EVALUATION IN INFORMATION RETRIEVAL

- There are many retrieval models/algorithms/systems.
- Which one is the best?
- Which is the best component for?
  - Ranking function (dot product, cosine, ..)
  - Term selection (stop word removal, stemming, ..)
  - Term weighting (TF, TF-IDF, ..)
- Effectiveness is related to the relevancy of retrieved items.
- Relevancy from a human stand point is,
  - Subjective → depends upon a specific user's judgment.
  - Situational → relates to user's current needs.
  - Cognitive → depends on human perception.
  - Dynamic → changes over time.
- Key utility measure is user happiness.
- Speed of response is a factor of user happiness.
- But blindingly fast with useless answers do not make a user happy.
- The standard approach to IR system evaluation revolves around the notion of relevant and non-relevant documents.
- With a user information need, a document in the collection is given a binary classification as either relevant or non-relevant.
- This decision is referred to as the gold standard or ground truth judgment of relevance.
- Relevance is assessed relative to an information need, not a query.
- For example: an information need might be "Information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine".
- This might be translated to a query such as "Wine AND red AND white AND heart AND attack AND effective".
- A document is relevant if it addresses the stated information need not because it just happens to contain all the words in the query.

- If a user type "Python" into a web search engine, they might want to know where they can purchase a pet python, or they might want information on the programming language "python".
- So, from a one word query, it is very difficult for a system to know what an information need is.

EVALUATION OF UNRANKED RETRIEVAL SETS

- There are two measures:
  - Precision (P)
    - Precision is the fraction of retrieved documents that are relevant.
    - Indicates what proportion of the retrieved documents is relevant.
    - Precision = $\frac{\#(relevant\ \ time\ \ retrieved)}{\#(retrieved\ \ items)}$ = P(relevant/retrieved)
  - Recall (R)
    - Recall is the fraction of relevant document that are retrieved.
    - Indicates what proportion of all the relevant documents have been retrieved from the collection.
    - Recall = $\frac{\#(relevant\ \ items\ \ retrieved)}{\#(relevant\ \ items}$ = P(retrieved/relevant)

QUESTION:

An IR system returns 8 relevant documents and 10 non-relevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search and what was its recall?

- Precision and recall can also be expressed in the following terms:

	Relevant	Non-Relevant
Retrieved	True Positive (tp)	False Positive (fp)
Not-retrieved	False Negative (fn)	True Negative (tn)

- Precision = $\frac{tp}{tp+fp}$ and Recall = $\frac{tp}{tp+fn}$
- An alternative way to judge an information retrieval system is by its accuracy.
- Accuracy is the fraction of its classification that is correct.

- Accuracy = $\frac{tp+tn}{tp+fp+fn+tn}$
- A single measure that trades off precision versus recall is the F-measure which is the weighted harmonic mean of precision and recall.

  i.e. F-measure = $\frac{2PR}{P+R}$
- *Why do we use harmonic mean rather than the simple arithmetic mean?*
  - It is because we can get 100% recall by just returning all documents and therefore we can always get a 50% arithmetic mean by the same process.
  - This strongly suggests that the arithmetic mean in an unsuitable measure to USP.
  - The harmonic mean is always less than or equal to the arithmetic mean and geometric mean.

EVALUATION OF RANKED RETRIEVAL RESULTS

- Number of relevant = 6

n	doc #	Relevant
1	588	X → P = ? R = ? P = 1/1 R = 1/6
2	589	X → P = ? R = ? P = 2/2 R = 2/6
3	576	
4	590	X → P = ? R = ? P = 3/4 R = 3/6
5	986	
6	592	X → P = ? R = ?
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	X → P = ? R = ?
14	990	

- In a ranked retrieval context, the set of retrieved documents are given by the top k retrieved documents.
- If $(k+1)^{th}$ document retrieved is non-relevant then recall is the same as for the top k documents but precision is dropped.

- If it is relevant, then both precision and recall increases.

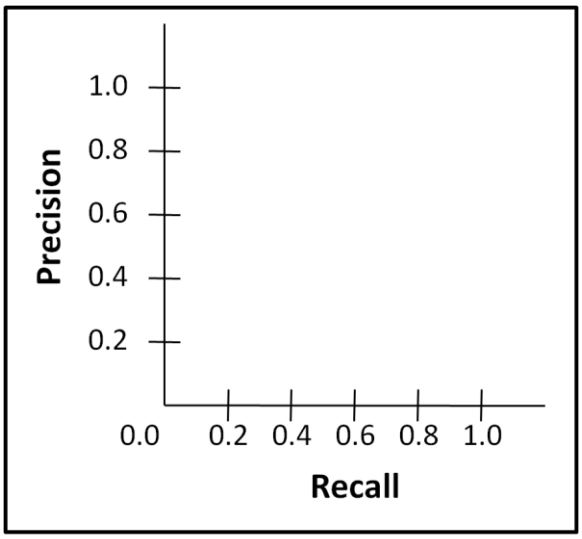

Fig: Precision/Recall Graph (Saw-Tooth Shape)

SYSTEM QUALITY

- There are many practical bench marks on which to rate an IR system beyond its retrieval quality, which includes:
  - How fast does it index? [i.e. how many documents per hour does it can index.]
  - How fast does it search?
  - How expressive is its query language? How fast is it on complex queries?
  - How large is its document collection?

USER UTILITY

- A way of quantifying user happiness is based on relevance, speed and user interface of a system.
- For a web search engine, happy search users are those who find what they want.
- One indirect measure of such users is that they tend to return the same engine.
- Advertisers are also users of modern web search engines and they are happy, if customer's clicked through to their sites and then make purchases.