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

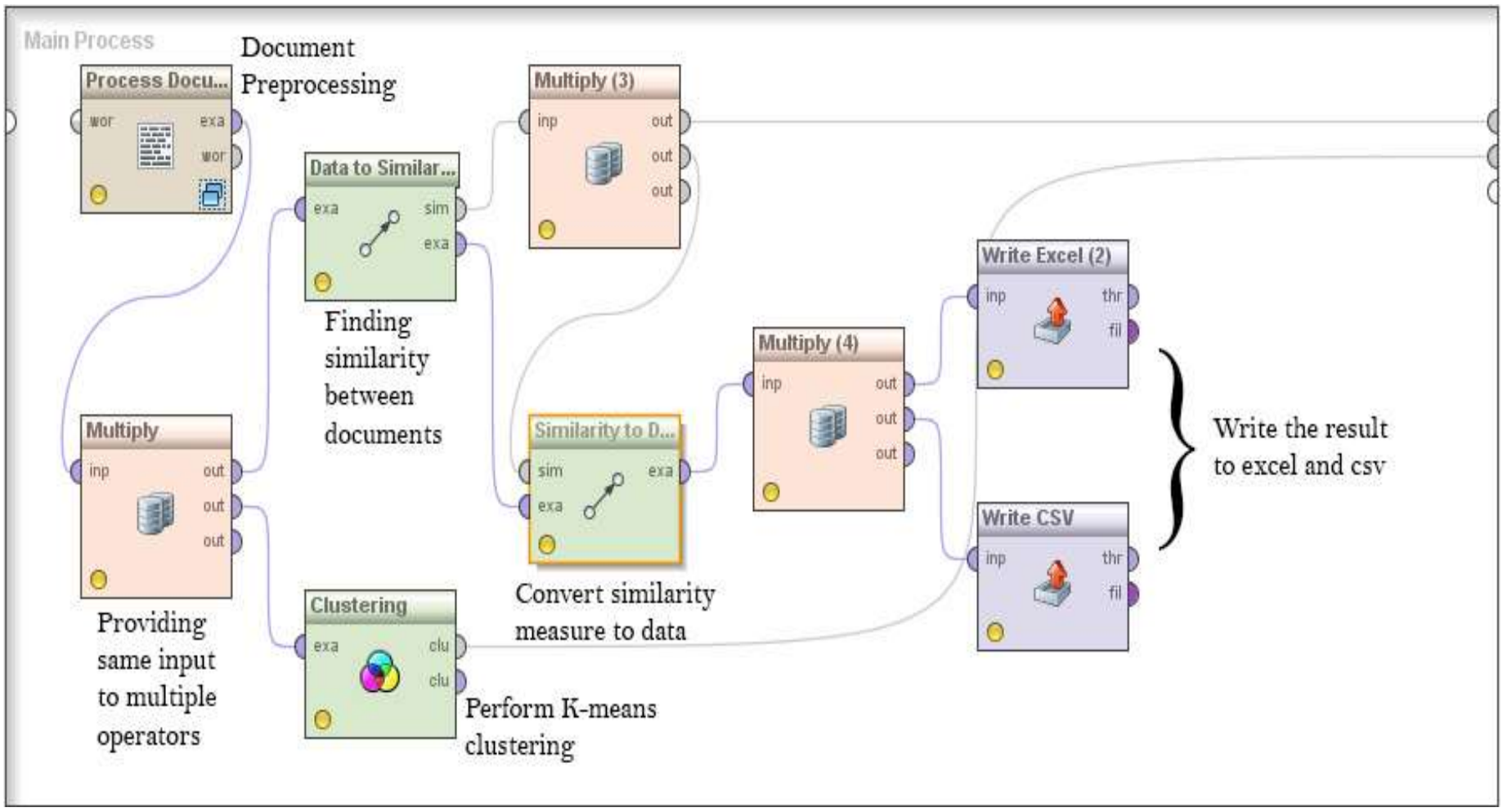

Fig 1.Complete test mining process to detect document similarity

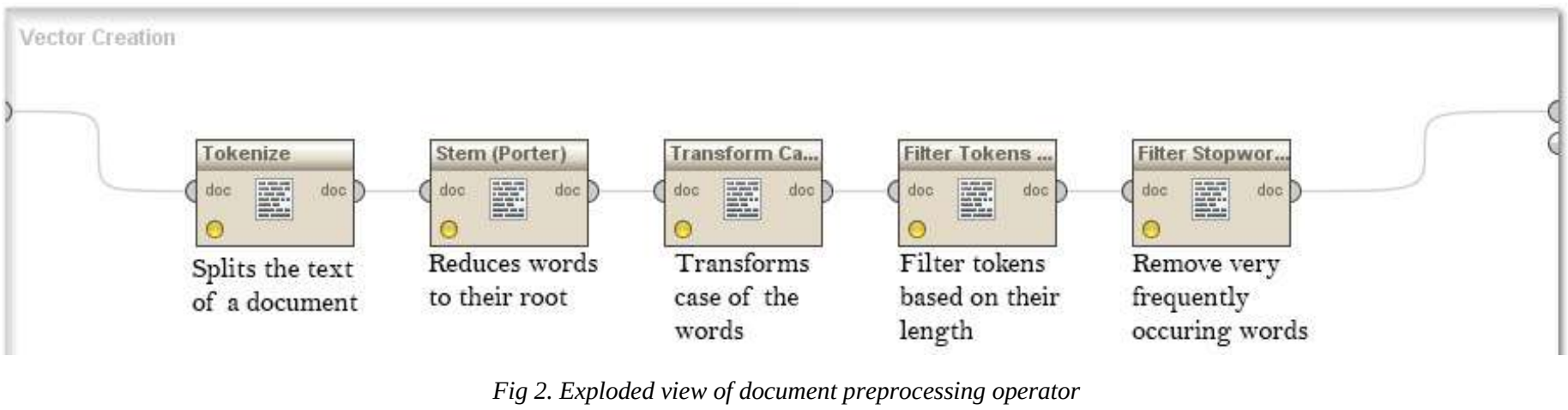

Fig 2. Exploded view of document preprocessing operator

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

Word	Attribute Name	Total Occurences	Document Occurences
task	task	2443	26
time	time	2313	28
scheduling	scheduling	1630	28
tasks	tasks	1510	24
core	core	1030	22
systems	systems	902	28
processor	processor	831	28
algorithm	algorithm	825	26
system	system	715	28
execution	execution	688	27
edf	edf	652	19
utilization	utilization	629	20
cache	cache	537	14
priority	priority	519	23
processors	processors	511	27
cores	cores	473	19
deadline	deadline	468	23
algorithms	algorithms	457	26
number	number	450	27
energy	energy	439	18

Fig 4. Word lists with their frequency

First	Second	Similarity
15.0	20.0	0.610
16.0	23.0	0.241
14.0	25.0	0.205
10.0	25.0	0.186
5.0	9.0	0.166
7.0	19.0	0.159
11.0	12.0	0.157
21.0	27.0	0.137
5.0	25.0	0.133
12.0	20.0	0.132
26.0	27.0	0.130
11.0	15.0	0.129
9.0	25.0	0.122
7.0	11.0	0.117
6.0	11.0	0.116
25.0	27.0	0.116
19.0	24.0	0.113
8.0	24.0	0.110
11.0	20.0	0.109
7.0	12.0	0.109

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.

