An Exemplary Survey Implementation on Text Mining with Rapid Miner

Kaltrina Nuredini[1], Özcan Asilkan[2], Atixhe Ismaili[3]

[1,3]Faculty of Contemporary Sciences and Technologies, SEE University, Tetovo-MACEDONIA
Email: kn11284@seeu.edu.mk, ai11780@seeu.edu.mk
[2]Epoka University, Department of Computer Engineering, Tirana, ALBANIA
Email: oasilkan@epoka.edu.al, Phone: +355(67)2026073

ABSTRACT

Recently many disciplines such as databases, statistics, and information retrieval have affected the growth of data mining. By this phenomenon, data mining had been allowing to extract information or knowledge from these data. Data mining employs diverse techniques based on this phenomenon, but most existing approach is data analysis and text mining. Text Mining allows various analyses for data representing the new framework known as Information Extraction. The extraction of information from unstructured resources has released new paths for analyzing, organizing and querying data. Adapting and implementing these patterns has been time-consuming in the past but with the use of some discovery tools now this task has been significantly easier. The first part of the study includes the mining of an unstructured text and the second part is visualization, both provided by Rapid Miner. The visualization part has some subparts which are directly related to the conducted survey dedicated to particular individuals. These subparts are; finding correlation between the attributes of people, determination of the most weighted attribute and clustering subpart which classifies a group of people that has more similarity between them. Each of these parts has been examined and shown by definitions, examples, analysis and conclusions.

Keywords: *Data mining text mining, text processing, tokenization, clustering, correlation*

INTRODUCTION

Generally, data mining is a powerful new technology that discovers hidden predictive information from a large database. From this large database there is a need to provide data summarization through visualization in case to identify important patterns and act upon the findings. The most subsidiary issues of data mining are: Data cleansing, Visualization and Warehousing. One way to aid users to understand the models is to visualize them. With the use of a powerful data mining tool, Rapid Miner, it is integrated the data analyzing and visualization very tightly. Models built in this case can be viewed and interacted.

On the other hand information is present on websites. Based on this fact, Information Extraction techniques are used in order to identify entities, relations and events in free text [2]. This issue actually is used to get some information out of unstructured data including written and spoken text, pictures, videos and all other unstructured versions of data. In case to analyze and extract the data from particular texts there is a need of a data mining tool. It is often a hassle when there is a need to get text from documents that are in different formats. By this approach text mining as a growing new technology attempts to discover meaningful information from natural language text. This involves the process of analyzing text to extract information that can be useful for a specified reason. The phrase "text mining" is used to denote any system that analyzes a large language text and extracts useful information [1]. This review involves the usage of Rapid Miner tool for implementation of a specific example for information extraction and text mining.

This paper is organized as follows. First section briefly explains text mining with Information Extraction and demonstrates the usage of the state-of-art data mining software Rapid Miner. Next section describes the key principles concerning IE and modelling how IE is practiced using Rapid Miner.. Finally, last section presents the conclusion.

TEXT DATA MINING AND INFORMATION EXTRACTION

The search process where the user is looking for something that is already known is totally different from text mining. The most important role of text mining is to discover unknown information, something that no one yet knows. Text mining falls into a category with data mining that is able to find particular patterns from large database [8]. The difference between these two is that in text mining the patterns are extracted from natural language text rather than from structured databases [1]. Information Extraction is a meaning which wends to an automatic extraction of ordered information such as entities, relationships between entities, and attributes describing entities from unstructured resources. This way of extracting information permits much richer forms of queries on the plentiful unstructured sources than possible with keyword searches alone. When structured and unstructured data co-exist, information extraction makes it promising to put together the two types of sources and present queries spanning them.

The extraction of structure from unstructured sources is a challenging task, which has occupied a real community of researchers for more than two decades. In most of the cases this activity worries about processing human language texts by means of natural language processing (NLP). The structure extraction has origins in NLP field, but now connects many different communities including information retrieval, database, document analysis, etc.

Therefore, with the root in NLP Community now the topic of information extraction holds a lot of researcher and communities who carry in methods from machine learning, databases, information retrieval, and computational linguistics for different sides of the information extraction problem. A bit previous extraction works have been concentrated on the identification of named entities, like people and company names and relationship among them from natural language text. This field was also influenced by Message Understanding Conference [3, 6], and Automatic Content Extraction [7] program. The invention of the Internet significantly improved the scope and variety of applications depending on different forms of information extraction.

Rapid Miner organization and implementation

In this part, an overview of some basic concepts on traditional DM will be exposed. One of the best data mining open source tools is Rapid Miner which supplies an environment for machine learning and data mining processes. Today, Rapid Miner is the world-wide popular open-source data mining solution and is broadly used by researchers and companies. GUI of Rapid Miner can be used to design the XML description of the operator tree, to interactively control and examine running processes, and to constantly observe the visualization of the process results. Break points can be used to verify transitional results and the data flow between operators. Clear interfaces define an easy way of applying single operators, operator chains, or complete operator trees on input data. A command line version and a Java API allow invoking Rapid Miner from other programs without using the GUI. Since Rapid Miner is entirely written in Java, it runs on any major platforms/operating systems. In order to install Rapid Miner plug-ins, it is enough to copy them to the lib/plug-ins subdirectory of the Rapid Miner installation directory. With the use of plug-ins, importing a content and tokenization becomes possible. Since texts or files are accessible in different forms like simple ASCII files, xml, html files and pdf files, some plug-ins it is needed [9]. Some text-formats are effortless to process and some need additional parsing effort.

