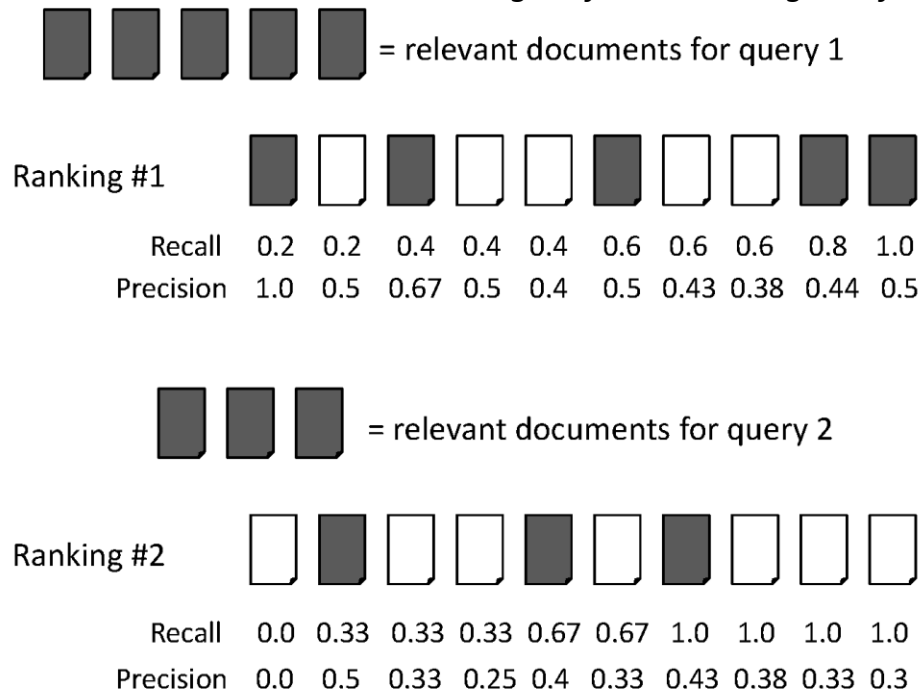
AvgPrec of the two rankings

$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6) / 6 = 0.78$$

$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6) / 6 = 0.52$$

MAP is about a system

Figure from Manning Stanford CS276, Lecture 8



Query 1, $\text{AvgPrec} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5) / 5 = 0.62$

Query 2, $\text{AvgPrec} = (0.5 + 0.4 + 0.43) / 3 = 0.44$

$\text{MAP} = (0.62 + 0.44) / 2 = 0.53$

MAP metric

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant document to be zero
- MAP is macro-averaging: each query counts equally
- MAP assumes users are interested in finding many relevant documents for each query
- MAP requires many relevance judgments in text collection

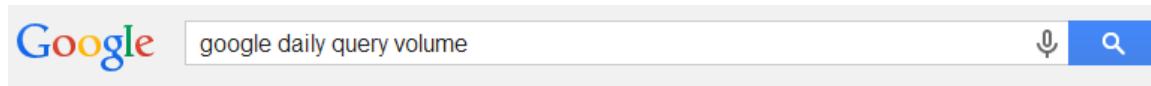
Mean Reciprocal Rank

- Measure the effectiveness of the ranked results
 - Suppose users are only looking for one relevant document
 - looking for a fact
 - known-item search
 - navigational queries
 - query auto completion
- Search duration \sim Rank of the answer
 - measures a user's effort

Mean Reciprocal Rank

- Consider the rank position, K , of the first relevant document
- Reciprocal Rank = $\frac{1}{K}$
- MRR is the mean RR across multiple queries

Beyond binary relevance



Google Search Statistics - Internet Live Stats

www.internetlivestats.com/google-search-statistics/

Historical search **volume**, growth rate, and Google's share of global search market. ... launched, Google was already answering 3.5 million search **queries** daily.

Excellent

Google Annual Search Statistics | Statistic Brain

www.statisticbrain.com/google-searches/

The first funding for Google was an August 1998 contribution of US\$100,000 from ... Year, Annual Number of Google Searches, Average Searches Per Day.

Good

Insight into Google Search Query Numbers and What It ...

getstat.com/google-search-queries-the-numbers/

Jul 27, 2012 - Insights into the true meaning of Google Search **Queries** and the numbers behind it. ... Google has had an immense impact on how we operate in everyday ... Each month, the sheer **volume** of **queries** it answers continues to ...

Fair

How many search queries does Google serve worldwide ...

www.quora.com/How-many-search-queries-does-Google-serve-w...

Answer 1 of 8: This is latest data that Matt Cutts update yesterday - Google has seen more than 30 trillion URLs and crawls 20 billion pages a day. 3 billion...

