Information Retrieval

- ***Information Retrieval (IR):*** retrieving desired information from textual data.
- Library Science
- Digital Libraries
- Web Search Engines
- Traditionally keyword based
- Sample query:

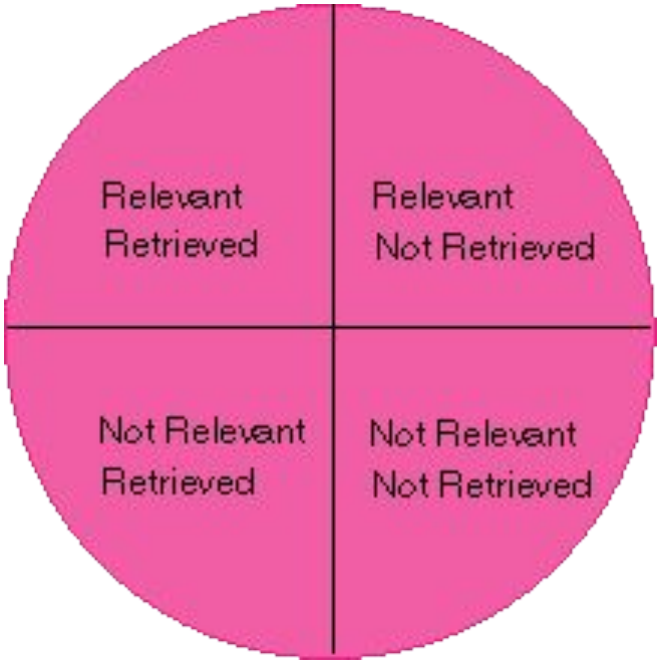
  Find all documents about “data mining”.

DM: Similarity measures;
Mine text/Web data.

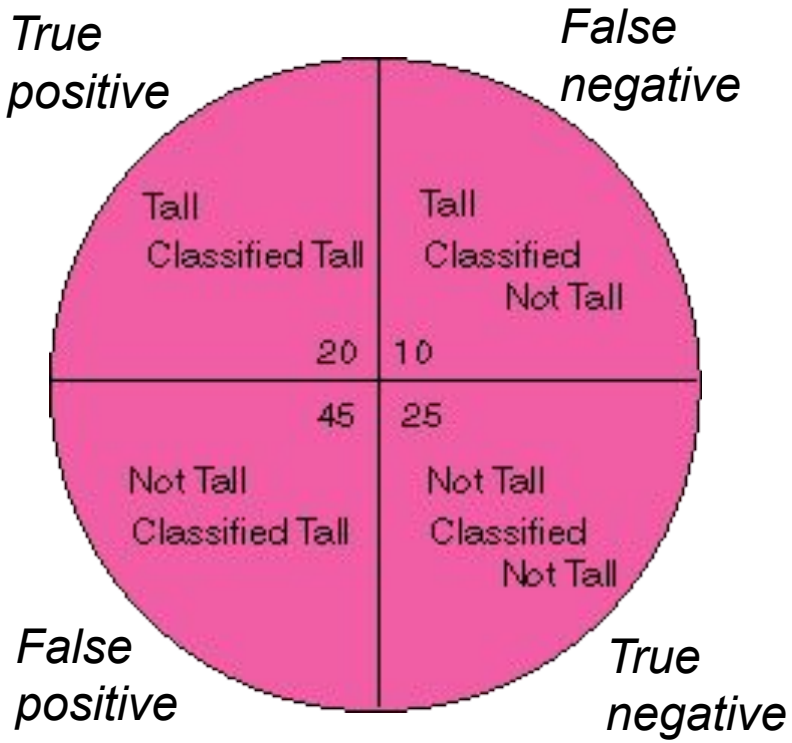
Information Retrieval (cont'd)

- ***Similarity:*** measure of how close a query is to a document.
- Documents which are "close enough" are retrieved.
- Metrics:
  - ***Precision*** = $\frac{|\text{Relevant and Retrieved}|}{|\text{Retrieved}|}$
  - ***Recall*** = $\frac{|\text{Relevant and Retrieved}|}{|\text{Relevant}|}$
- ***F1 score*** = 2*Precision*Recall / (Precision+Recall)

IR Query Result Measures and Classification



IR



Classification

Precision = *TP / (TP + FP)*

Recall = *TP / (TP + FN)*

Project III

- Cancer diagnosis using machine learning
- Data set:
  - Labels: B (benign), M (Malignant)
  - Features: 20 clinical variables
  - 569 samples
  - On brightspace
- Select two machine learning algorithms (e.g. Random Forest) to train and evaluate prediction models and compare their performances
- Using Spark MLLib
  - https://spark.apache.org/mllib/

Project III

- Due: 11:59pm, 12/18/25
- Codes
- Readme
- Report
  - Justification of the algorithms that you choose
  - Your training procedure
    - Data splits, cross-validation etc.
  - Your testing results
    - Report using F1 score, precision, recall,and accuracy(% #correct predictions/#all samples).
    - Comparison of two algorithms.
  - Discussion
    - Limitations and future improvement.
- Email to book a zoom meeting for any questions.

Project III

- Rubric
  - Algorithm justifications - 30
  - Training procedure - 30
  - Coding -20
  - Reporting -20

Teacher's Evaluation

- phone: www.hunter.cuny.edu/mobilete
- Computer: www.hunter.cuny.edu/te