TEXT MINING AND INFORMATION EXTRACTION WITH RAPID MINER

This section presents a short overview and usage of the two important techniques in text mining and information extraction. In case to represent and show both of the phases Rapid Miner is used. As mentioned above, in order to do text mining and information extraction there is a need of installing some additional extensions. In this case Tex Processing is installed for further analyzes. The important points are how to load data in Rapid Miner, how to do a pre- processing, how to do some classification models based on it and how to extract useful information. To load document from files there is a process named as "Read Document". For further processing of that specified document a process named as

"Process Documents" is used [9]. When the text is processed in that process they are taken as an unstructured list of terms and it is transformed into a back of words model or a vector model. The documents are represented as vectors where each position corresponds to fixed word over a whole corpus of the document and for each document is counted how many times the words occur in that particular document. The most advanced schema is called TF-IDF which stands for Term Frequency Inverse Document Frequency. With higher Term Frequency, a word occurs more often in the document. The inverse document frequency is based on the document frequency which is a number of documents that the word occurs in and if it occurs in too many documents it's not so characteristic any more. After loading the specified document, in this case is an html format, there is a need for assigning a vector value (see figure 1).

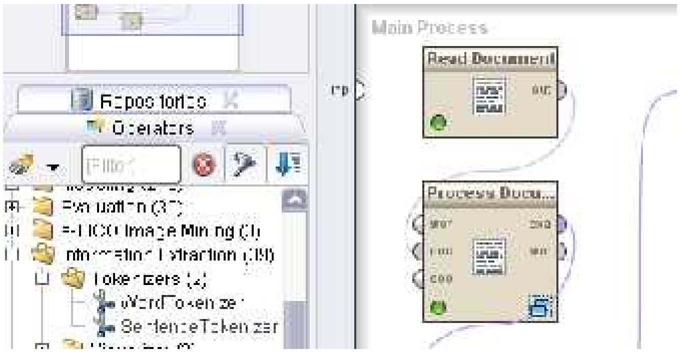


Figure 1. Load a process for an html file.

Next step is to determine which stages should be used inside the process loop. In the loop there can be a need for porter stemming which removes morphological and inflectional endings from words in English Language. After stemming, a filter process is used to filter and remove stop words. The point here is that we can eliminate stop words because they do not play any important role at all in document. Another process "Information Extraction" is used in this stage which determines more important information that needs to be extracted from the specified document [9] (See figure 2).

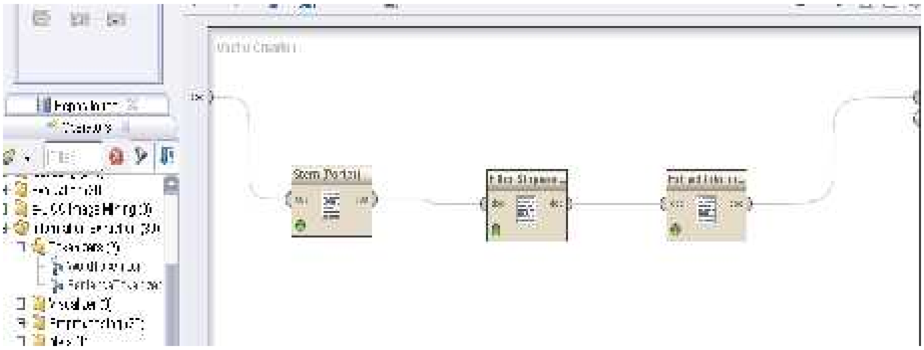


Figure 2. Looping through the process

After loading the document the next step is to tokenize the text. The first process is "Sentence Tokenizer" which allows the text to be tokenized into sentences [9]. During this process we have used the attributes with the name "sentence" that displays all sentences in the document file. The second step is tokenizing these sentences into tokens. The process "Word Tokenizer" is used with the attribute word [9]. Figure 3 displays the processes, figure 4 displays the results.

