ARABIC DOCUMENT SIMILARITY ANALYSIS USING N-GRAMS AND SINGULAR VALUE DECOMPOSITION

This work is a part of the Preliminary Research Project No. PRP2012.R12.2 funded by the Information Technology Academic Collaboration (ITAC) program, Information Technology Industry Development Agency (ITIDA), Ministry of Communication and Information Technology, Egypt.

Ashraf S. Hussein

ashrafh@acm.org

FCIT, Arab Open University, HQ, Kuwait
www.arabou.edu.kw

FCIS, Ain Shams University, Cairo, Egypt
www.asu.edu.eg

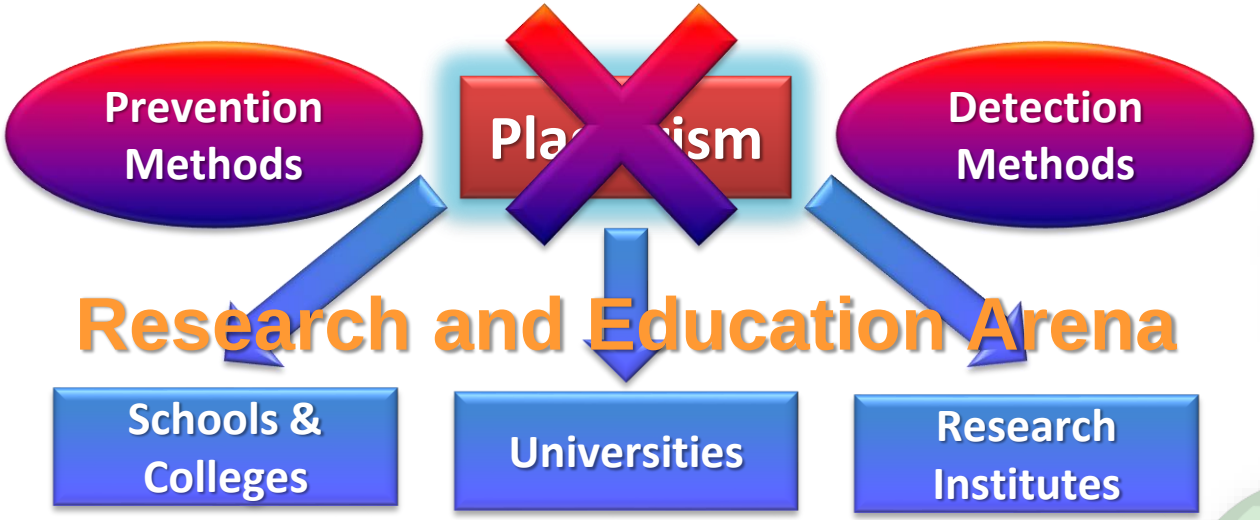
Outline

- What is Plagiarism ?
- Plagiarism Taxonomies
- Problem Statement
- Challenges in Arabic Language
- Plagiarism Detection Methods and Existing Products
- The Proposed Solution versus Existing Ones
- Research Road Map
- Detailed Research Objectives
- Proposed Solution Overview
- Arabic Document Similarity Estimation Method
- Results and Discussions
- Conclusions

What is plagiarism?

"Plagiarism is the act of presenting words, ideas, images, sounds, or the creative expression(s) of the others as your own."

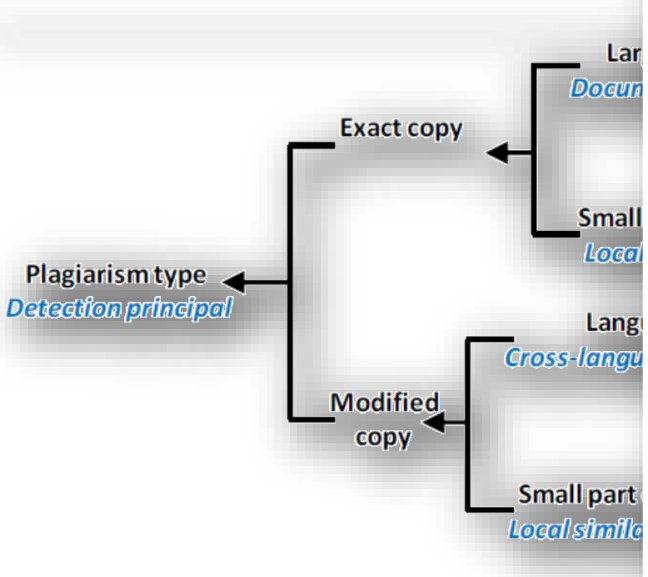
(Research resources at plagiarism.org, 2014)



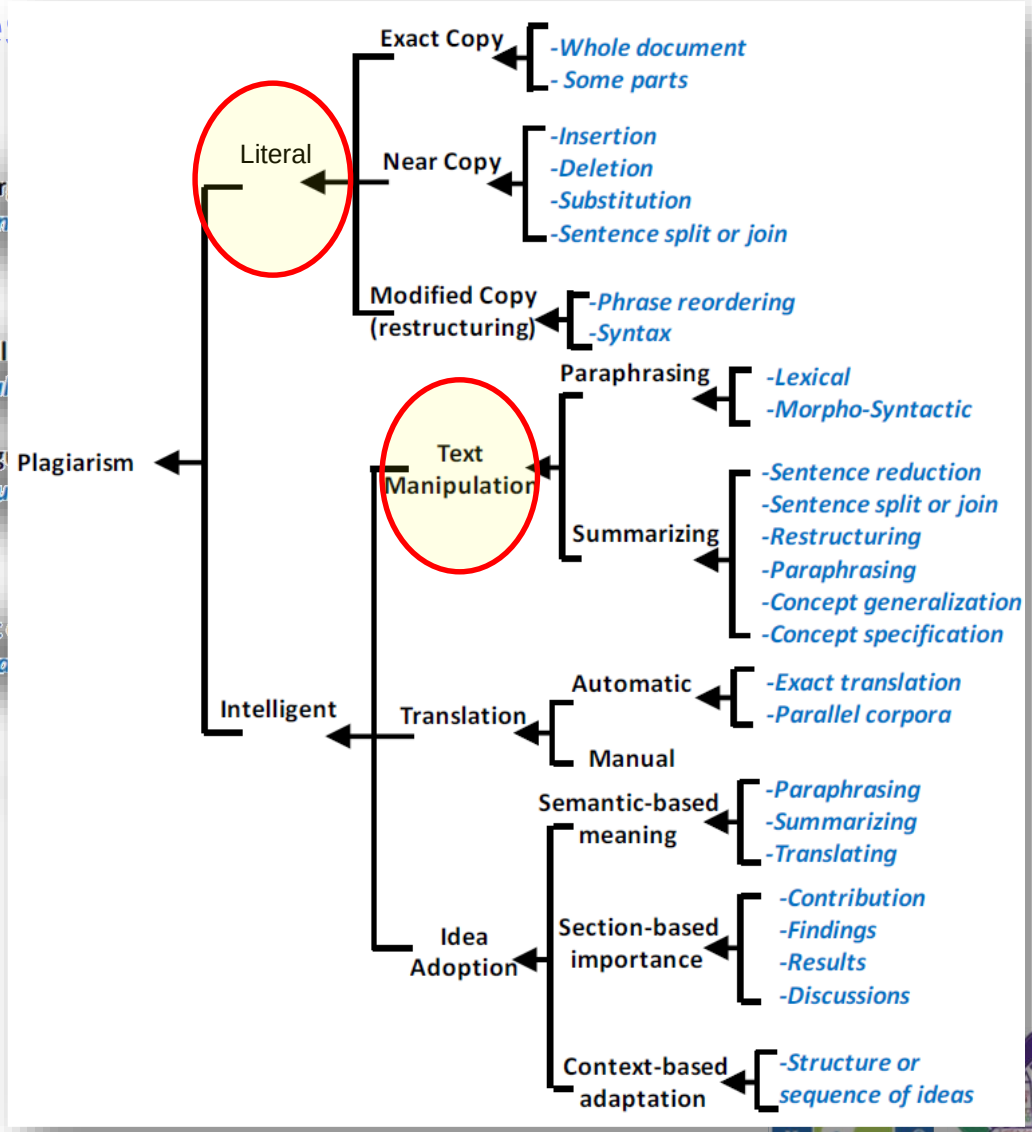
ROI = f(Reputation, Credibility, Dignity)

(Anderson, 2009)

Plagiarism Taxonomies



(Eissen, Stein, & Kulig, 2007)



(Alzahrani, Salim, & Abraham, 2012)

Problem Statement

- Most of plagiarism detection m… o English language in… s such as German and Chines…
- Most of the… mainly on "literal" plagiarism.
- Although th… t techniques for plagiarism de… ms is not satisfactory, … mplicated languages su…
- Most of the f… detection in Arabic text file… into their equivalent… structure.

Our objective is to develop a Proof-of-Concept (PoC) for Document Similarity Estimation Method devoted to Arabic language.

Challenges in Arabic Language

- Literature show that Arabic is highly inflectional, as there are about **five possible different morphological analyses per word on average**.
- Prefixes and suffixes can be attached to words in a concatenative manner.
- A single string can comprise verb inflections, prepositions, pronouns, and connectives. Therefore, word lexical disambiguation in Arabic text is a challenging task.

For instance, the word ، المكاتب transliterated “al-makatib” and meaning offices, is derived from the stem ، مكتب transliterated “maktab” and meaning office, which is derived from the root كتب transliterated “katab” and meaning to write.