Fair

Google Trends

www.google.com/trends/

50,000+ searches. Image Source - New York **Daily** News · David Wilson. 50,000+ searches. Image Source - NBCSports.com · Marilyn Burns. 20,000+ searches.

Bad

Google Trends - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Google_Trends

Google Trends also allows the user to compare the **volume** of searches between ... the information provided by Google Trends daily; Hot Trends is updated hourly ... Because the relative frequency of certain **queries** is highly correlated with the ...

Bad

Same P@6?!

Same MAP?!

Relevant

Nonrelevant

Beyond binary relevance

- The level of documents' relevance quality with respect to a given query varies
 - Highly relevant documents are more useful than marginally relevant documents
 - The lower the ranked position of a relevant document is, the less useful it is for the user, since it is less likely to be examined
 - ***Discounted Cumulative Gain***

Discounted Cumulative Gain

- Uses graded relevance as a measure of usefulness, or gain, from examining a document
- Gain is accumulated starting at the top of the ranking and discounted at lower ranks
- Typical discount is $1/\log(\text{rank})$
 - With base 2, the discount at rank 4 is $1/2$, and at rank 8 it is $1/3$

Discounted Cumulative Gain

- DCG is the total gain accumulated at a particular rank position p :

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

Relevance label at position i

- Alternative formulation

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(1 + i)}$$

- Standard metric in some web search companies
- Emphasis on retrieving highly relevant documents

Normalized Discounted Cumulative Gain

- Normalization is useful for contrasting queries with varying numbers of relevant results
- Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
 - The ideal ranking is achieved via ranking documents with their relevance labels

How about $P@4$, $P@5$, MAP and MRR?

NDCG - Example

5 documents: d_1, d_2, d_3, d_4, d_5

i	Ground Truth		Ranking Function ₁		Ranking Function ₂	
	Document Order	rel _i	Document Order	rel _i	Document Order	rel _i
1	d5	4	d3	2	d5	4
2	d4	3	d4	3	d3	2
3	d3	2	d2	1	d4	3
4	d2	1	d5	4	d1	0
5	d1	0	d1	0	d2	1

$$DCG_{GT} = \frac{2^4-1}{\log_2 2} + \frac{2^3-1}{\log_2 3} + \frac{2^2-1}{\log_2 4} + \frac{2^1-1}{\log_2 5} + \frac{2^0-1}{\log_2 6} = 21.35$$

$$DCG_{RF1} = \frac{2^2-1}{\log_2 2} + \frac{2^3-1}{\log_2 3} + \frac{2^1-1}{\log_2 4} + \frac{2^4-1}{\log_2 5} + \frac{2^0-1}{\log_2 6} = 14.38$$

$$DCG_{RF2} = \frac{2^4-1}{\log_2 2} + \frac{2^2-1}{\log_2 3} + \frac{2^3-1}{\log_2 4} + \frac{2^0-1}{\log_2 5} + \frac{2^1-1}{\log_2 6} = 20.78$$

What does query averaging hide?