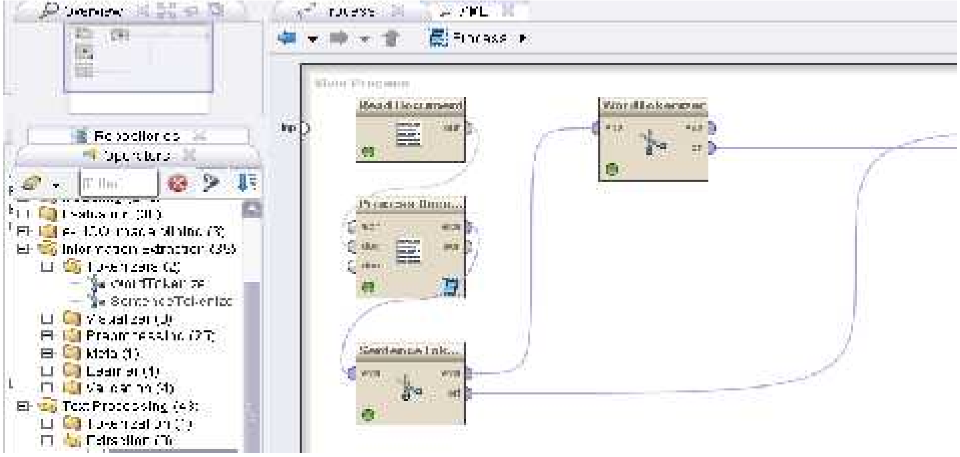


Figure 3. Tokenizer's processes

After word Tokenizer there is a need for text pre-processing. This is the most important part in IE plug-in that is used to enrich the document and its tokens with syntactical information which will be used for further extraction. The columns represent the attributes defined for each word. There are two other attributes used for the word itself and the prefix of the current word of length 3.

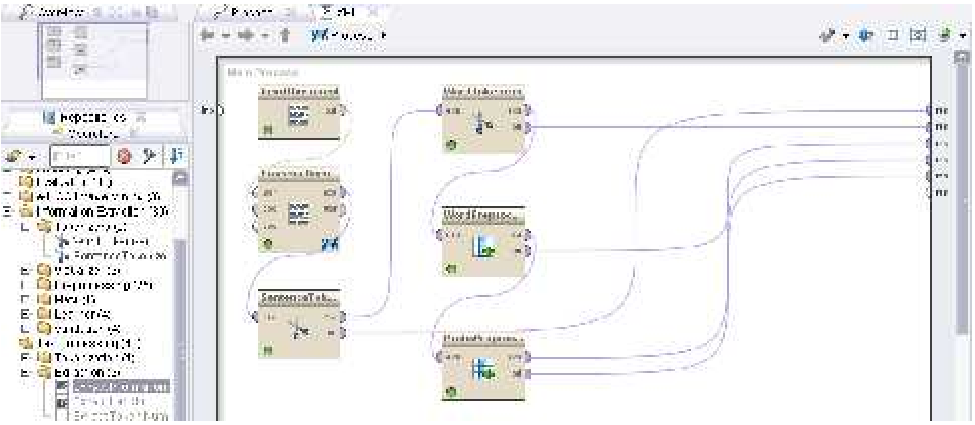


Figure 4 Example set after preprocessing

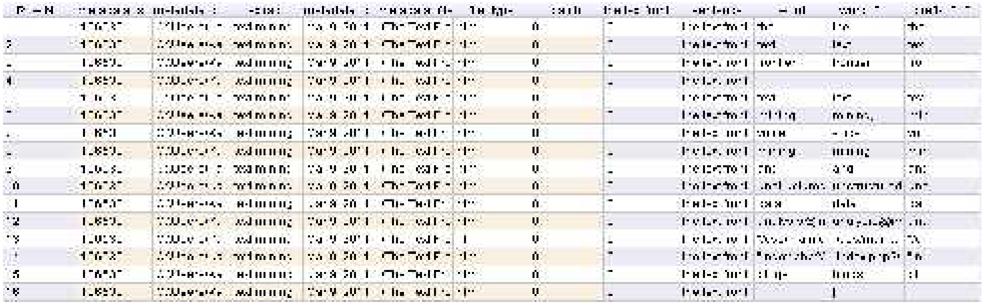


Figure 5. Results after preprocessing

The extract attribute shows the text extracted from the particular html file that shows the key word text mining as an entity. Attribute Sentence shows the sentences tokenized from the text. Word attribute depicts word tokenization process. By figure 5, we can see that the text is tokenized into sentences and forms a sentence into a word that helps us the identifiy some important information.

DATA VISUALIZATION

This section depicts some visualization capabilities of Rapid Miner. There is a need for a large dataset in case to provide real results. Based on this fact, the first important step in data mining towards building a productive program is the gathering the data [4]. In case to collect these data, we have employed a particular survey. This survey is implemented in Google docs; The following survey was designed to collect data about our case study to categorize individuals treating to similar behaviors that are presented below. This survey aims to collect information on particular individuals' characteristics. So the **focus group** is the individual of *Balkan Part* because of the distribution of the survey along different cities. The number of participants is 40.

Questionnaire contains thirteen questions and is not divided into parts but all questions has to do with the person itself, so there is no need to divide it. Questions are asking the person to fulfill the text with a characteristic of their body for example the eye color, the height, etc., hobbies or profession.

