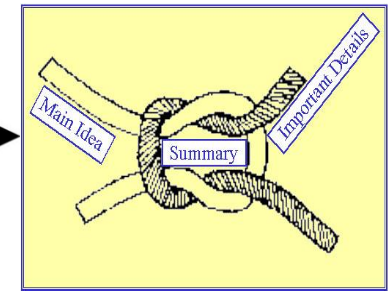
Document analysis by means of data mining techniques

Saima Jabeen
Matr. No: s168999
PhD-XXVI Cycle
Tutor: Prof. Elena Baralis

DAUIN – Department of Control and Computer Engineering,
Politecnico di Torino

Funded By: HEC-Higher Education Commission of Pakistan

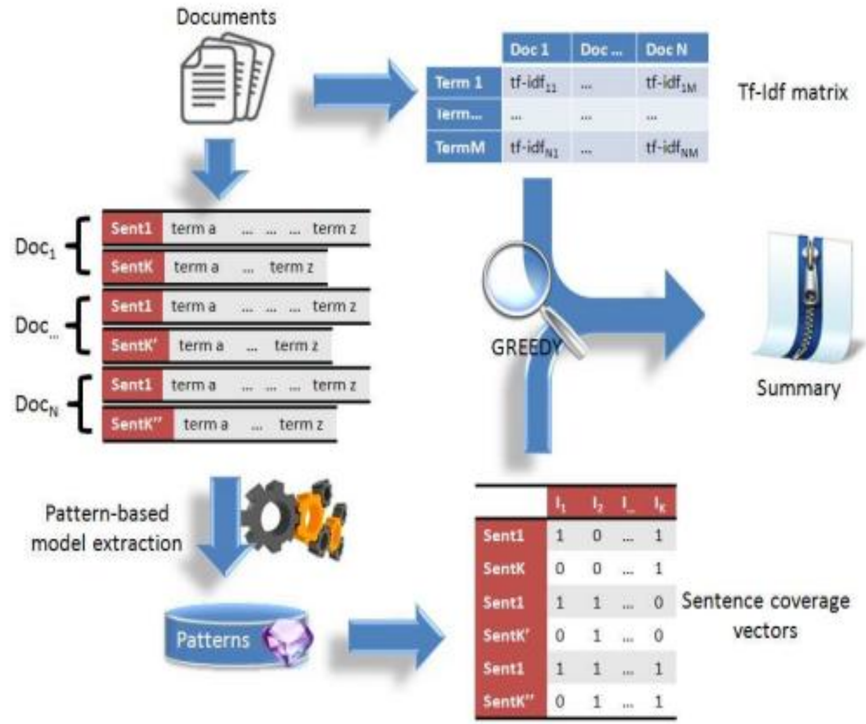
Motivation and Research Context



- Information overload
- Unstructured form
- Different aspects of information

- Data mining and IR techniques
- **Automatic** Summary
- Text Summarization

ItemSum



- **Research Problem(s)**
  - Previous approaches typically focus **single-word** significance
- Need for capturing correlations among multiple-words
- **Novel Contributions**
  - Frequent itemset mining to discover correlations in TS
  - The usage of an itemset-based model to represent the most relevant and not redundant correlations among document terms
  - The selection of minimal set of representative sentences based on:
    - Sentence coverage – Transactional data format
    - Sentence relevance score – Score based on tf-idf statistics

Performance Comparison–ItemSum

Performance comparison in terms of ROUGE-2 score

dataset	PatTexSum				OTS			TexLexAn		
	p	R	Pr	F	R	Pr	F	R	Pr	F
Natural Disaster	16	**0.116**	**0.288**	**0.141**	0.040	0.120	0.053	0.038	0.114	0.045
Royal Wedding	12	**0.036**	**0.215**	**0.058**	0.034	0.174	0.054	0.030	0.150	0.047
Technology	5	**0.141**	**0.465**	**0.210**	0.042	0.208	0.067	0.042	0.172	0.065
Sports	10	**0.145**	**0.297**	**0.189**	0.055	0.133	0.075	0.071	0.149	0.093
Education	8	**0.039**	**0.241**	**0.064**	0.036	0.170	[illegible]	[illegible]	[illegible]	[illegible]

