Persian Plagdet 2016

A Text Alignment Algorithm Based on Prediction of Obfuscation Types Using SVM Neural Network

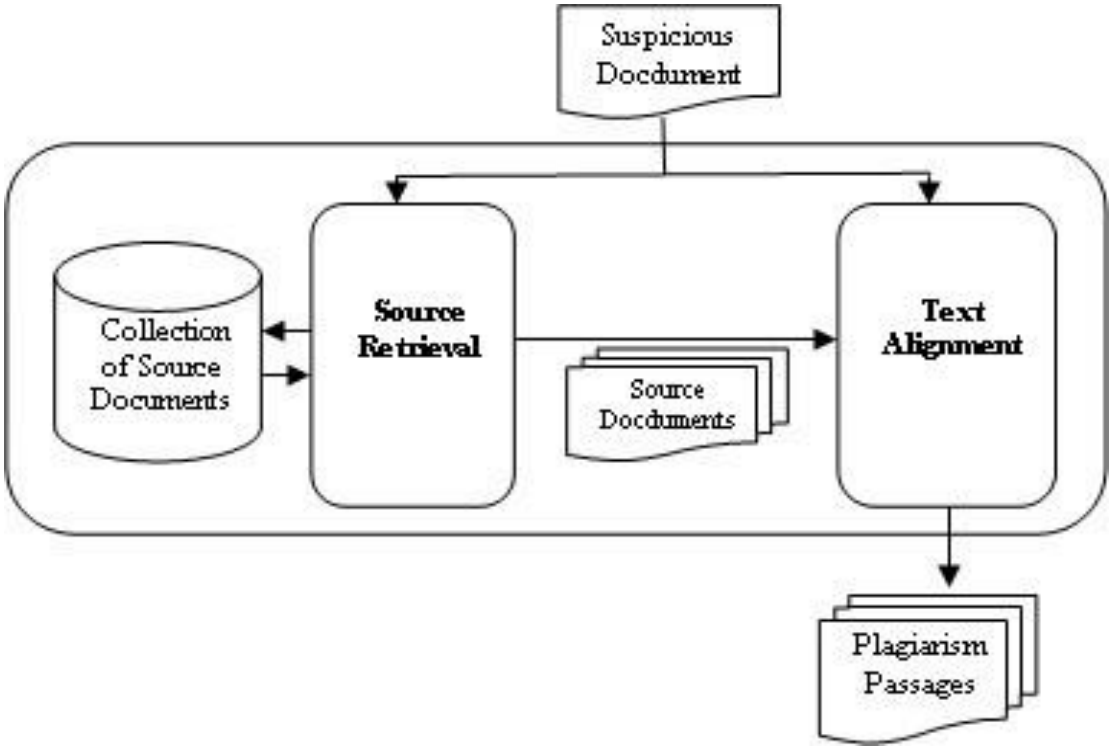
Fatemeh Mashhadirajab, Mehrnoush Shamsfard

NLP Research Lab, Faculty of Computer Science and Engineering, Shahid Beheshti University

Outline

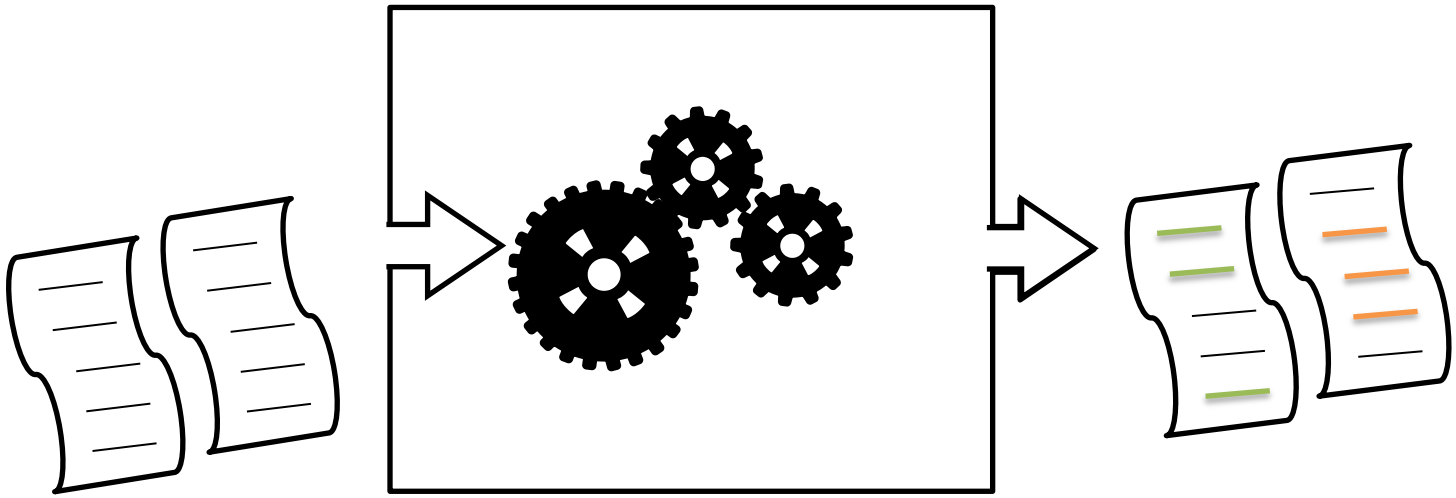
- Introduction
- The Proposed Approach
- Experiments
- Conclusions and Future Work
- References