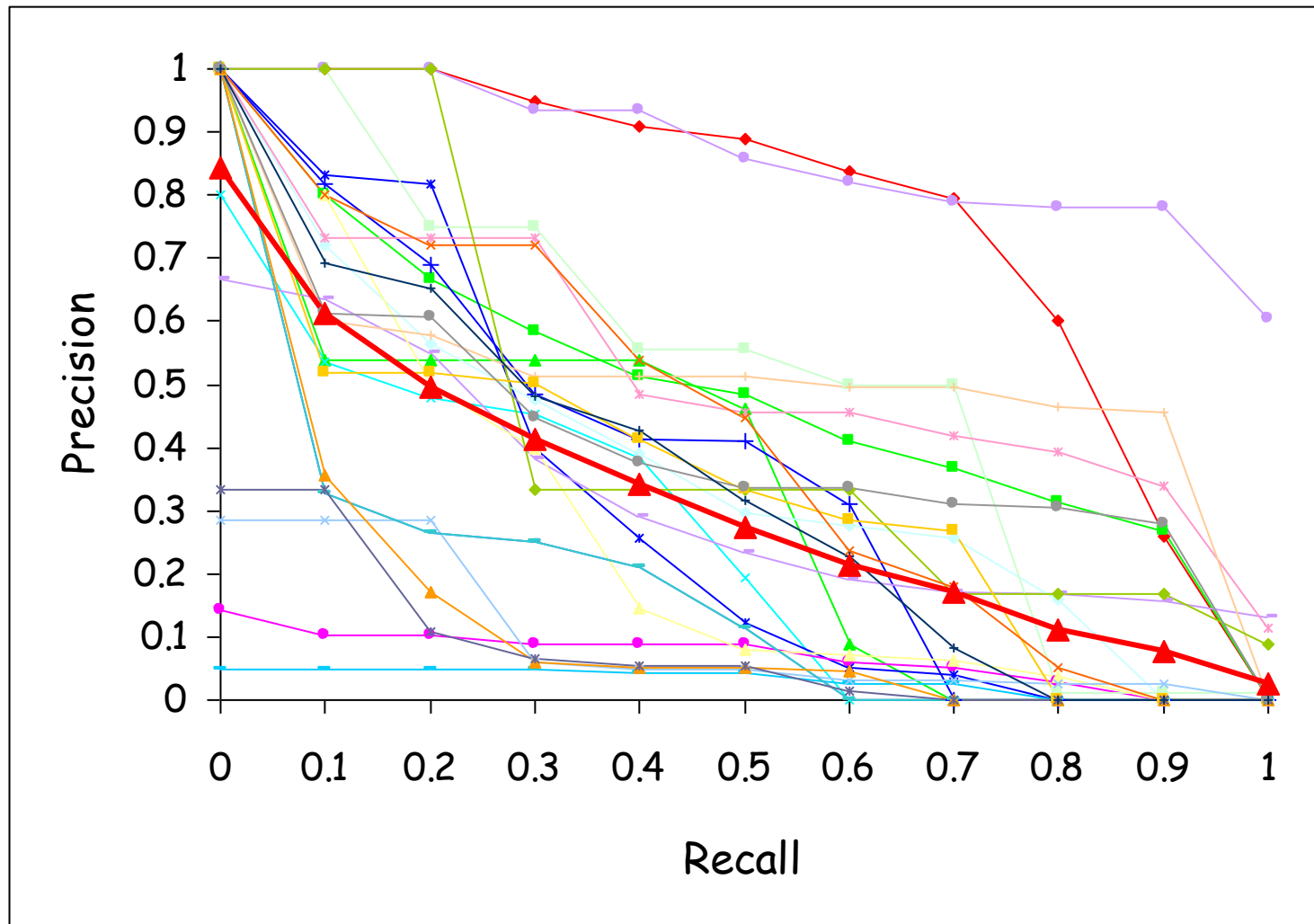


Figure from Doug Oard's presentation, originally from Ellen Voorhees' presentation

Statistical significance tests

- How confident you are that an observed difference doesn't simply result from the particular queries you chose?

Experiment 1			Experiment 2		
<u>Query</u>	<u>System A</u>	<u>System B</u>	<u>Query</u>	<u>System A</u>	<u>System B</u>
1	0.20	0.40	11	0.02	0.76
2	0.21	0.41	12	0.39	0.07
3	0.22	0.42	13	0.26	0.17
4	0.19	0.39	14	0.38	0.31
5	0.17	0.37	15	0.14	0.02
6	0.20	0.40	16	0.09	0.91
7	0.21	0.41	17	0.12	0.56
Average	0.20	0.40	Average	0.20	0.40

Background knowledge

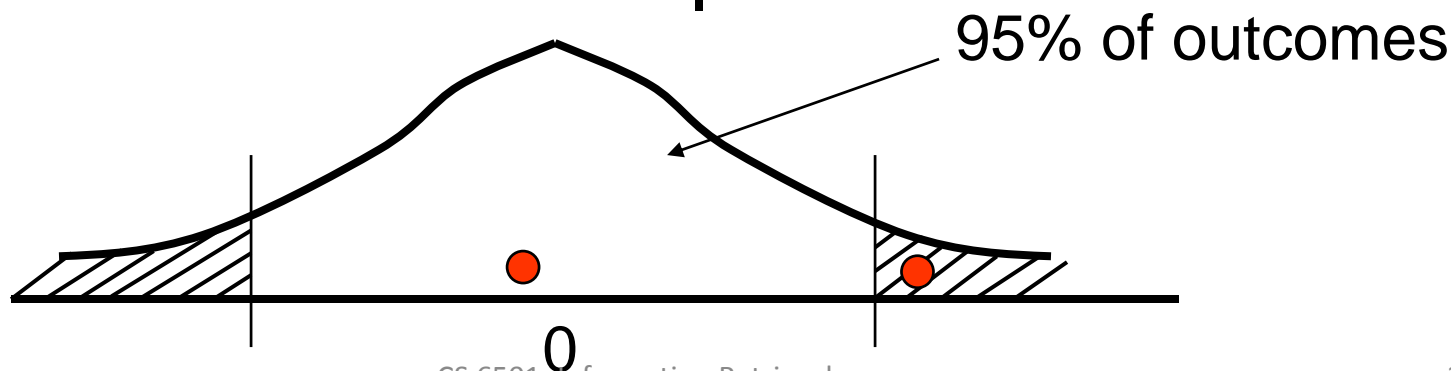
- p -value in statistic test is the probability of obtaining data as extreme as was observed, if the null hypothesis were true (e.g., if observation is totally random)
- If p -value is smaller than the chosen significance level (α), we reject the null hypothesis (e.g., observation is not random)
- We seek to reject the null hypothesis (we seek to show that the observation is a random result), and so small p -values are good