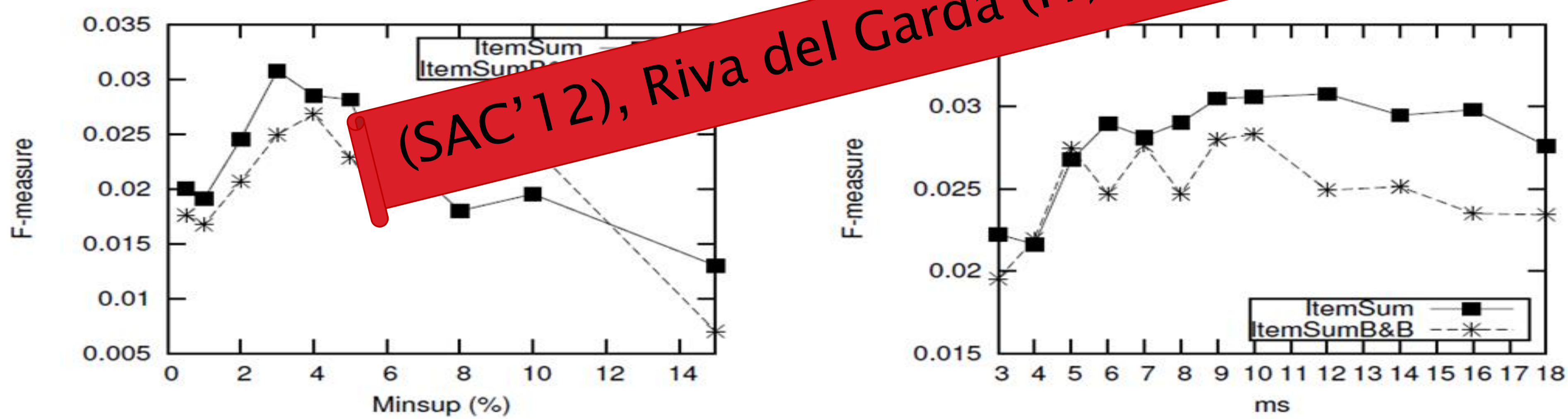
MCP 2011, (AI*AI'11), Palermo (Italy)

dataset	PatTexSum				OTS			TexLexAn		
	p	R	Pr	F	R	Pr	F	R	Pr	F
Natural-Disaster	16	**0.060**	**0.125**	**0.068**	0.005	0.012	0.006	0.005	0.011	0.006
Royal-wedding	12	**0.009**	**0.082**	**0.015**	0.003	0.018	0.005	0.003	0.018	0.005
Technology	5	**0.113**	**0.356**	**0.167**	0.009	0.065	0.016	0.003	0.011	0.005
Sports	10	**0.059**	**0.112**	**0.077**	0.004	0.010	0.006	0.022	0.036	0.027
Education	8	**0.017**	**0.141**	**0.030**	0.003	0.012	0.005	0.003	0.009	0.004

Performance Comparison–ItemSum

Summarizer		ROUGE-2			ROUGE-3			ROUGE-4		
		R	Pr	F	R	Pr	F	R	Pr	F
OTS		0.0746*	0.0740*	0.0743*	0.0236*	0.0234*	0.0235*	0.0087*	0.0086*	0.0087*
TexLexAn		0.0655*	0.0643*	0.0649*	0.0197*	0.0193*	0.0195*	0.0071*	0.0069*	0.0070*
DUC'04 competitors	peer102	0.0840	0.0846	0.0843	0.0264	0.0267	0.0265	0.0103*	0.0104*	0.0104*
	peer103	0.0774*	**0.0896**	0.0826	0.0243*	0.0284	0.0260*	0.0092*	0.0108	0.0099*
	peer140	0.0685*	0.0692*	0.0688	0.0218*	0.0220*	0.0219*	[illegible]	0.0094*	0.0093*
ItemSum		**0.0864**	0.0869	**0.0866**	**0.0307**	**0.0309**	[illegible]	[illegible]	**0.0136**	**0.0135**

(SAC'12), Riva del Garda (IT), 2012



(a) ms=12. Impact of the support threshold.

(b) min_sup=3%. Impact of the model size.

