An Empirical Comparison of Text Categorization Methods

Ana Cardoso-Cachopo and Arlindo L. Oliveira

acardoso@gia.ist.utl.pt and aml@inesc-id.pt

Instituto Superior Técnico / ALGOS-INESC-ID

Outline

- Data sets
- Information Retrieval methods
- Evaluation
- Experimental setup
- Results
- Conclusions

Data sets

- C10 (in Portuguese)
  - 461 help desk messages, with answers
  - 10 classes, 34 to 58 messages each
- mini20 (in English)
  - #2000 subset of 20Newsgroups
  - 100 messages for each newsgroup
- pre-processing
  - Discard words shorter than 3 characters
  - Discard words longer than 20 characters
  - Remove numbers and non-letter characters
  - Case and special character unification

IR methods

- Vector model
- Latent Semantic Analysis/Indexing (LSA)
- Support Vector Machines (SVM)
- k-NN Vector
- k-NN LSA

IR methods — Vector

- Words $\sim$ terms
- Docs are vectors in an N-dimensional space
- Similarity between docs is the cosine of the angle formed by the vectors representing the docs
- A doc's class is the class of the most similar doc

IR methods – LSA

- Words $\sim$ terms
- Docs are vectors in an N-dimensional space
- Apply Singular Value Decomposition
- (M $\ll$ N)-dimensional space representing concepts
- Similarity between docs is the cosine of the angle formed by the vectors representing the docs in this lower-dimensional space
- A doc's class is the class of the most similar doc

