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.

<table>
<thead>
<tr>
<th>Operation #</th>
<th>Operation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selection of top 80% long sentences (based on length in chars)</td>
</tr>
<tr>
<td>2</td>
<td>Selection of top 80% sentences (based on number of nouns)</td>
</tr>
<tr>
<td>3</td>
<td>Selection of top three sentences (based on average tf.idf1 values)</td>
</tr>
<tr>
<td>4</td>
<td>Selection of top three sentences (based on number of words with highest tf.idf1 and tf.idf2 values)</td>
</tr>
</tbody>
</table>

- 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.

<table>
<thead>
<tr>
<th>Team</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
<th>Queries</th>
<th>Downloads</th>
<th>No Detection</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rafiei15</td>
<td>0.12</td>
<td>0.08</td>
<td>0.41</td>
<td>43.5</td>
<td>183.3</td>
<td>1</td>
<td>0:32:37</td>
</tr>
<tr>
<td>Han15</td>
<td>0.36</td>
<td>0.55</td>
<td>0.32</td>
<td>194.5</td>
<td>11.8</td>
<td>12</td>
<td>20:43:02</td>
</tr>
<tr>
<td>Kong15</td>
<td>0.38</td>
<td>0.45</td>
<td>0.42</td>
<td>195.1</td>
<td>38.3</td>
<td>3</td>
<td>17:56:55</td>
</tr>
<tr>
<td>Ravi15</td>
<td>0.43</td>
<td>0.61</td>
<td>0.39</td>
<td>90.3</td>
<td>8.5</td>
<td>8</td>
<td>09:17:20</td>
</tr>
<tr>
<td>Suchomel15</td>
<td>0.09</td>
<td>0.06</td>
<td>0.43</td>
<td>42.4</td>
<td>359.3</td>
<td>4</td>
<td>161:51:26</td>
</tr>
</tbody>
</table>

- 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 documents while keeping the complete set of related documents for query filtering.