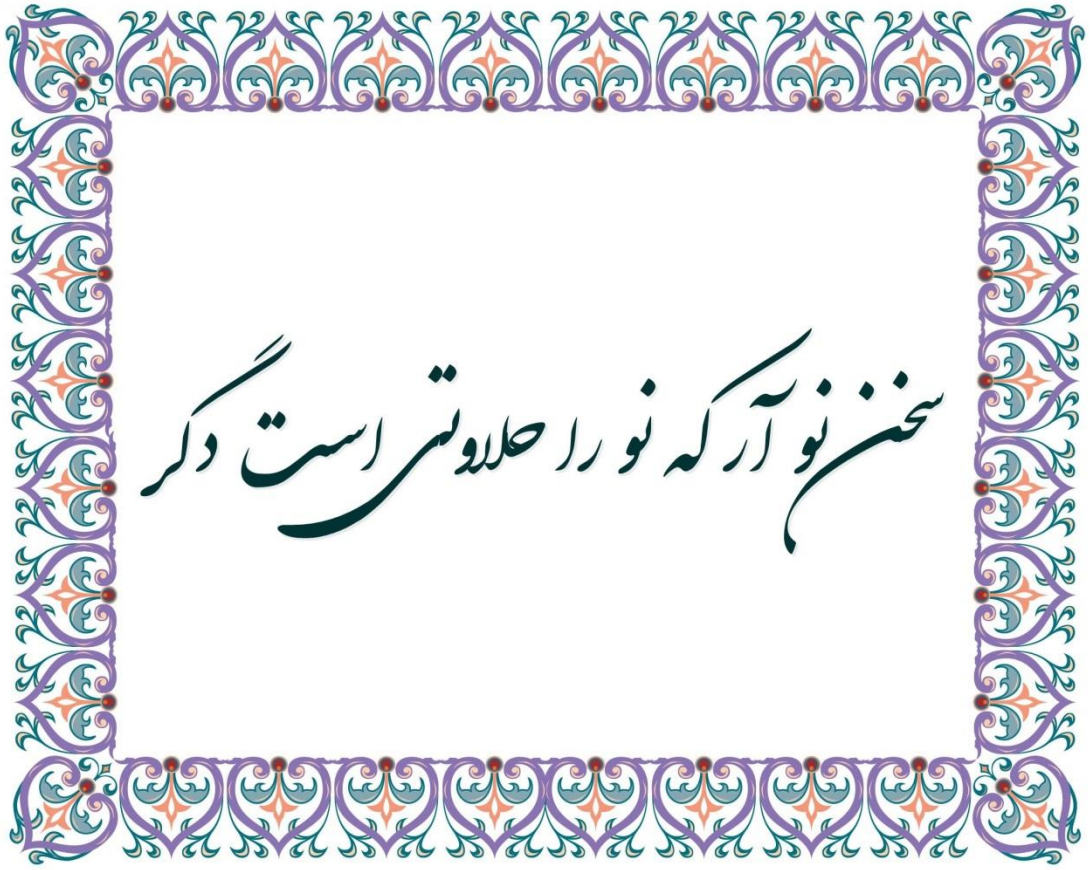
PERSIAN PLAGIARISM DETECTION USING SENTENCE CORRELATIONS

Muharram Mansoorizadeh and Taher Rahgooy
Bu-Ali Sina University
Hamedan, Iran



Outline

- Plagiarism Detection
- The Proposed Approach
- Results and Discussion

The Problem

- Plagiarism: Publishing someone else's words/works/ideas as one's own words/works/ideas.

- Scientific Plagiarism: Plagiarism activities targeting scientific publications
  - Usually works and ideas are plagiarized.

- Our Focus: Scientific Plagiarism in Persian Documents

Scientific Plagiarism is Really Hard!

