Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation

Khadijeh Khoshnavataher, Vahid Zarrabi, Salar Mohtaj and Habibollah Asghari

ICT Research Institute

Academic Center for Education, Culture and Reseach (ACECR), Iran

Text Alignment Corpus Construction

Plagiarism Corpus Construction Approaches:

- **Collection:** Find Real-World instances of text reuse or plagiarism, and annotate them.
- **Generation:** Given pairs of documents, generate passages of reused or PLAGIARIZED TEXT between them. Apply a means of obfuscation of your choosing.
  - ▷ Simulated
  - ▷ Artificial

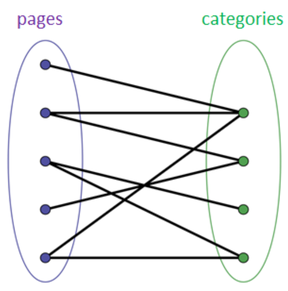
Our Approach

Preprocessing

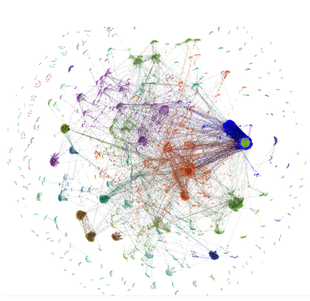
Unification of letters to Unicode characters designed for Persian and using zero-width non-joiner space (Normalization) , Tokenization, Stemming and Part of Speech (POS) tagging.

Documents Clustering

- In this step, collection of Wikipedia documents clustered into different topically related groups
- A bipartite graph of documents-categories was created to cluster the documents
- In the next step, the info- map community detection algorithm was applied to the graph and all communities were detected
- Finally, Documents within a community are considered as one cluster



The Bipartite Graph of Pages and Categories



The Clustered Graph of Pages

Fragments Extraction

The task of the fragment extraction is to extract fragments with different length from source documents

Fragment lengths in words.

Type	Length (Word)	Ratio (%)
Short	30-50	38
Medium	150-250	35
Long	300-500	27

Fragments Obfuscation

- **Artificial Obfuscation**
  - . None (No Obfuscation)
  - . Random Change of Order
  - . POS-preserving Change of Order
  - . Synonym Substitution
  - . Addition / Deletion

Insert Plagiarism Cases in Suspicious Documents

In this step, according to suspicious documentâs length, one or more plagiarism cases which are in the same cluster of suspicious documents are selected. Then, each of them inserted at random positions in suspicious document.

Ratio of Plagiarism fragments in Documents.

Type	Percentage (%)
Little	5-20
Medium	20-50
Much	50-80
Very Much	80-100

Results

- In this section, the result and statistics of the corpus is presented.

Documents

The number of source documents:	1057
The number of suspicious documents:	
- With plagiarism:	529
- No plagiarism:	528

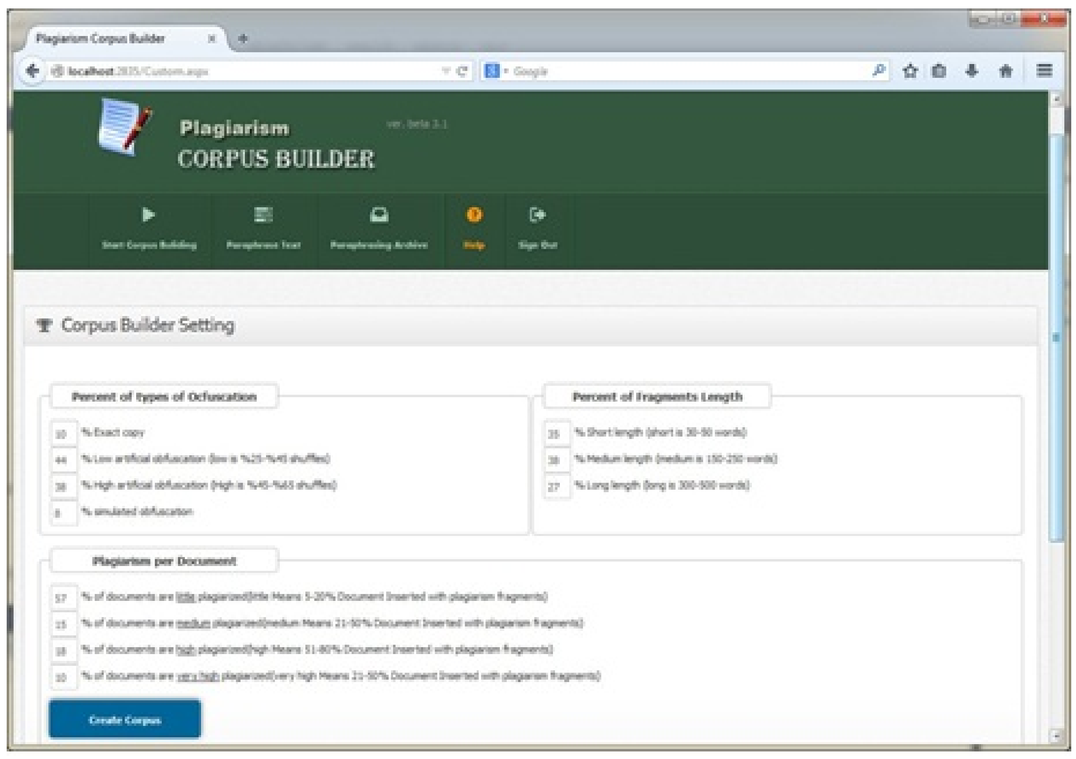
Plagiarism cases

The number of plagiarism cases:	
- No obfuscation cases:	259
- Artificial Obfuscation	
- Low Obfuscation	225
- Medium Obfuscation	215
- High Obfuscation	124

Plagiarism per Document

The number of Little plagiarized documents:	301
The number of Medium plagiarized documents:	80
The number of Much plagiarized documents:	96
The number of Very much plagiarized documents:	52

- For developing this corpus and other corpora in our laboratory, we have developed a web based application that can process the input documents and construct various plagiarism corpora based on corpus builder settings.



Snapshot of our Plagiarism Corpus Builder

Conclusion

- We have discussed our approach to the task of text alignment in the context of PAN 2015 competition.
- Developing the first **Persian** plagiarism detection corpus.
- The corpus have been constructed based on **Artificial** obfuscation.
- The open access **Wikipedia** documents have been used for compiling the corpus.
- We plan to improve our corpus by implementing obfuscation techniques such that **Simulated** obfuscation and other obfuscation strategies using plagiarism corpus builder.

This work has been accomplished under the support of Vice Presidency for Science and Technology of Iran grant No. 1164331

{khadijeh.khoshnava, vahid.zarrabi, salar.mohtaj,habibollah.asghari}@ictrc.ir