Plagiarism Detection Methods

There are mainly two categories of meth ty in free texts (Alzahrani, Salim, & Abraha methods and content-based metho ection systems works in

Technique	Tasks		IR		Language(s)	Plagiarism Type(s)								
						Literal			Intelligent					
	extrinsic	intrinsic	Mono-lingual	Cross-lingual		copy	near copy	restructuring	paraphrasing	summarising	Translating	Idea (section)	Idea (context)
Char-Based (CNG)	●		●		Any	●	●	●					
Vector-Based (VEC)	●		●		Any	●	●	●			●		
Syntax-Based (SYN)	●		●		Specific	●	●	●	●		●		
Semantic-Based (SEM)	●		●		Specific	●	●	●	●		●		
Fuzzy-Based (FUZZY)	●		●		Specific	●	●	●	●		●		
Structural-Based (STRUC)	●		●		Specific	●	●	●	●		●		
Stylometric-Based (STYLE)		●		●	Specific	●	●	●		●		●	●
Cross-Lingual (CROSS)	●				Cross								

The notions in the table indicate the following: ● means include/support by evidence from research publications, ● means possibility to include/support but need further research for proof.

Existing Products

The most prominent product

- Tur
- Ess
- Plag
- How
- Goo
- Plag
- EduT
- Glatt ... ection (GPSD
- Copy
- Word
- Script
- iThenti
- Sa ... .com)

- **Turnitin is still considered the most prominent software, and it has more than 70% of the market share!!**
- **Unfortunately, most of the aforesaid products/solutions do not support the Arabic language, except at the level of word to word plagiarism (e.g. Google.com).**

The Proposed Solution *versus* Existing Ones

For Arabic text, there are few research prototypes like Arabic Plagiarism Detection tool (APD) (Alzahrani & Salim, 2009), Arabic Plagiarism Checker (Menai & Bagais, 2011) and Iqtebas 1.0 (Jadalla & Elnagar, 2012).

System	Reference(s)	Tasks		IR		Language(s)	Plagiarism Type(s)							
							Literal			Intelligent				
		extrinsic	intrinsic	Mono-lingual	Cross-lingual		Copy	near copy	restructuring	Paraphrasing	summarising	translating	Idea (section)	Idea (context)
APD	(Alzahrani & Salim, 2009)	●		●		Arabic	●	●						
APlag	(Menai & Bagais, 2011)	●		●		Arabic	●	●					●	
Iqtibas 1.0	(Jadalla & Elnagar, 2012)	●		●		Arabic	●	●	●					
Proposed one		●		●	●	Arabic	●	●	●	●	●		●	●

The notions in the table indicate the following: ● means include/support by evidence from research publications, ● means there is no proof, ● means a new feature to be supported, ● means there is a potential to support.

The Research Roadmap

- Morphological Analysis and Syntactic Parsing
- Use of synonym thesaurus
- Latent Semantic Analysis
- "Fingerprinting" Authors
- Reference and Citation Tracking

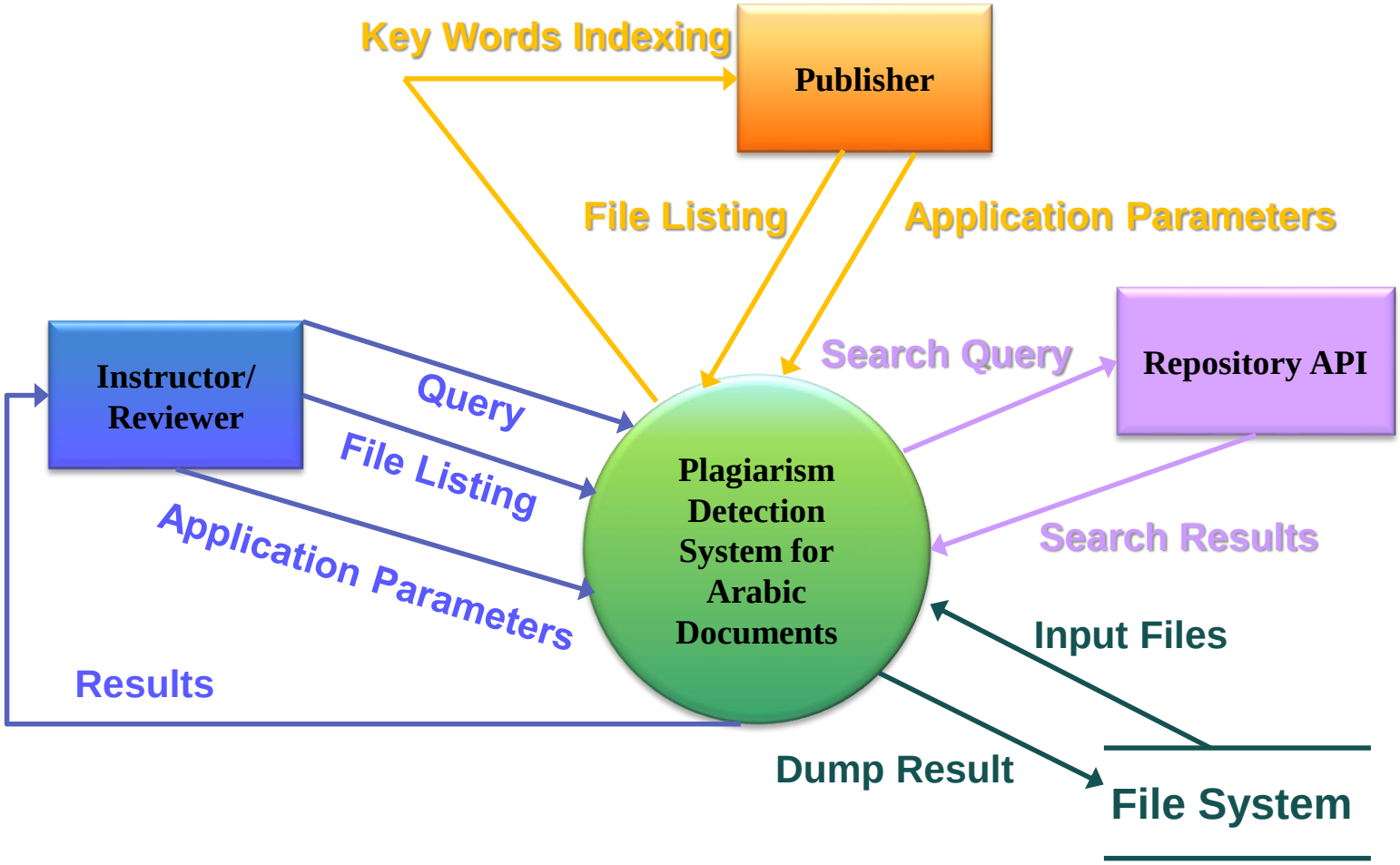
Detailed Research Objectives

This research work is aiming at carrying on rese and development to wards Pla Text , con

- by
- .
- lingual text, in fu phases.

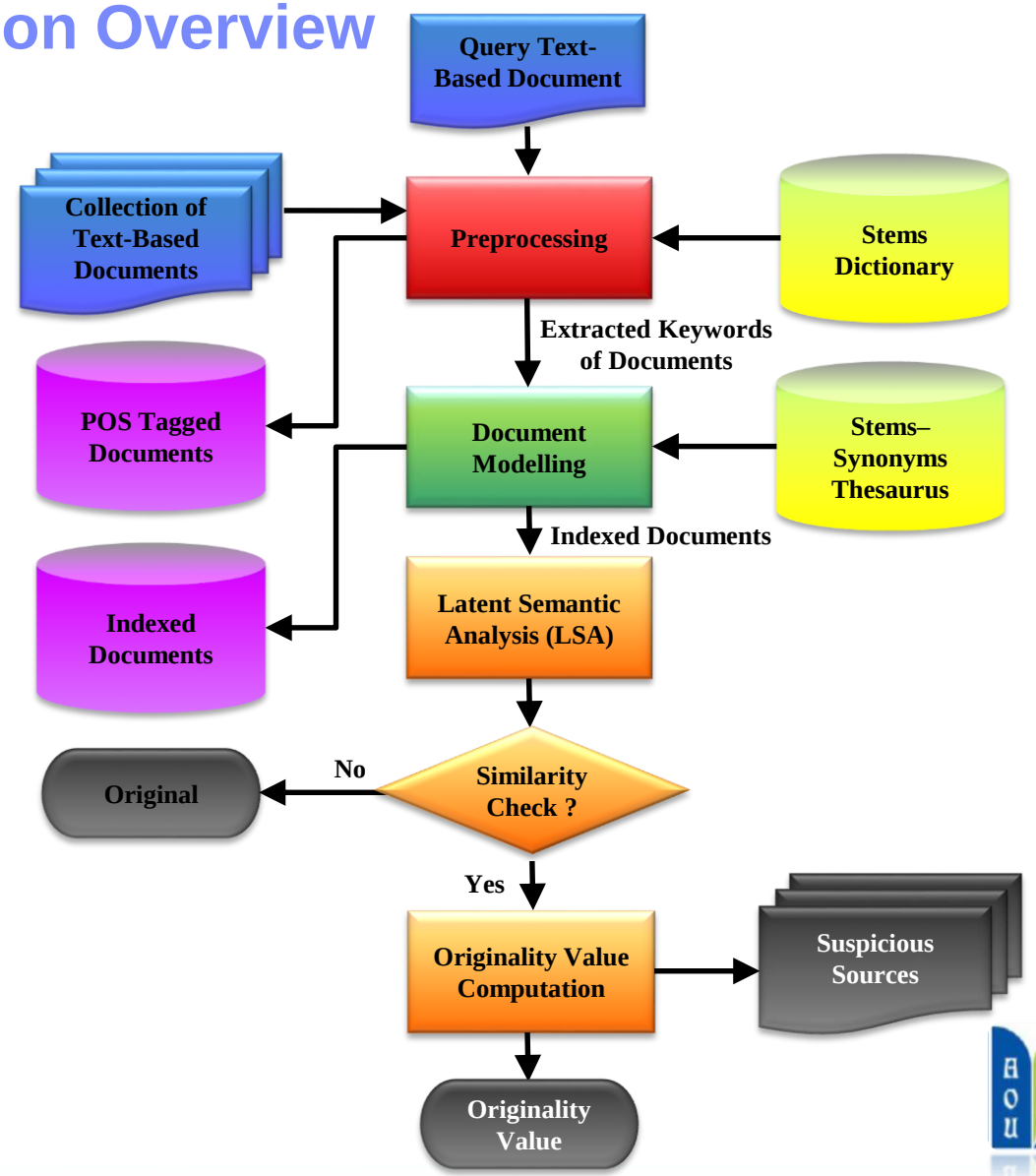
"This PoC will contribute positively to the ICT-for-education industry, as it has a wide range of beneficiaries, ranging from schools, universities and research institutes to publishers and e-Learning Learning Management System (LMS) providers.