Introduction



Plagiarism detection systems

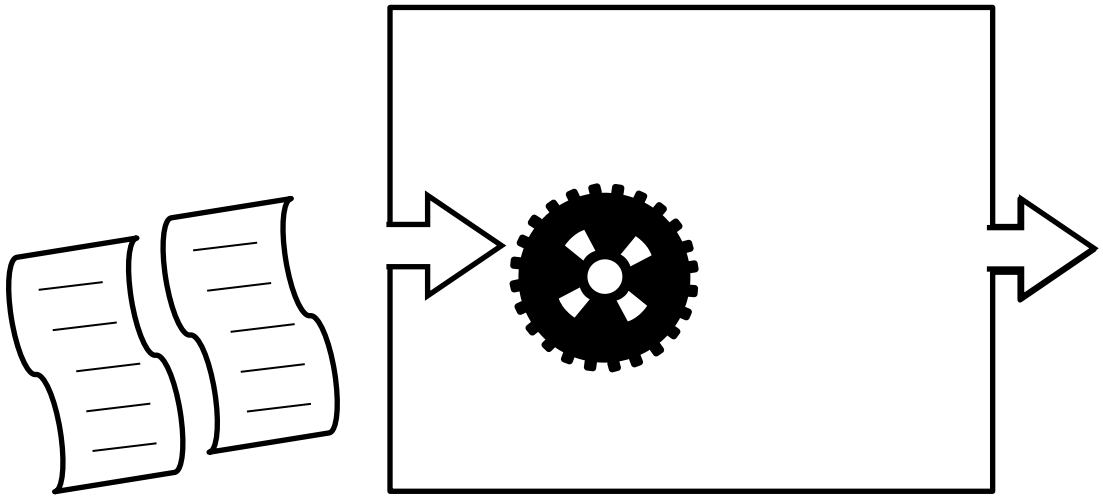
Introduction



A Text Alignment Algorithm

Introduction

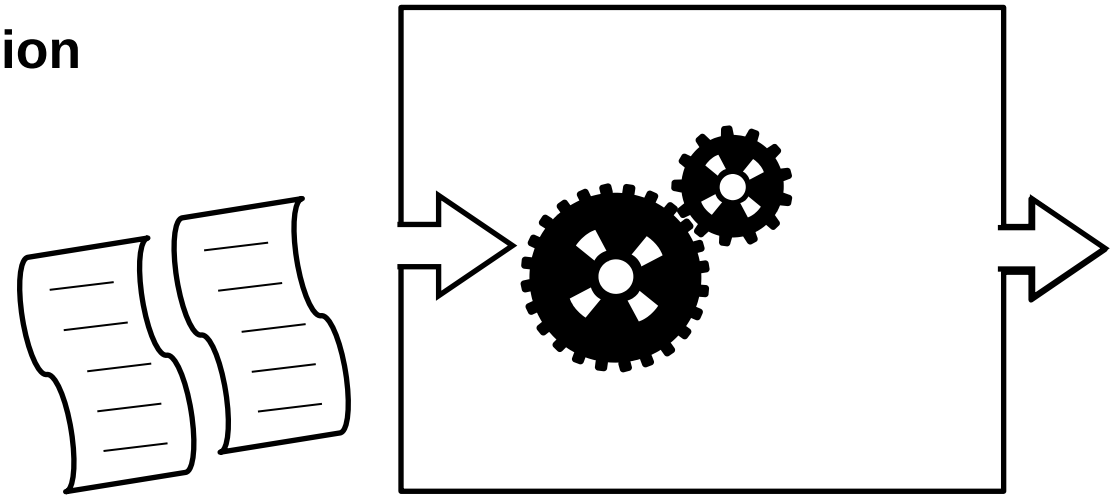
- **Seeding**



A Text Alignment Algorithm

Introduction

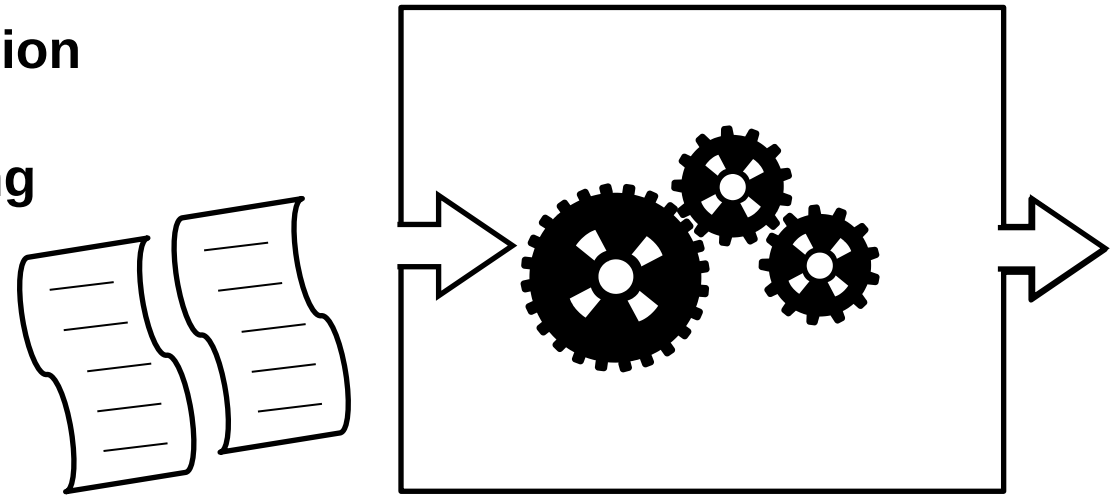
- **Seeding**
- **Extension**



A Text Alignment Algorithm

Introduction

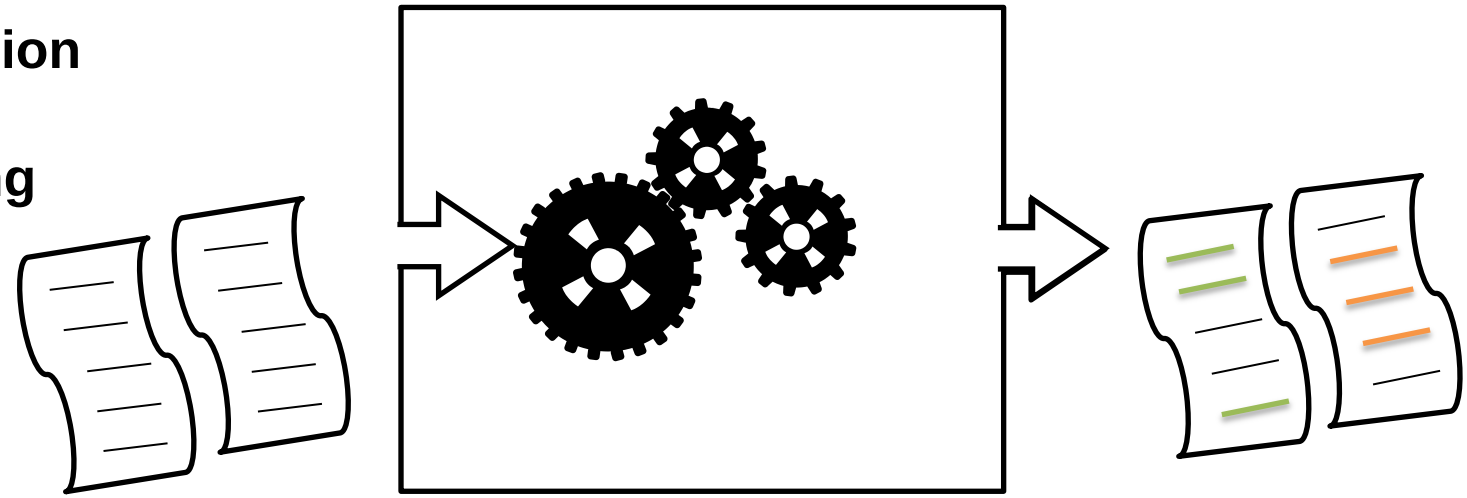
- **Seeding**
- **Extension**
- **Filtering**



