Event Detection in Newspaper Texts

Nikola Ljubešić, senior research assistant
Department of Information Sciences
University of Zagreb

JOTA, 28 October 2010

Overview

1. What is event detection
2. Document clustering
3. The gold standard
4. Experimental setup
5. Results
6. Further steps

Event detection

- event - a particular thing that happens at a specific time and place (TDT, 2004)
- event detection - process of detecting an event description in a piece of information
- part of the topic detection and tracking problem set
- document : event == 1 : 1?

Classification problem

- events are categories - classification task

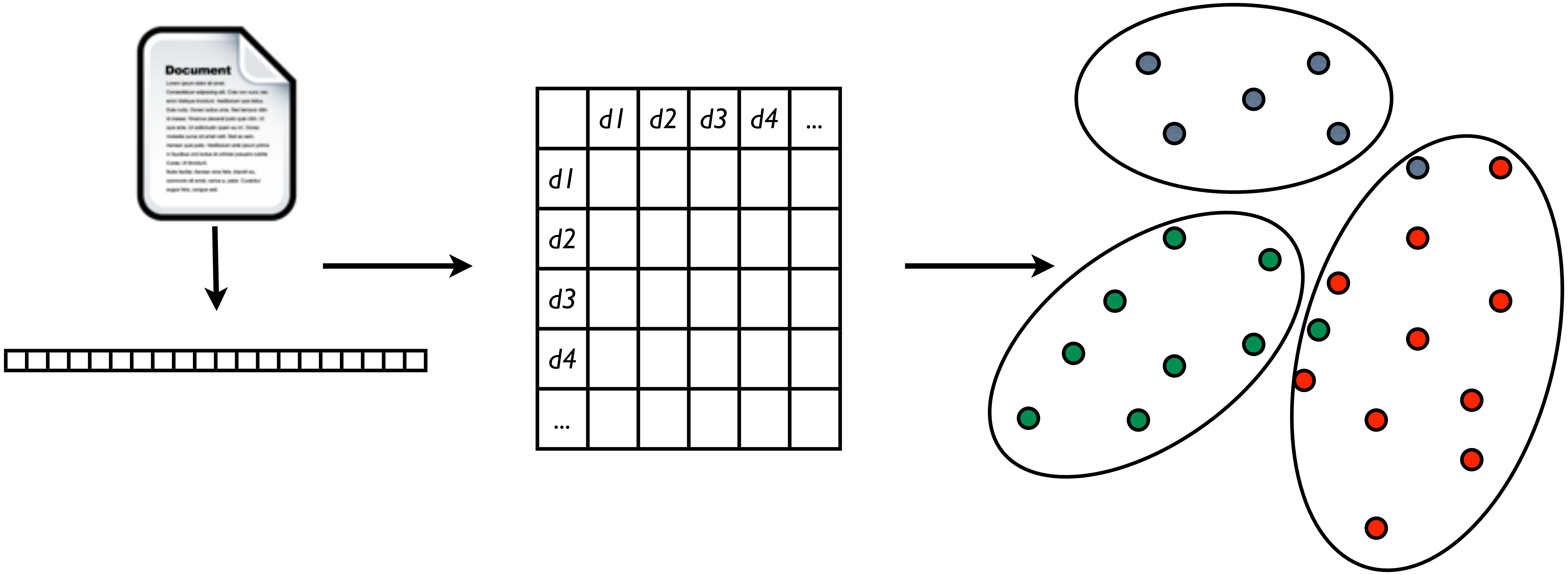
1. unknown classification schema - solvable only by unsupervised classification - clustering
2. unknown number of events - unknown number of classes - hierarchical clustering

Document clustering

document formalization

distance matrix

clustering



	d1	d2	d3	d4	...
d1					
d2					
d3					
d4					
...					