Potential Beneficiaries

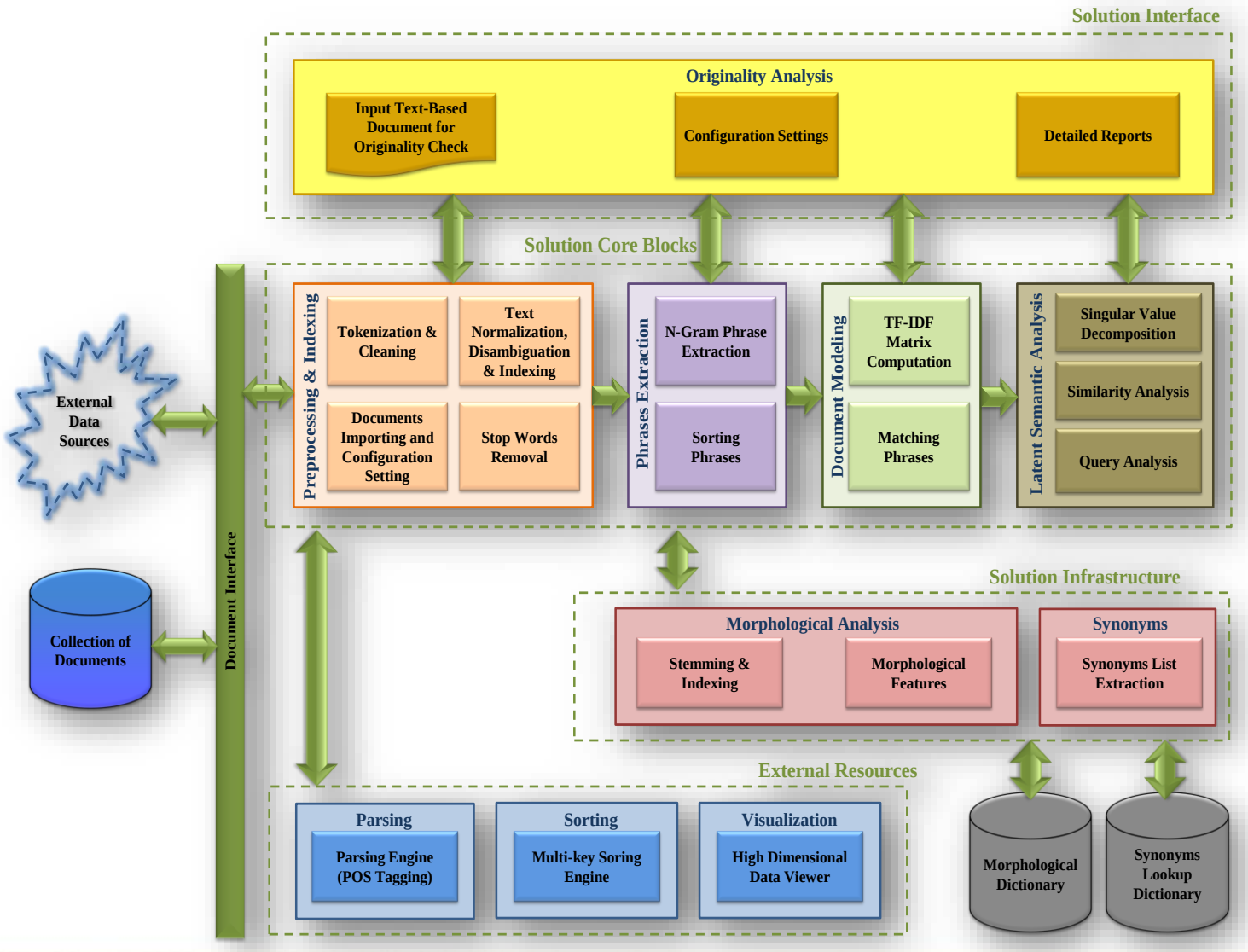


Arabic document similarity analysis.

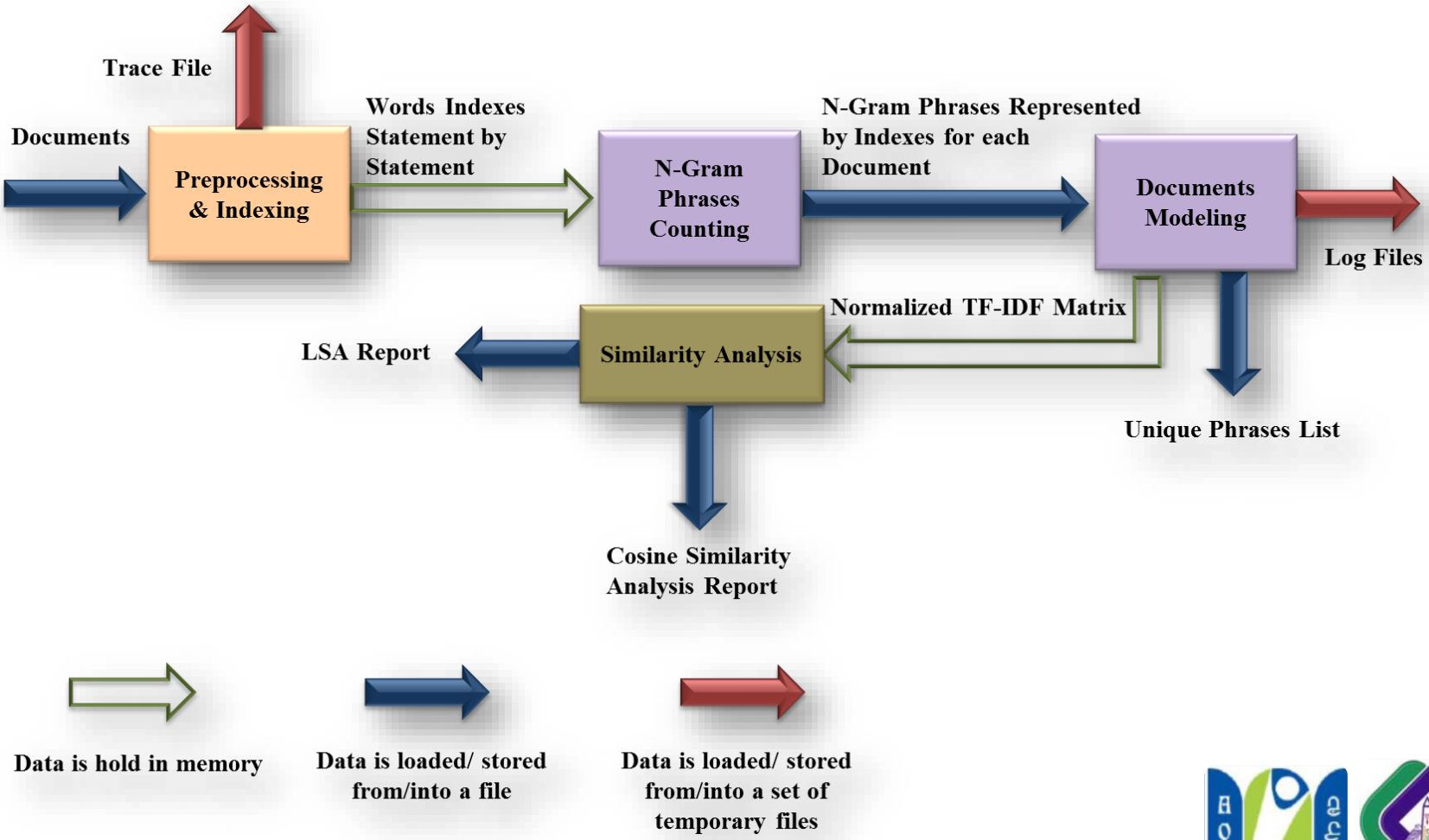
Proposed Solution Overview



The Solution Architecture



The Solution Main Data Flow Diagram



Document Similarity Estimation Method
Pre-processing and Indexing

- **PoS Tagging**

  (Stanford Parser http://nlp.stanford.edu/software/lex-parser.shtml)