A Text Alignment Algorithm

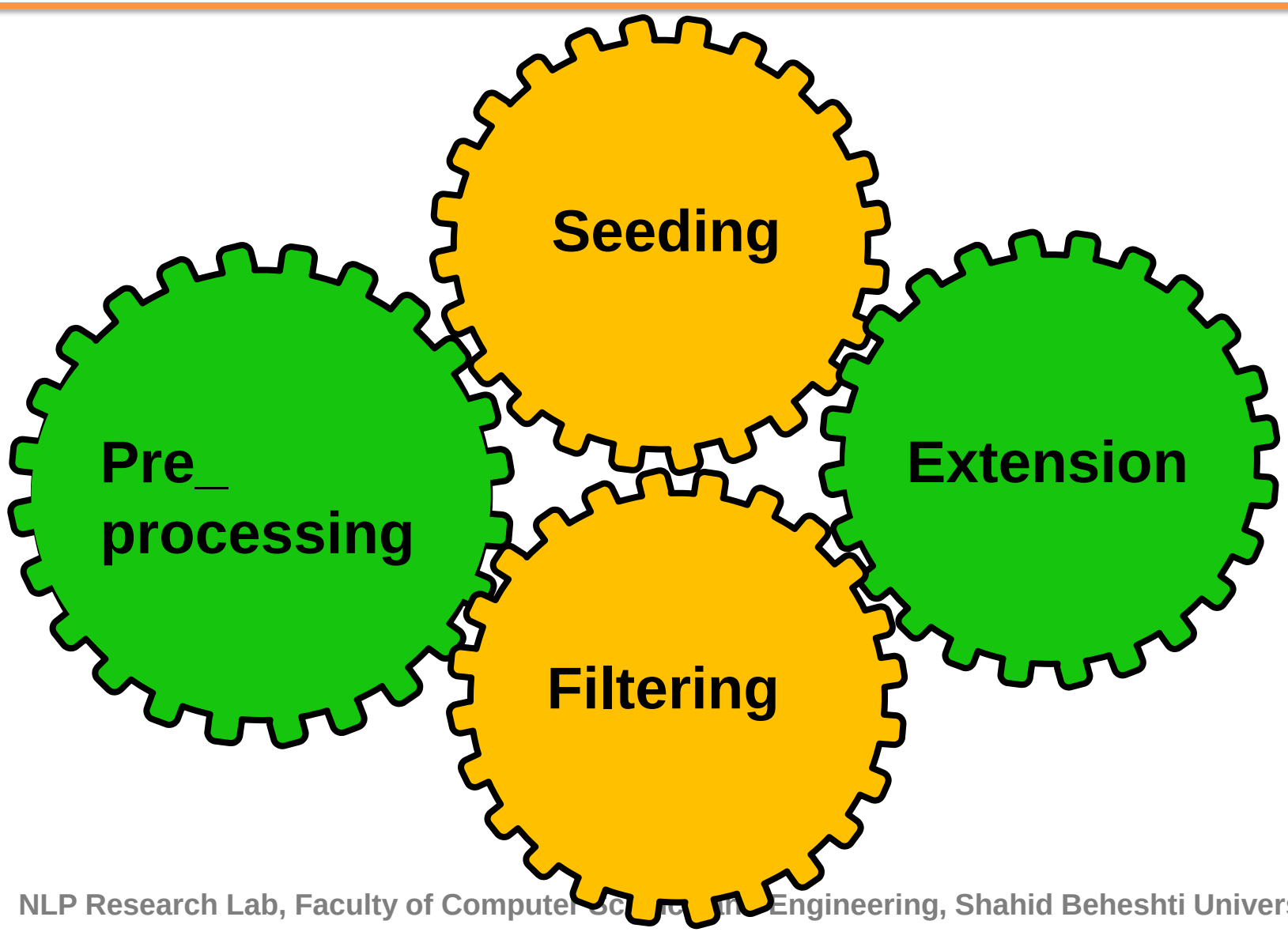
Introduction

- Preprocessing
- Seeding
- Extension
- Filtering

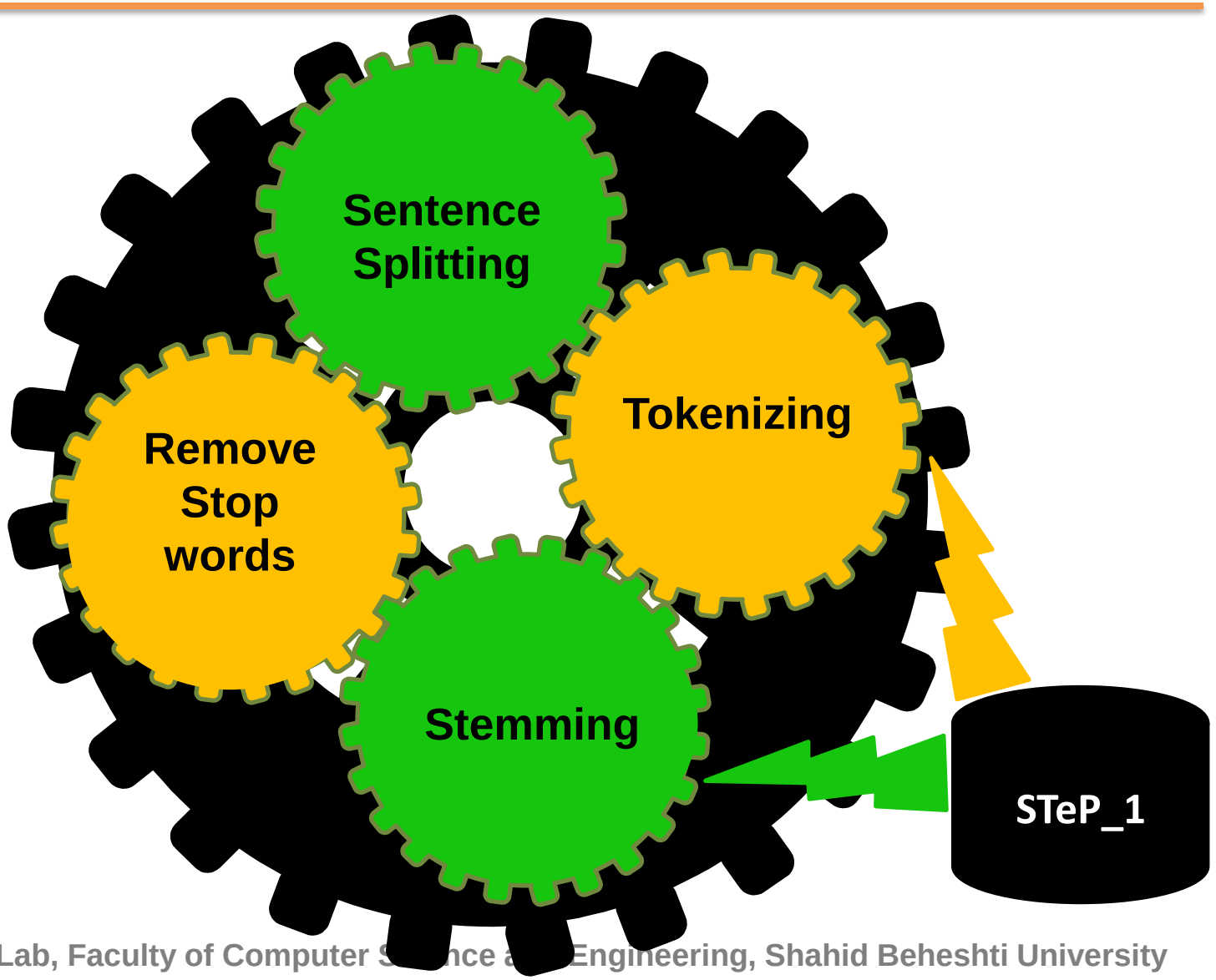


A Text Alignment Algorithm

The Proposed Approach



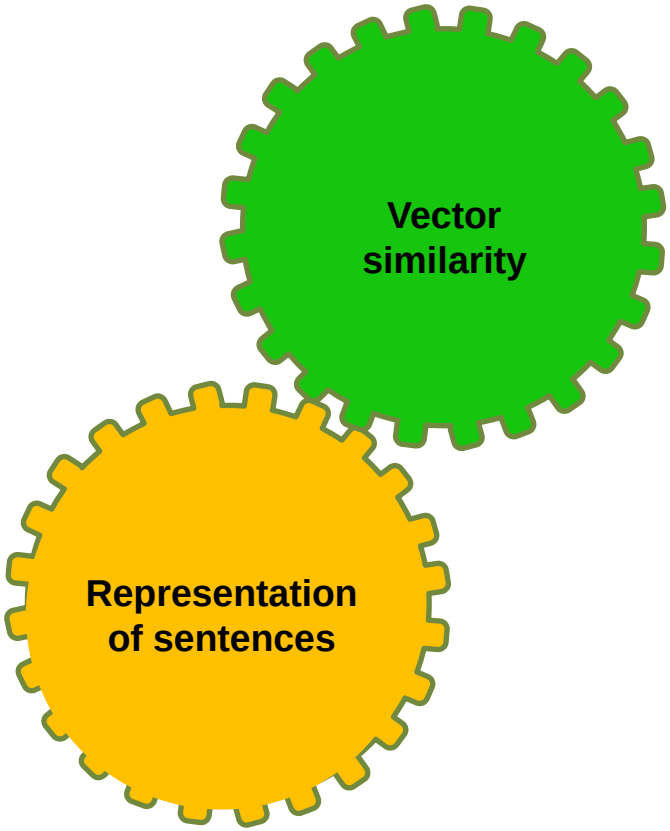
Preprocessing



Seeding

- **Vector representation of sentences:** **VSM**

