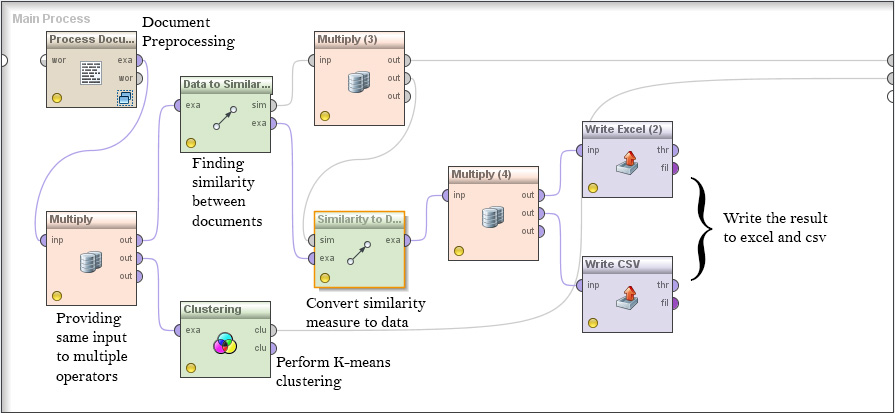
**DETECTING DOCUMENT SIMILARITY**

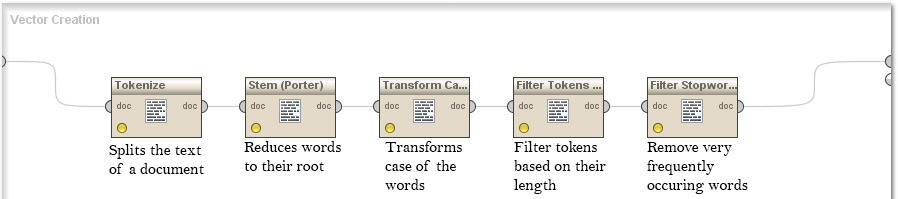
Aim: To find the degree of similarity between any two documents from among a cluster of documents.

Approach: Collecting data, preprocessing data, application of similarity measure, result presentation.

Tools Used: RapidMiner



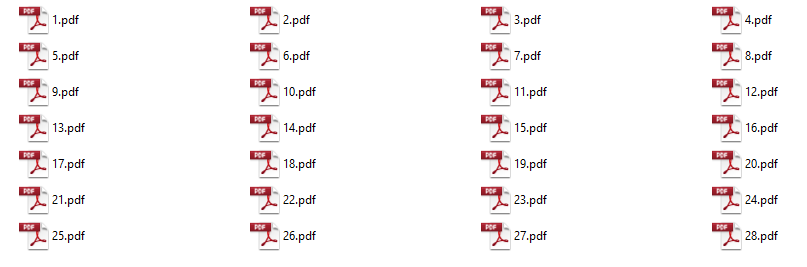
*Fig 1.Complete test mining process to detect document similarity*



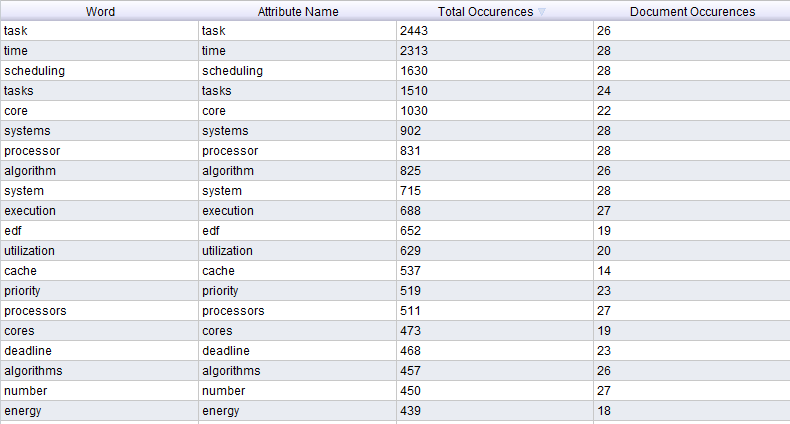
*Fig 2. Exploded view of document preprocessing operator*

Keywords:

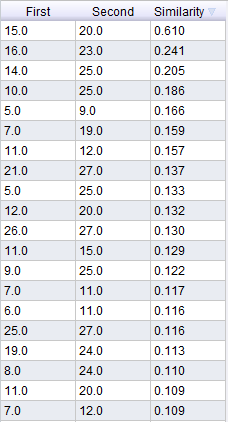
* Cosine Similarity: Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.
* K-Means Clustering: Clustering is concerned with grouping objects together that are similar to each other and dissimilar to the objects belonging to other clusters. Clustering is a technique for extracting information from unlabeled data. k-means clustering is an exclusive clustering algorithm i.e. each object is assigned to precisely one of a set of clusters.
* Stop words: Frequently occurring words like a, an, the, in, at, on, of, for, from etc which do not contribute much to the text mining process.