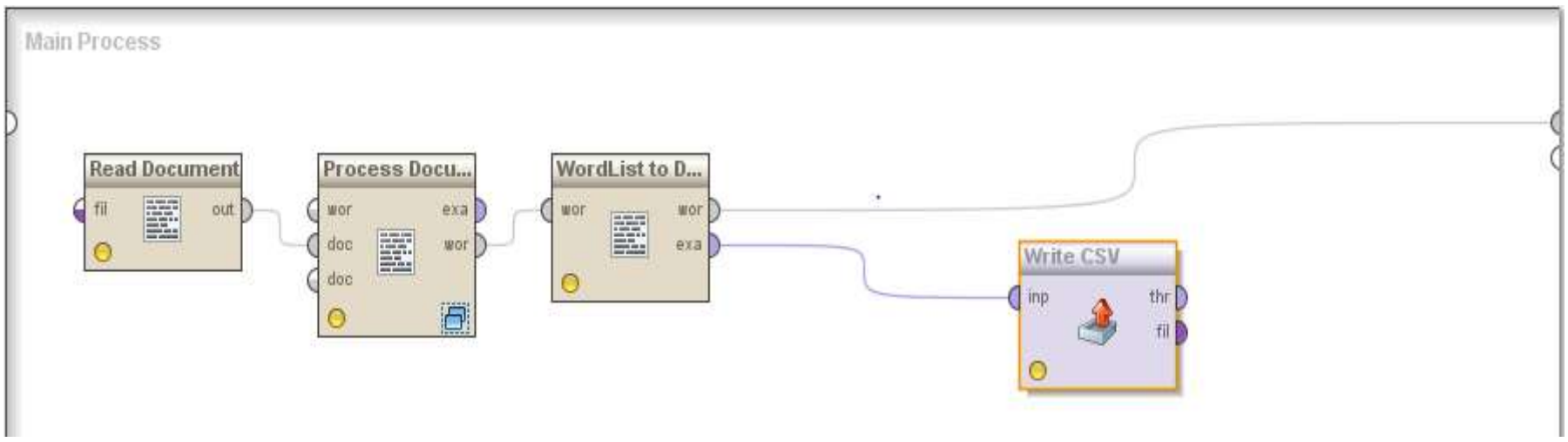

Fig 1. Complete text mining process for keyword generation

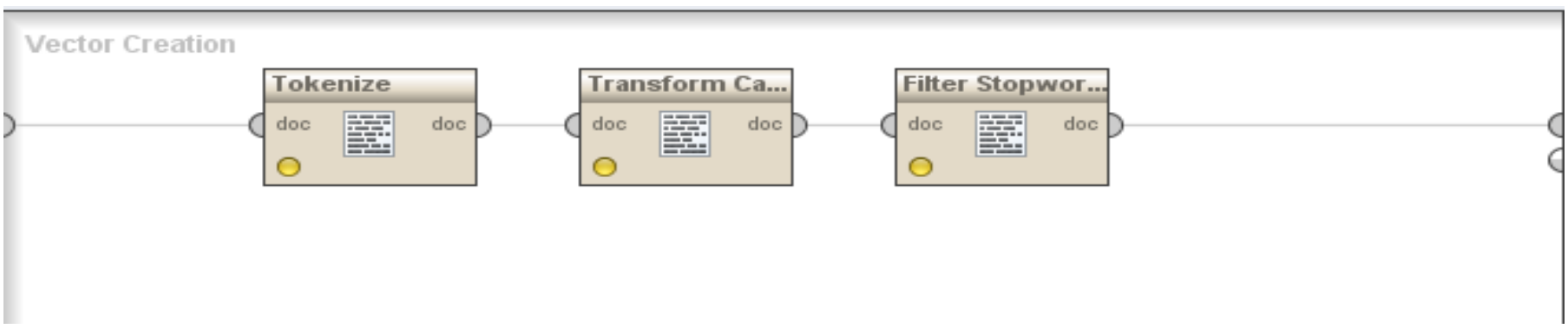

Fig 2. Exploded view of preprocessing operator

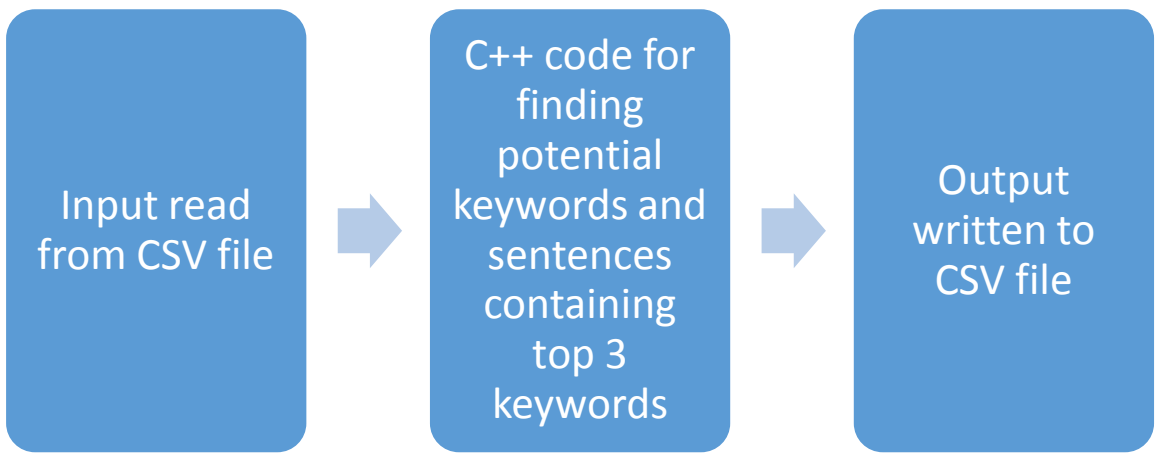

Fig 3. Flow of process to identify significant sentences

```
1  The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of ``Big Data.'' While the promise of
   Big Data is real -- for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 -- there is currently a wide gap
   between its potential and its realization.
2  Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The
   problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and
   what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and
   blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming
   such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data
   integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence
   the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modeling are other
   foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the
   complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to
   extracting actionable knowledge.
3  During the last 35 years, data management principles such as physical and logical independence, declarative querying and cost-based optimization have led, during
   the last 35 years, to a multi-billion dollar industry. More importantly, these technical advances have enabled the first round of business intelligence
   applications and laid the foundation for managing and analyzing Big Data today. The many novel challenges and opportunities associated with Big Data necessitate
   rethinking many aspects of these data management platforms, while retaining other desirable aspects. We believe that appropriate investment in Big Data will lead
   to a new wave of fundamental technological advances that will be embodied in the next generations of Big Data management and analysis platforms, products, and
   systems.
4  We believe that these research problems are not only timely, but also have the potential to create huge economic value in the US economy for years to come.
   However, they are also hard, requiring us to rethink data analysis systems in fundamental ways. A major investment in Big Data, properly directed, can result not
   only in major scientific advances, but also lay the foundation for the next generation of advances in science, medicine, and business.
5  We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based
   on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of
   our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.