- **Tokenization**
- **Stop-words Removal**

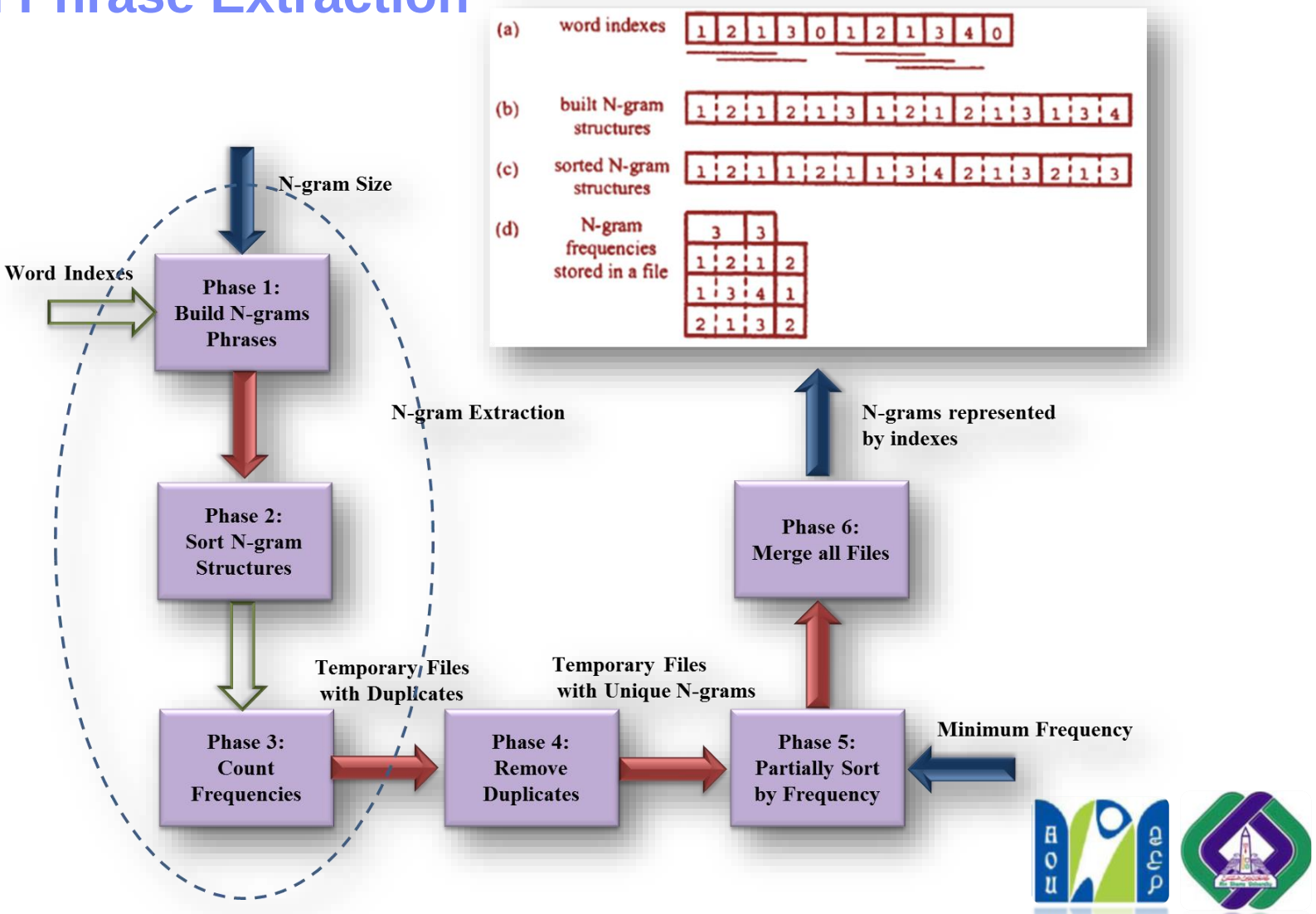
  (Rule-based method using morphological analysis + Lookup table)

- **Stemming**

  (Morphological analyzer + Arabic lexical lookups). This morphological analyzer was developed based on a linguistic approach.

- The stems' lexical indices stored in the dictionary are used to **index the inflected words**, according to the chosen stems.
- Morphological analyses are **disambiguated**, employing the **associated PoS tag** to each inflected word. If there are still more than one possible analysis, **Levenshtein** edit distance is then used to choose the most probable analysis.

Document Similarity Estimation Method
n-gram Phrase Extraction



Document Similarity Estimation Method
Document Modelling

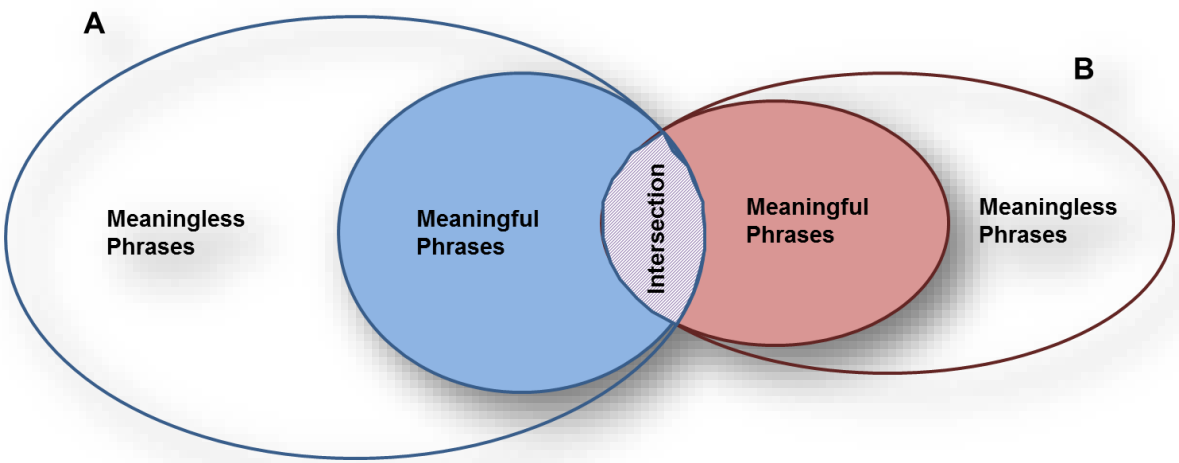
TF-IDF Matrix A is an n-by-m rectangular matrix which is composed of m vectors [A1, A2, …, Am], where the vector Aj represents n-gram phrases contained in document j.

$$a_{i,j} = \begin{cases} \frac{1}{2} + \frac{PF_{i,j} \cdot \log\left(\frac{|N|}{DF_i}\right)}{2 \cdot \max_j\left(PF_{i,j}\right) \cdot \log(|M|)}, & \text{if phrase i occurs in document j} \\ 0, & \text{otherwise} \end{cases}$$

Where each vector Aj is composed of n elements $a_{i,j}$ representing the weighted occurrence frequency of phrase i in document j. $PF_{i,j}$ represents the occurrence frequency of phrase i in document j, DF_i represents the number of documents where phrase i occurs, and finally |M| is the number of all documents

Document Similarity Estimation Method
Phrase Analysis and Reduction/Filtering

- The **phrases existing just in one document are removed right away** since they are not plagiarized in any other document.
- We propose to **remove such phrases that are contained in more than μ+σ documents**, where μ is the mean document frequency and σ is the standard deviation from the mean document frequency. In other words, it removes all common phrases from the documents.



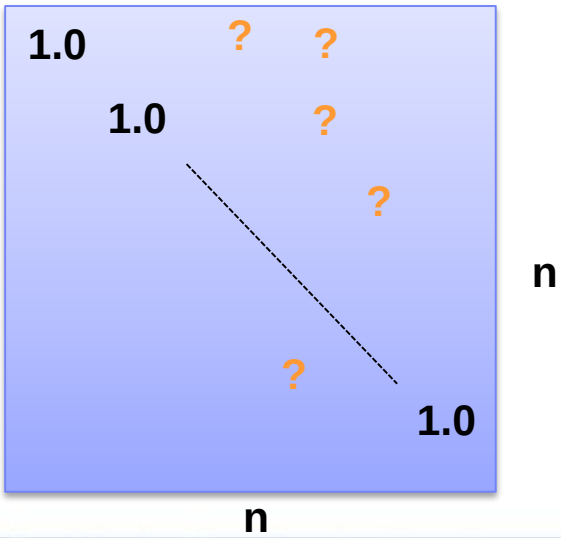
Document Similarity Estimation Method
Phrase Pair-wise Matching

- **Matching Cost Matrix**
- **Bipartite Matching**
- **Heuristic Matching**

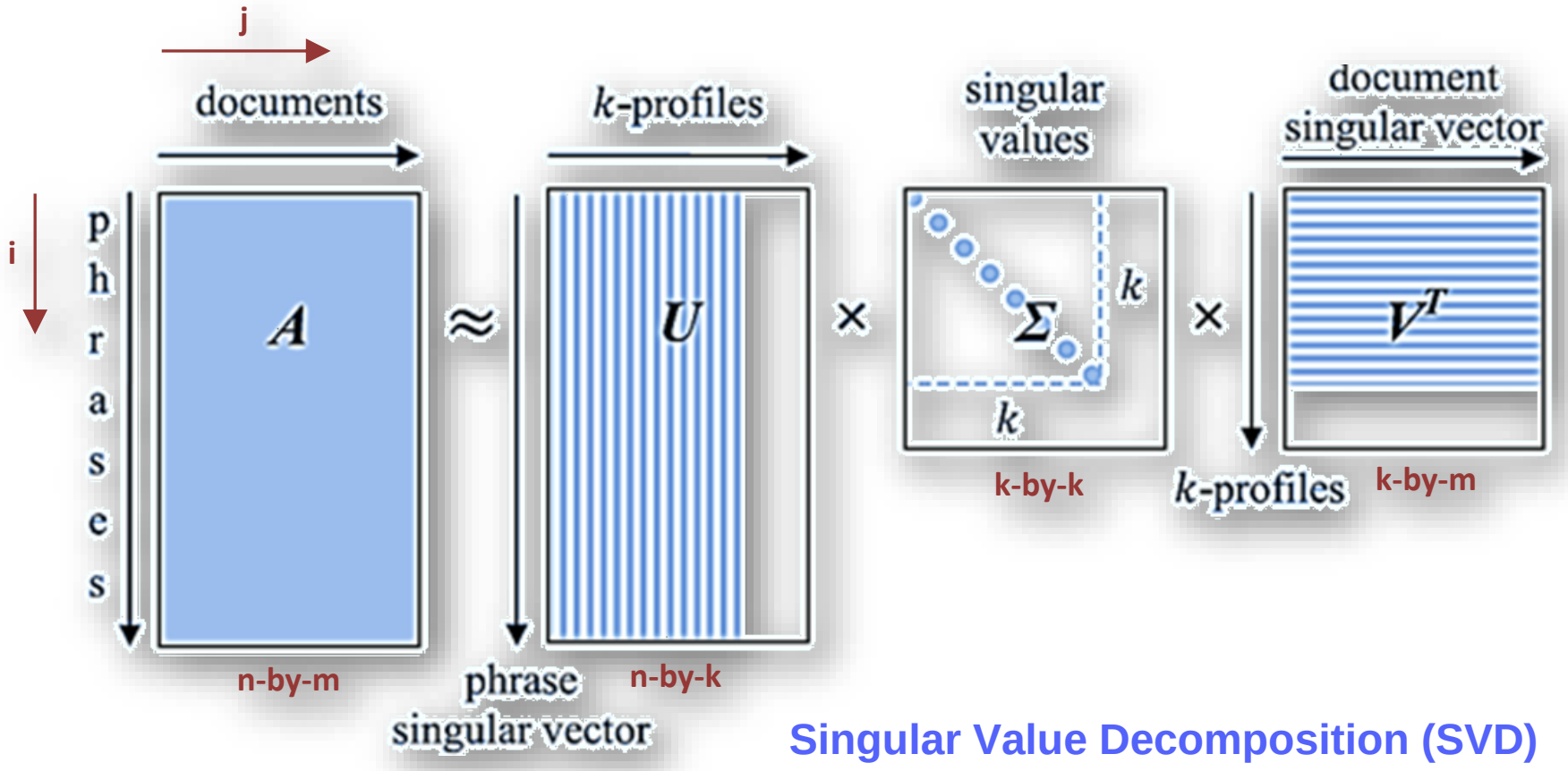
Dice Coefficient $\frac{2\times|X\cap Y|}{|X|+|Y|}$

