EDGAR

Jonathan Lung

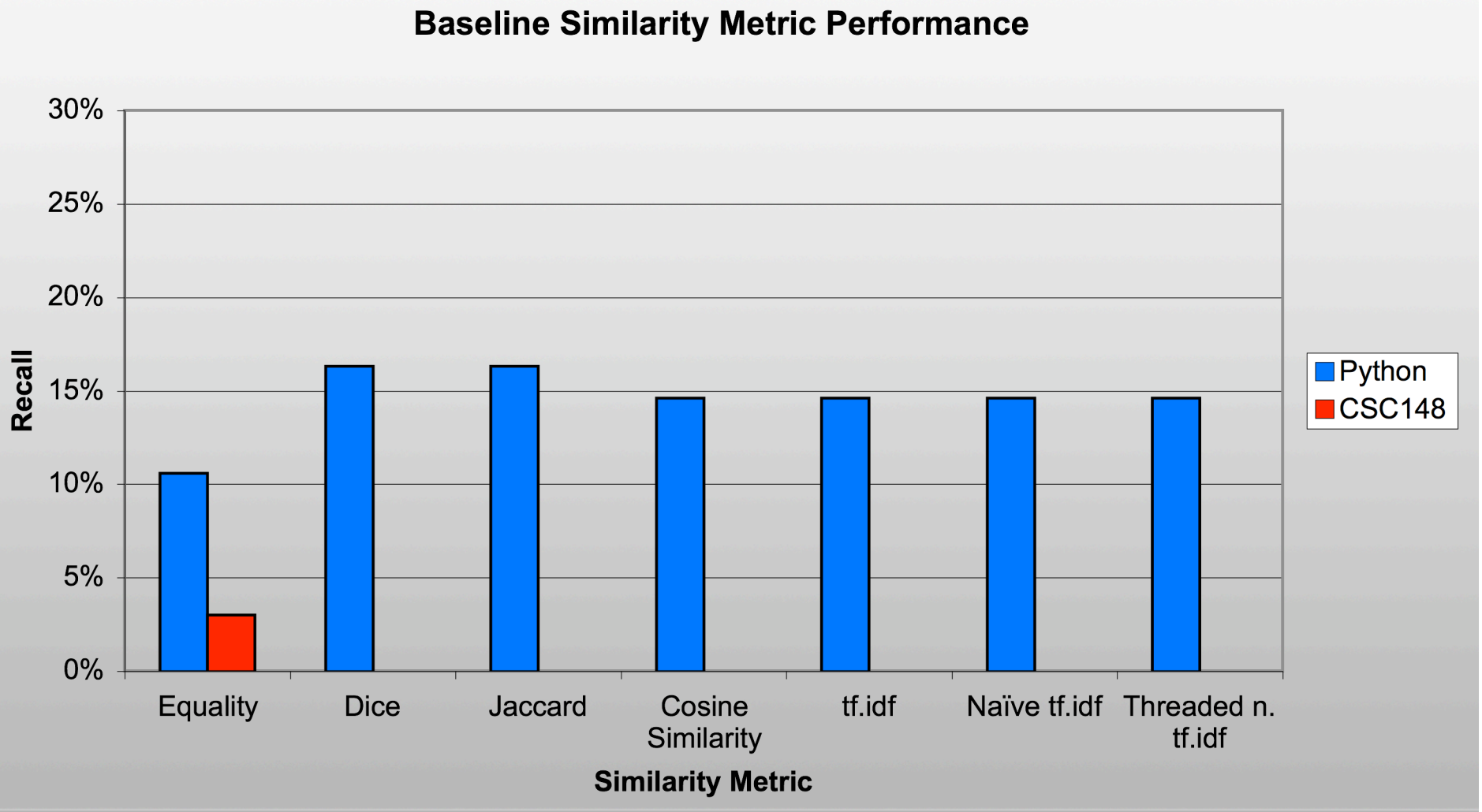
EDGAR

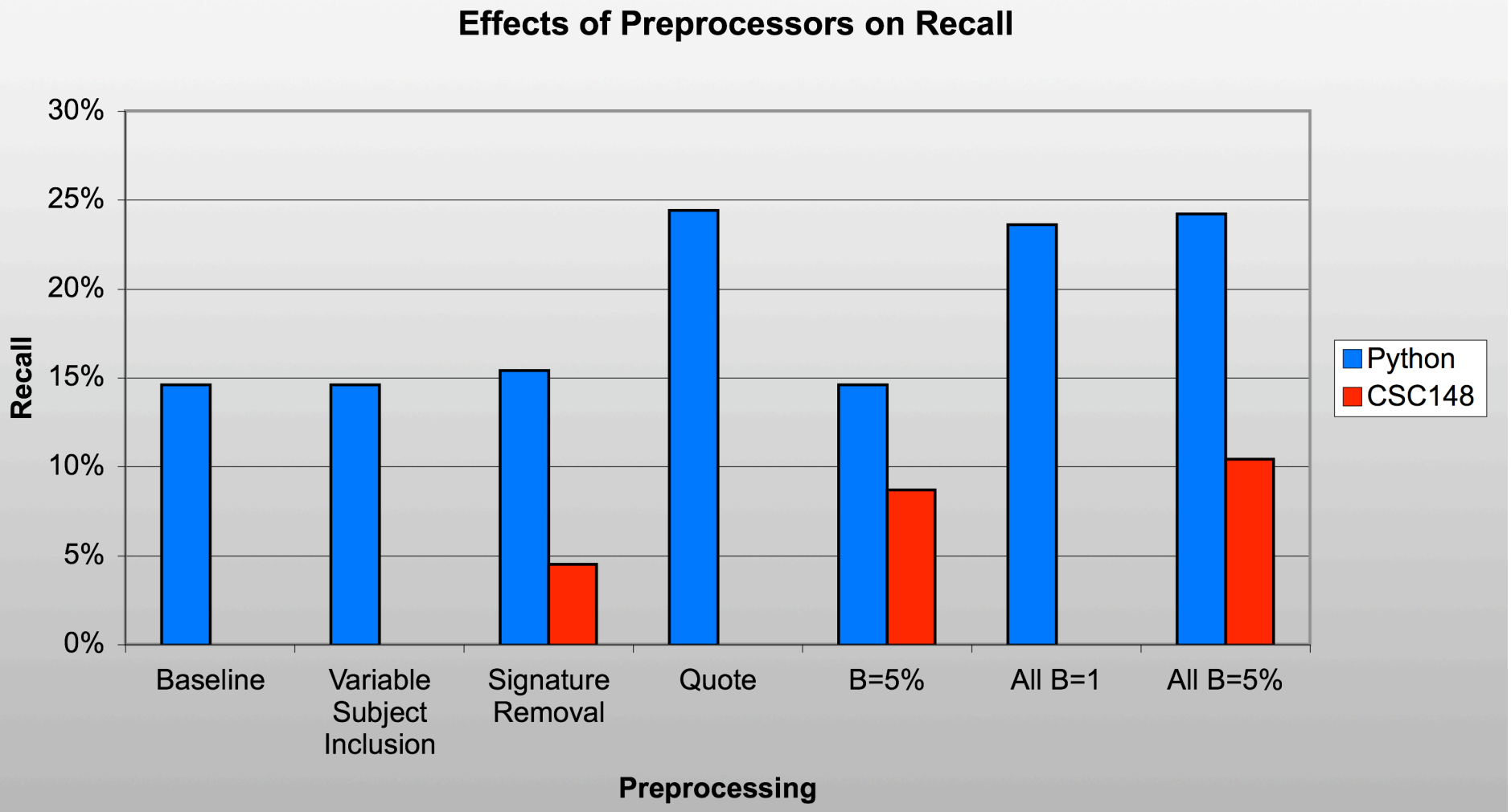
