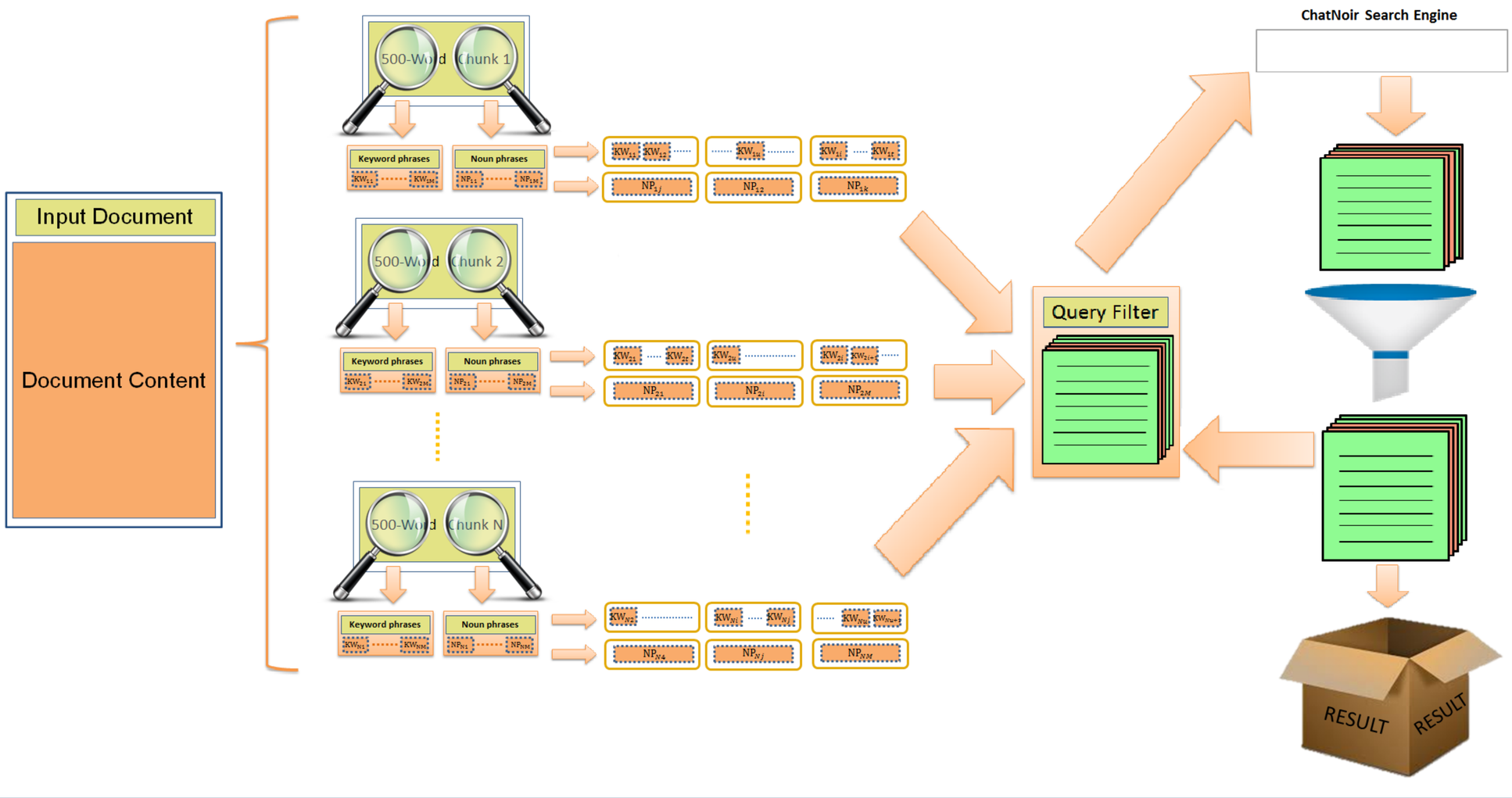
Source Retrieval Plagiarism Detection based on Noun Phrase and Keyword Phrase Extraction

Javad Rafiei, Salar Mohtaj, Vahid Zarrabi, Habibollah Asghari

ICT Research Institute

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

Source Retrieval system



Our Approach

Our approach has been divided into five steps as follows:

▶ Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks
- Sufficient length of chunks, In order to comprise
  - . At least one plagiarism fragment per chunk
  - . And Maximum numbers of extracted queries from the chunks
- Individual sentences sets of 500 words Chunks as results

▶ Noun phrase and keyword phrase Extraction

Multiple Operations on sentences in keywords extraction.

Operation #	Operation Description
1	Selection of top 80% long sentences (based on length in chars)
2	Selection of top 80% sentences (based on number of nouns)
3	Selection of top three sentences (based on average tf.idf1 values)
4	Selection of top three sentences (based on number of words with highest tf.idf1 and tf.idf2 values)

- Scenario1:Operation 1 –> Operation 2 –> Operation 3 for noun phrase extraction
- Scenario2:Operation 1 –> Operation 2 –> Operation 4 for keyword phrase extraction

▶ Query Formulation

- From each selected sentence, one query is extracted
- Selection of high weighted terms to reach the ChatNoir limitation
- The terms are placed next to each other based on the order in sentence

▶ Search Control

Drop a query when at least 60% of its terms are contained in downloaded documents

▶ Document Filtering and Downloading

- The query is divided into two sub-queries:
  - . Snippet with the length of 500 characters are extracted as a sub-query
  - . Snippets are combined with each other and make a passage
- If the resulted passage contains at least 50% words of the query
  - . The related document is downloaded
  - . The document is maintained for search control operation

Evaluations

- Using python programming language and NLTK package for text processing operations.
- the following parameters have been optimized during the training phase:
  - Chunk length
  - Number of queries in each chunk
  - Returned results for each query
  - Similarity threshold between a query and resulted snippets
  - Similarity threshold between a query and downloaded documents

Source retrieval results with respect to retrieval performance and cost-effectiveness.

Team	F1	Precision	Recall	Queries	Downloads	No Detection	Runtime
Rafiei15	**0.12**	**0.08**	**0.41**	**43.5**	**183.3**	**1**	**8:32:37**
Han15	0.36	0.55	0.32	194.5	11.8	12	20:43:02
Kong15	0.38	0.45	0.42	195.1	38.3	3	17:56:55
Ravi15	0.43	0.61	0.39	90.3	8.5	8	09:17:20
Suchomel15	0.09	0.06	0.43	42.4	359.3	4	161:51:26

- According to **No Detection** score, our software has achieved highest rank in this measure. In other words, for only one plagiarized document, the no true positive detection was made.
- The number of queries that is used as input to ChatNoir search engine is one the best among other participants.
- The software has achieved the best rank in software Runtime measure among the participants.

Conclusion

- We have discussed our approach to the task of Source Retrieval in the context of PAN 2015 competition.
- This process has achieved second highest rank in **Query Number** and first in **No Detection** score.
- For future works, we will try to decrease the number of downloaded source doc uments while keeping the complete set of related documents for query filtering.

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

{javad.rafiei, salar.mohtaj, vahid.zarrabi, habib.asghari}@ictrc.ir