The Yago-based summarizer

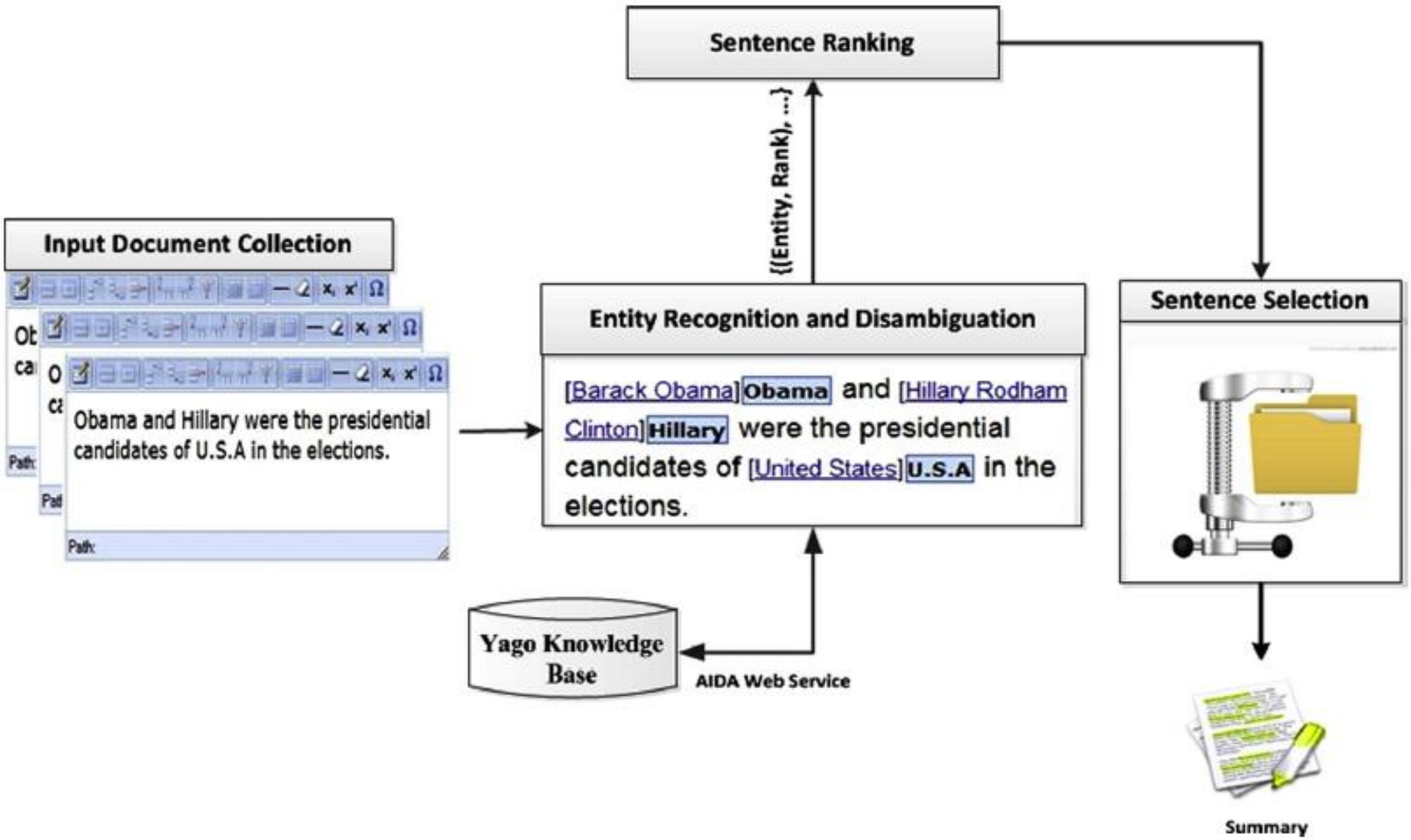
- Research Problem(s)
  - Unsatisfactory soundness and readability of summaries
  - Semantic models were used only in preprocessing
  - Domain specific ontologies/taxonomies, dictionaries were used
- Need of a generic summarizer based on a generic knowledgebase

- Novel Contribution(s):
  - Integration of a semantics-based model into document summarization
  - Use Yago ontology to **evaluate** and **select** document sentences

- Results:

  YagoSum outperforms many state-of-the-art summarizers in terms of ROUGE scores on benchmark collections