Gold standard

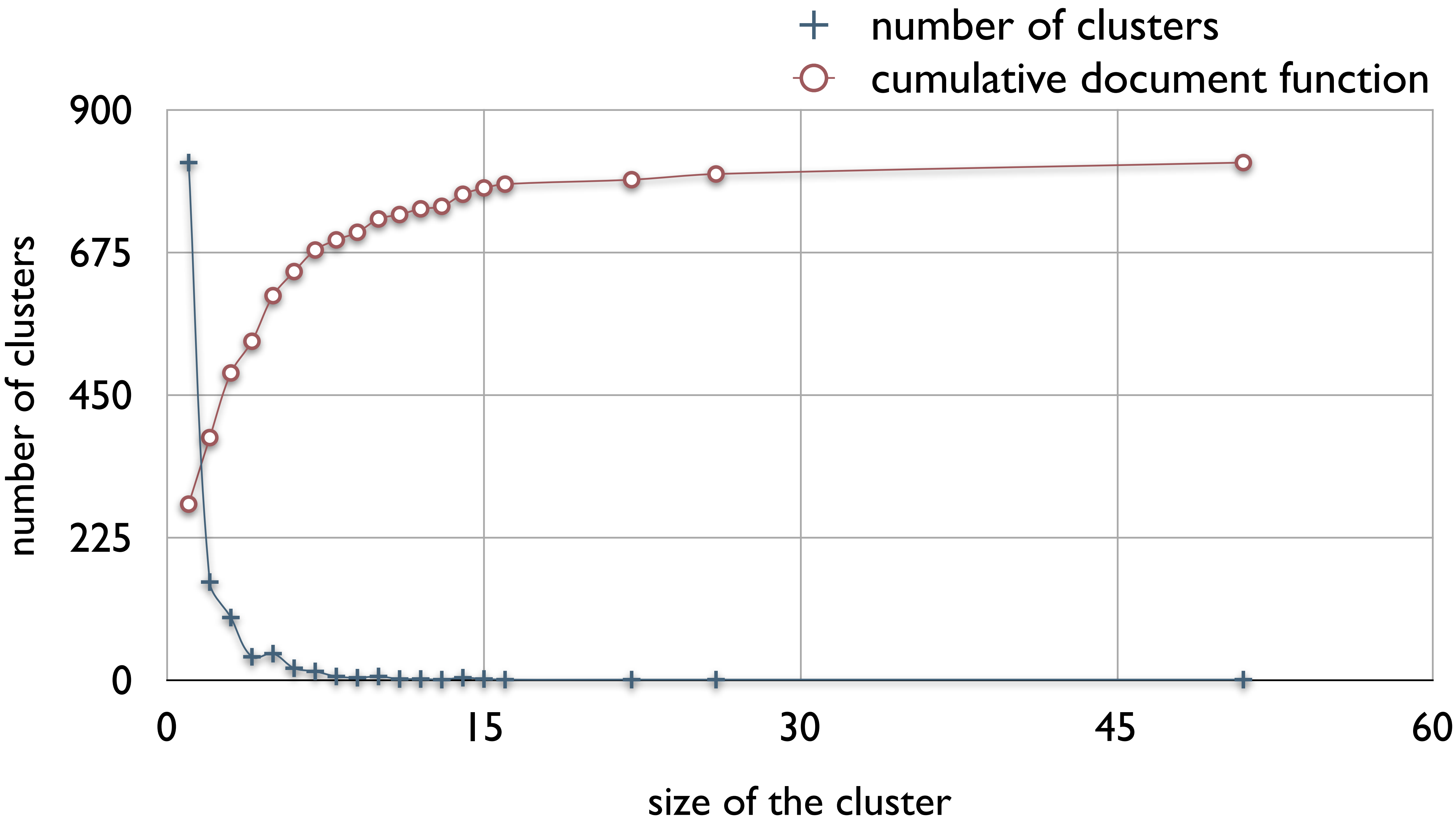
- 2,398 documents published on 17 Croatian news portals in three days
- two annotators, application developed for that purpose
- pooling - using a combination of all similarity metrics to obtain a candidate list
- built 1,214 and 955 clusters