Goals

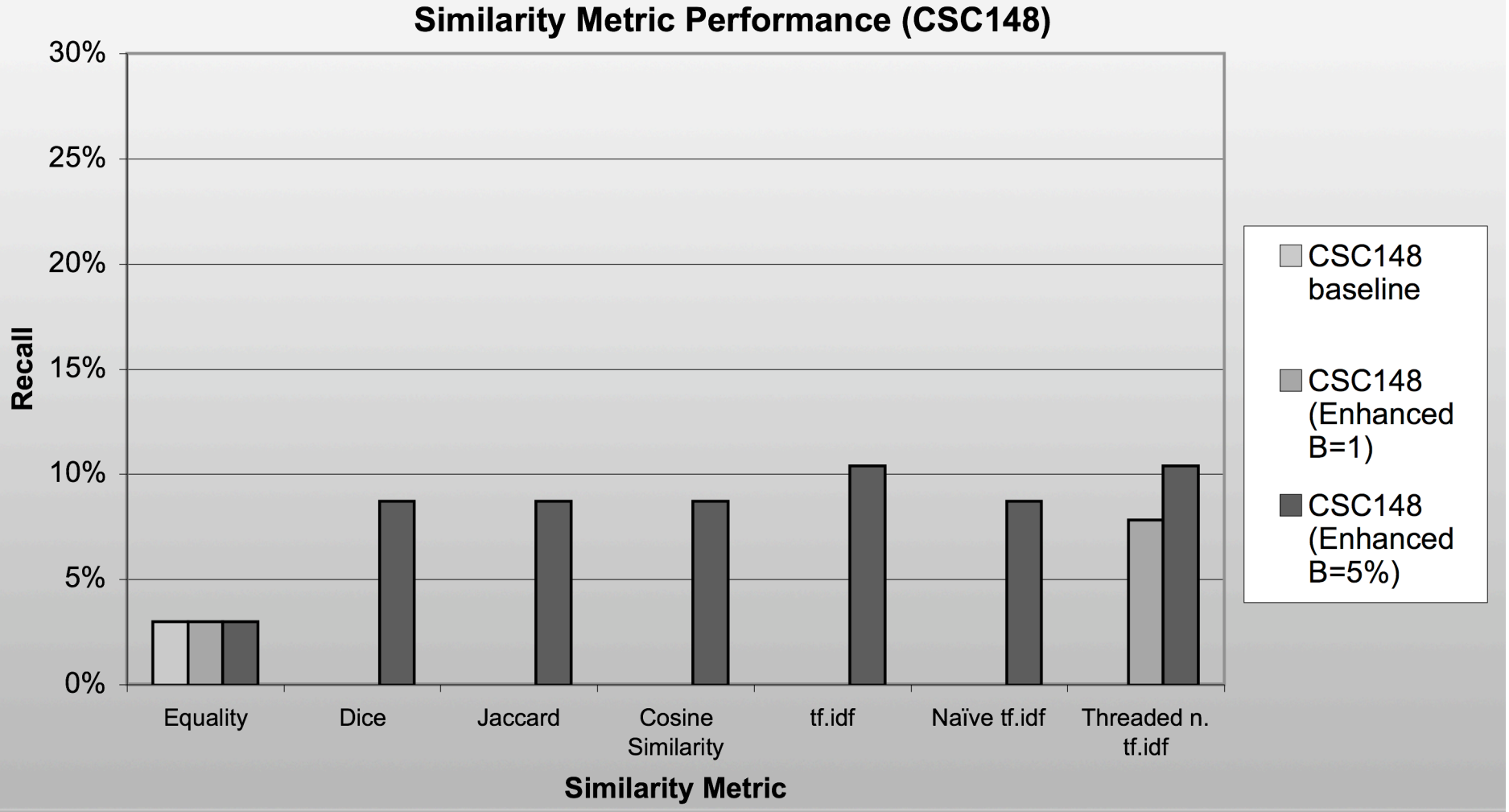
- Reduce noise in electronic fora
  - Filtering
  - Throttling
- No human intervention
- Not domain-specific
- Can be incorporated into existing software