The Yago-based summarizer



Performance Comparison–YagoSum

Summarizer		ROUGE-2			ROUGE-4		
		R	Pr	*F*	*R*	Pr	*F*
TOP RANKED DUC'04 PEERS	peer120	0.076*	**0.103**	0.086*	0.014*	**0.019**	0.016
	peer65	0.091*	0.090*	0.091*	0.015*	0.015	0.015*
	peer19	0.080*	0.080*	0.080*	0.010*	0.010*	0.010*
	peer121	0.071*	0.085*	0.077*	0.012*	0.014*	0.013*
	peer11	0.070*	0.087*	0.077*	0.012*	0.015*	0.012*
	peer44	0.075*	0.080*	0.078*	0.012*	0.013*	[illegible]
	peer81	0.077*	0.080*	0.078*	0.012*	[illegible]	[illegible]
	peer104	0.086*	0.084*	0.085*	[illegible]	[illegible]	[illegible]
	peer124	0.083*	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	peer35	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.011*
DUC'04 HUMANS	[illegible]	[illegible]	[illegible]	[illegible]	0.009*	0.010*	0.010*
	[illegible]	[illegible]	0.096	0.092	0.013*	0.013*	0.013*
	[illegible]	0.094	0.102	0.098	0.011*	0.012*	0.012*
	D	0.100	**0.106**	0.102	0.010*	0.010*	0.010*
	E	0.094	0.099	0.097	0.011*	0.012*	0.012*
	F	0.086*	0.090*	0.088*	0.008*	0.009*	0.009*
	G	0.082*	0.087*	0.084*	0.008*	0.008*	0.007*
	H	**0.101**	0.105	**0.103**	0.012*	0.013*	0.012*
OTS		0.075*	0.074*	0.074*	0.009*	0.009*	0.009*
texLexAn		0.067*	0.067*	0.067*	0.007*	0.007*	0.007*
ItemSum		0.083*	0.085*	0.084*	0.012*	0.014*	0.014*
Baseline		0.092*	0.091*	0.092*	0.014*	0.014*	0.014*
Yago-based Summarizer		**0.095**	0.094	**0.095**	**0.017**	0.017	**0.017**

