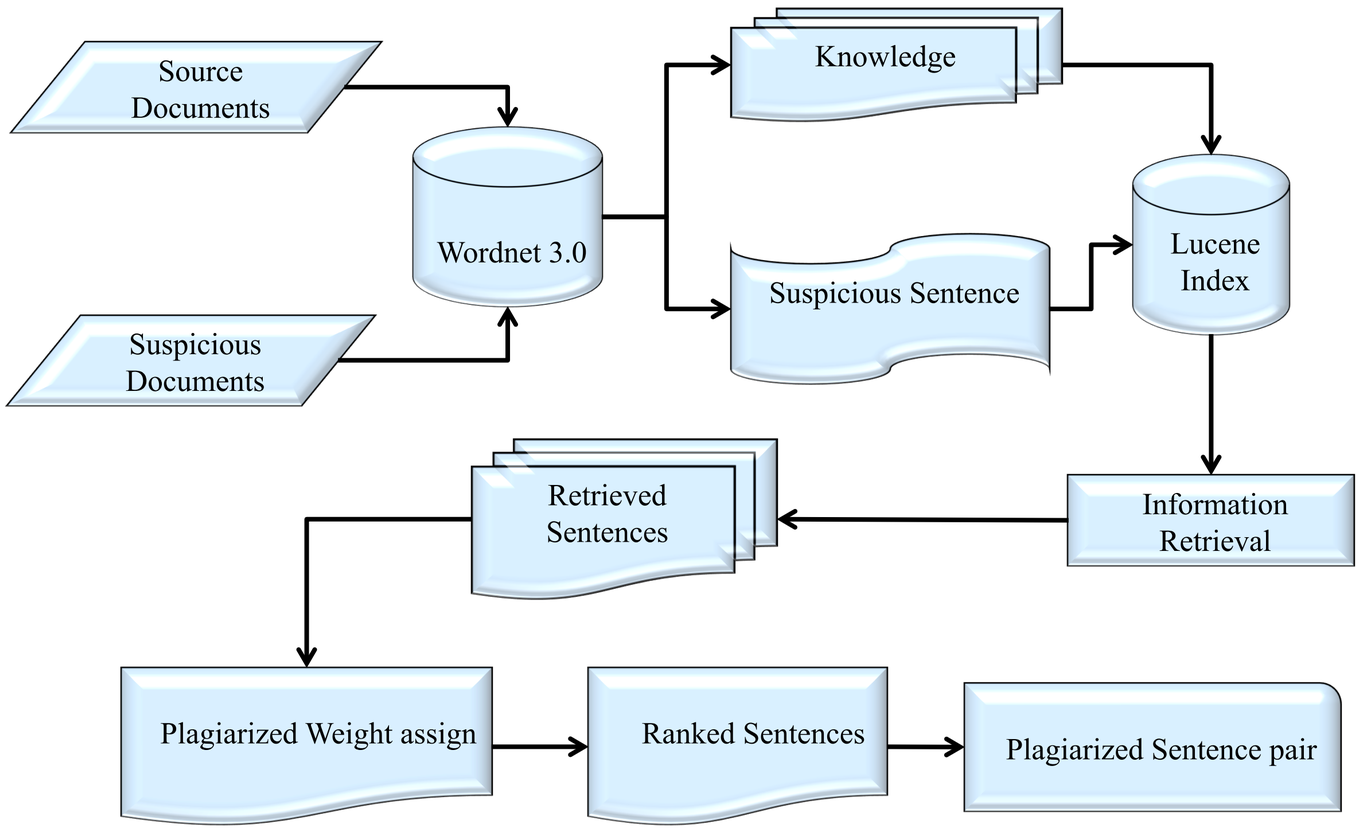
# Rule Based Plagiarism Detection using Information Retrieval

Aniruddha Ghosh, Pinaki Bhaskar, Santanu Pal, Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University, Kolkata – 700032, India

## System Architecture




## Knowledge Preparation

- Each line of the source documents is divided in to text files, containing one line per text file.
- The file names of knowledge files are created in such a manner that the source sentence in the original source document can be tracked.
- Stop words are removed.
- The knowledge of each sentence in the knowledge file is stored in the form of stems, synonyms, hyponyms, hypernyms and synsets of each word (after removal of the stop words) that are extracted from WordNet 3.0.
- Duplicate words are removed to get the set of identical sense unique words.
- These knowledge files are indexed using Lucene

## Candidate Retrieval

- Each sentence of suspicious documents are treated as query .
- Stop words are removed.
- Words are being stemmed using WordNet 3.0 stemmer .
- After generating the query from the suspicious sentences, the query is fired to Nutch to retrieve the probable set of source sentences corresponding to each suspicious sentence.
- A set of probable candidate source sentences is identified by Nutch in ranked order for each suspicious sentence.
- Nutch provides the similarity score between a suspicious sentence and the corresponding candidate source sentence.

## Plagiarism Detection

- Calculate dissimilarity score between the suspicious sentence and its corresponding retrieved candidate sentences.
- Generate fine-grained score = similarity score - dissimilarity score
- Ranked according to this fine-grained score.
- The top ranked candidate source sentence is identified as the source sentence for the plagiarized sentence in the suspicious document.

## Evaluation Score

| Measurement | Precision | Recall | Granularity | Pladget |
|---|---|---|---|---|
| Score | 0.0011829 | 0.0050052 | 2.0028818 | 0.0012063 |

## Conclusion

The Present task is our first attempt in Plagiarism Detection. We have tested the plagiarism at the sentence level but phrase level experimentation is still left for investigation. In future, An algorithm has to be developed to test the relevance of the candidate source sentences retrieved by Nutch and choose the most relevant plagiarized part. The knowledge files for the source documents will also have to be updated.

## Acknowledgement

(DIT), Govt of India funded Project Development of “Cross lingual Information Access (CLIA)” system Phase II