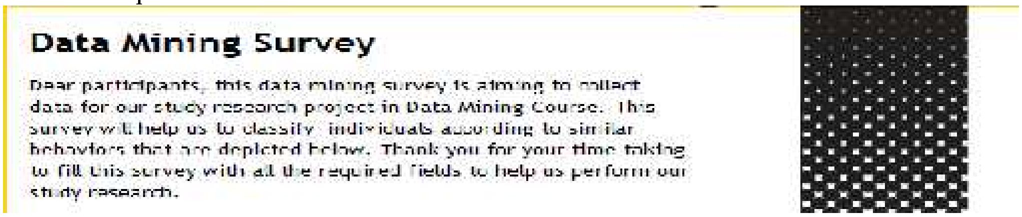
Gathered data was loaded into the Rapid Miner for further analyzing. There would be some analysis between attributes of the peoples which one has more correlation, which is the attribute with more weight. Rapid Miner will help visualize this data in a kind of histograms, scatter plots and so on but there is the other part which it should be analyzed which means that we should come in some conclusions about the similar behaviors, and correlated attributes.

In addition to what was putted as a question, in the importing phase of the gathered data in Rapid Miner, it was needed to determine the type of the attributes.

Because there were different types of attributes and the almost all number of questions were text and number asking so all of them were determined as *nominal*. The nominal attribute has to do with the distinctness of the attributes which means that we gather information from distinct attributes and distinct people and there in no any order in the people and attribute part.

Survey

The questions that were asked to individuals are shown below.

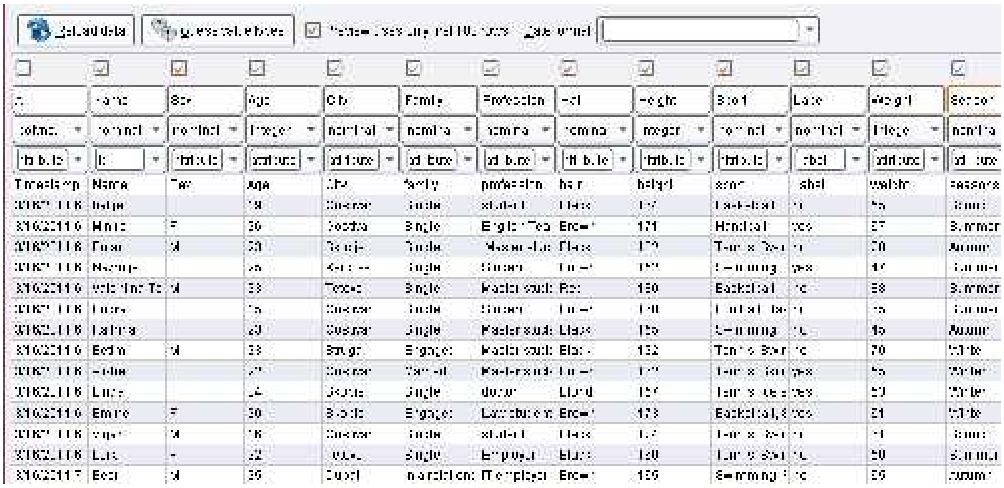


Data Mining Survey

Dear participants, this data mining survey is aiming to collect data for our study research project in Data Mining Course. This survey will help us to classify individuals according to similar behaviors that are depicted below. Thank you for your time taking to fill this survey with all the required fields to help us perform our study research.

1. Please specify the name.
2. Please specify your age.
3. Please specify your sex.
4. Please specify your city.
5. Select your family status:
   a) Single,
   b) In a relationship,
   c) Engaged,
   d) Married,
   e) Divorced.
6. Specify your profession.
7. Specify the color of your hair:
   a) Black,
   b) Red,
   c) Brown,
   d) Blond,
   e) Grey,
   f) White.
8. Specify the color of your eyes:
   a) Blue,
   b) Green,
   c) Black,
   d) Brown.
9. Specify your height in centimeters.
10. Specify your weight in kg.
11. Specify your favorite sport.
12. Specify your favorite season.
    a) Spring,
    b) Summer,
    c) Autumn,
    d) Winter.
13. Specify your favorite perfume.
    a) Soft,
    b) Citrus,
    c) Floral,
    d) Dry
    e) Oriental
    f) Fougere

The collected results were saved in Excel spreadsheet. This spreadsheet will be used as a data model for that particular analysis. This file is imported into Rapid Miner from where there are defined data types of the attributes. Rapid Miner assigns roles to attributes defining what they can be used for individual operators. As can be seen in figure 6 depicts the Meta data there are two columns with integer and all other data are nominal. As id is categorized the Name of the participant whereas the other columns are attributes.