	X	*Y*	
n-gram phrase 1	$S1_1$	$S2_1$	n-gram phrase 2
	$S1_2$	$S2_2$	
	$S1_3$	$S2_3$	
	.	.	
	.	.	
	$S1_n$	$S2_n$	

Cost =



Document Similarity Estimation Method
Latent Semantic Analysis (LSA)



Singular Value Decomposition (SVD)

Ceska, Z.: Plagiarism Detection Based on Singular Value Decomposition. In: A. Ranta, & B. Nordström, (Eds.), Lecture Notes in Computer Science, vol. 5221 (Advances in Natural Language Processing), pp. 108-119, Springer, Heidelberg (2008)

Document Similarity Estimation Method
Document Similarity Normalization

$$B = \Sigma \times V^T$$

$$\mathrm{sim}_{\mathrm{SVD}} = \|B\|^T \times \|B\|$$

$$\mathrm{sim}(R,S) = \mathrm{sim}_{\mathrm{SVD}}(R,S) \cdot \frac{\sqrt{|N_{red}(R)|.|N_{red}(S)|}}{\min\left(|N_{orig}(R)|.|N_{orig}(S)|\right)}$$

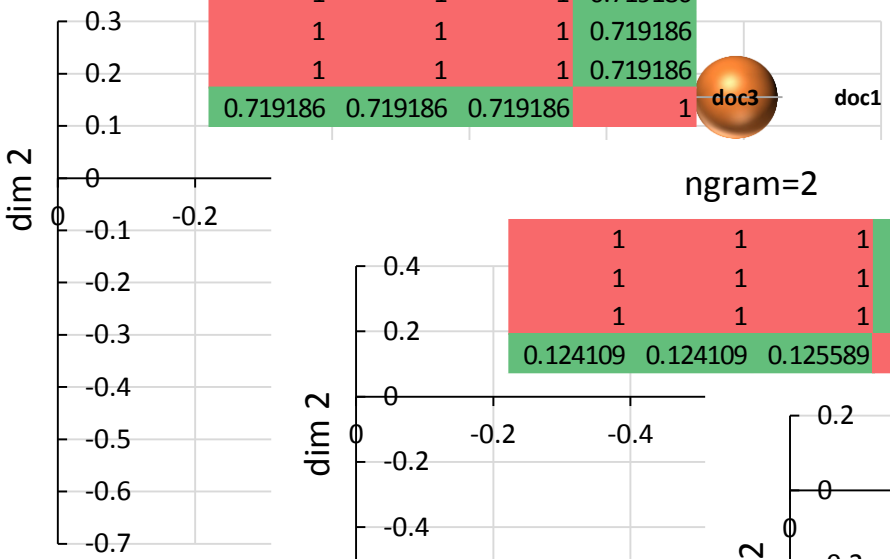
Ceska, Z., Fox, C.: The Influence of Text Pre-processing on Plagiarism Detection. In: Re-cent Advances in Natural Language Processing, RANLP 2009, pp. 55-59, Borovets, Bul-garia (2009)

$$\|A - A_k\|_F = \min_{rank(B) \le k} \|A - B\|_F = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_{r_A}^2}$$

Results and Discussions
Fundamental Experiment

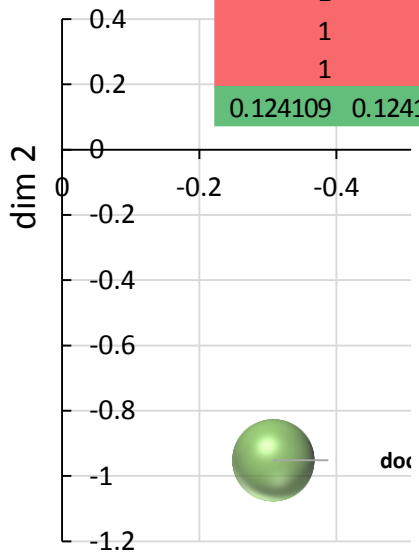
ngram=1

1	1	1	0.719186
1	1	1	0.719186
1	1	1	0.719186
0.719186	0.719186	0.719186	1



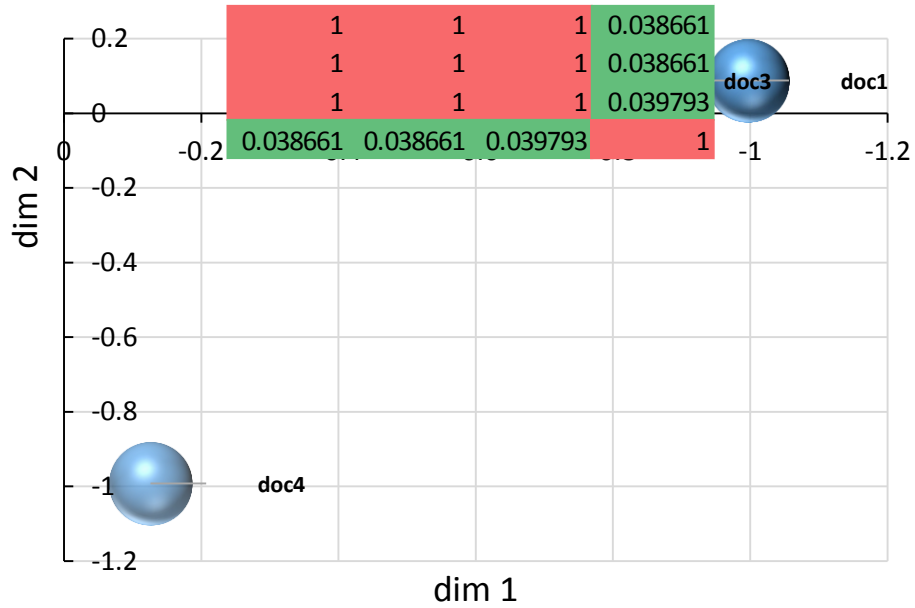
ngram=2

1	1	1	0.124109
1	1	1	0.124109
1	1	1	0.125589
0.124109	0.124109	0.125589	1



ngram=3

1	1	1	0.038661
1	1	1	0.038661
1	1	1	0.039793
0.038661	0.038661	0.039793	1



- تشير الإحصاءات التجارية في أسواق الخليج والديكور حيث يبلغ مجموع الاستثمارات التي دولار سنويا وذلك يبين حجم النشاط التى يتمتع
- تشير الإحصاءات التجارية ه الديكو ، حيث يبلغ مجموع
- جالي الديكور ط هذا السوق

Results and Discussions
Fundamental Experiment

		ngram = 1		
1	1	1	0.764057	0.764057
1	1	1	0.764057	0.764057
1	1	1	0.764057	0.764057
0.764057	0.764057	0.764057	1	1
0.764057	0.764057	0.764057	1	1

- أشارت الإحصاءات إلى أن منطقة الخليج لديها أسواق والأثاث ويتعدى إجمالي الاستثمارات خمسة مليارات السوق في الخليج .

		ngram = 2		
1	1	1	0.233244	0.233244
1	1	1	0.233244	0.233244
1	1	1	0.237958	0.237958
0.233244	0.233244	0.237958	1	1
0.233244	0.233244	0.237958	1	1

- أشارت التقييمات إلى أن بقعة تفوقة في مجالي الزخرفة والـ في ات دولار سنويا مما يدل على حجم نشاط هذا السوق

		ngram = 3		
1	1	1	0.094971	0.094971
1	1	1	0.094971	0.094971
1	1	1	0.099419	0.099419
0.094971	0.094971	0.099419	1	1
0.094971	0.094971	0.099419	1	1