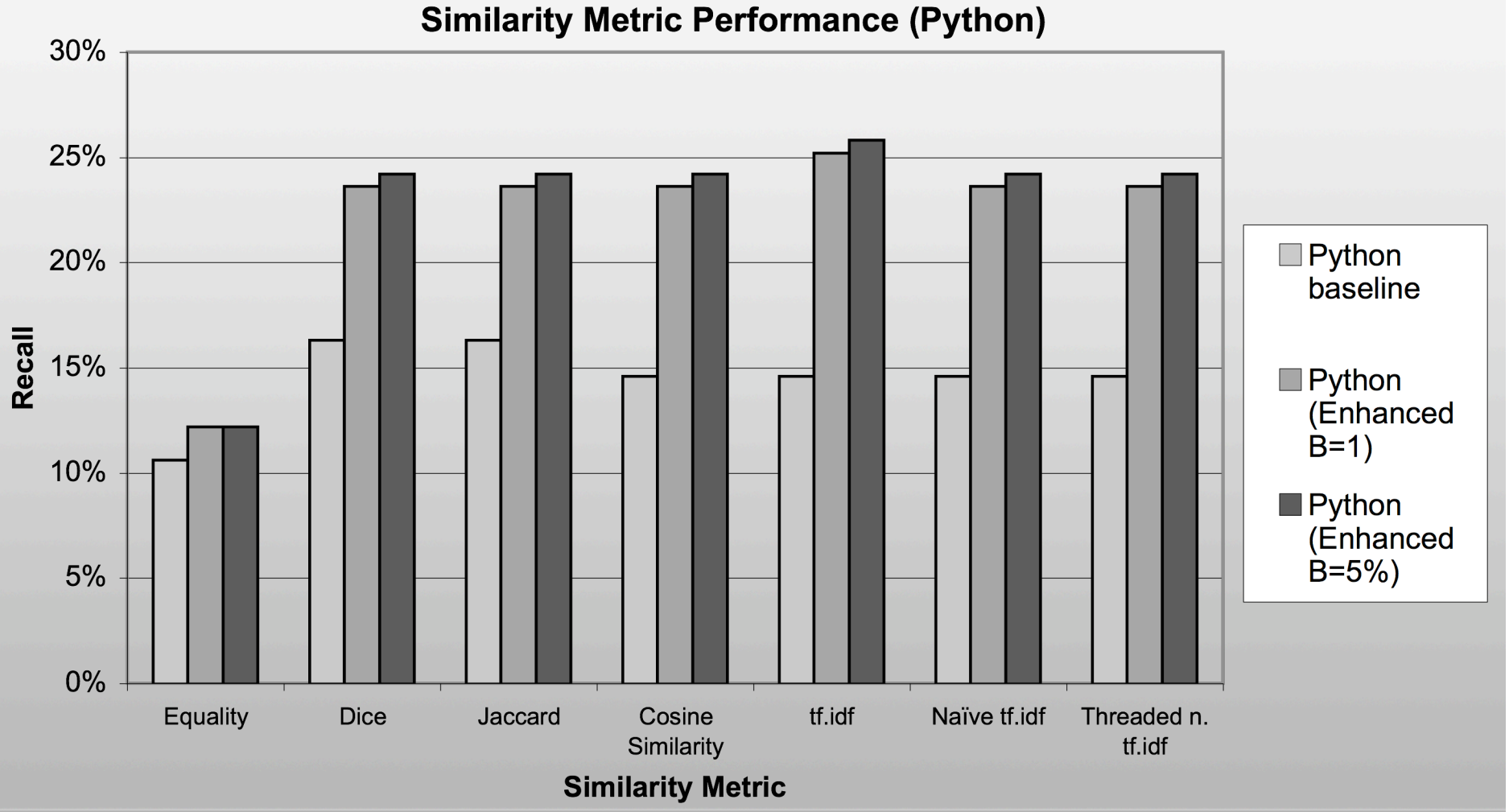
Comparing Algorithms

- Precision and recall
- Parameter adjustment
- Corpora
  - Python mailing list
  - CSC148 newsgroups