□	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
A	Name	Sex	Age	City	Family	Profession	Hair	Height	Sport	Label	Weight	Season
polynominal	nominal	nominal	integer	nominal	nominal	nominal	nominal	integer	nominal	nominal	integer	nominal
attribute	id	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	label	attribute	attribute
Timestamp	Name	Sex	Age	City	family	profession	hair	height	sport	label	weight	seasons
[illegible]	[illegible]		19	[illegible]	Single	student	Black	[illegible]	Basketball	no	[illegible]	[illegible]
3/16/2011 6	[illegible]	F	30	[illegible]	Single	English Tea	Brown	171	Handball	yes	57	Summer
[illegible]	[illegible]	M	20	[illegible]	Single	Master stu	Black	[illegible]	Tennis, Swi	no	70	Autumn
[illegible]	[illegible]		[illegible]	[illegible]	Single	Singer	Brown	[illegible]	Swimming	yes	[illegible]	Summer
3/16/2011 6	[illegible]	M	33	[illegible]	Single	Master stud	Red	180	Basketball	no	58	Summer
[illegible]	[illegible]		[illegible]	[illegible]	Single	Singer	Brown	[illegible]	Football, Ba	no	[illegible]	Summer
3/16/2011 6	[illegible]		20	Durres	Single	Master stud	Black	155	Swimming	no	45	Autumn
3/16/2011 6	[illegible]	M	33	Struga	Engaged	Master stud	Black	182	Tennis, Swi	no	70	Winter
[illegible]	[illegible]		[illegible]	[illegible]	Married	Master stud	Brown	[illegible]	Tennis, Ski	yes	[illegible]	Winter
3/16/2011 6	[illegible]		24	[illegible]	Single	doctor	Blond	[illegible]	Tennis, Vo	yes	50	Winter
3/16/2011 6	[illegible]	F	30	[illegible]	Engaged	Law student	Brown	173	Basketball, S	yes	51	Winter
[illegible]	[illegible]	M	18	[illegible]	Single	student	Black	[illegible]	Tennis, Swi	no	[illegible]	Summer
3/16/2011 6	[illegible]	F	22	[illegible]	Single	Employee	Black	160	Tennis, Swi	no	50	Summer
3/16/2011 7	[illegible]	M	35	Durres	in a relation	IT employee	Brown	189	Swimming	no	59	Autumn

Figure 6. The definition of data types

After retrieving the data from repository there is a need to create attribute weights that indicates which the most important attributes are. After that it will go through clustering for individual segmentation and finally it calculates correlations between all attributes. Figure 7 shows the processes involved in this program in case to visualize the issues emphasized above.

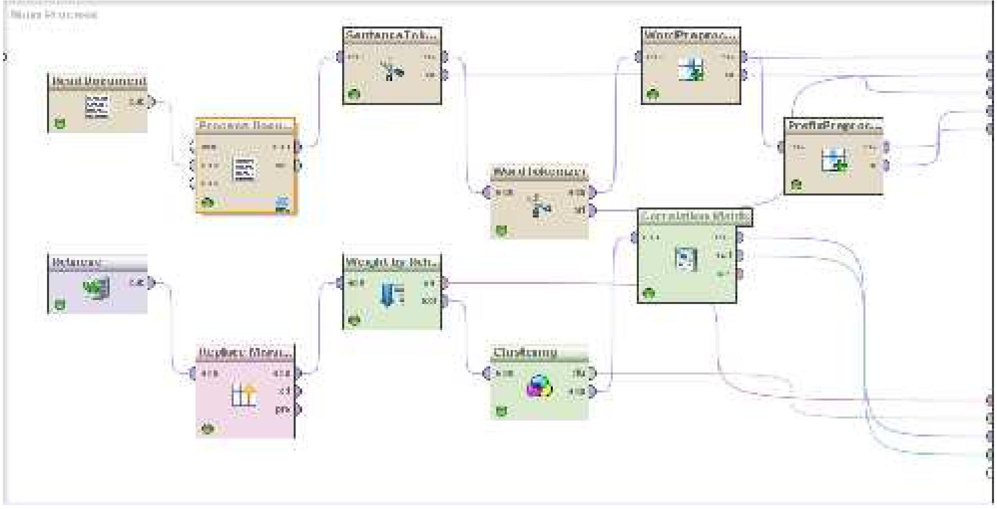


Figure 7 Main processes.

Attribute weights

For each attribute, a weight is stored. The weight with value 0 means that the attribute was not used at all and 1 means the opposite. Table 1 shows the importance of attributes in the modelling task.

Table 1. Attribute weight.

Attribute Name	Attribute Value
Height	1
Perfume	0.386
Family	0.328
City	0.320
Profession	0.315
Eye	0.268
Sex	0.257
Hair	0.253
Weight	0.252
Sport	0.233
Season	0.097
Age	0

The visualization of the table gives a hint that shows which might be the most important factors affecting the response rate. As can be seen from figure 8, the attribute Height has an important role in which can affect the data. Height as attribute can be correlated with Age, Weight, Sex, and Sport.