Inter-annotator agreement

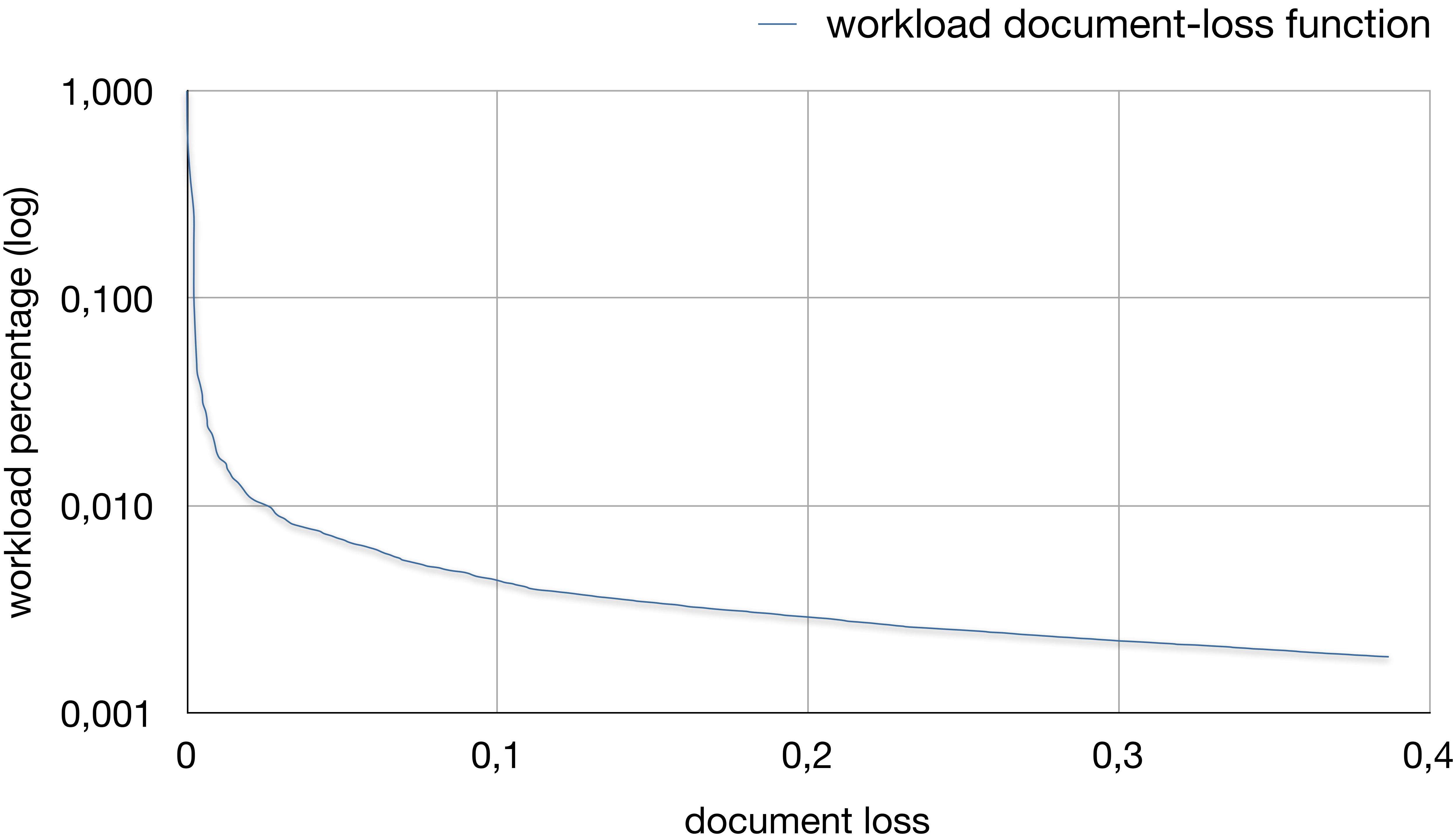
kappa	$\varkappa=\frac{2\cdot\vert A_1\cap A_2\vert}{\vert A_1\vert+\vert A_2\vert}$	0,684
modified kappa	$\varkappa_{\mathrm{mod}}=\frac{\vert A_1\cap A_2\vert}{\min(\vert A_1\vert,\vert A_2\vert)}$	0,91