*Fig 3. Documents used for analysis*



*Fig 4. Word lists with their frequency*



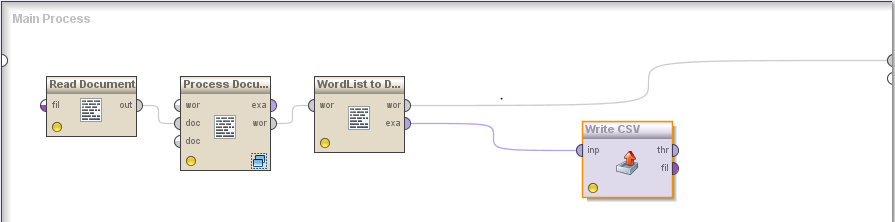
*Fig 5. Similarity measure between the documents*

**IDENTIFICATION OF SIGNIFICANT SENTENCES**

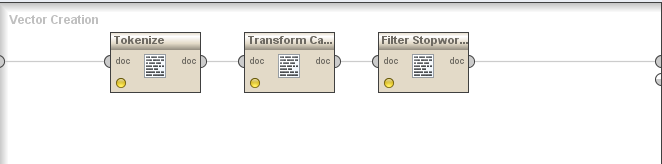
**BASED ON KEYWORDS**

Aim: To find significant sentences from a long document based on the occurrence of keywords in the document.

Approach: Collecting data, preprocessing data, keyword generation, sentence identification, result presentation.

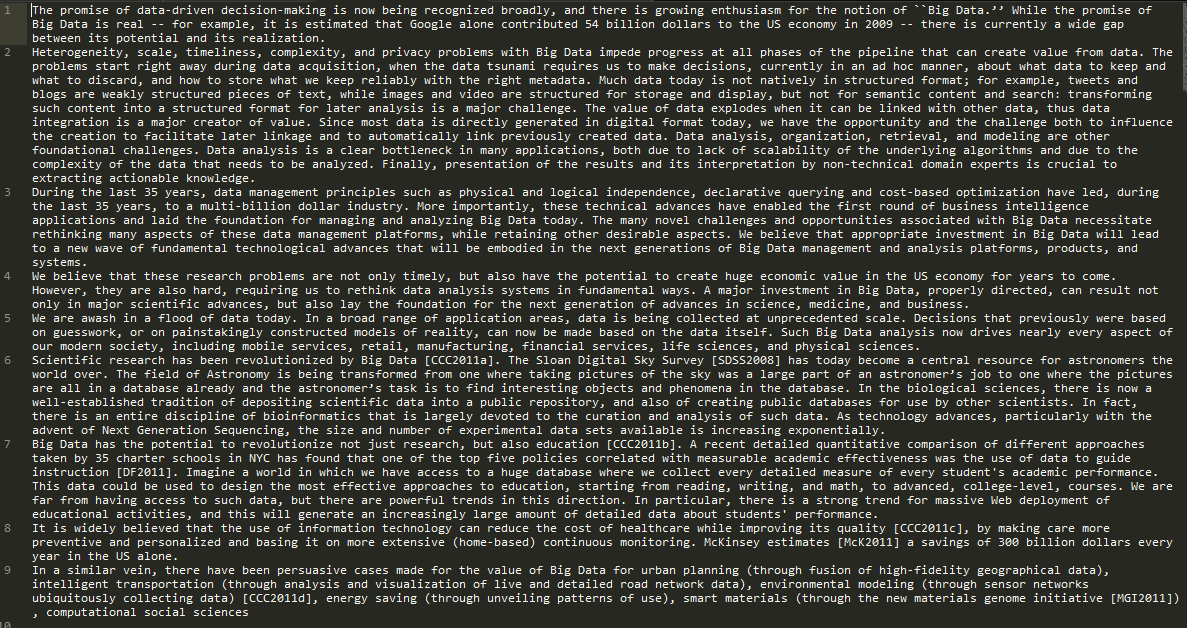


*Fig 1. Complete text mining process for keyword generation*

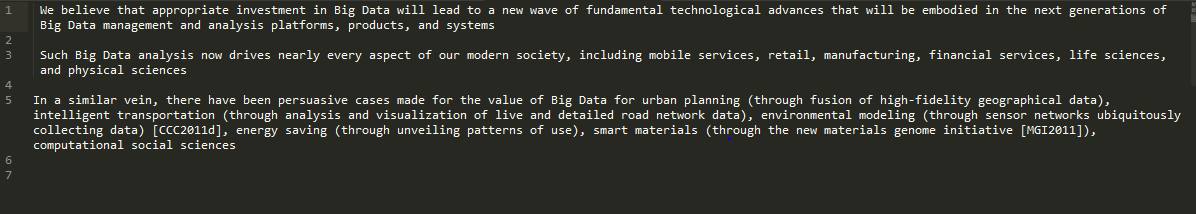


*Fig 2. Exploded view of preprocessing operator*

*Fig 3. Flow of process to identify significant sentences*



*Fig 4. Text input to the program*

**

*Fig 5. Rough summarization of the input text*