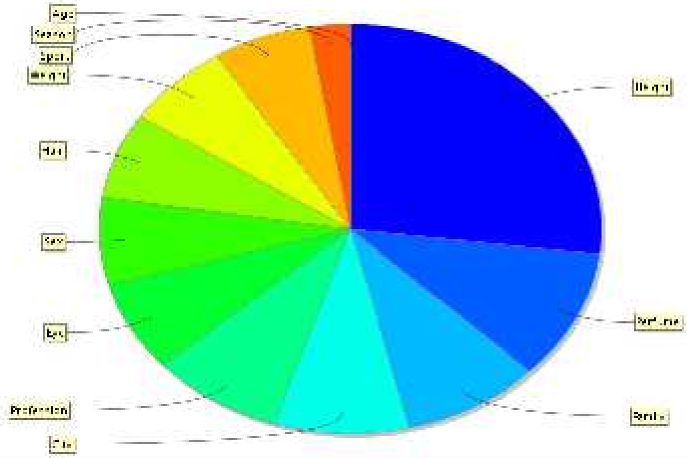


Figure 8. Visualization of attribute weight.

Clustering

Clustering in general shows individual segments where several individuals with similar behaviors are grouped together in a complete individual hierarchy. If we click on several segmentations we can properly identify common properties of individuals in the segment. The goal of clustering is simply to explore the structure of the data. Cluster Analysis sorts the data into similar clusters that share certain characteristics. In this statistic data miner is used K-means as a cluster option. The K-mean tool works toward on maximizing the distance between means [5]. In K-mean there is selected Mixed Euclidian distance. Data points in one cluster are more similar to another and are grouped together. As seen from figure 9, the cluster with number 70 when clicked contains the listed names of other individuals in the right side of the figure. They are grouped according the similarities between them based on the attributes.

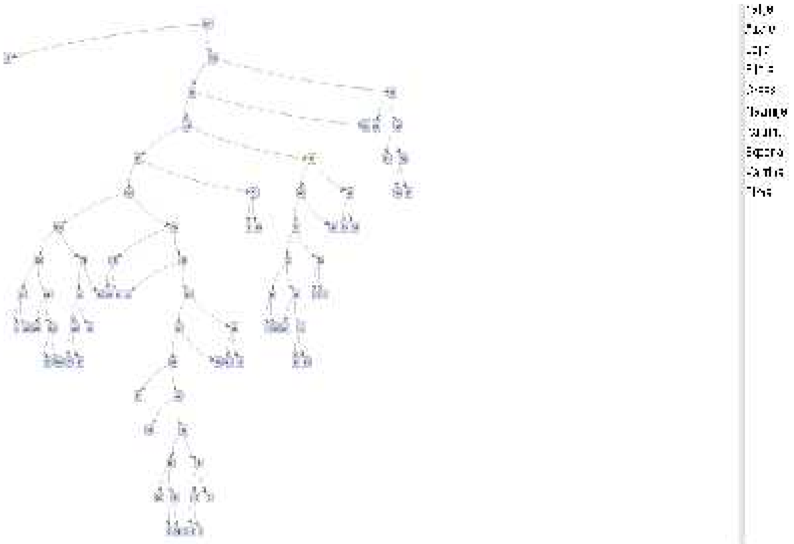


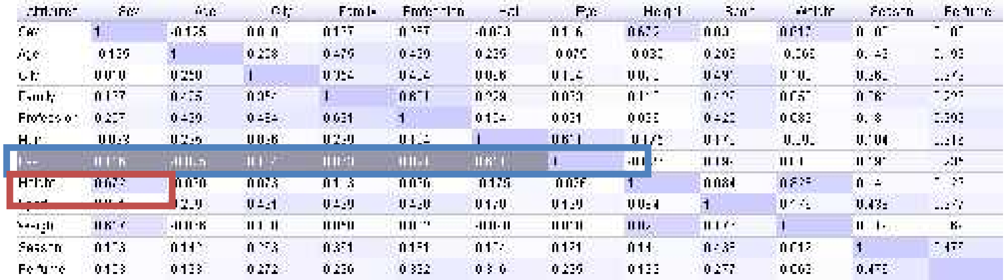
Figure 9. Clustering from dataset

CORRELATION

Correlation matrix

Correlation matrix calculates and visualizes the rank correlation coefficient of the columns in a data set. This matrix is symmetric because correlation between Age and City is the same as correlation between City and Age. Its elements lie in the [-1, 1] interval. As the correlation of a variable is 1, the diagonal elements of the matrix are all 1 which means each variable is perfectly correlated with itself. With value 1, there is a perfect correlation and with 0 there is no correlation. If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher the value stronger is the correlation. $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated (Han, Hiswei & Kamber, Micheline, 2006).

If there is a question "what colors are your eyes?" and "what color is your hair", there is a need to see if people who have the same hair color have can have the same eye color as well. According to the result below, eye and hair colors are almost correlated with each other according to the dataset with a value of 0.611. Also there is a different correlation between Height and Sex, which means that the height depends on what sex the individual is. So if the participant was male than has a highest height and for females it is opposite.