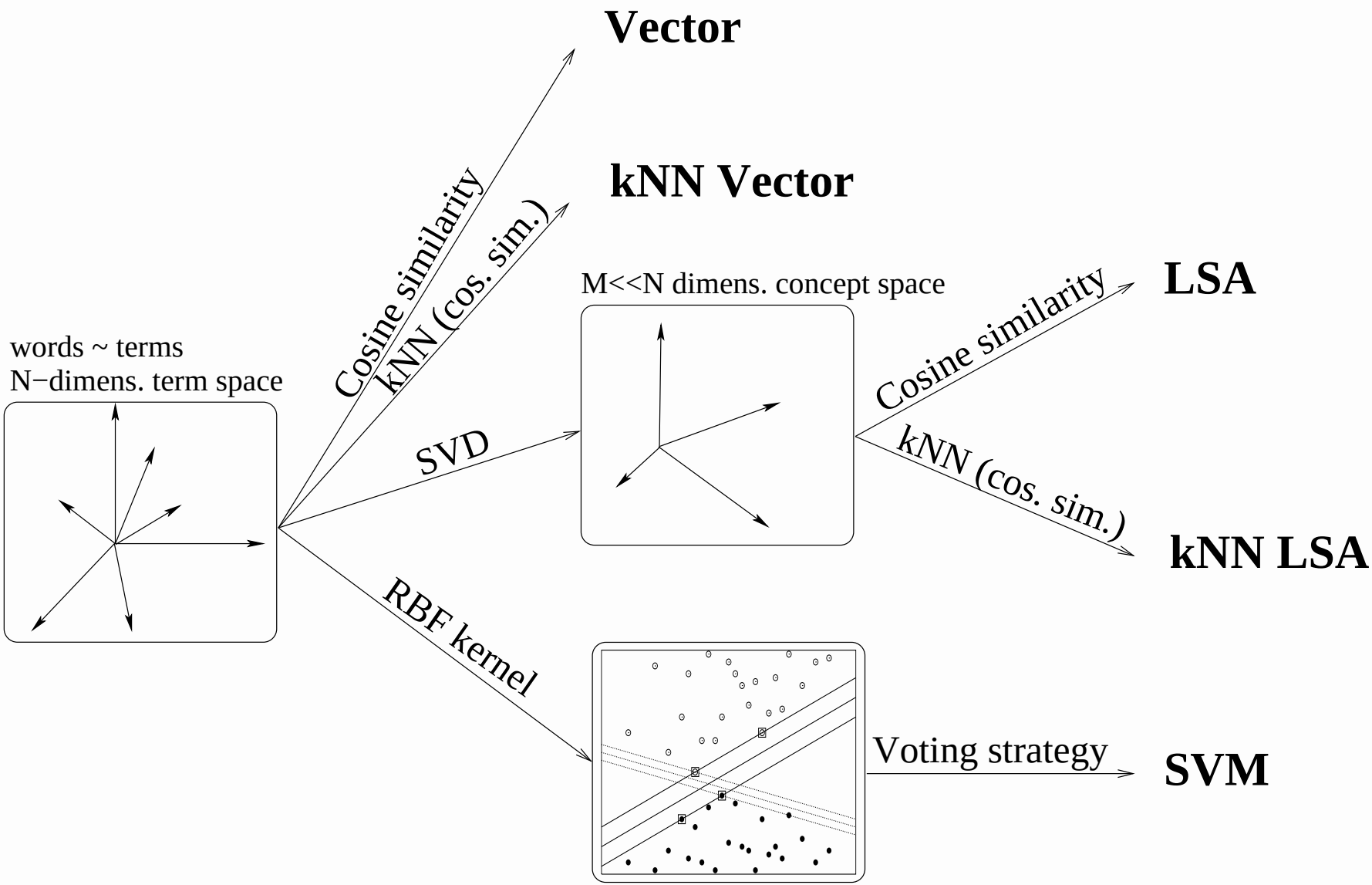
IR methods – SVM

- Words $\sim$ terms
- Docs are vectors in an N-dimensional space
- Transform the space using a kernel function
- Find a decision surface for each class that separates it from the others
- One-against-one or one-against-all approach for multiclass problems
- A doc belongs to the class that had more "belongs" votes

IR methods — k-NN Vector / k-NN LSA

- Words $\sim$ terms
- Docs are vectors in an N-dimensional space
- A doc's class is the most wheighted class among its k neighbours
- The weight is the cosine similarity in the Vector/LSA space