Seeding

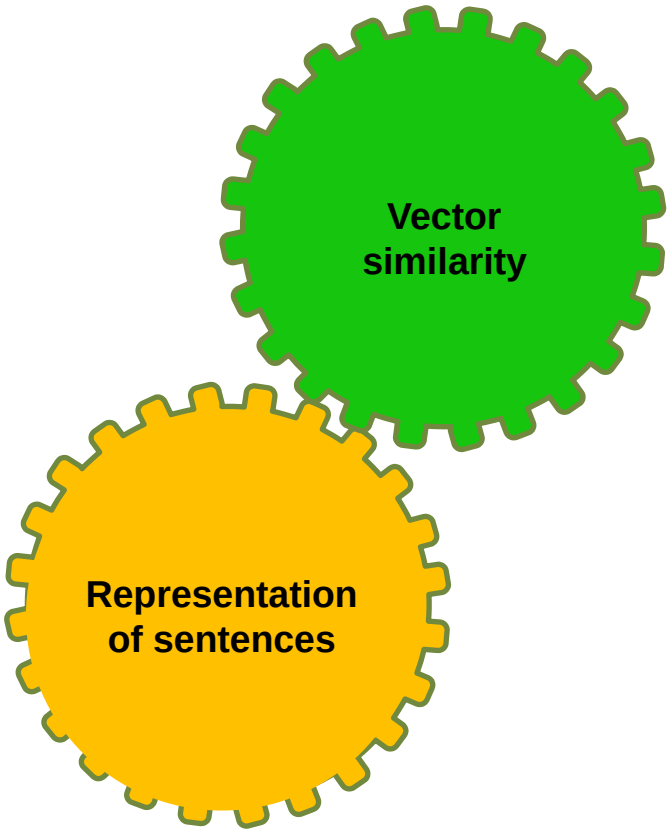


- **Vector similarity:** **Cosine Measure**
  Dice Coefficient

- **If** **Cosine> Threshold &**
  Dice > Threshold

Seed

Seeding

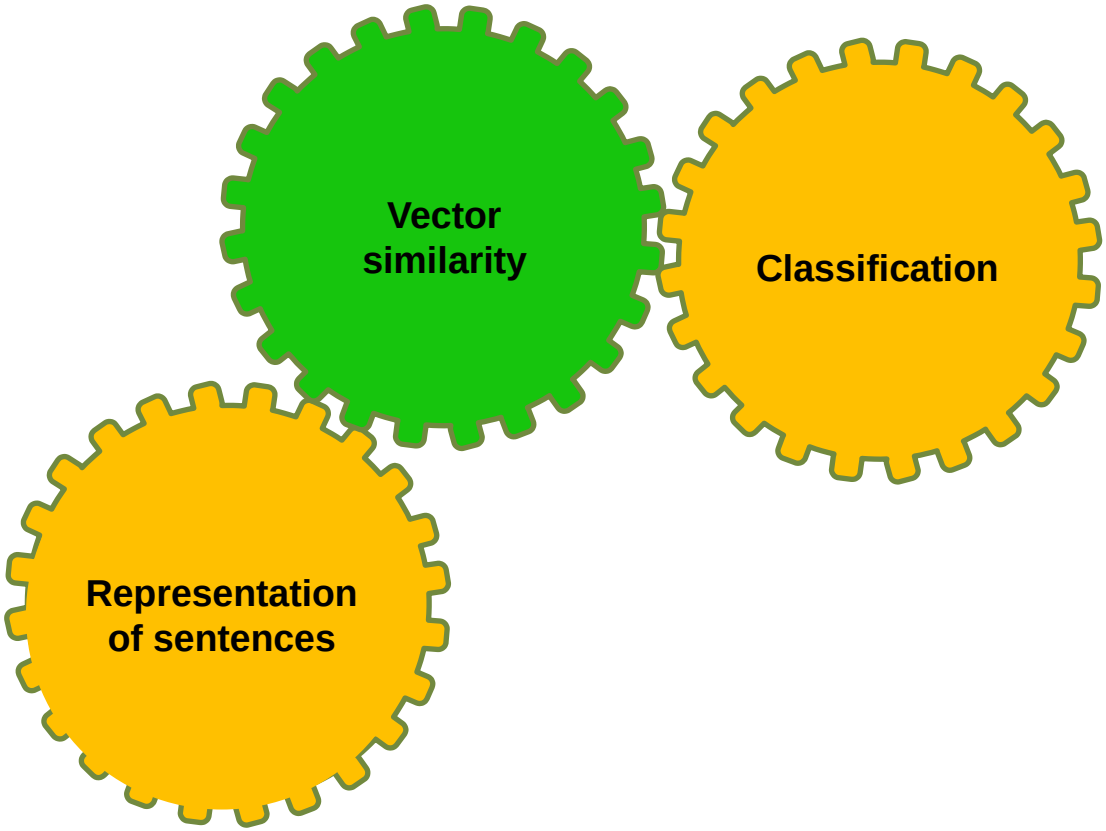


- **Vector similarity:** **Cosine Measure**
  Dice Coefficient
  - **Otherwise:**
  - **If threshold1<Cosine< threshold2**

