# **IS421: Knowledge Management & IR (10 marks) *CIS Program, King Faisal University***

Created by: Shaheen Khatoon

Course Project: Text Mining with RapidMiner (Due Date: 8th Week).

**Objectives** of this course project is to give the students hands-on experience on:

* **Text Retrieval:** Learn how to index collection of documents to effectively retrieve document matching with user query
* **Text Analysis/Mining:** How to extract knowledge from plain text by identifying interesting patterns across large set of text collection
* **Text Organization:** Classify a text object into one or several of the predefined categories where the categories can vary depending on applications. (organize knowledge to facilitate Text Access and Text Analysis

# **General Steps:**

* **Make Teams:** Work on project in group of max **three** students (group work is strongly encouraged but not mandatory).
* **Choose Topic:** You are required to pick an interesting topic where extracting knowledge from plain text could be useful. While choosing, a topic ask yourself, what problem you are going to solve and search for related articles on topic selected. Please consult appendix at the end of this document on “where and how to find useful research paper”. You can seek help of instructor to get approval of the topic and related papers you identified.
* **Work on Project:** Collect the required dataset and use any existing tool for indexing, searching and extracting from text collection, classifying and clustering documents. Evaluate results to identify reliability of proposed model.
* **Write Project Report:** Prepare a project report and submit before presentation. Follow guidelines from section 3 to write a project report
* **Present** **Project**: Follow guidelines from section 4 to prepare presentation

# **Write a Project Report**

You are required to write a research report like conference paper in two column IEEE Format (IEEE format is attached). Quality requirement is the same as research papers, i.e., in formal written English and rigorous paper format. The report should contain following elements:

**Title and Author:** Select an informatics title of your selected topic e.g. “Automatic text categorization using RapidMiner”, “Mining and summarizing customer reviews”. Please also mention your Author detail, including name, affiliation and email address.

**Abstract:** a brief summary of what you did and what you found.

**Introduction:**  Outlines context, background and purpose Briefly state the background/motivation, what has been done, what is missing, how do you plan to solve it, how do you plan to prove the usefulness of your method, and summarize your contribution(s).

**Related Work:** In this section provide a summary of related work in your own words. This survey will help you to identify whether the problem you would like to tackle has already been solved. You are encouraging to leverage existing work using RapidMiner text processing workflow as it allows you to minimize amount of work you have to do. Follow guidelines from Appendix on “how to find useful papers”

**Methodology:** What technique/algorithm/ will you use/ develop to solve the problem. How data will be collected? This section should have detailed description of entire work flow of text mining process with screen capture from RapidMiner.

For more detail please follow sample paper available with this document “Supervised Learning Methods for Sentiment Classification with RapidMiner”

**Experiments and Results:** how will you evaluate your work? Briefly explain what you did, and what the results were. If you used multiple algorithms compare result to investigate which one has better accuracy.

**Conclusions:** Summary of analysis, your findings, mention limitation of your research and future research idea.

**References:** follow IEEE template for references

## Tips on report Writing:

# Advice on Research and Writing: <http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mleone/web/how-to.html>

* [What is Research in Computer Science](http://www.dcs.gla.ac.uk/~johnson/teaching/research_skills/research.html): <http://www.dcs.gla.ac.uk/~johnson/teaching/research_skills/research.html>

# Basic Research Skills in Computing Science

<http://www.dcs.gla.ac.uk/~johnson/teaching/research_skills/basics.html>

# **Present Project**

You are required to prepare 15 minutes’ presentation. Try to show screen shots and/or plots of your experimental results and demonstrate your project.

## Tips on Presentation:

## watch [TedTalk](http://www.ted.com/talks/roger_antonsen_math_is_the_hidden_secret_to_understanding_the_world#t-85451) [http://www.ted.com/talks/roger\_antonsen\_math\_is\_the\_hidden\_secret\_to\_understanding\_the\_world#t-85451](http://www.ted.com/talks/roger_antonsen_math_is_the_hidden_secret_to_understanding_the_world%23t-85451)

# **Items to Submit:**

You are required to submit following items due on last week of term before presentation:

* Project Report Hard copy
* CD containing ***soft copy of project report***, ***dataset*** and ***RapidMiner Process work flow***

# **Project Grading**

Your project will be graded based on the following required components:

Project report (75%): Detailed written report of your project strictly following required element and template

Presentation: (15% ): presentation about what you have done for this course project including slides presentation and system demo.

Questioning answering (10 %): Each member will be graded individually based upon response to question.

It is preferred to publish outcome of project as paper in conference/journal. **Bonus points (5)** will be given to the groups who meets this requirement.

## APPENDIX: Where and how to find useful research papers

1. Try out **Google Scholar** at <https://scholar.google.com/> using appropriate keywords.

2. Search the **Saudi Digital Library,** <https://www.kfu.edu.sa/en/Deans/Library/pages/sdl.aspx>   
(Use your KFU user ID/Password and use Firefox Browser for SDL)

3. Search SpringerLink, <http://link.springer.com/> (if you use KFU network, you can download most books and papers from here FREE. SpringerLink includes Conference proceedings as well).