Results and Discussions
Real Experiment 1

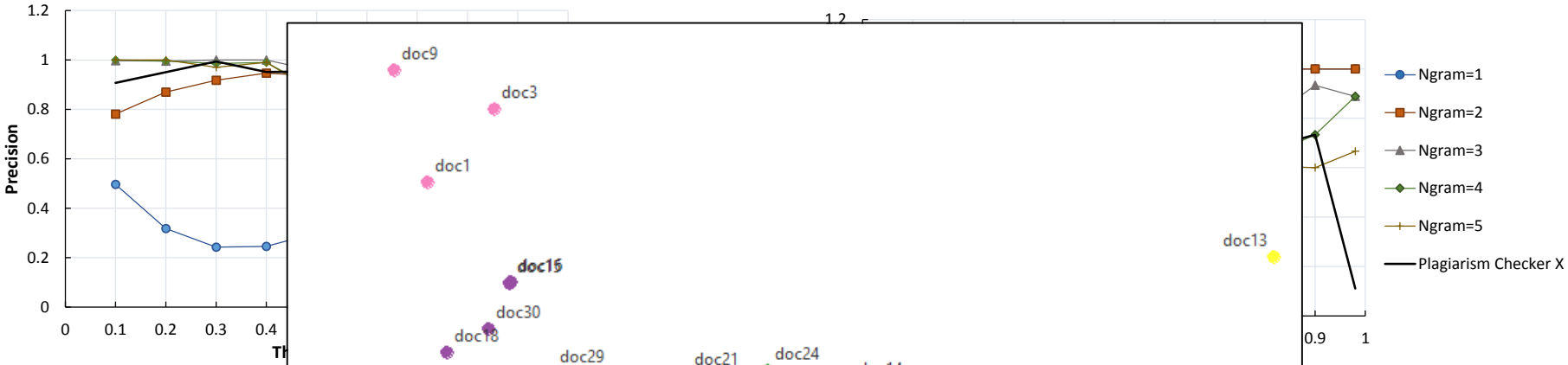
A real data set consists of 30 Arabic documents was

Files	1.txt	2.txt	3.txt	4.txt	5.txt	6.txt	7.txt	8.txt	9.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt	16.txt	17.txt	18.txt	19.txt	20.txt	21.txt	22.txt	23.txt	24.txt	25.txt	26.txt	27.txt	28.txt	29.txt	30.txt
1.txt	N/A	3%	59.00%	4%	5%	5%	8%	4%	51.00%	3%	7%	10%	2%	21%	80.00%	80.00%	21%	72.00%	8%	8%	57.00%	2%	18%	4%	7%	3%	27%	1%	70.00%	80.00%
2.txt	4.00%	N/A	5%	1%	25%	25%	15%	3%	15%	41.00%	30.00%	6%	3%	4%	6%	5%	5%	16%	6%	16%	5%	14%	18%	10%	14%	1%	2%	13%	5%	5%
3.txt	70.00%	5%	N/A	2%	7%	7%	11%	3%	52.00%	3%	9%	15%	3%	6%	75.00%	75.00%	6%	72.00%	9%	11%	43.00%	5%	22%	5%	13%	8%	31%	3%	66.00%	76.00%
4.txt	6%	2%	2%	N/A	17%	17%	2%	2%	3%	16%	16%	68.00%	1%	4%	5%	5%	3%	3%	95.00%	2%	10%	1%	13%	87.00%	5%	2%	2%	1%	8%	6%
5.txt	4%	21%	6%	8%	N/A	100.00%	26%	34.00%	9%	24%	93.00%	60.10%	3%	5%	5%	5%	11%	11%	57.00%	26%	7%	7%	33.00%	15%	21%	6%	10%	23%	4%	4%
6.txt	5%	21%	6%	8%	100.00%	N/A	26%	34.00%	9%	23%	93.00%	59.00%	3%	5%	5%	5%	11%	11%	57.00%	26%	7%	7%	33.00%	15%	21%	6%	10%	23%	4%	4%
7.txt	7%	13%	9%	2%	27%	27%	N/A	7%	16%	6%	30.00%	15%	3%	4%	7%	7%	5%	18%	15%	99.78%	6%	14%	23%	11%	17%	2%	9%	20%	6%	7%
8.txt	3%	2%	2%	2%	32.00%	31%	7%	N/A	2%	4%	34.00%	34.00%	1%	18%	21%	21%	12%	21%	32.00%	7%	3%	2%	14%	19%	26%	15%	9%	12%	2%	2%
9.txt	53.00%	18%	47.00%	2%	11%	11%	20%	3%	N/A	4%	7%	16%	3%	6%	63.00%	63.00%	7%	81.00%	8%	21%	40.00%	16%	25%	13%	22%	8%	21%	14%	77.00%	68.00%
10.txt	4%	63.00%	4%	14%	48.00%	46.00%	9%	8%	4%	N/A	66.76%	24%	3%	2%	9%	8%	7%	5%	25%	9%	12%	3%	24%	14%	6%	3%	3%	5%	7%	8%
11.txt	5%	21%	6%	7%	78.00%	76%	20%	32.00%	5%	31.00%	N/A	63.00%	3%	4%	5%	5%	11%	7%	60.00%	21%	7%	3%	26%	9%	16%	5%	10%	17%	5%	5%
12.txt	7%	4%	9%	27%	47.00%	43.00%	11%	29%	9%	11%	56.00%	N/A	2%	5%	9%	9%	10%	9%	81.00%	12%	7%	3%	22%	25%	12%	8%	9%	16%	10%	9%
13.txt	3%	4%	5%	1%	6%	5%	5%	3%	4%	2%	6%	4%	N/A	4%	3%	3%	4%	5%	4%	6%	3%	4%	4%	3%	3%	2%	2%	3%	3%	3%
14.txt	43.00%	7%	10%	5%	10%	10%	8%	41.00%	9%	2%	10%	13%	4%	N/A	56.00%	56.00%	47.00%	48.00%	13%	9%	24%	7%	8%	55.00%	67.00%	10%	3%	6%	8%	7%
15.txt	52.00%	4%	44.00%	3%	4%	4%	6%	22%	40.00%	5%	5%	10%	1%	22%	N/A	100.00%	4%	73.00%	6%	6%	42.00%	2%	22%	37.00%	25%	17%	19%	1%	56.00%	67.00%
16.txt	52.00%	4%	44.00%	3%	4%	4%	6%	22%	40.00%	4%	5%	10%	2%	22%	100.00%	N/A	4%	70.00%	6%	6%	40.00%	2%	22%	36.00%	25%	17%	19%	1%	57.00%	65.00%
17.txt	36.60%	8%	9%	3%	23%	23%	9%	30.00%	9%	7%	30.00%	26%	4%	43.00%	7%	7%	N/A	10%	33.00%	11%	23%	6%	18%	10%	38.00%	46.00%	6%	12%	8%	8%
18.txt	62.00%	14%	55.00%	2%	10%	10%	17%	23%	58.00%	3%	8%	12%	3%	23%	85.00%	85.00%	6%	N/A	8%	18%	36.00%	13%	25%	27%	39.00%	11%	22%	11%	58.00%	63.00%
19.txt	6%	4%	6%	40.00%	49.00%	49.00%	12%	34.00%	6%	13%	61.00%	89.00%	2%	5%	7%	7%	13%	7%	N/A	13%	8%	3%	24%	35.00%	12%	5%	10%	18%	5%	6%
20.txt	7%	14%	9%	2%	31.00%	30.00%	99.00%	7%	17%	6%	26%	15%	3%	4%	7%	7%	6%	19%	15%	N/A	6%	13%	23%	10%	16%	2%	9%	19%	6%	7%
21.txt	67.00%	5%	48.00%	8%	8%	8%	8%	4%	40.00%	9%	9%	11%	2%	15%	70.00%	70.00%	15%	49.00%	11%	8%	N/A	2%	24%	8%	8%	3%	37.00%	2%	64.00%	70.01%
22.txt	2%	21%	6%	1%	13%	13%	24%	3%	21%	2%	6%	6%	4%	5%	3%	3%	5%	22%	5%	23%	3%	N/A	23%	16%	22%	1%	2%	21%	2%	3%
23.txt	16%	15%	18%	8%	35.00%	35.00%	22%	17%	19%	15%	36.00%	32.00%	2%	4%	28%	32.00%	9%	26%	29.00%	22%	19%	13%	N/A	27%	19%	13%	11%	18%	16%	16%
24.txt	3%	9%	4%	44.00%	16%	16%	11%	22%	10%	9%	13%	35.00%	2%	22%	47.00%	48.00%	6%	32.00%	40.00%	10%	6%	10%	31.00%	N/A	32.00%	16%	2%	9%	3%	3%
25.txt	9%	17%	14%	4%	33.00%	33.00%	23%	43.00%	23%	5%	26%	22%	3%	42.00%	43.00%	44.00%	24%	55.00%	19%	22%	9%	18%	26%	43.00%	N/A	20%	9%	22%	10%	10%
26.txt	2%	2%	6%	2%	6%	6%	2%	18%	6%	2%	7%	11%	1%	5%	21%	21%	20%	11%	6%	2%	2%	1%	13%	16%	15%	N/A	3%	2%	6%	6%
27.txt	44.00%	3%	44.00%	2%	16%	16%	14%	16%	25%	3%	18%	17%	2%	2%	38.00%	37.00%	5%	37.00%	17%	14%	42.00%	2%	17%	3%	10%	5%	N/A	14%	39.00%	39.00%
28.txt	2%	20%	5%	1%	43.00%	44.00%	35.00%	23%	19%	4%	37.00%	38.00%	2%	5%	2%	2%	10%	21%	37.00%	37.33%	2%	20%	34.00%	15%	30.00%	3%	15%	N/A	2%	2%
29.txt	80.00%	6%	72.00%	6%	4%	4%	7%	4%	70.00%	6%	7%	16%	2%	6%	97.00%	96.00%	6%	80.00%	7%	7%	67.00%	2%	20%	4%	9%	8%	36.00%	2%	N/A	99.00%
30.txt	80.00%	5%	76.00%	5%	4%	4%	7%	3%	59.00%	6%	5%	13%	1%	4%	100.00%	100.00%	5%	81.00%	7%	7%	67.00%	2%	19%	4%	8%	8%	34.00%	1%	90.00%	N/A