Semantic Similarity

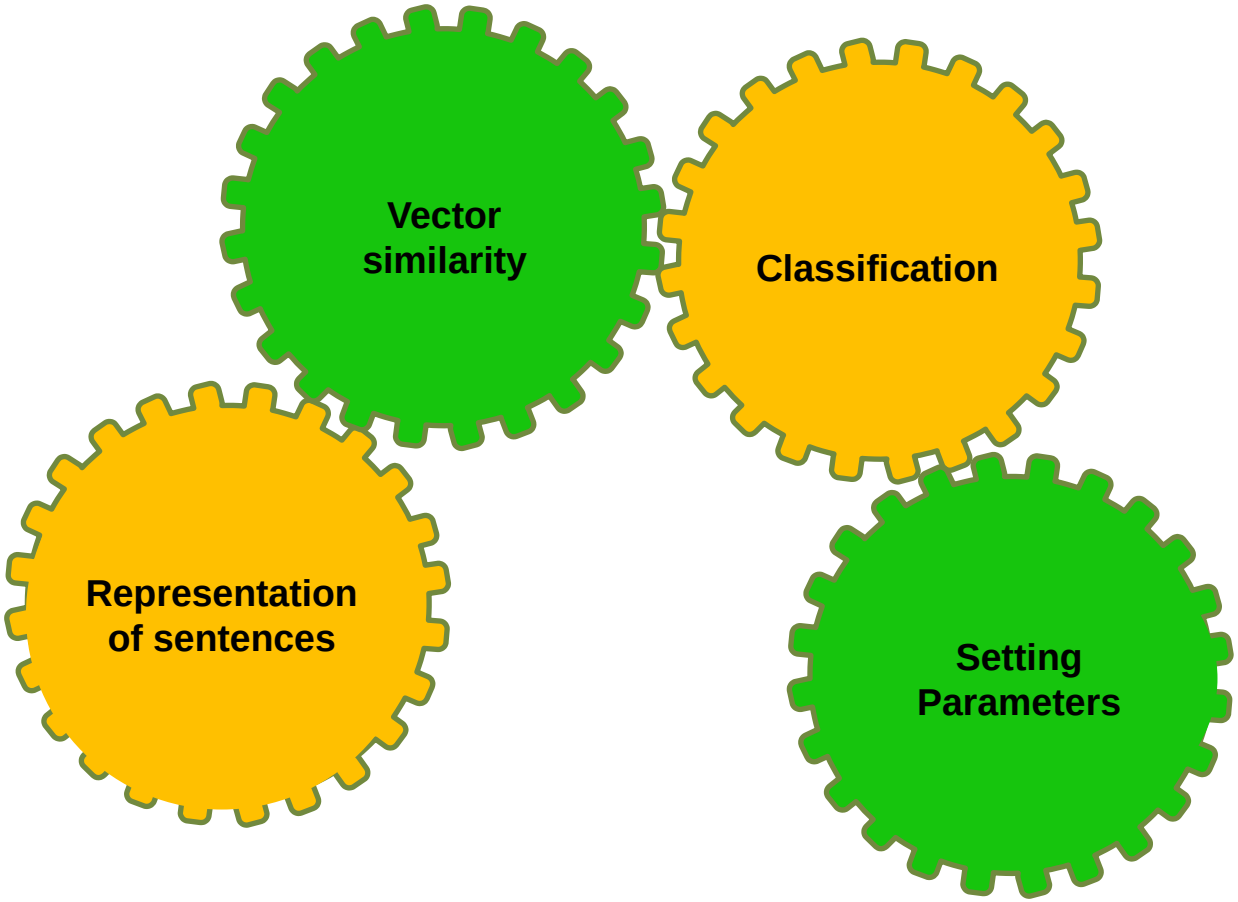
Seeding

- **Classification:** **SVM neural network**



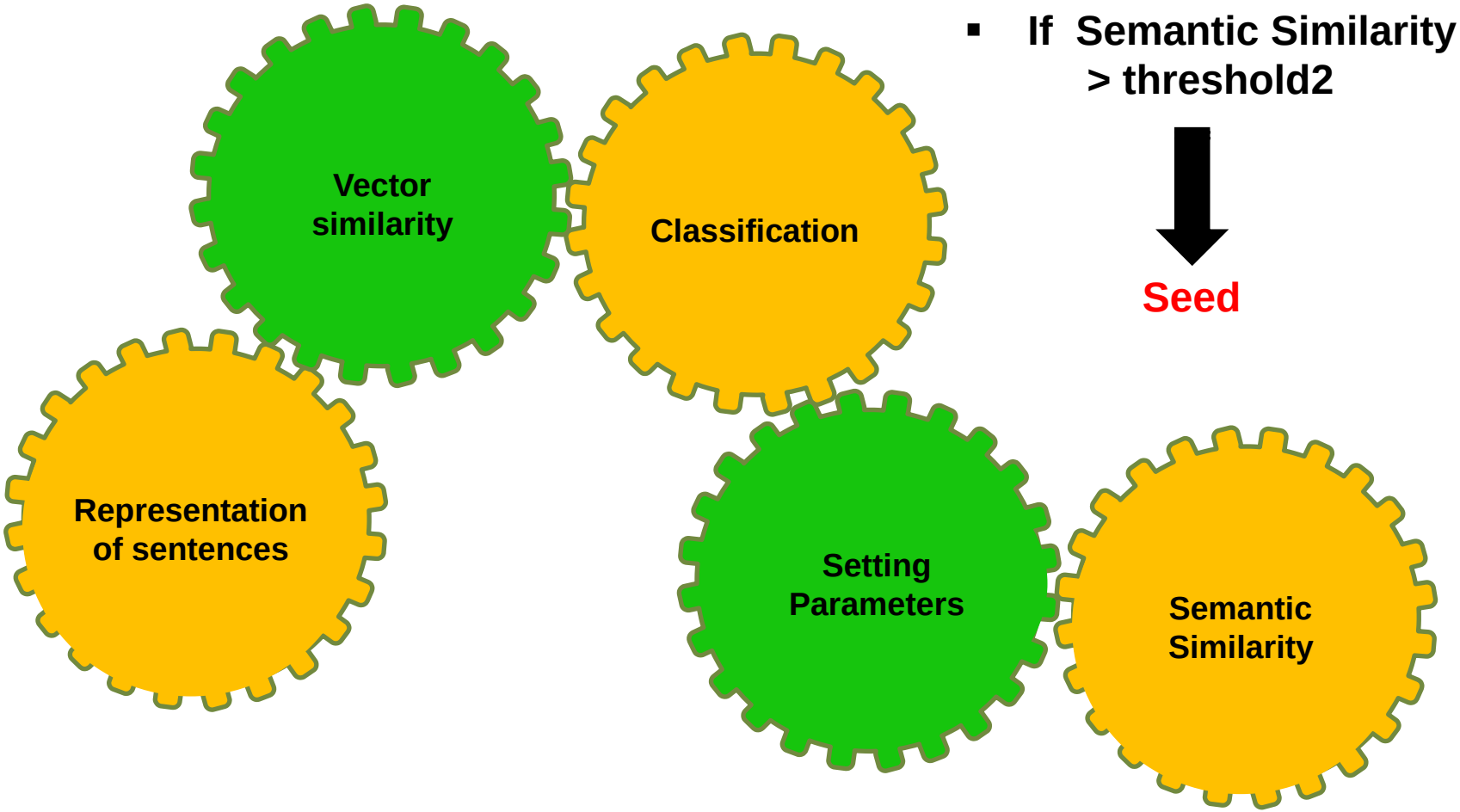
Seeding

- **Setting Parameters**



Seeding

- Semantic Similarity: FarsNet
- If Semantic Similarity > threshold2



Seed

Extension

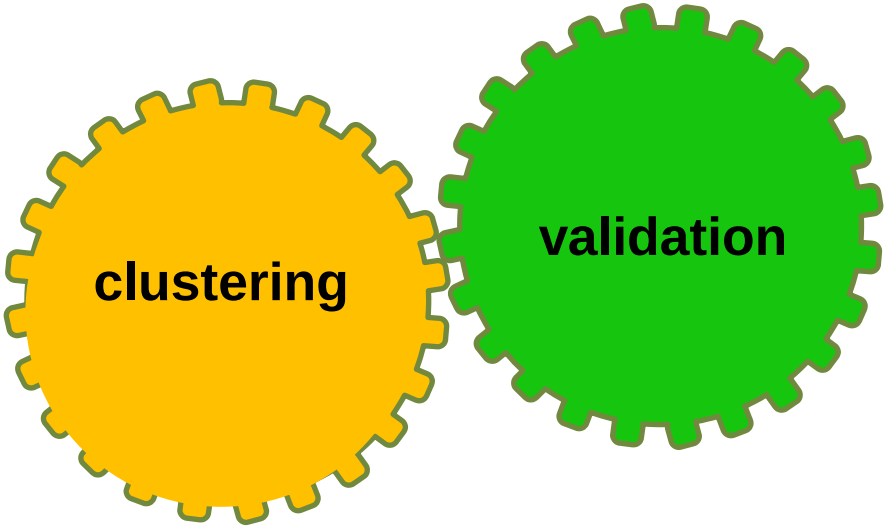
Clustering:

the seeds are clustered into passages.
In each passage, the seeds are not separated by more than a *maxgap* number of sentences.