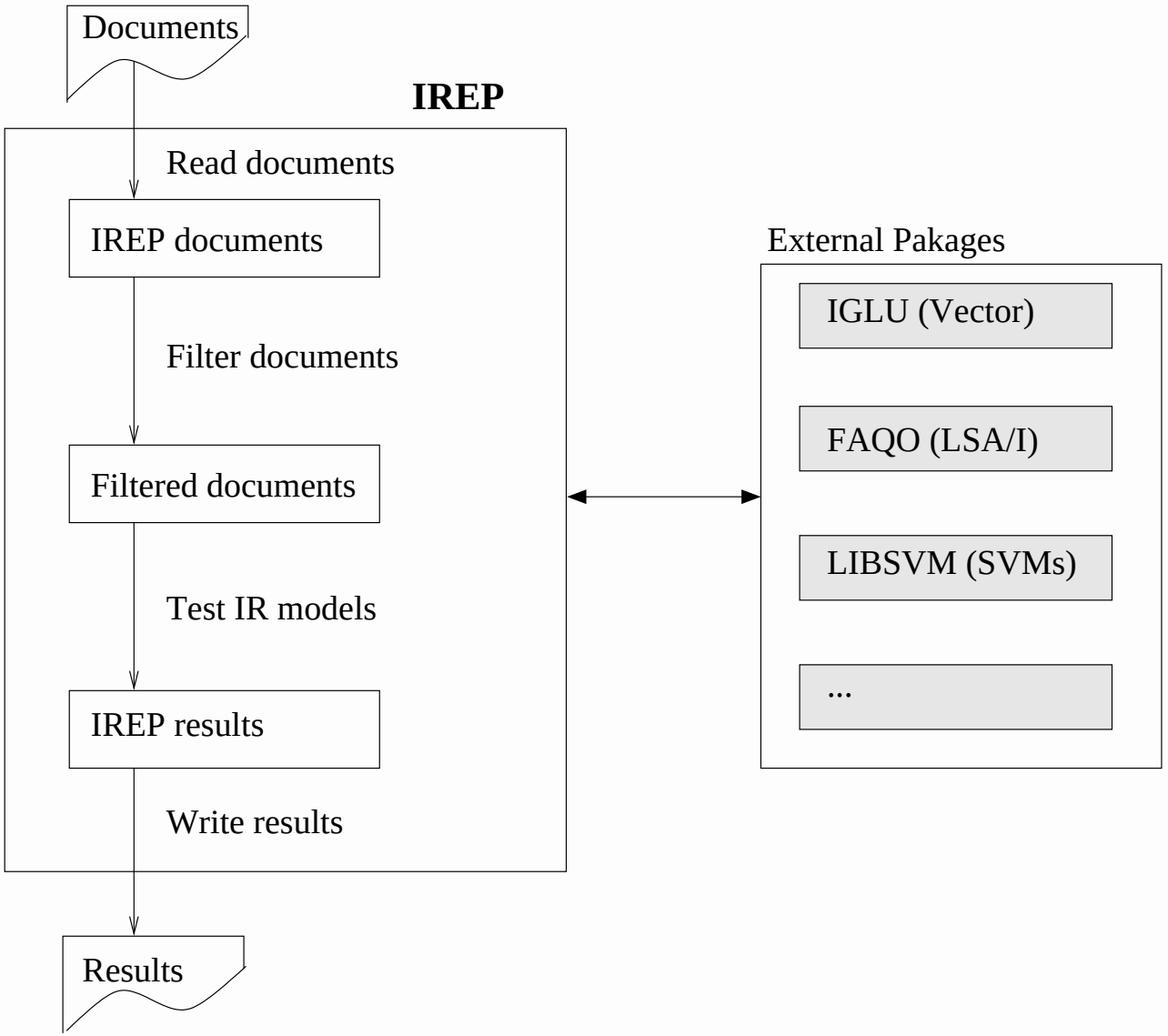
IR methods – overview



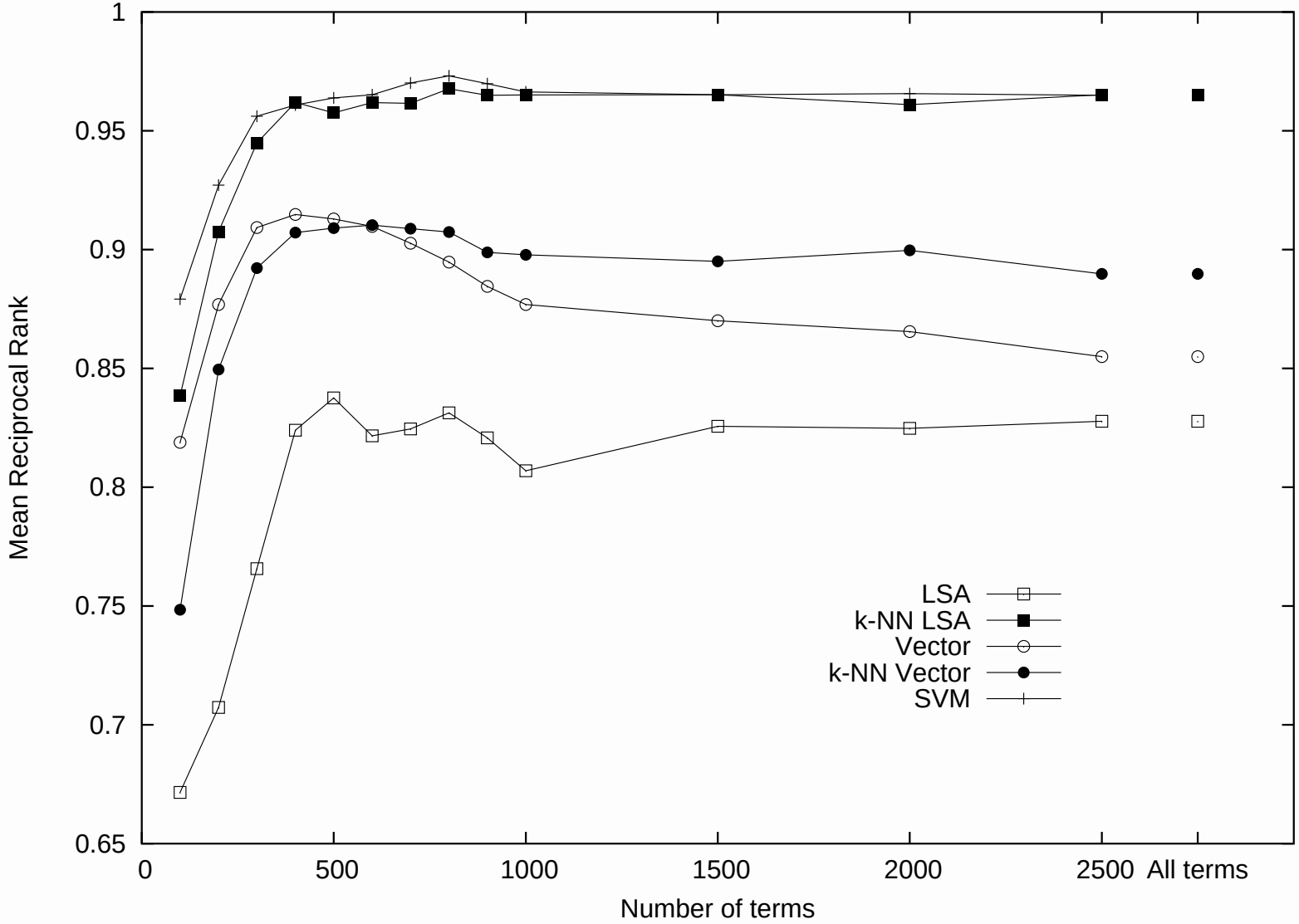
Evaluation

- Text Categorization task
- Each document has ONE category
  (Recall is not important)
- The rank of the first correct answer is important
  (Precision is not enough)
- Preferably one single number
- Mean Reciprocal Rank (MRR)
  The MRR of each individual query is the reciprocal of the rank at which the first correct response was returned, or 0 if none of the first N responses contained a correct answer.
  The score for a sequence of queries is the mean of the individual query's reciprocal ranks.