Expert Systems With Applications. ISSN 0957-4174

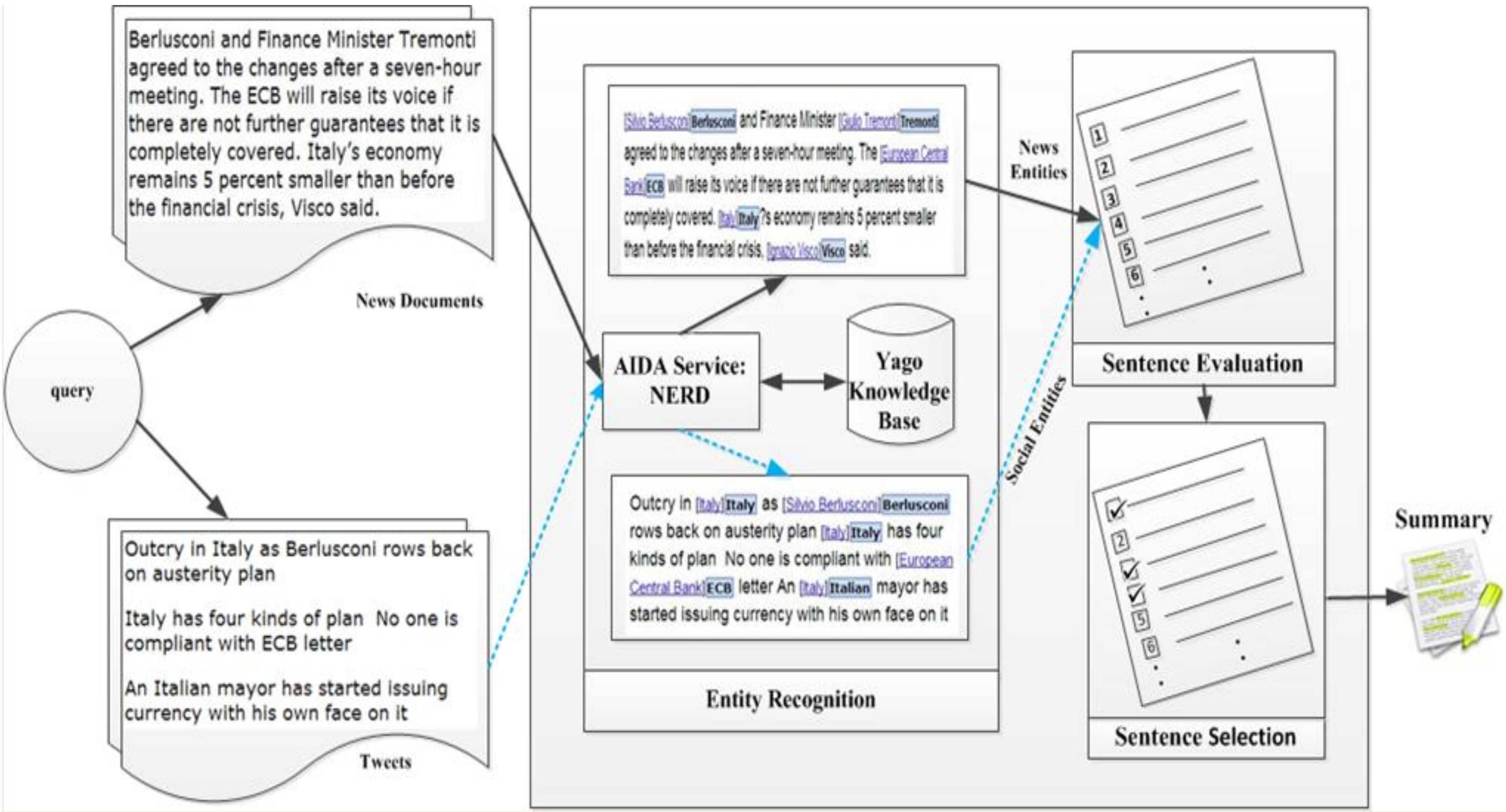
SocioNewSum

- Research Problem(s)
  - correlations among relevant concepts are missed
  - Reader's expectations and current interests
- With the advent of social media, user's interest can easily be identified

- Novel Contribution(s)
  - Combining Semantic and social knowledge for news TS
  - Preliminary analysis of Twitter messages to discover the actual user's interests
  - Identifying ontological concepts in both genres
  - Summarization of news collections driven by social knowledge to generate appealing news summaries

- Results:
  - Integration of social knowledge substantially improves news document summarization performance

Overview of SocioNewSum



Performance Comparison–SocioNewSum

Dataset (Tuned setting)	TexLexAn			OTS			Baseline			SociONewSum Standard result (Tuned result)		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
Debt Crisis (K=50, δ =0.1, α =0.6)	0.068	0.541	0.121	0.077	**0.581**	0.135	0.070	0.518	0.121	**0.081* (0.082*)**	0.569 **(0.582)**	**0.141* (0.143*)**
Irene (K=30, δ =0.1, α =0.2)	0.062	0.577	0.110	0.064	0.602	0.116	0.055	0.532	0.099	**0.066* (0.070*)**	**0.612* (0.646*)**	**0.119* (0.127*)**
Steve Jobs (K=40, δ =0.2, α =0.8)	0.057	0.533	0.102	0.060	0.573	0.109	0.068	0.615	0.122	**0.072* (0.072*)**	**0.625* (0.645*)**	**0.126* (0.128*)**
Terrorism (K=70, δ =0.1, α =0.7)	0.047	0.447	0.086	0.052	0.479	0.093	0.045	0.444	0.082	**0.055* (0.058*)**	**0.517* (0.544*)**	**0.099* (0.104*)**
UK riots (K=10, δ =0.1, α = 0.2)	0.060	0.522	0.107	**0.065**	**0.586**	**0.117**	0.052	0.456	0.092	**0.065 (0.067)**	0.578 (0.578)	**0.117 (0.120*)**
US Open (K=40, δ =0.1, α =0.5)	0.076	0.651	0.136	0.065	0.545	0.116	0.065	0.492	0.114	**0.080* (0.082*)**	**0.654 (0.663*)**	**0.142* (0.145*)**

SocioNewSum: Performance comparison in terms of ROUGE-1

Dataset (Tuned setting)	TexLexAn			OTS			Baseline			SociON[illegible]		
	R	P	F1	R	P	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Debt Crisis (K=50, δ =0.1, α =0.6)	0.021	0.175	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**[illegible] (0.048*)**
Irene (K=30, δ [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**0.023 (0.024*)**	**0.221* (0.230*)**	**0.041 (0.043*)**
Stev[illegible] (K=40, δ [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.037	0.024	0.225	0.043	**0.025 (0.026)**	**0.232* (0.244*)**	**0.045* (0.047*)**
Terr[illegible] (K=70, δ [illegible]	[illegible]	0.147	0.028	**0.017**	**0.161**	**0.031**	0.014	0.142	0.025	0.014 **(0.019*)**	0.142 **(0.188*)**	0.025 **(0.035*)**
UK [illegible]ts (K=10, δ =0.1, α =0.2)	0.022	0.195	0.039	0.022	0.203	0.040	0.015	0.139	0.027	**0.023 (0.024)**	**0.211* (0.211*)**	**0.043* (0.043*)**
US Open (K=40, δ =0.1, α =0.5)	0.025	0.221	0.045	0.021	0.182	0.038	0.025	0.198	0.044	**0.032* (0.034*)**	**0.264* (0.278*)**	**0.056* (0.059*)**

SocioNewSum: Performance comparison in terms of ROUGE-SU4

Concluding remarks

- frequent itemsets
- semantic-based
- Social document analysis
- Effectiveness of the proposed approaches
- Future developments
  - Incremental Updated summaries
  - Extension to multi-lingual document collections

Thanks for the attention!