Tests usually used in IR evaluations

- Sign test
 - Hypothesis: the difference median is zero between samples from two continuous distributions
- Wilcoxon signed rank test
 - Hypothesis: data are paired and come from the same population
- Paired t -test
 - Hypothesis: difference between two responses measured on the same statistical unit has a zero mean value
- One-tail v.s. two-tail?
 - If you aren't sure, use two-tail

Statistical significance testing

<u>Query</u>	<u>System A</u>	<u>System B</u>	<u>Sign Test</u>	<u>paired t-test</u>
1	0.02	0.76	+	+0.74
2	0.39	0.07	-	-0.32
3	0.26	0.17	-	-0.09
4	0.38	0.31	-	-0.07
5	0.14	0.02	-	-0.12
6	0.09	0.91	+	+0.82
7	0.12	0.56	+	+0.44
Average	0.20	0.40	$p=0.9375$	$p=0.2927$



Where do we get the relevance labels?

- Human annotation
 - Domain experts, who have better understanding of retrieval tasks
 - Scenario 1: annotator lists the information needs, formalizes into queries, and judges the returned documents
 - Scenario 2: given query and associated documents, annotator judges the relevance by inferring the underlying information need

Assessor consistency

- Is inconsistency of assessors a concern?
 - Human annotators are idiosyncratic and variable
 - Relevance judgments are subjective
- Studies mostly concluded that the inconsistency didn't affect relative comparison of systems
 - Success of an IR system depends on how good it is at satisfying the needs of these idiosyncratic humans
 - Lesk & Salton (1968): assessors mostly disagree on documents at lower ranks, but measures are more affected by top-ranked documents

Measuring assessor consistency

- *kappa* statistic

- A measure of agreement between judges

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- $P(A)$ is the proportion of the times judges agreed
 - $P(E)$ is the proportion of times they would be expected to agree by chance

- $\kappa = 1$ if two judges always agree

- $\kappa = 0$ if two judges agree by chance

- $\kappa < 0$ if two judges always disagree

Example of *kappa* statistic

		judge 2 relevance		
		Yes	No	Total
judge 1 relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

$$P(A) = \frac{300 + 70}{400} = 0.925$$

$$P(E) = \left(\frac{80 + 90}{400 + 400} \right)^2 + \left(\frac{320 + 310}{400 + 400} \right)^2 = 0.2125^2 + 0.7875^2 = 0.665$$

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.925 - 0.665}{1 - 0.665} = 0.776$$

Prepare annotation collection

- Human annotation is expensive and time consuming
 - Cannot afford exhaustive annotation of large corpus
 - Solution: pooling
 - Relevance is assessed over a subset of the collection that is formed from the top k documents returned by a number of different IR systems

Does pooling work?

- Judgments cannot possibly be exhaustive?
 - Relative rankings among the systems remain the same
- This is only a subset of people's opinion about relevance
 - Relative rankings among the systems remain the same
- What about documents beyond top k ?
 - Relative rankings among the systems remain the same
- A lot of research work can be done here
 - Effective pool construction
 - Depth v.s. diversity

Rethink retrieval evaluation

- Goal of any IR system
 - Satisfying users' information need
- Core quality measure criterion
 - *“how well a system meets the information needs of its users.” – wiki*

What we have considered

- The ability of the system to present all relevant documents
 - Recall-driven measures
- The ability of the system to withhold non-relevant documents
 - Precision-driven measures

Challenging assumptions in classical IR evaluations

- Assumption 1
 - Queries sent to an IR system would be the same as those sent to a librarian (i.e., sentence-length request), and users want to have high recall
- Assumption 2
 - Relevance = independent topical relevance
 - Documents are independently judged, and then ranked (that is how we get the ideal ranking)

What we have not considered

- The physical form of the output
 - User interface
- The effort, intellectual or physical, demanded of the user
 - User effort when using the system
- Bias IR research towards optimizing relevance-centric metrics

What you should know

- Core criterion for IR evaluation
- Basic components in IR evaluation
- Classical IR metrics
- Statistical test
- Annotator agreement