- biggest story of May 3 2009 - the Myanmar cyclone
- annotator 1 - one cluster with 52 documents
- annotator 2 - three clusters - the catastrophe, first rescue operations, Croatian Red cross reaction

Event cluster distribution



Workload-recall trade-off



Experimental setup

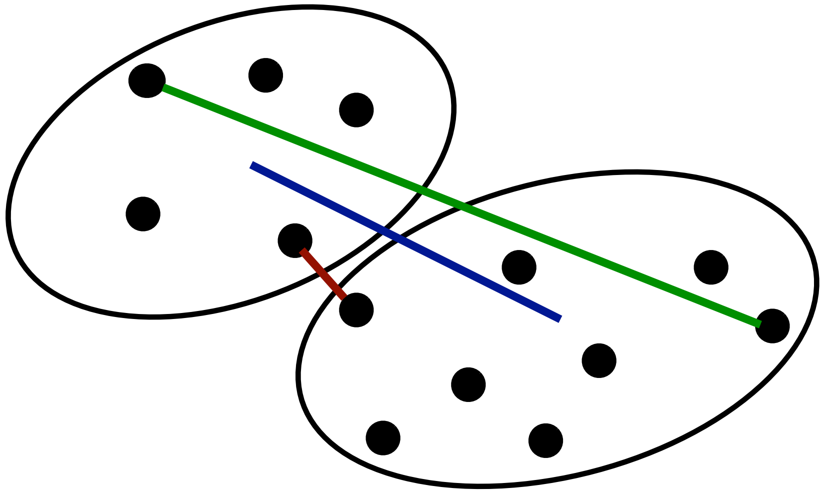
- 14 categorical variables with 2-6 levels - 2,073,600 experiments
- huge search space - independence assumption
- variable categories:
  - clustering algorithm
  - distance metrics
  - feature weight measures
  - feature selection and extraction methods
  - reference corpus significance

Evaluation measures

purity	$purity(\Omega,\mathrm{C})=\frac{1}{N}\sum_{k}\max\|\omega_{k\cap \mathrm{c}_j}\|$
normalized mutual information	$NMI(\Omega,\mathrm{C})=\frac{I(\Omega;\mathrm{C})}{[H(\Omega)+H(\mathrm{C})]\cdot 0.5}$
rand index (accuracy)	$RI=\frac{TP+TN}{TP+FP+TN+FN}$
precision, recall	$P=\frac{TP}{TP+FP}\quad R=\frac{TP}{TP+FN}$
F$_\beta$	$F_\beta=\frac{(\beta^2+1)PR}{\beta^2P+R}$