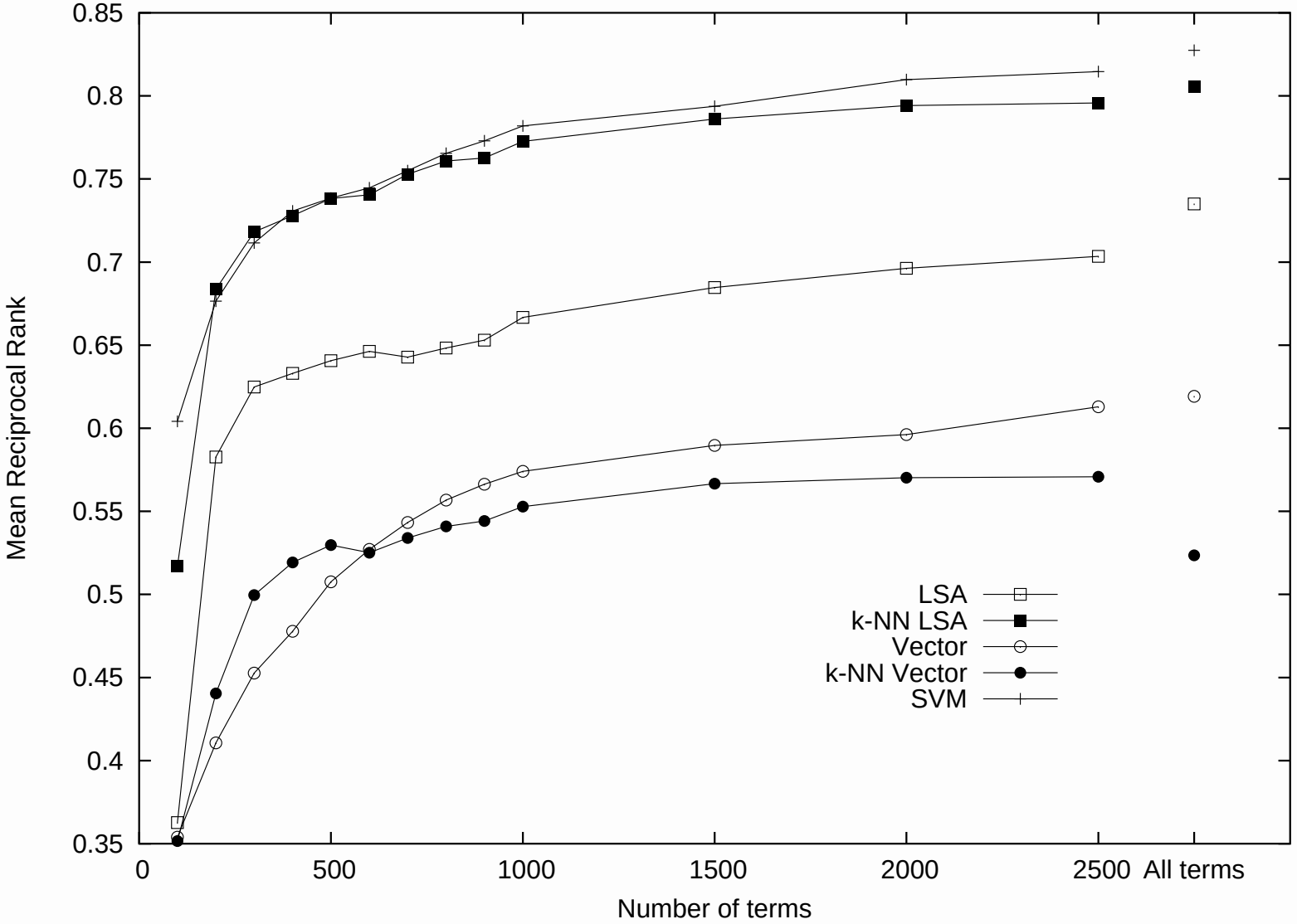
Experimental setup



Results – C10


Mean Reciprocal Rank
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
500
1000
1500
2000
2500
All terms
Number of terms
LSA
k-NN LSA
Vector
k-NN Vector
SVM

Results – mini20



Significance Tests – C10

Results of the t-test for dataset C10

	SVM	k-NN LSA	LSA	k-NN Vector	Vector
SVM	–	$\sim$, 0.1581	$\gg$, 0.0001	$\gg$, 0.0004	$\gg$, 0.0012
k-NN LSA		–	$\gg$, 0.0001	$\gg$, 0.0007	$\gg$, 0.0020
LSA			–	$\ll$, 0.0026	$\ll$, 0.0038
k-NN Vector				–	$\sim$, 0.7221
Vector					–

$$\{\text{k-NN LSA}, \text{SVM}\} \gg \{\text{Vector}, \text{k-NN Vector}\} \gg \text{LSA}$$

Significance Tests – mini20

Results of the t-test for dataset mini20

	SVM	k-NN LSA	LSA	k-NN Vector	Vector
SVM	–	$\gg$, 0.0010	$\gg$, 0.0001	$\gg$, 0.0000	$\gg$, 0.0001
k-NN LSA		–	$\gg$, 0.0001	$\gg$, 0.0000	$\gg$, 0.0000
LSA			–	$\gg$, 0.0004	$\gg$, 0.0000
k-NN Vector				–	$\ll$, 0.0079
Vector					–

SVM $\gg$ k-NN LSA $\gg$ LSA $\gg$ Vector $\gg$ k-NN Vector

Conclusions

- 2500 is a good upper-bound for the number of terms
- k-NN LSA $\sim$ SVM
- Both are significantly better than the others
- MRR is useful for one-class Text Categorization tasks