Extension

Validation:

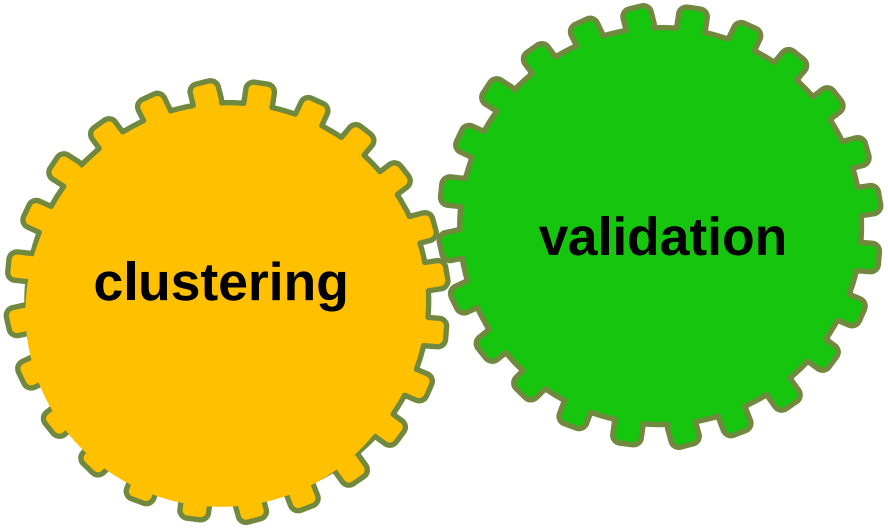
This stage assesses the resulting clusters from the clustering stage



Extension

Validation:

If semantic similarity in a pair of passages is less than a given threshold, then

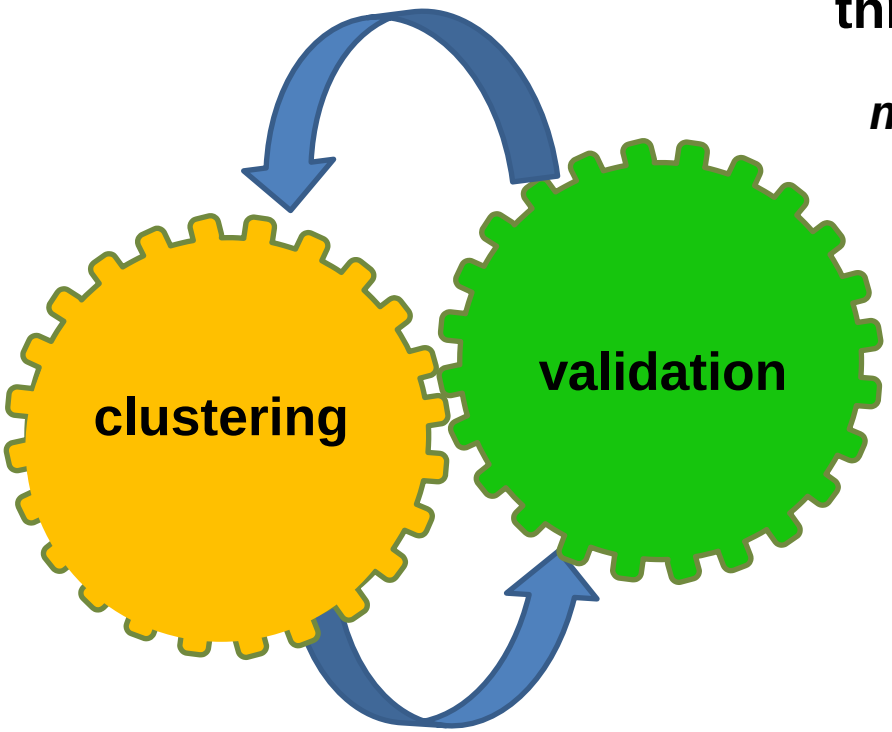


Extension

Validation:

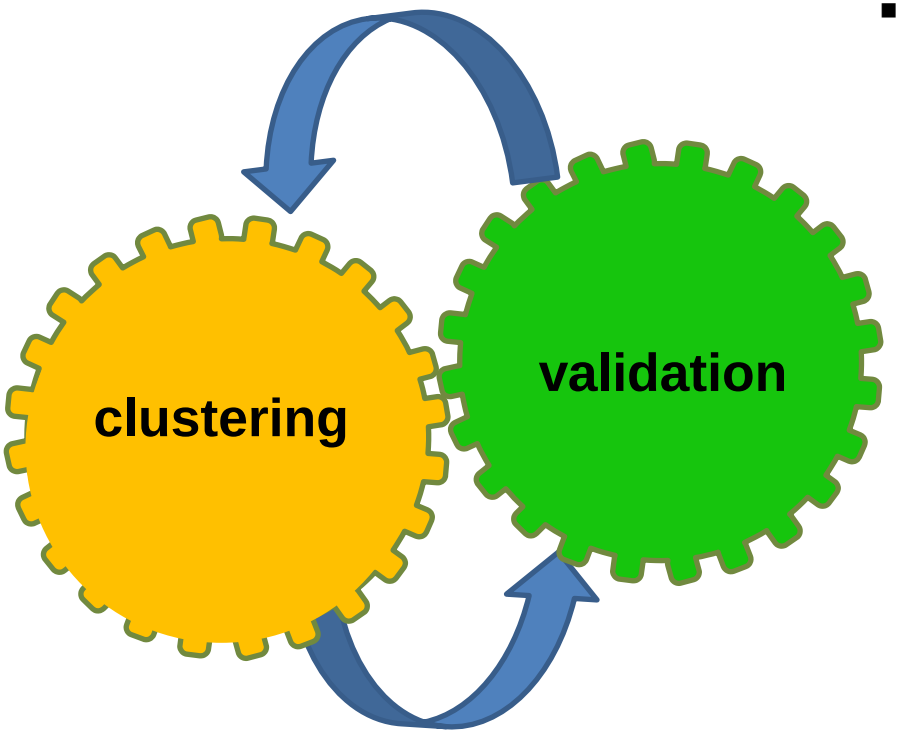
If semantic similarity in a pair of passages is less than a given threshold, then