- Every scientific field has a specialized terminology
  - Shared vocabulary of related research communities
  - Published as specialized glossaries and dictionaries

- Authors must adopt this vocabulary to get their works published
  - Using uncommon words and phrases would make reviewers suspect plagiarism
    - An example in machine learning community:
      - Feature selection, ~~Attribute elicitation~~, ~~Choosing attributes~~, ~~Characteristics extraction~~

- Automatic text analysis tools detect out of subject documents
  - Automatic topic detection, keyword extraction, and document clustering

Social Insights

- Mostly, lazy people do plagiarize or cheat
- They just alter first few paragraphs and sentences of each section
- Algorithms, formulas, and equations are hard to change!
- References and bibliography remain the same with minor changes.

The Proposed Approach

- Motivation: The plagiarized document would share important words, phrases and symbols with the original document
- The Idea: Use text similarity estimation and matching algorithms to retrieve susceptive cases
- Documents are mapped to TF-IDF vector space and analyzed

TF-IDF Representation of Documents

- Document set(corpus) D ={d_1, d_2 , …, d_N}, d_i is a document, N= |D|
- Vocabulary V ={ t_1, t_2, …, t_M}, the set of distinct terms in D, M=|V|
- Term Frequency of t_i in document d, $\boldsymbol{TF_i = \frac{F_i}{|d|+1}}$
- Inverse Document Frequency of t_i , $IDFi = \log(\frac{N}{N_i+1})$, N_i documents contain t_i
- TF and IDF combined as $TFIDF_i = TF_i \,.\, IDF_i$
- Document d is represented by vector v_{1xM}, where v(i) =$TFIDF_i$
- Similarity of two document vectors u and v is $\cos(u, v) = \frac{u.v}{\|u\|\|v\|}$

The Proposed Approach

Text Normalization

Split Sentences

Tokenize

Vector Space Representation

Map to TFIDF space

Decision

Construct similarity matrix and Threshold

Evaluation Metrics

- S : Plagiarism Cases, R: Set of Detections, $S_R \subseteq S$ are cases detected by detections in R, and $R_s \subseteq R$ are the detections of a given s.

- $\text{precision} = \frac{|S \cap R|}{|R|}$, $\text{recall} = \frac{|S \cap R|}{|S|}$, $\text{f\_measure} = 2\frac{precision \,.\, \text{recall}}{precision + recall}$

- $granularity(S,R) = \frac{1}{|S_R|}\sum_{s \in S_R}|R_s|$, $plagdet(S,R) = \frac{f\_measure}{\log_2(1+gran(S,R))}$

Detection Results on Main Corpus

- The corpus: 5830 Documents, 4118 plagiarism cases
  - Simulated and artificially generated samples

Threshold	Precision	Recall	F-Measure	Granularity	Plagdet
0.4	91	81	86	3.86	0.39
0.5	82	93	87	4.48	0.35

Detection Results on User Corpora

- Five independent corpora
  - Diverse dimensions and qualities

	Niknam	Samim	Mashhadirajab	ICTRC	Abnar
Documents	3218	4707	11089	5755	2470
Plags	2308	5862	11603	3745	12061
PlagDet	0.3	-	0.13	-	0.27

Discussion and Conclusion

- Straightforward approach for plagiarism detections
- Motivated by the vocabulary limitations in scientific contexts
- Reasonable performance in terms of precision and recall
- Easily scalable
  - Follows the architecture of modern information retrieval systems

Future Directions

- More advanced preprocessing and filtering
- Semantic normalization of documents
  - Context vocabulary normalization
- Topic based analysis

Selected References

- Asghari, Habibollah, et al. "Algorithms and Corpora for Persian Plagiarism Detection.“, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

- Potthast, Martin, et al. "An evaluation framework for plagiarism detection." *Proceedings of the 23rd international conference on computational linguistics: Posters*. Association for Computational Linguistics, 2010.

- Professors against plagiarism, http://pap.blog.ir/ [last visited: jan 22 2017]