Clustering

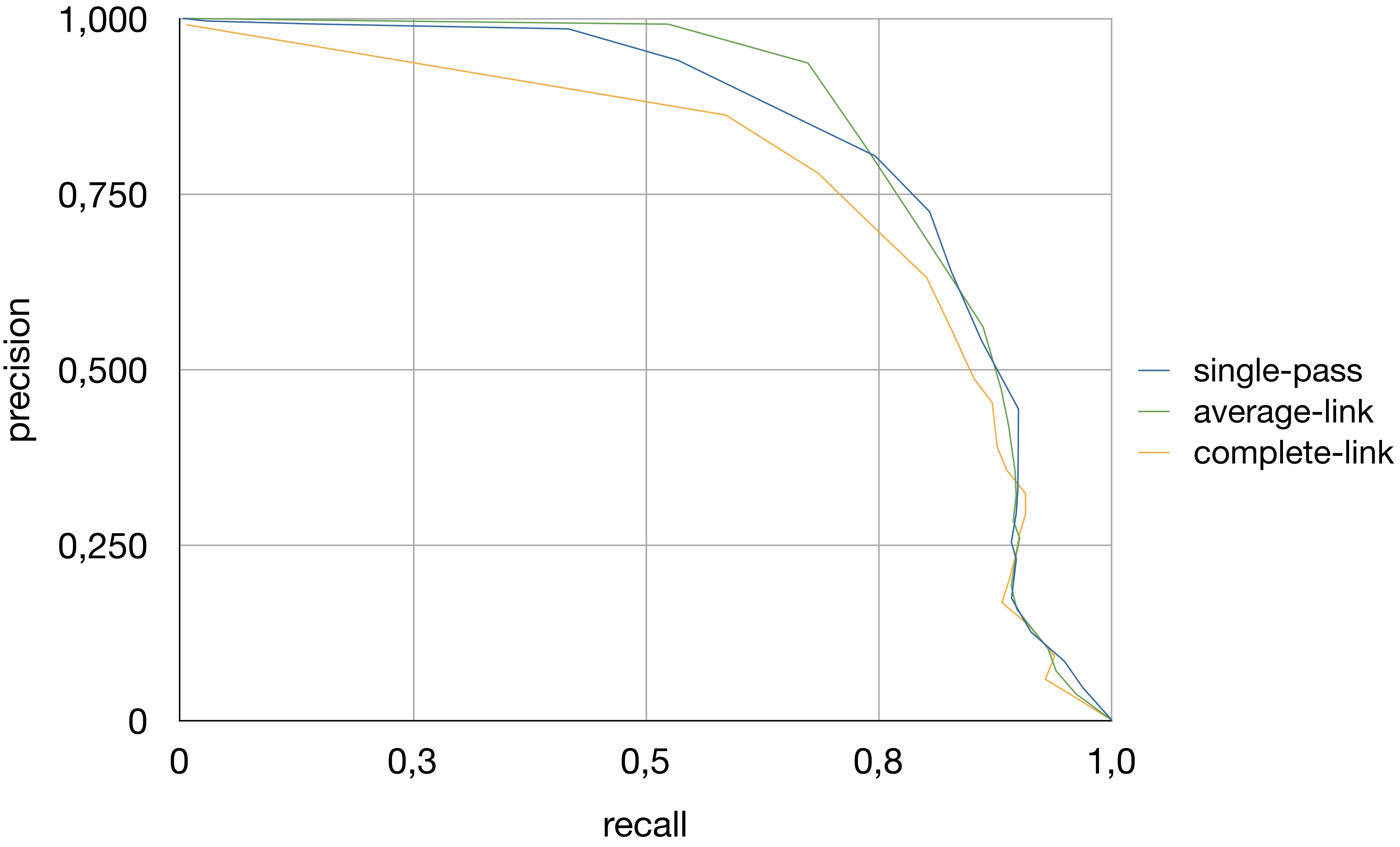
- partitional vs. hierarchical
- retrospective vs. on-line
- linkage criterion in hierarchical algorithms
    - maximum - complete-link
    - minimum - single-link
    - mean - average-link



