

# Soft Document Clustering Analysis Using One Pass Algorithm

Mislav Cimperšak and Marija Tkalec

Mentor: prof. dr.sc. Damir Boras

Faculty of Humanities and Social Sciences

University of Zagreb

25.5.2009.

# Introduction

- Constant increase of textual documents available
  - Internet
  - Digital libraries
  - Private networks etc.
- Finding relevant information by minimizing time and cost
- Research and development of different kinds of techniques for information retrieval

# Goal

- Determining the optimal threshold in one pass soft document clustering
- Vector features as objects of observation
  - Digraphs
  - Trigraphs
  - Quadrigraphs
  - Tokens

# Document clustering

- Sorting documents into clusters with specific characteristics
- Document placed in one cluster has similar or same characteristics with other documents in the cluster
- Soft and hard clustering

# Soft clustering

- Documents in one or more clusters
- Certain degree of belonging to every cluster
- Higher degree → greater relevancy for the cluster

# Similarity threshold

- Measurement on which is determined belonging of a certain document to a certain cluster
- Cosine similarity
- 6 observed thresholds

# Corpus

- 4 days
  - May 1st 2008 → May 5th 2008
- 2532 articles
- 18 Croatian news portals

# Corpus

NEWS PORTAL	NEWS PORTAL
business.hr	business.hr
dnevnik.hr	dnevnik.hr
glas-slavonije.hr	glas-slavonije.hr
index.hr	index.hr
javno.com	javno.com
jutarnji.hr	jutarnji.hr
liderpress.hr	liderpress.hr
mojportal.hr	mojportal.hr
nacional.hr	nacional.hr
net.hr	net.hr
poslovni.hr	poslovni.hr
rtl.hr	rtl.hr
seebiz.eu	seebiz.eu
sutra.hr	sutra.hr
totalportal.hr	totalportal.hr
tportal.hr	tportal.hr
vecernji.hr	vecernji.hr
vijesti.hrt.hr	vijesti.hrt.hr
<b>TOTAL</b>	<b>TOTAL</b>
1	2
3	5
6	2
3	3
0	8
4	5
4	7
3	4
0	2
4	5
1	8
5	5
3	4
0	3
0	7
7	2
8	4
8	5
4	3
2	3
5	3
2	2

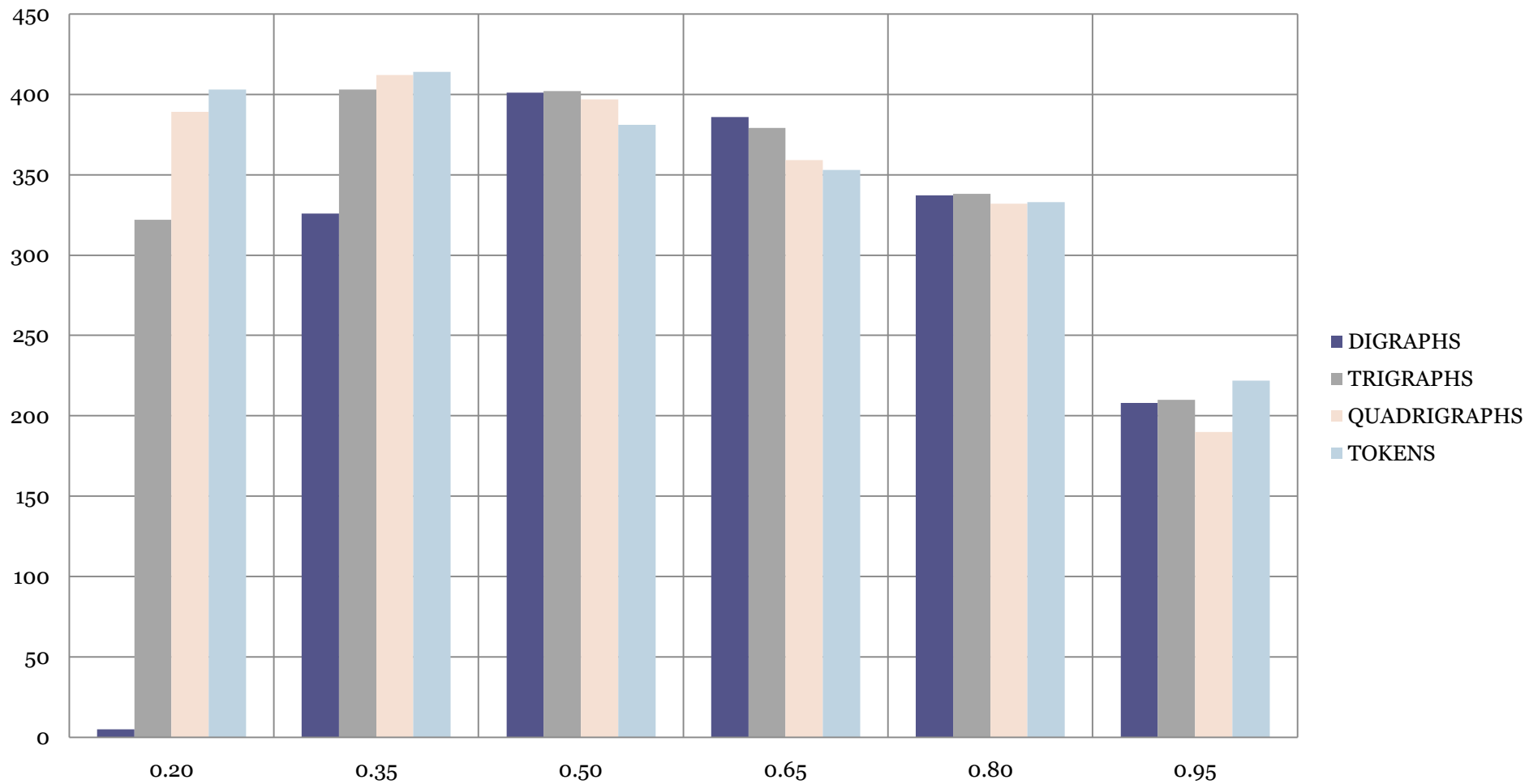
# Algorithm

- Written in Python
- Tokenization
- Tf-Idf measure
- Cosine values
- Vector space model
- Soft clustering based on similarity threshold

# Assumption

- Clustering based on trigraphs would give optimal results
- Clustering based on quadrigraphs would be least usable

# Results



# Results

- Example: An article about celebration of 1st May in Maksimir
- Key phrases
  - 1st May
  - Zagreb
  - Maksimir
  - Milan Bandić
  - beans

# Results

- Low thresholds
  - Clusters contain articles of a too broad subject spectrum
- Medium thresholds
  - Optimal results
    - Different vector features
- High thresholds
  - Tendency towards hard clustering

# Conclusion

- Digraphs  $\rightarrow$  not precise enough
- Quadrigraphs  $\rightarrow$  high vector space complexity
- Optimal similarity threshold  $\rightarrow$  0.5 when using tokens as vector features

# Conclusion

- Future research could include lemmatized and/or stemmatized words
- Possible algorithm optimization could give better results for trigraphs