Attributes	Sex	Age	City	Family	Profession	Hair	Eye	Height	Sport	Weight	Season	Feature
Sex	1	-0.125	[illegible]	0.177	0.257	-0.053	0.116	0.622	0.03	0.612	[illegible]	[illegible]
Age	0.135	1	0.228	0.475	0.439	0.235	0.075	0.032	0.203	0.203	[illegible]	[illegible]
City	0.070	0.250	1	0.054	[illegible]	0.058	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Family	0.177	0.475	[illegible]	1	0.871	0.329	0.073	[illegible]	[illegible]	0.057	[illegible]	[illegible]
Profession	0.227	0.439	0.454	0.031	1	0.124	0.031	0.035	0.422	0.083	[illegible]	0.393
Hair	0.053	0.235	0.058	0.329	0.124	1	0.611	0.175	[illegible]	[illegible]	[illegible]	[illegible]
Eye	0.116	[illegible]	[illegible]	[illegible]	[illegible]	0.611	1	[illegible]	0.194	[illegible]	[illegible]	[illegible]
Height	0.622	0.050	0.073	[illegible]	0.036	0.175	0.036	1	0.084	0.825	[illegible]	[illegible]
Sport	[illegible]	0.209	0.421	0.429	0.420	0.170	0.129	0.054	1	[illegible]	0.435	[illegible]
Weight	0.612	-0.018	[illegible]	0.050	[illegible]	-0.040	[illegible]	[illegible]	[illegible]	1	[illegible]	[illegible]
Season	0.173	[illegible]	[illegible]	0.351	0.151	[illegible]	0.121	0.11	[illegible]	0.012	1	0.477
Feature	0.133	0.133	0.272	0.236	0.332	[illegible]	0.235	0.133	0.277	0.063	0.475	

Figure 10. Correlation Matrix

Figure 11 shows that Eye and Hair have the highest correlation value, which in the graph is depicted with a red dot emphasized on the color bar depicted for correlations.

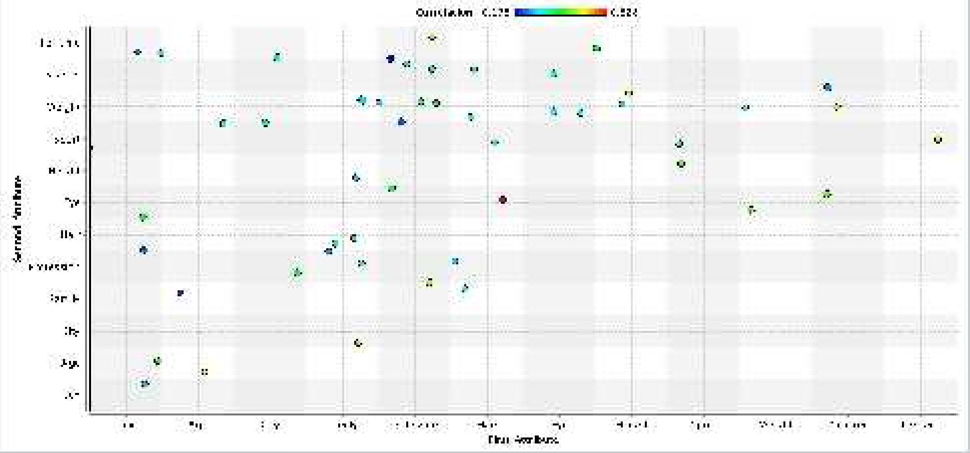


Figure 11 Scatter Plot for EYE and HAIR attributes.

Correlation data

In case to visualize the data there is a need of a Scatter plot. This graph allows to know how much the variables are affected by another. The relationship between two variables is called a correlation. The closer the data points the higher the correlation between two variables or the stronger the relationship. When a scatter plot shows a relationship between two variables, there is not necessarily a cause and effect relationship. Both variables could be related to some third variable that explains their variation or there could be some other cause.

Figure 12 evaluates three attributes, Family, Age and Sex. Based on this Plot, the data are positively correlated but weak because they are not so close to each other. Until the age of 30 there are participants with single, engaged, married and in a relationship status. After this age there are almost married and divorced participants.

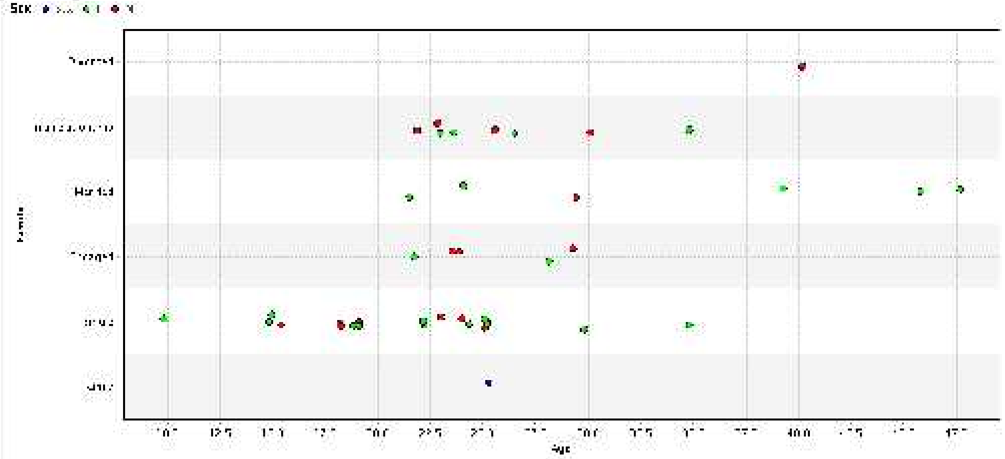


Figure 12. Scatter Plot for Age, Family and Sex.