maxgap-1 **and go back to clustering stage [1].**



Extension

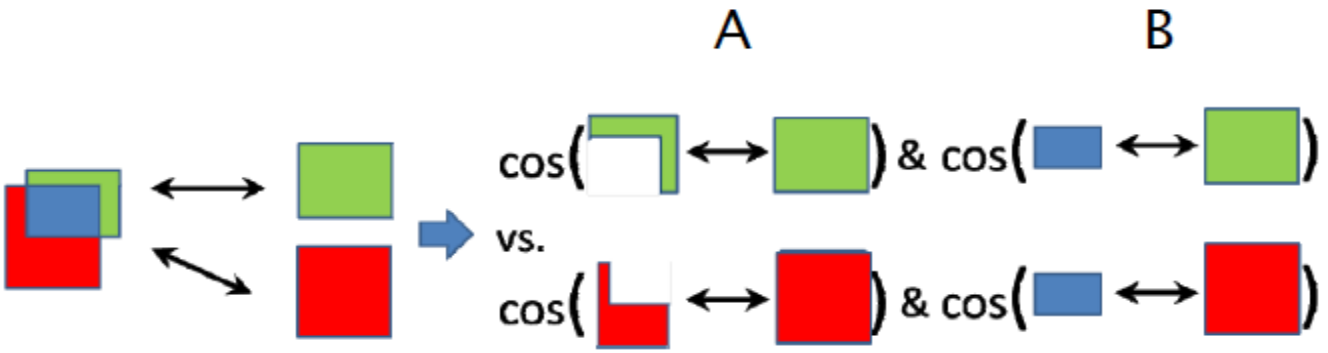
Validation:



- **If the similarity on each pair of cluster > threshold ➡ Filtering.**
- **If the cluster has less than *minsize* seeds, then it is discarded.**

Filtering

Resolving Overlapping [1]:

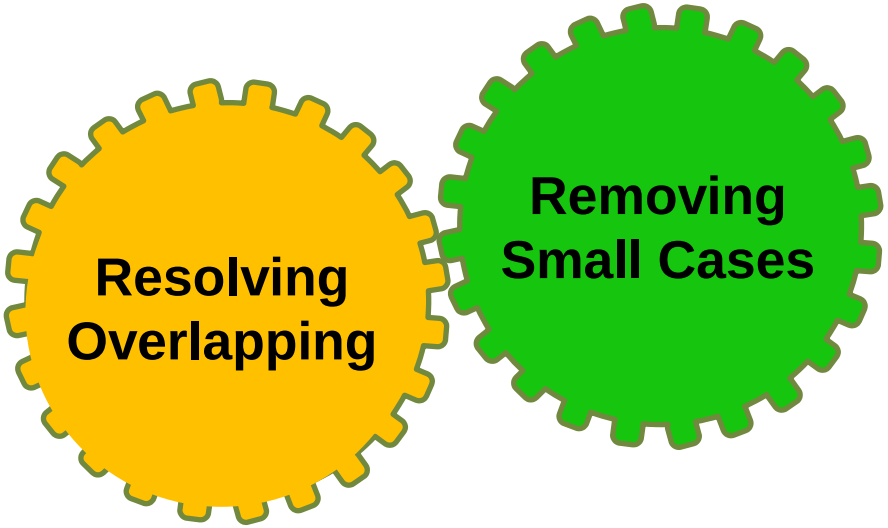


$$score = B + (1 - B) \times A,$$

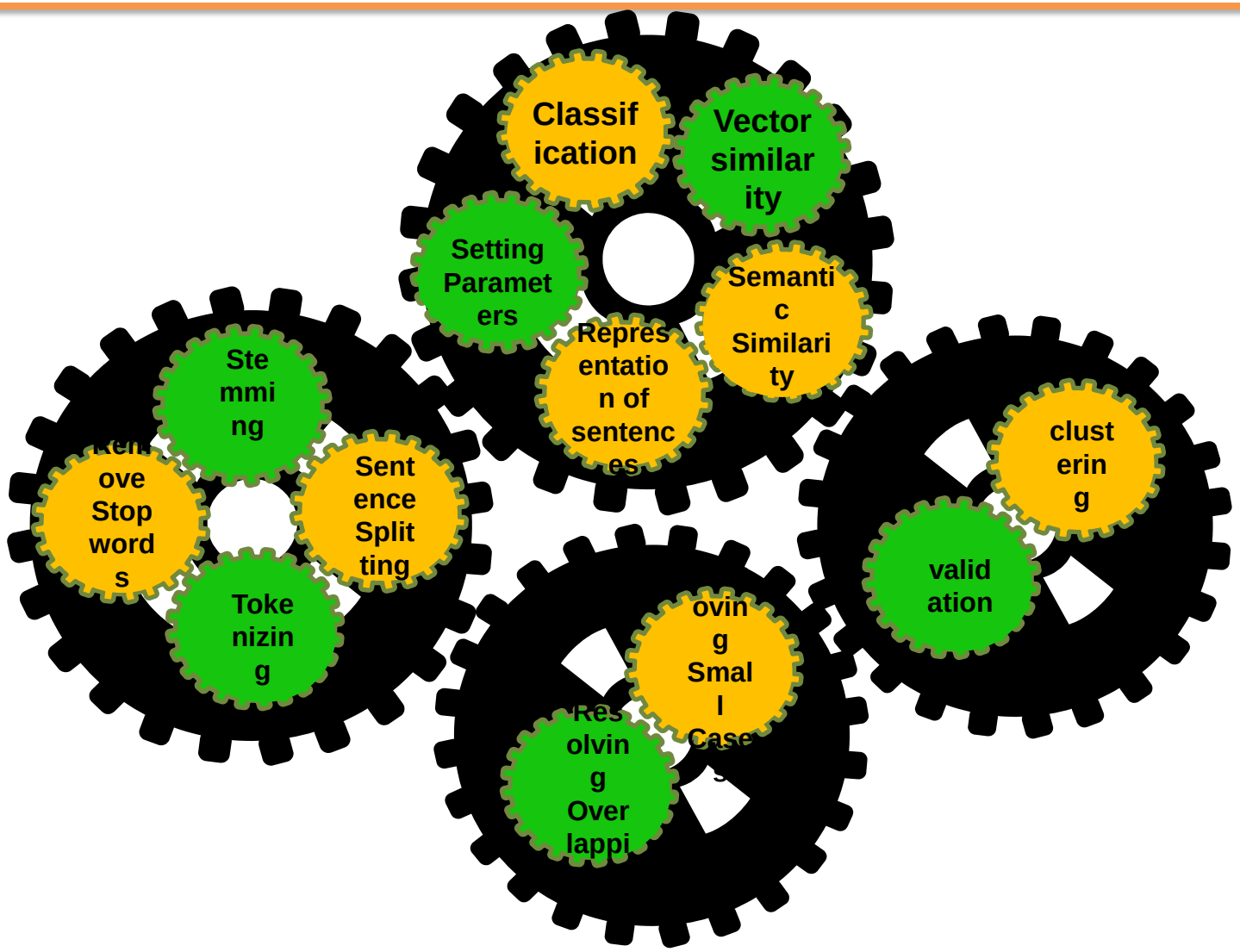
Filtering

Removing Small Cases:

If a plagiarism case has length in characters < threshold, then the case is discarded.