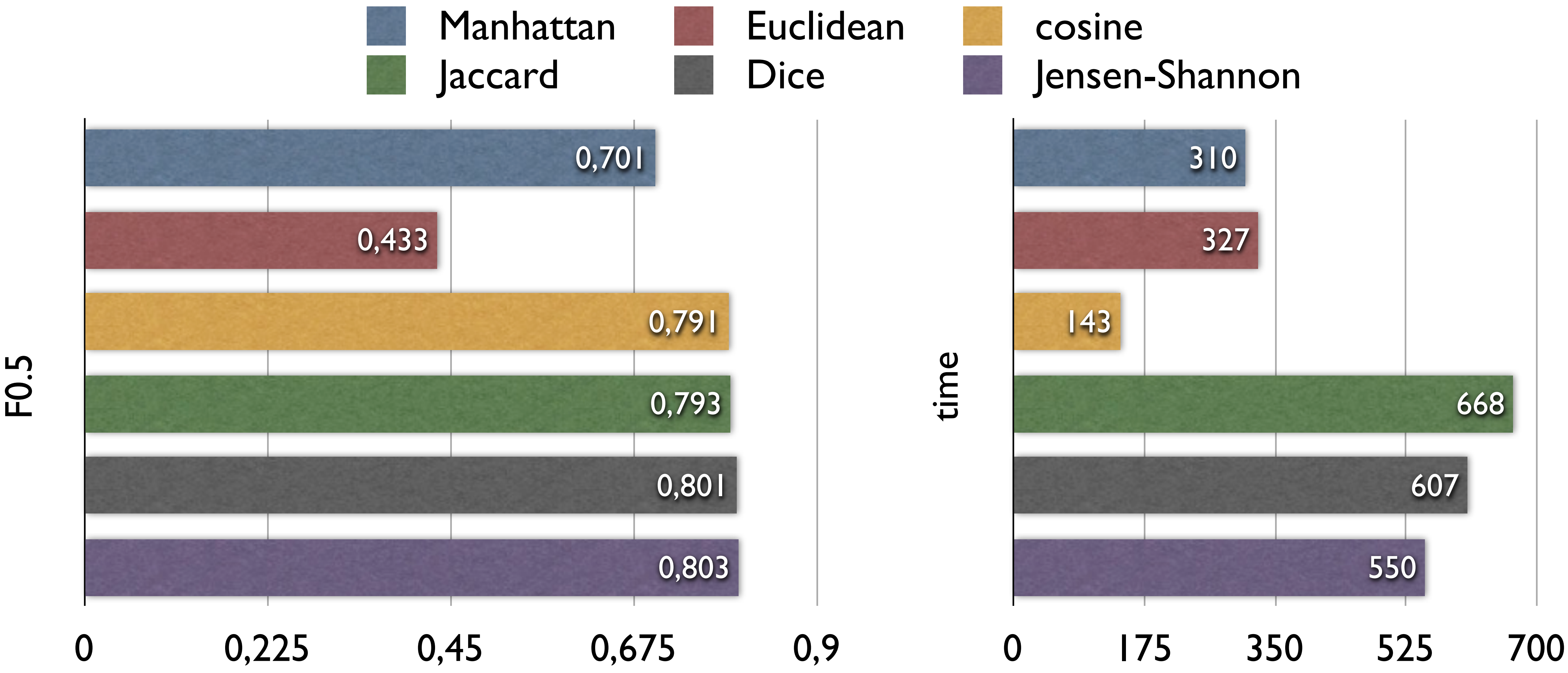
Clustering algorithms

algorithm	linkage criterion	time complexity	on-line
hierarchical agglomerative	complete	$O(n^2 \log n)$	no
hierarchical agglomerative	average	$O(n^2 \log n)$	no
single-pass	single	$O(n)$	yes