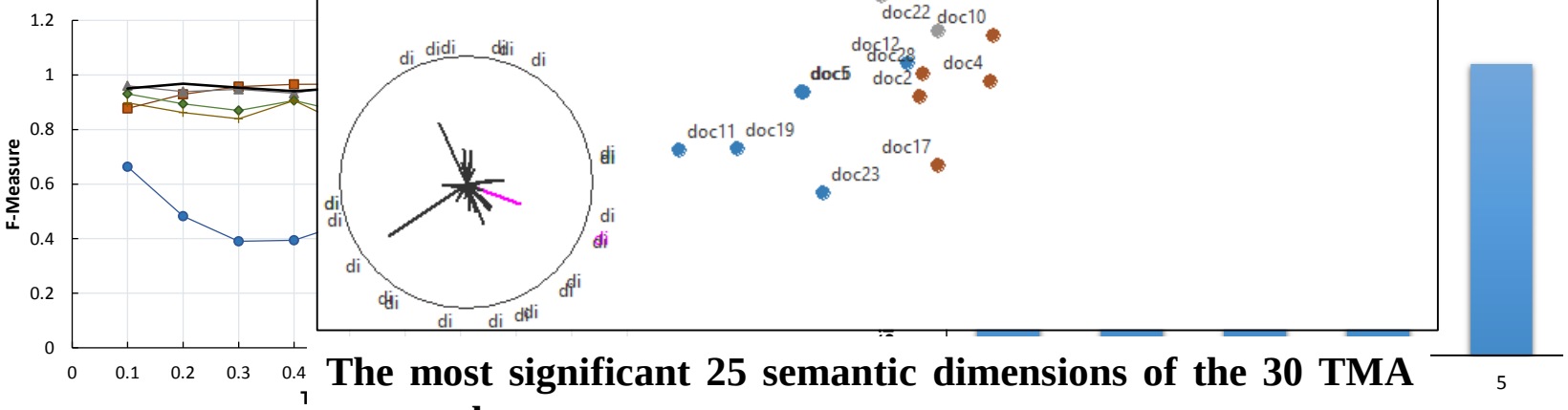
The subset-measure ground-truth data for the set of 30 TMA answer documents

the Course (131 Kuwait Civilization History, which is offered as a Level 1 Course for the fresh students.

Results and Discussions
Real Experiment 1

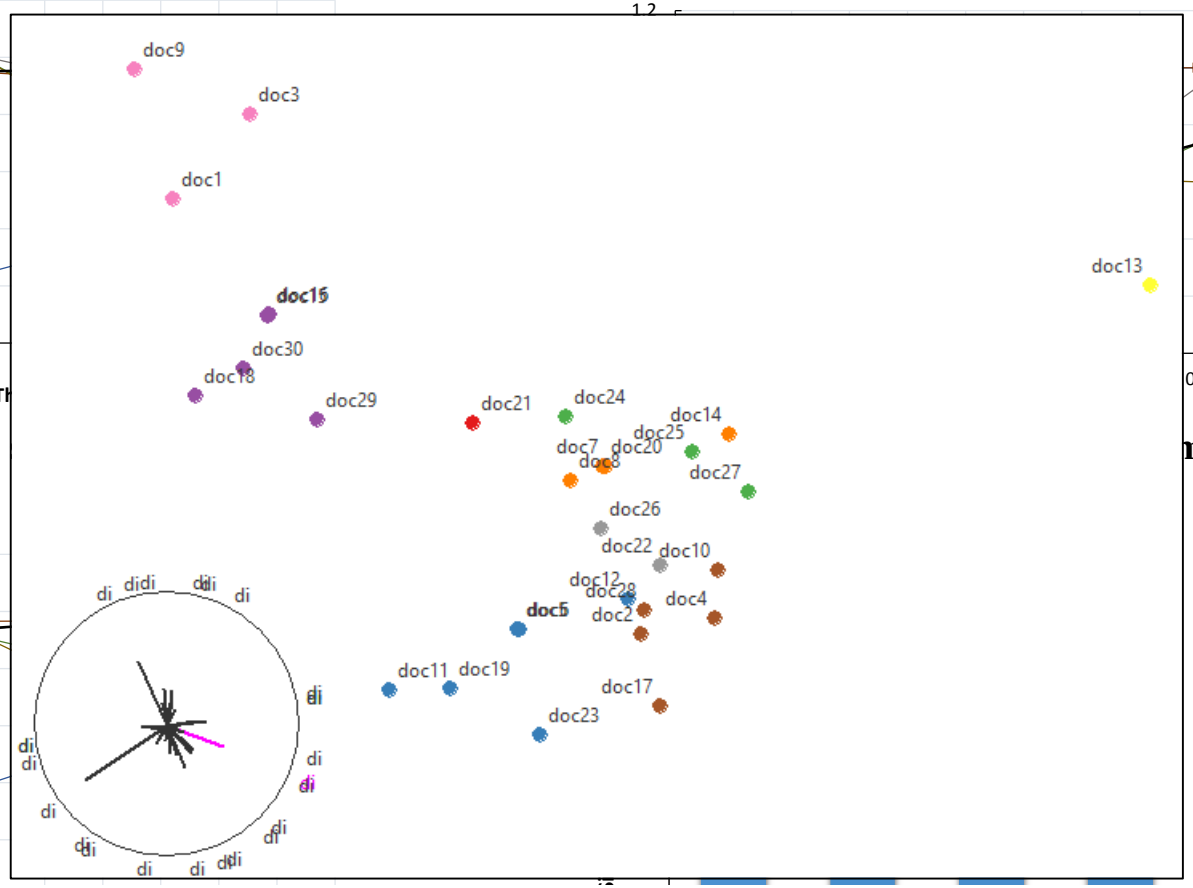


Precision of the ... mation



F-Measure of the subset measure estimation

Computing time for various ngram sizes



The most significant 25 semantic dimensions of the 30 TMA answer documents

Results and Discussions
Real Experiment 2

The first 5 documents are original while the 6th one is

	L1.txt	L2.txt	L3.txt	L4.txt	L5.txt	L6.txt	L7.txt	L8.txt	L9.txt
L1.txt	N/A	3%	4%	6%	3%	4%	6%	6%	7%
L2.txt	2%	N/A	3%	3%	3%	3%	5%	5%	5%
L3.txt	2%	3%	N/A	3%	5%	100%	45%	45%	45%
L4.txt	5%	4%	3%	N/A	3%	3%	70%	70%	70%
L5.txt	2%	3%	5%	2%	N/A	5%	5%	4%	5%
L6.txt								45%	44%
L7.txt								100%	100%
L8.txt	4%						100%	N/A	100%
L9.txt	5%						100%	100%	N/A

Maximum and average difference in estimating similarity among set of 9 documents, activating the synonyms component

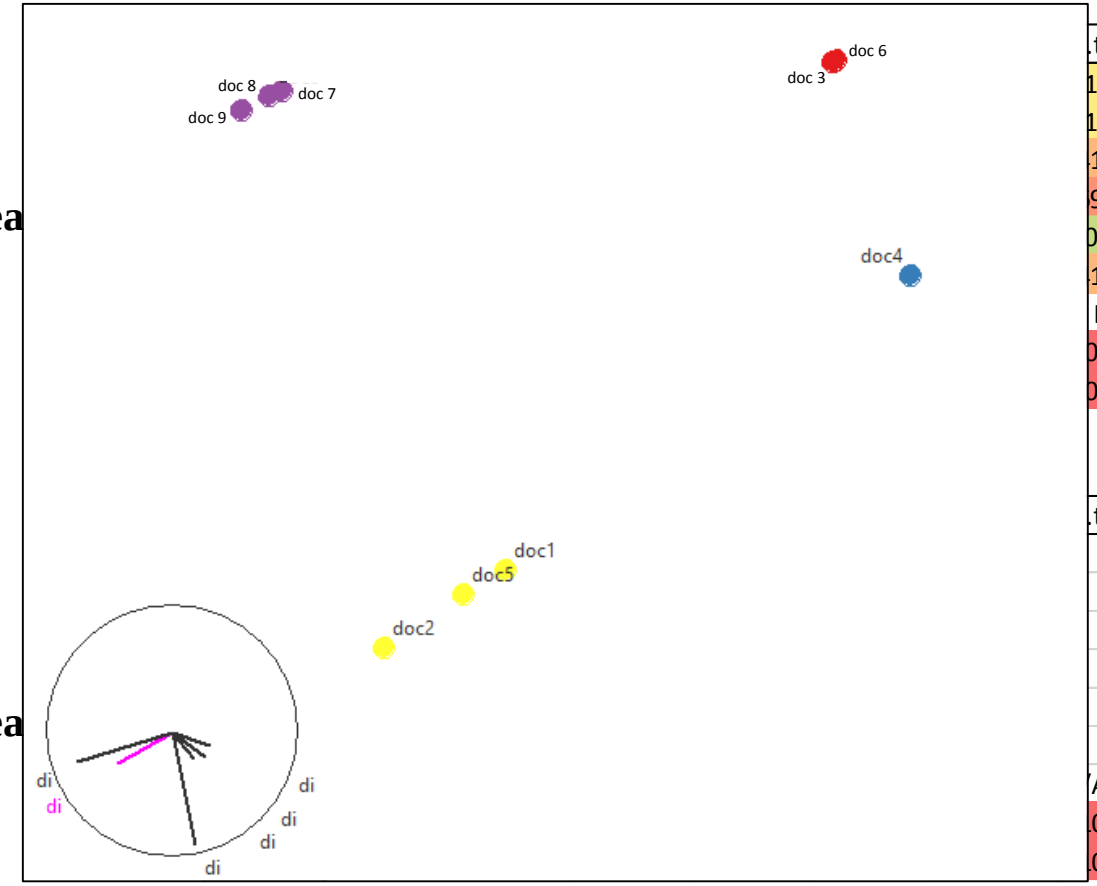
Method	Max |Diff|	Average |Diff|
Plagiarism Checker X	57.00%	8.02%
ngram = 1	48.74%	32.13%
ngram = 2	4.78%	2.53%
ngram = 3	4.78%	2.35%
ngram = 4	10.86%	3.59%
ngram = 5	17.40%	4.85%

The **documents**

document is ge
from the 7th one
restructuring 50% of the
statements.