6  Scientific research has been revolutionized by Big Data [CCC2011a]. The Sloan Digital Sky Survey [SDSS2008] has today become a central resource for astronomers the
   world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures
   are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database. In the biological sciences, there is now a
   well-established tradition of depositing scientific data into a public repository, and also of creating public databases for use by other scientists. In fact,
   there is an entire discipline of bioinformatics that is largely devoted to the curation and analysis of such data. As technology advances, particularly with the
   advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially.
7  Big Data has the potential to revolutionize not just research, but also education [CCC2011b]. A recent detailed quantitative comparison of different approaches
   taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide
   instruction [DF2011]. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance.
   This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are
   far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of
   educational activities, and this will generate an increasingly large amount of detailed data about students' performance.
8  It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality [CCC2011c], by making care more
   preventive and personalized and basing it on more extensive (home-based) continuous monitoring. McKinsey estimates [McK2011] a savings of 300 billion dollars every
   year in the US alone.
9  In a similar vein, there have been persuasive cases made for the value of Big Data for urban planning (through fusion of high-fidelity geographical data),
   intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks
   ubiquitously collecting data) [CCC2011d], energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative [MGI2011])
   , computational social sciences
10
```

Fig 4. Text input to the program

```
1  We believe that appropriate investment in Big Data will lead to a new wave of fundamental technological advances that will be embodied in the next generations of
   Big Data management and analysis platforms, products, and systems
2
3  Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences,
   and physical sciences
4
5  In a similar vein, there have been persuasive cases made for the value of Big Data for urban planning (through fusion of high-fidelity geographical data),
   intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously
   collecting data) [CCC2011d], energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative [MGI2011]),
   computational social sciences
6
7
```

Fig 5. Rough summarization of the input text