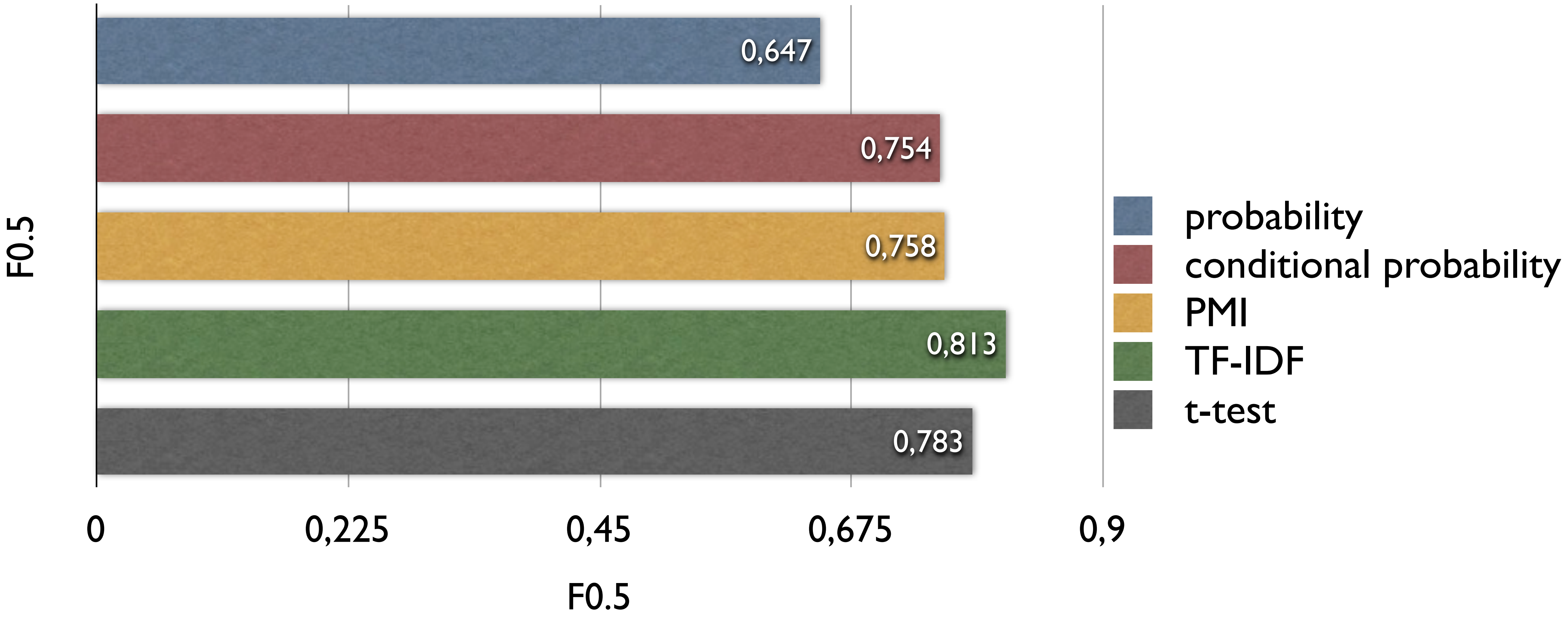
Precision-recall curve



Distance metrics



Feature weight measures



Feature selection

- character case and punctuation obsolete
- information in title more relevant, optimal repetition rate is four
- function words (IDF) - minor decrease in model and memory complexity
- hapax legomena - decreases number of dimensions drastically, memory 5-10%