4. Try ACM Digital Library, <http://dl.acm.org/> (You can search and download many relevant papers in the ACM digital library from KFU network using your KFU account)

4. Also try CiteSeerX, <http://citeseerx.ist.psu.edu/> and Arxiv, <http://arxiv.org/>

**SOME HELPFUL TIPS:**When doing a project or assignment that requires reading, remember, one paper will lead you to many relevant papers if you check the Reference section wisely. First try to skim through (quickly browse) as many papers as you can and then come back to read some of those papers more carefully.

# **Guidelines to work on Project**

Among many existing tools for text analysis RapidMiner is an excellent tool to build a text analysis process without writing a single line of code.

Here are the main steps that you can follow to work with RapidMiner.

## Install RapidMiner Studio <https://my.rapidminer.com/nexus/account/index.html#downloads>

You can use free version to work with 10,000 rows.

# Install the Text Processing extension

To perform text processing in RapidMiner, you need to have the Text Processing extension installed. The Text mining extension contains tasks specially designed to assist on the preparation of text documents for mining tasks, such as tokenization, stop word removal and stemming.

You can check whether you have it installed by following these steps:

* Open RapidMiner. On the main menu, click Extensions > Marketplace> Top Download.
* Make sure Text Processing extension is listed and checked. If not select and click Install Package.
* Other packages worth to install are Web Mining and Text Analysis AYLIEN, to work with web and text analysis.

# Getting the Data: (Download/Prepare the dataset)

You can prepare own dataset based upon use case you have selected. However, you may download dataset from [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/) from collection of text data. On this site, browse to Text data type, select required dataset and click the Data Folder link to download.

Twitter: Twitter is currently open to public, twitter streams can be accessed via their [APIs](https://dev.twitter.com/), and also there are some crawled twitter available: e.g., Stanford SNAP twitter [data set](https://dev.twitter.com/) <https://dev.twitter.com/>, and TREC [microblog collection](http://trec.nist.gov/data/tweets/) <http://trec.nist.gov/data/tweets/>.

[Yelp Dataset Challenge](http://www.yelp.com/dataset_challenge): A large set of Yelp reviews and entities provided by Yelp <https://www.yelp.com/dataset_challenge>.

# Load Data into RapidMiner.

Create a new RapidMiner process and load data. RapidMiner has variety of operators to load data depend upon input data source e.g. text from Excel, databases, the web, or folders on your computer.

Most commonly used is **Data import Wizard**, through which you can import data into rapid miner repository.

Other common operators are:

**Create Document**: creates a document containing the ***Text*** given as parameter

**Read Document:** creates a document from the specified ***Text file***

**Read Excel**: reads an ExampleSet from the specified ***Excel file***

**Process Document from files*:*** Generates word vectors from a text collection stored in multiple files.

**Others:**

## Create Document Corpus and Build Term Document Matrix

In text mining the main structure for managing document is called **Corpus**, which is collection of text documents. Once *corpus* or *a collection of documents* is created, you can apply all text analysis/retrieval techniques from text processing extension in RapidMiner. However, almost all existing learning and classification techniques require vectors of (real) numbers as input. They cannot work directly on documents (text). Therefore, vector representations of documents should be constructed to make these methods applicable.

To do this, use the ***Process Documents from Data or* Process Document from files (if data is read from multiple text files)**which receives document and generate **document vector** called **Term Document Matrix. *TDM or Document vector* is two-dimensional** matrix with each document name (doc Id or doc number) as row and each individual word found in that document as column. Each entry in this matrix can be a simple **term count**, term frequency or TF-IDF scores.

However, to generate TDM we should identify number of unique feature (in this case unique words in corpus) from input document set which determine the dimension of document vector. Hence before generating TDM we should apply some preprocessing step to generate unique features (words) from document.

To do this double click on ***Process Documents from Data*** operator, it will open sub process, where you can apply all the preprocessing steps to reduce document into ***TDM*** which can then be used to apply different data mining tasks. Following are the basic preprocessing steps you can use:

***Creating Tokens:*** Use ***Tokenize operator,*** which splits the text of a document into a sequence of tokens by removing punctuation, numbers and white spaces. There are several options how to specify the splitting points. Either you may use all **non-letter character**, what is the default settings. This will result in tokens consisting of one single word, what's the most appropriate option before finally building the word vector.

***Transform to Lower Case:*** Should “cat", “Cat", and “CAT" be counted as the same word? Use ***Transform Cases***, which transforms all characters in a document to either lower case or upper case, respectively.

***Remove Stop Word:*** Use ***Filter Stopwords*** operatorwhichfilters English stopwords from a document by removing every token which equals a stopword from the built-in stopword list. Please note that, for this operator to work properly, every token should represent a single English word only. To obtain a document with each token representing a single word, you may tokenize a document by applying the Tokenize operator beforehand.