Results and Discussions
Real Experiment 2

Subset mea



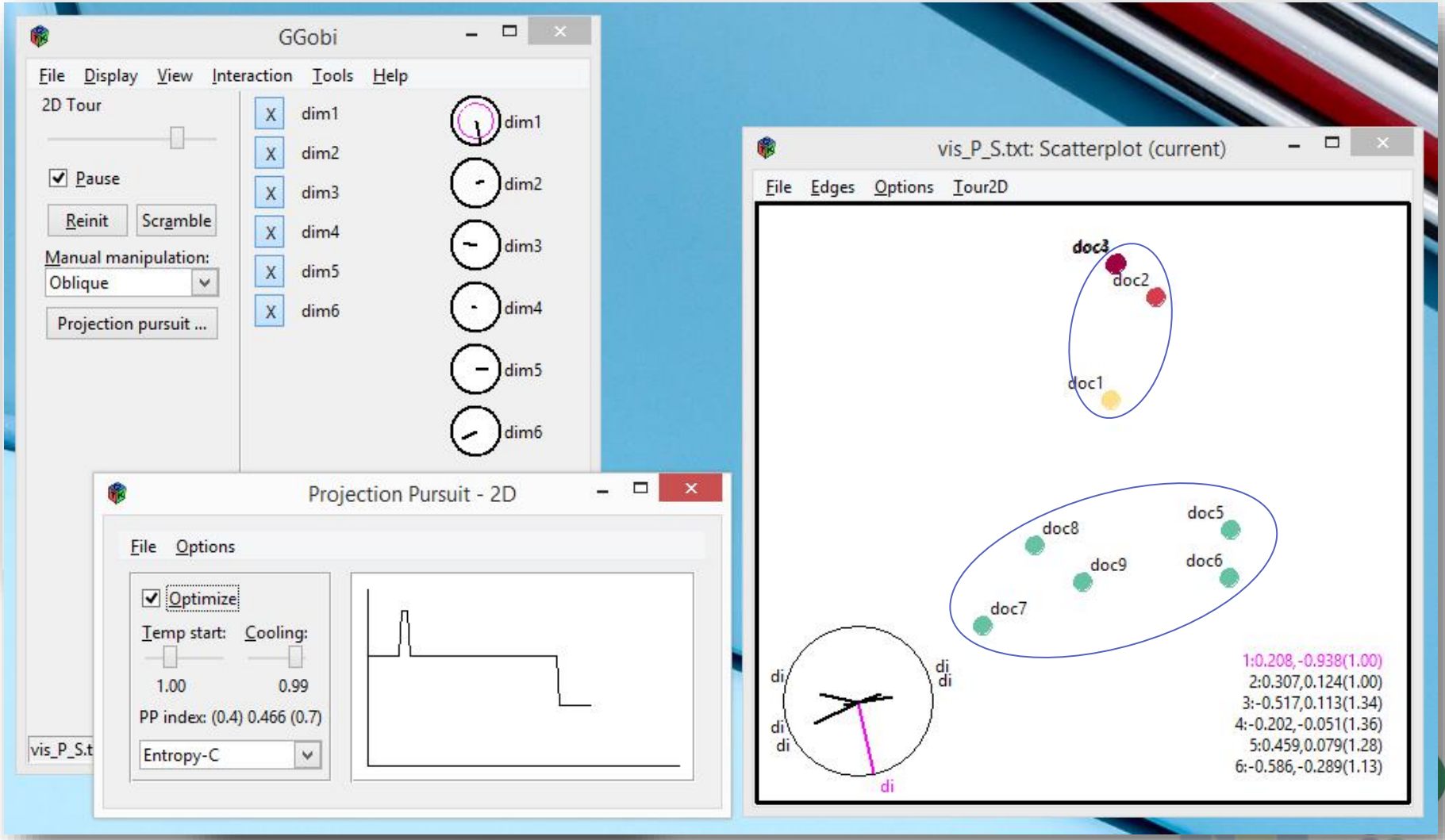
.txt	L8.txt	L9.txt
1.60%	1.66%	2.06%
1.55%	1.61%	1.61%
1.03%	40.80%	39.44%
9.63%	68.08%	65.96%
0.96%	0.96%	0.96%
1.23%	41.00%	39.62%
N/A	99.46%	99.02%
00.00%	N/A	99.59%
00.00%	100.00%	N/A

Similarity mea

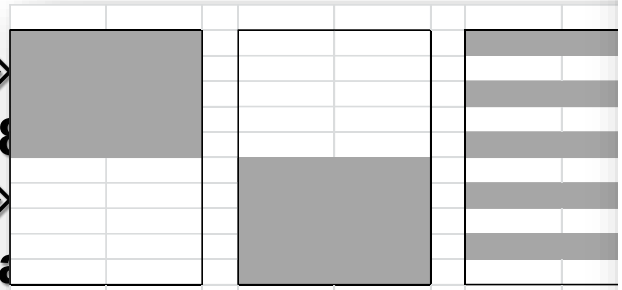
.txt	L8.txt	L9.txt
/A		
.00.00%	N/A	
.00.00%	100.00%	N/A

The most significant 6 semantic dimensions of the 9 documents used to estimate intelligent similarity

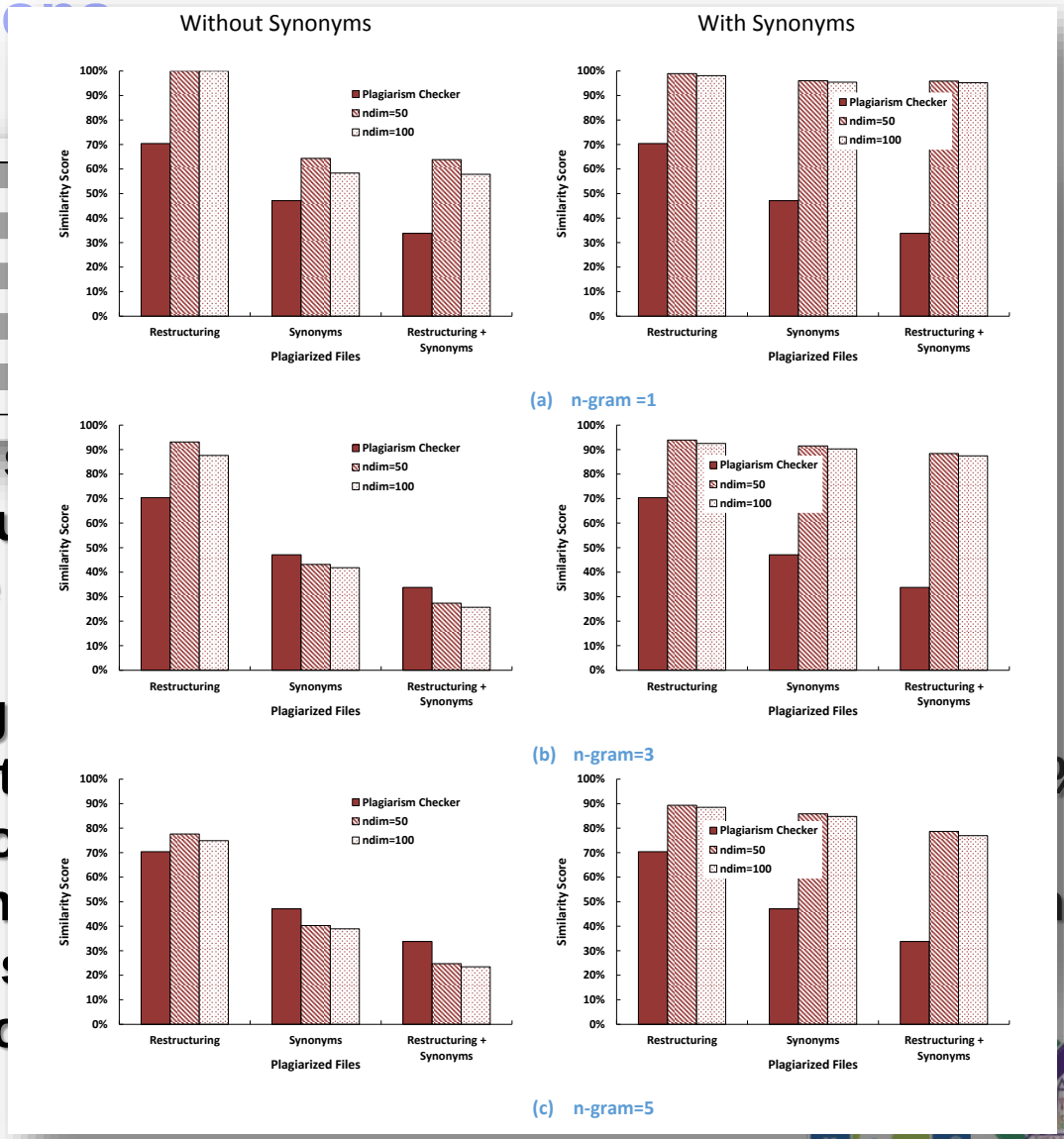
Results and Discussions
Real Experiment 3



Results and Discussions
Real Experiment 4



- to the Egyptian political
- From each original docu
  restructuring 50% of the
  schema shown.
- For the same target seg
  documents were generat
  of the words per each do
- The last group of docun
  50% of the words of the
  documents) to their sync



Conclusions

- A new plagiarism detection system for Arabic text documents is proposed based on modeling the relation between documents, under consideration, and their n-gram phrases.
- POS tagging is applied on the examined documents to support in resolving the morphological ambiguity during text normalization.
- Heuristic pairwise phrase matching algorithm is introduced to build the documents TF-IDF model, considering substitution of words with their synonyms.
- Finally, the hidden associations of the n-gram phrases contained in text documents are investigated using the LSA, employing the SVD.
- The proposed system exhibited strong capabilities in discovering literal plagiarism, and it could be considered as a serious step to-wards detecting intelligent plagiarism.

Questions

Thank You

Ashraf S. Hussein
ashrafh@acm.org