The Proposed Approach



Experiments

The algorithm submitted based on types of obfuscation

Team	No Obfuscation				Artificial Obfuscation				Simulated Obfuscation			
	Recall	Percision	Granularity	PlagDet	Recall	Percision	Granularity	PlagDet	Recall	Percision	Granularity	PlagDet
Mashhadirajab	**0.9939**	0.9403	**1**	0.9663	**0.9473**	0.9416	1.0006	**0.9440**	**0.8045**	0.9336	1.0047	**0.8613**
Gharavi	0.9825	0.9762	**1**	**0.9793**	0.8979	0.9647	**1**	0.9301	0.6895	**0.9682**	**1**	0.8054
Momtaz	0.9532	0.8965	**1**	0.9240	0.9019	0.8979	**1**	0.8999	0.6534	0.9119	**1**	0.7613
Minaei	0.9659	0.8663	1.0113	0.9060	0.8514	0.9324	1.0240	0.8750	0.5618	0.9110	1.1173	0.6422
Esteki	0.9781	0.9689	**1**	0.9735	0.7758	0.9473	**1**	0.8530	0.3683	0.8982	**1**	0.5224
Talebpour	0.9755	**0.9775**	**1**	0.9765	0.8971	**0.9674**	1.2074	0.8149	0.5961	0.9582	1.4111	0.5788
Ehsan	0.8065	0.7333	**1**	0.7682	0.7542	0.7573	**1**	0.7557	0.5154	0.7858	**1**	0.6225
Gillam	0.7588	0.6257	1.4857	0.5221	0.4236	0.7744	1.5351	0.4080	0.2564	0.7748	1.5308	0.2876
Mansourizadeh	0.9615	0.8821	3.7740	0.4080	0.8891	0.9129	3.6011	0.4091	0.4944	0.8791	3.1494	0.3082

Experiments

The text alignment algorithms performance on Persian Plagdet corpus 2016

Rank/Team	Runtime (h:m:s)	Recall	Percision	Granularity	F-Measure	PlagDet
1 Mashhadirajab	02:22:48	**0.9191**	0.9268	1.0014	**0.9230**	**0.9220**
2 Gharavi	**00:01:03**	0.8582	0.9592	**1**	0.9059	0.9059
3 Momtaz	00:16:08	0.8504	0.8925	**1**	0.8710	0.8710
4 Minaei	00:01:33	0.7960	0.9203	1.0396	0.8536	0.8301
5 Esteki	00:44:03	0.7012	0.9333	**1**	0.8008	0.8008
6 Talebpour	02:24:19	0.8361	**0.9638**	1.2275	0.8954	0.7749
7 Ehsan	00:24:08	0.7049	0.7496	**1**	0.7266	0.7266
8 Gillam	21:08:54	0.4140	0.7548	1.5280	0.5347	0.3996
9 Mansourizadeh	00:02:38	0.8065	0.9000	3.5369	0.8507	0.3899

Conclusions and Future Work

- ✓ The proposed method consists of four stages used to aligned the passages of a given document pair
- ✓ The SVM neural network was used to identify the type of obfuscation and set the parameters on the basis of obfuscation.
- ✓ The results showed that this was effective for improving precision and recall.
- ✓ Although the proposed approach ranked first for performance compared with other participants, but the runtime should be decreased.
- ✓ Future study will focus on improving the runtime and the semantic similarity measure in the seeding stage.

References

1. Sanchez-Perez, M. A., Gelbukh, A. F., Sidorov, G. 2015. Dynamically adjustable approach through obfuscation type recognition. In: *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, (Toulouse, France, September 8-11, 2015). CEUR Workshop Proceedings, vol. 1391. CEUR-WS.org.
2. Shamsfard, M., Kiani, S. and Shahedi, Y. STeP-1: standard text preparation for Persian language, *CAASL3 Third Workshop on Computational Approaches to Arabic Script-Languages.*
3. Shamsfard, M. 2008. Developing FarsNet: A lexical ontology for Persian. *proceedings of the 4th global WordNet conference*.
4. Davarpanah, M. R., sanji, M. and Aramideh, M. 2009. Farsi lexical analysis and StopWord list. *Library Hi Tech*, vol. 27, pp 435–449.
5. FIEDLER, R. and KANER, C. 2010. Plagiarism Detection Services: How Well Do They Actually Perform. *IEEE Technology And Society Magazine,* pp. 37-43.
6. Alzahrani, M., Salim, N. and Abraham, A. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Trans. SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS*, vol. 42, no. 2.
7. Ali, A. M. E. T., Abdulla, H. M. D. and Snasel, V. 2011. Survey of plagiarism detection methods. *IEEE Fifth Asia Modelling Symposium (AMS)*, pp. 39_42.

References

8. Potthast, M., Göring, S. 2015. Towards data submissions for shared tasks: first experiences for the task of text alignment. Working Notes Papers of the CLEF 2015 Evaluation Labs, CEUR Workshop Proceedings, ISSN 1613-0073.
9. Sanchez-Perez, M., Sidorov, G., Gelbukh, A. 2014. The Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014. In: Notebook for PAN at CLEF 2014. (15-18 September, Sheffield, UK). CEUR Workshop Proceedings, ISSN 1613-0073, Vol. 1180, CEUR-WS.org, pp. 1004–1011.
10. Glinos, D. 2014. A Hybrid Architecture for Plagiarism Detection—Notebook for PAN at CLEF 2014. CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, (15-18 September, Sheffield, UK). CEUR-WS.org. ISSN 1613-0073.
11. Palkovskii, Y. and Belov, A. 2014. Developing High-Resolution Universal Multi-Type N-Gram Plagiarism Detector—Notebook for PAN at CLEF 2014. CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, (15-18 September, Sheffield, UK). CEUR-WS.org. ISSN 1613-0073.
12. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B. 2014. Overview of the 6th International Competition on Plagiarism Detection. In: Working Notes for CLEF 2014 Conference, (Sheffield, UK, 15-18 September). CEUR Workshop Proceedings, vol. 1180, pp. 845-876. CEUR-WS.org.
13. Smith, T., Waterman, M. 1981. Identification of common molecular subsequences. Journal of molecular biology. Vol. 147(1), pp. 195–197.

References

14. Asghari, H., Mohtaj, S., Fatemi, O., Faili, H., Potthast, M. and Rosso, P. 2016. Overview of the PAN@FIRE2016 Shared Task on Persian Plagiarism Detection and Text Alignment Corpus Construction, *Notebook Papers of FIRE 2016,* FIRE-2016, CEUR-WS.org.
15. Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. 2010. An Evaluation Framework for Plagiarism Detection. In *23rd International Conference on Computational Linguistics (COLING 10)*, pp. 997-1005.
16. Gollub, T., Stein. B. and Burrows, S. 2012. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pp. 1125-1126. ACM. ISBN 978-1-4503-1472-5.
17. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., and Stein, B. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*. *5th International Conference of the CLEF Initiative (CLEF 14)*, pp. 268-299, Berlin Heidelberg New York. Springer. ISBN 978-3-319-11381-4

Thanks for your attention.