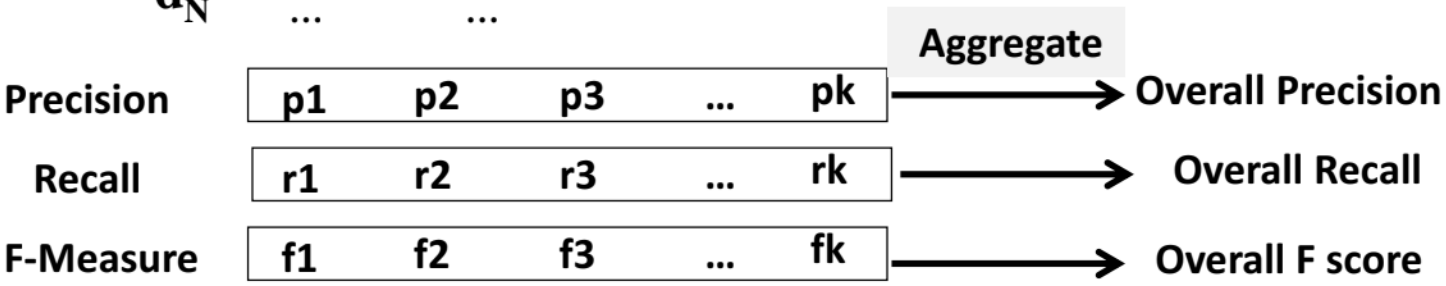
Text Categorization: Evaluation

Part 2

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

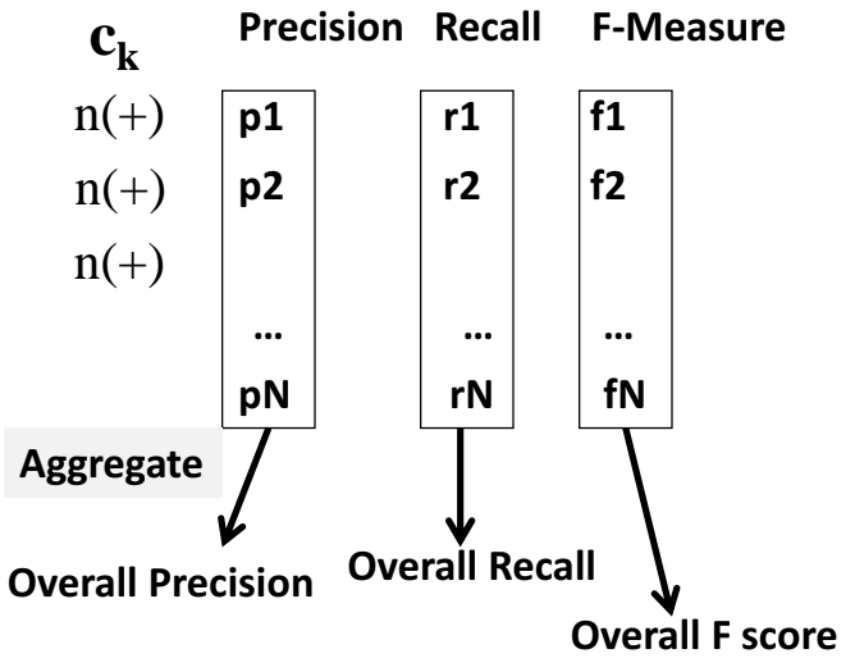
(Macro) Average Over All the Categories

	c_1	c_2	c_3	…	c_k
d_1	y(+)	y(-)	n(+)		n(+)
d_2	y(-)	n(+)	y(+)		n(+)
d_3	n(+)	n(+)	y(+)		n(+)
…					
d_N	…	…			



(Macro) Average Over All the Documents

	c_1	c_2	c_3	…	c_k
d_1	y(+)	y(-)	n(+)		n(+)
d_2	y(-)	n(+)	y(+)		n(+)
d_3	n(+)	n(+)	y(+)		n(+)
…					
d_N					



Micro-Averaging of Precision and Recall

	c_1	c_2	c_3	…	c_k
d_1	y(+)	y(-)	n(+)		n(+)
d_2	y(-)	n(+)	y(+)		n(+)
d_3	n(+)	n(+)	y(+)		n(+)
…					
d_N	…		…		

First pool all decisions, then compute precision and recall

	System ("y")	System ("n")
Human (+)	True Positives (TP)	False Negatives (FN)
Human (-)	False Positives(FP)	True Negatives(TN)

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

Sometimes Ranking Is More Appropriate

- The categorization results are often passed to a human for
  - further editing (e.g., correcting system mistakes on news categories)
  - prioritizing a task (e.g., routing an email to the right person for processing)
- In such cases, we can evaluate the results as a ranked list if the system can give scores for the decisions
  - E.g., discovery of spam emails (➔ rank emails for the “spam” category)
  - Often more appropriate to frame the problem as a ranking problem instead of a categorization problem (e.g., ranking documents in a search engine)

Summary of Categorization Evaluation

- Evaluation is always very important, so get it right!
- Measures must reflect the **intended use** of the results for a particular application (e.g., spam filtering vs. news categorization)
  - Consider: How will the results be further processed (by a user)?
  - Ideally associate a different cost with each different decision error
- Commonly used measures for **relative** comparison of different methods:
  - Accuracy, precision, recall, F score
  - Variations: per-document, per-category, micro vs. macro averaging
- Sometimes **ranking** may be more appropriate

Suggested Reading

- Manning, Chris D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2007. (Chapters 13-15)
- Yang, Yiming. 1999. An Evaluation of Statistical Approaches to Text Categorization. *Inf. Retr.* 1, 1-2 (May 1999), 69-90. DOI=10.1023/A:100998222029O