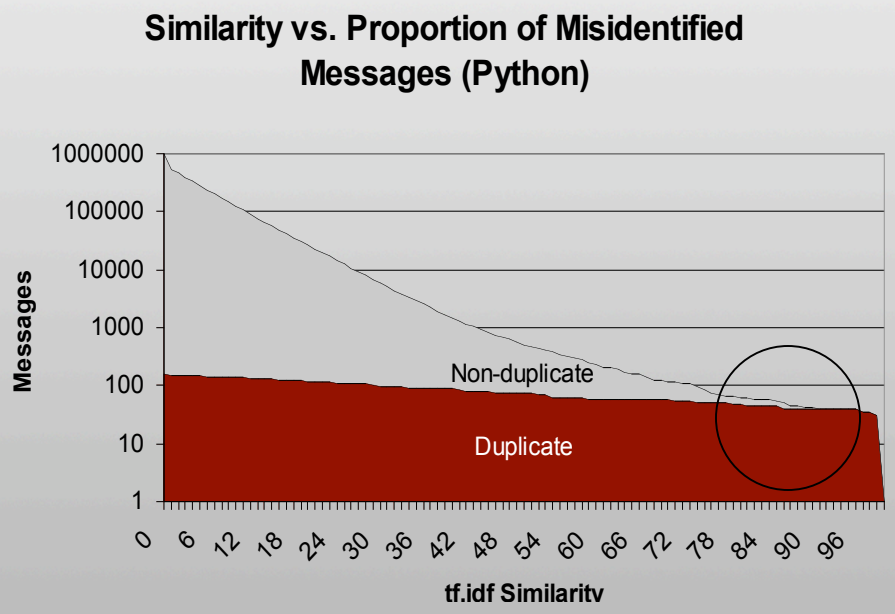

Baseline Similarity Metric Performance
30%
25%
20%
15%
10%
5%
0%
Recall
Equality
Dice
Jaccard
Cosine Similarity
tf.idf
Naïve tf.idf
Threaded n. tf.idf
Similarity Metric
Python
CSC148

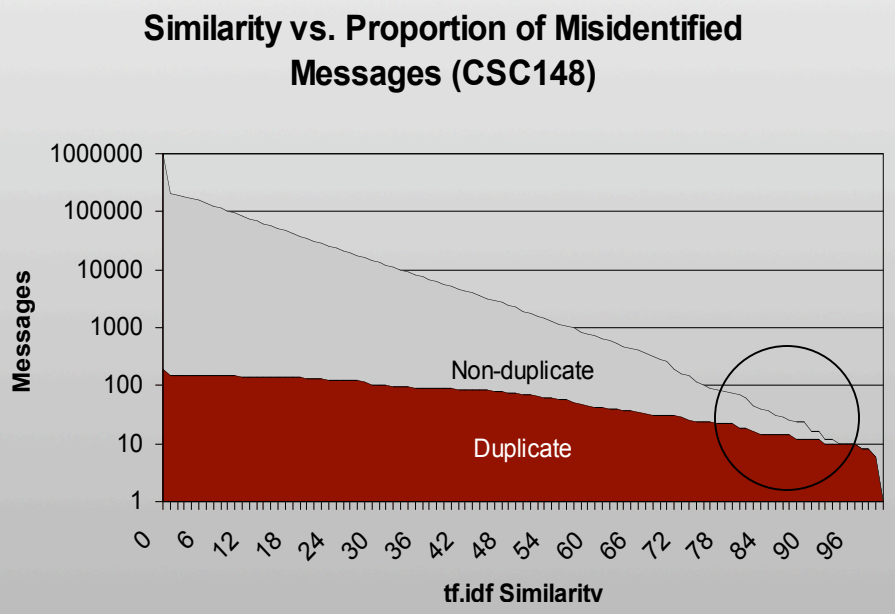

Effects of Preprocessors on Recall
30%
25%
20%
15%
10%
5%
0%
Recall
Baseline
Variable Subject Inclusion
Signature Removal
Quote
B=5%
All B=1
All B=5%
Preprocessing
Python
CSC148


Similarity Metric Performance (CSC148)
30%
25%
20%
15%
10%
5%
0%
Recall
Equality
Dice
Jaccard
Cosine Similarity
tf.idf
Naïve tf.idf
Threaded n. tf.idf
Similarity Metric
CSC148 baseline
CSC148 (Enhanced B=1)
CSC148 (Enhanced B=5%)


Similarity Metric Performance (Python)
30%
25%
20%
15%
10%
5%
0%
Recall
Equality
Dice
Jaccard
Cosine Similarity
tf.idf
Naïve tf.idf
Threaded n. tf.idf
Similarity Metric
Python baseline
Python (Enhanced B=1)
Python (Enhanced B=5%)

Comparing the Corpora





Conclusion

- Best ROI for improving pre-processors
- Room for improvement
- Many possible enhancements
- More data still required

Questions