Feature extraction

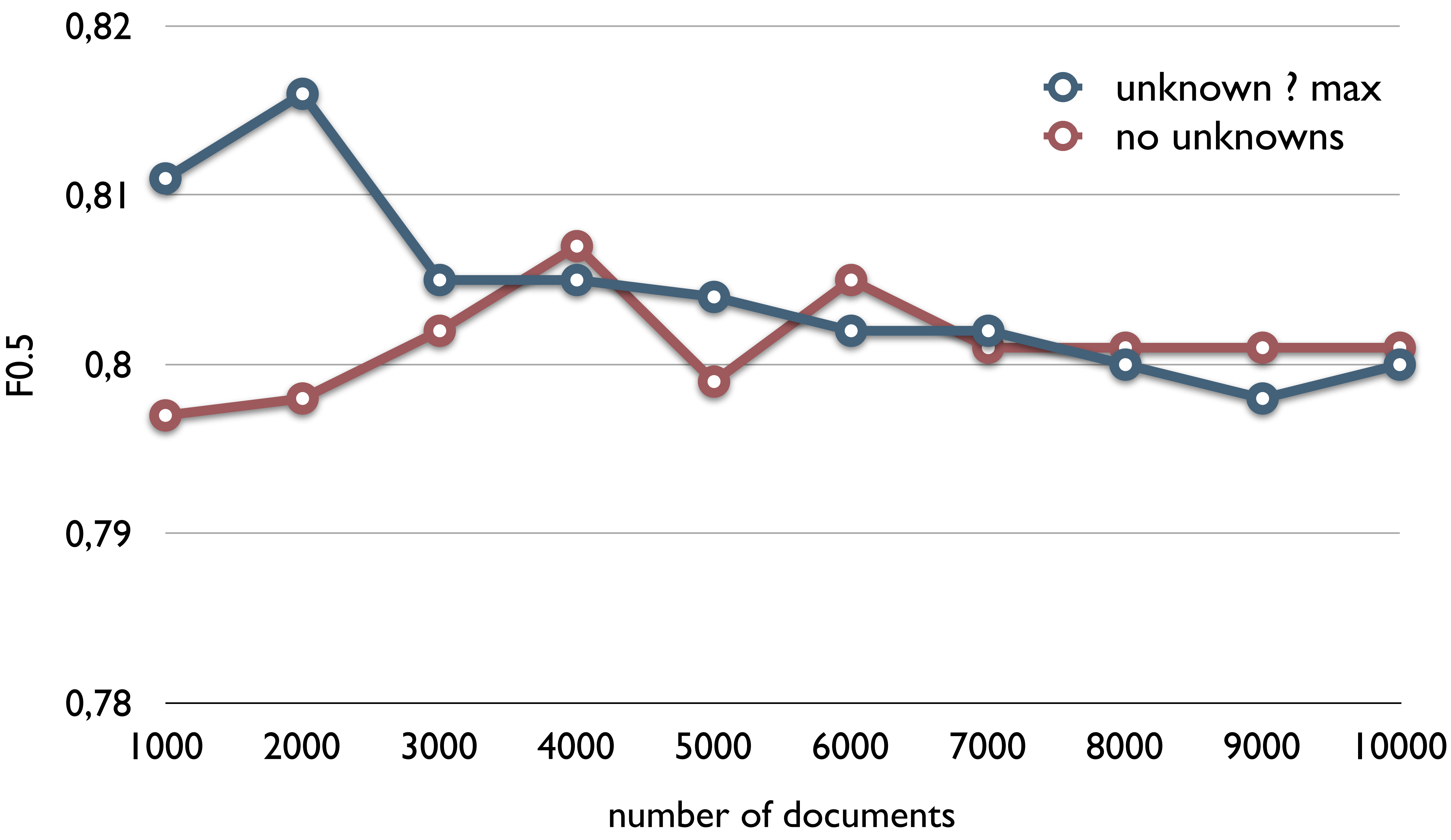
- stemming, POS tagging, lemmatization (two stemmers, TnT, HML)
- multi-word expressions (chi-square)
- named entity recognition (person and business entities)
- no significant improvement

Heuristics

1. an event ranges on a one-day time span true in 83% of documents (non-singleton events)
2. one source reports only once about an event - true in 86% of documents (non-singleton events)

- implementing heuristics increases $F_{0.5}$, first heuristic simplifies calculation drastically

Reference corpus



Output example

```
Primorac saslušao studente (vijesti.hrt.hr)
Studenti nakon sastanka s Primorcem ipak ne odustaju od prosvjeda (index.hr)
Primorac pokušava izbjeći studentske prosvjede razgovorom s Rektorskim
zborom (business.hr)
Primorac primio organizatore studentskog štrajka (javno.com)
Studenti ne odustaju od najavljenog prosvjeda (dnevnik.hr)

VIDEO: Istukla i opljačkala susjedu zbog ljubomore (javno.com)
U stanu ju udarila palicom po glavi i opljačkala (index.hr)
Prijateljicu nevjenčanog supruga pretukla palicom i opljačkala (vecernji.hr)
Opalila je palicom u stanu i opljačkala (index.hr)

Sindikalna košarica u travnju 0,17 posto skuplja nego u ožujku (vecernji.hr)
Sindikalna košarica u travnju 0,17 posto skuplja (tportal.hr)
Sindikalna košarica u travnju 0,17 posto skuplja nego u ožujku (poslovni.hr)
Životni troškovi četveročlane obitelji 6206 kuna (business.hr)
```

Further steps

- windowing technique $\Longleftrightarrow$ decay function
- feature position - features found at the beginning (in the first sentence?) should be given more weight
- write a Java API (Apache license)
- events $\Longrightarrow$ topics;

  event : document relationship

Thank you!