***Stemming Words:*** Should \organize", \organizes", and \organized" be counted as the same word? This is usually a good idea in most text mining processes.In many applications, words need to be stemmed to retrieve their radicals, so that various forms derived from a stem would be taken as the same when counting word frequency. Word stemming can be done with the various stemming algorithm for different languages such as snowball stemmer, porter stemmer etc. You can use the ***Stem (Porter)*** operator to intelligently strip the suffixes of words. The stem operators put all words into lower case.

***Find Phrase in document***: Should short sentence fragments be counted as distinct items? For example, along with individually using “quick", “brown", and “fox", we could include the fragment “quick brown fox", and count that as its own attribute. Perhaps the fragment is a better predictor than the individual words. You can use the ***Generate N-Grams*** (Terms) operator to do this.

***Others:*** Range of other transformation can be used based on objective in hand.

## Generate Term Document Matrix [Indexing]

In next step we generate TDM to index terms based on a set of tokenized documents. To do this follow below steps:

Go back to Process Document from Data operator by clicking the blue \up" arrow to exit the inner process. Check ***Create Word Vector*** option, which create a table where each document is the row and each column is a unique word in corpus. The values inside the table depend on the type of word vector you are creating E.g.:

**Term Occurrences**: a value in a cell represents the number of times that word appeared in that document.

**Binary Term Occurrences**, value in the cell will be zero if the word did not appear in that document, and one if the word appeared one or more times in that document.

**Term Frequency:** Value is based on the normalized number of occurrences of term in document.

**TF-IDF:** Calculated by using the product of term frequency and inverse document frequency (IDF = # documents in the collection / # documents that contain the feature)

This process is also called ***feature generation***. It defines the building blocks of document objects and gives a meaningful ways to compare them.

# Apply Descriptive and Predictive Analytic Techniques

Once we define how to conceptualize documents in previous step, we can index them, cluster them, and classify them, among many other text mining tasks.

# Experimental Results and Model Evaluation

Explain what dataset was used, how many experiments were conducted? how many documents (example set) were used to trained model, how many to test model. What is the accuracy of model, e.g. how many documents were correctly classified/grouped ((e.g. in case of clustering a match was found between the manually generated clusters and the automatic clusters).

To decide whether a model is accurately capturing a pattern, evaluate model by using **Accuracy, precision and recall** matrix. Try out different technique and compare result based on Accuracy, precision and recall.

# **Practice Exercises**

## Automatic Document Classification

The aim of classifying documents is to assign one or more category or labels to a document, making it easier to manage and sort.

To classify document, collect set of text document, convert them into Term Document Matrix (TDM) by using any word vector (term frequency, term occurrence, TF-IDF, binary term occurrence). Try out different clustering algorithm (K mean, DBSCAN etc.) to identify clusters of document that occurs in similar context. You will observe number of topical clusters will created based upon similarity measure selected.

What if you want to categorize document in a particular way according to set of predefined categories such as technology, sport, entertainment etc. To categorized document into predefined category you can use classification algorithm (KNN, SVM, naive Bayes, Decision tree etc.) to automatically classify document into set of predefined document categories. In this case, you must prepare dataset where each category has given label such as Business, Finance, Sport etc. called training dataset and train the classifier model on training dataset. Then Test the model on unlabeled dataset called test dataset to identify how correctly it classify unlabeled document into appropriate category. In subsequent phase store model and apply it to new document to automatically classify new documents into appropriate document category e.g. Deciding whether an email is spam or not. Deciding what the topic of a news article is, from a fixed list of topic areas such as "sports," "technology," and "politics."

## Sample Text Mining Process

This assignment contains pre-built text mining process and data reside in folder “**LocalRepository**”. The dataset used is from [tripadvisor.com](file:///C:\Users\Dr.%20Shaheen\Google%20Drive\Knowledge%20Management\2016-II\Assignment2\tripadvisor.com) containing review and rating of customer for 100 hotels. The original data was extracted by The Database and Systems Information Laboratory at the University of Illinois at Urbana-Champaign, and is available under <http://sifaka.cs.uiuc.edu/~wang296/Data/>.

The folder contains two subfolders: Process and Data. Process folder has two sample process, Process01 and Process02.

Process 01 is using word association to identify association among word in given corpus which could help to find semantic or syntactic relationship between words. The results can help in navigation between related term, since related term can facilitate linking of objects with each other.

Process 02 using cluster to group similar term together, which can reveal naturel semantic structure in data. E.g. clustering customer reviews can reveal some major complaint about hotels. Cluster results also can support in navigation of relevant subsets of the data since the structures can facilitate linking of objects inside a cluster and linking of related clusters.

Import process to RapidMiner by following steps from “***Text Mining with RapidMiner G. Ertek, D. Tapucu, and I. Arın)***” paper provided with assignment.

It is highly recommended to load given process in RapidMiner and run to intuitively understand text retrieval and analytical process.