Figure 13 also positively correlated with analysis of attributes eye and sex. This shows that many participants evaluated in this survey fall in the group with black and green eyes. In the group of black eyes both females and males are similar. With green eyes there are more females then males.

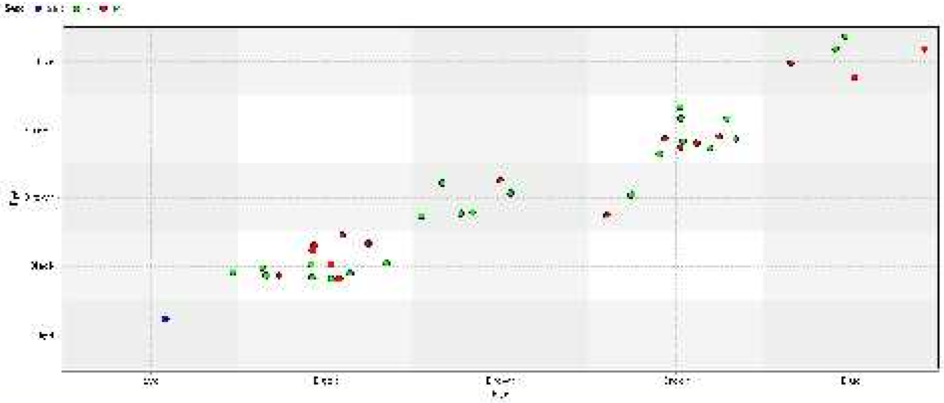


Figure 13 Scatter Plot for Eye and Sex.

It is important to emphasize the correlation between family and sex. In this case there are many singles and in a relationship participants. The graph is correlated positively because Family attribute on X increases and Family attribute on Y increases as well. They are strongly correlated to each other.

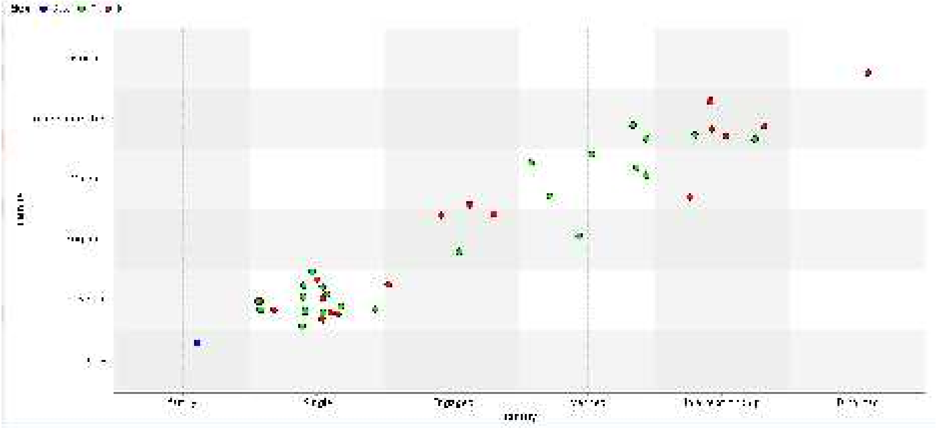


Figure 14. Scatter Plot for Family and Sex.

CONCLUSION

Text mining is an important part of extracting information by means of tokenizing, and removing the stop words and implementing other processes.

This study demonstrated the implementation of these techniques by using open source Rapid Miner software and Information Extraction plug-in.

Study might be useful for some students, developers and researchers to learn how to implement text mining by using a popular software, e.g. Rapid Miner.

REFERENCES

[1] Han, Jiawei. Kamber, Micheline (2006). Data Mining Concepts and Techniques, Second Edition

[2] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in Proceedings *of the 16th Conference on Computational Linguistics*, pp. 466–471, USA.

[3] Hand.D, Mannila H, Smyth.P. Principles of Data Mining (2001). Massachusetts Institute of Technology.

[4] Berry Michael, Linoff Gordon, Data Mining Techniques (2004), Revised Edition.

[5] Moore, Andrew. Statistical Data Mining Tutorials (2006) [http://www.autonlab.org/tutorials/ Last retrieved on: 16.03.2011]

[6] N. A. Chinchor, Overview of MUC-7/MET-2 (1998).

[7] ACE. Annotation guidelines for entity detection and tracking (2004). Last retrieved in January 2011.

[8] Witten, H. Ian. Text Mining, Computer Science, University of Waikato, Hamilton, New Zealand

[9] Rapid Miner 4.4. User Guide, Last retrieved in March 14, 2011.