Abstract—The aim of this article is to present an application whose purpose is to make suggestions of restaurants to users. The application uses as input the descriptions of restaurants, reviews, user reviews available on the specialized Internet sites and blogs. In the application there are used processing techniques of natural language implemented using parsers, clustering algorithms and techniques for data collection from the Internet through web crawlers.

Keywords: Web crawler, parser, stemmatizer, cluster, lemmatizer, data mining, machine learning.

I. INTRODUCTION

The present paper aims to design an application that allows the suggestion of restaurants based on a clustering algorithm. The algorithm is called SimpleKMeans. Briefly, the application is first able to download data from the Internet using a crawler visiting several websites which are copied and then downloaded. Subsequently these rough data are processed with a parser to extract information relevant to the application such as descriptions, reviews and comments. Parsing data involves their normalization and this translates to extracting stems from words [6]. A stem is the root of a word. Stems are extracted using stemmer algorithms. Data normalization introduces a new stage required in clustering: calculating the scores of occurrence data in documents. This score is called the TF / IDF (term frequency / inverse document frequency) and is a measure of frequency of words for each document (the most frequent words in most documents have a score of IDF equal to zero, therefore will be considered frequent and they will not be taken into account in the clustering process). Having the scores TF / IDF it can be used a clustering algorithm such as SimpleKMeans to generate a predefined number of clusters and to assign a sample data to each cluster.

II. THEORETICAL ASPECTS

A) Collection of data from the Internet using a crawler

The first objective of the project was to obtain information about restaurants on which the classification was to be conducted. This information consisted in describing the restaurant, reviews and comments of the user. In order to obtain these data available on the Internet, it was used a web crawler that performs systematic searches on the site indicated.

A Web crawler is a program that browses systematically the WWW (World Wide Web) with the primary purpose to allow indexing documents on the Web. Web search engines and other sites use these crawlers to update their content or content indexes of other Web sites. These programs allow copying all the pages visited for further processing by a search engine that indexes downloaded pages so that users can search them much faster. These crawlers can also be used to collect data on the Internet for further processing. The behavior of a Web crawler is determined by a number of combinations of policies regarding its operation:

• A selection policy specifying which pages on the Internet are to be brought and saved;
• A re-visitation policy specifying when to check whether there have been made changes to the pages;
• A coordinating policy specifying how to coordinate distributed Web crawlers;
• An avoidance policy in order to avoid overloaded websites.

B. SnowballStemmer - stem extraction algorithm

The second objective was the process of extracting the roots from words. The process of extracting the root of a word or the stem is called stemming algorithm. Roots obtained from the stemming process need not necessarily to be identical to the morphological root of the word.

The role of a stem extraction algorithm [1] is to enable the identification of variations and inflections that come from the same word: for example, the words go, goes come from the variation of the same word: the verb to go.

In order to calculate the normalized scores tf / idf all the variations of the same word must be indexed once at the root of that word, and this is something the extraction algorithm of a word’s root ensures of.

Using an algorithm for extracting [3] the stem together with a list of stop words is a good solution that provides a control method to remove very common words in speech: conjunctions, prepositions, verbs, state adverbs very common nouns.

Using the algorithm of extracting the stem allow the normalization of words in a text, ie removing suffixes, prefixes, articles and stop words.
In this case it was chosen an adapted version of SnowballStemmer which in turn is an adaptation of Porter Stemmer, the standard recommended by the Stanford University in terms of algorithms of extracting stem.

It was preferred the use of a stemmatizer instead of a lemmatizer because stemmatizers are easier to implement, easier to use and have a relatively high processing speed for large data sets. The advantage of lemmatizer is that it analyzes the context in which the word occurs and the root extraction process takes place taking into account the words that surround it. But, the dependence on context of a lemmatizer makes it difficult to implement and harder to use.

3. Calculating scores Tf / Idf

The third objective was on the basis of information obtained through Jsoup and the roots obtained from the process of extracting stem of a word, to calculate the scores of occurrence TF / IDF (Term Frequency / Inverse Document Frequency) [4].

In order to achieve a classification is important to know how often a concept appears in a document and how is this reflected in the entire collection of documents. Scores tf / idf [5] represent a statistical indicator that reflects how important a word is to a document within a collection of documents. Most times it is used as a weight factor to process existing information. This value increases with the number of times a word appears in a document.

4. Clustering algorithm SimpleKMeans

This clustering algorithm [2] is part of the algorithms based on the calculus of the center of each cluster. To assign instances to clusters, first it is calculated the center of each generated cluster, then the instances are assigned using a mathematical function of minimum distance (Euclidean distance is preferred). Clustering allows classification of instances. K-means is a vector quantization method and aims to partition n observations or instances into k clusters, except that each instance belongs to only one cluster. This leads to a partitioning of the space given into Voronoi cells. This algorithm is used in the data mining process and in the machine learning algorithms. The figure below, Figure 1, illustrates the partitioning algorithm of k-means. K-means algorithm has an important parameter represented by the number of clusters we want to achieve in the process of clustering. Usually this parameter is preset depending on how many instances we have to classify in order to obtain a number of clusters that is relatively balanced.

III. PRACTICAL RESULTS

Based on the theoretical aspects described above, an application with a modular structure was designed to categorize restaurants. The structure of modules is as follows:

1. The module of collecting data
2. The parsing module of data comprising the data filtering module and the extraction of stems module
3. The calculation module of scores tf / idf
4. The clustering and assignment of instances to generated clusters module
5. The display module

The modular structure of the application has several advantages:

1. It offers scalability to the application: the property to be extended easily by adding new modules with minimal modification to existing modules
2. Increase application portability through the fact that the modular structure makes it easier to adapt to the requirements and functional structures of different operating systems, so that the transition from a desktop application to cell one can be achieved with minimum effort
3. It offers re-usability: modules can be reused in other similar applications.

It was made an application which through a graphical user interface allows the user to track in real time how the data is downloaded and processed for their clustering. The application works with files that are dynamically generated during program execution.

These files are suitable for the modules and contain input data and the output data for the operation of modules.

To calculate scores TF / IDF and for clustering it was used the Weka library (stable version 3.6.10). This library provides both a graphical interface that can perform various operations from text parsing to classification and association, and also an API that provides methods that can be integrated into the Java language used to develop the application.

For k-means algorithm must be taken into account four important parameters: the number of clusters is intended to be generated, the number of seeds, the function used to calculate distance and the maximum number of iterations of the algorithm.

After using the application for the 51 data sets (restaurants) there were obtained seven clusters with the following distribution:
1. In cluster 1 were found a number of 2 instances with a total of 4%.
2. In cluster 2 were found a number of 7 instances with a total of 14%.
3. In cluster 3 were found a number of 15 instances with a total of 29%.
4. In cluster 4 were found a number of 2 instances with a total of 4%.
5. In cluster 5 were found a number of 2 instances with a total of 4%.
6. In cluster 6 were found a number of 7 instances with a total of 14%.
7. In cluster 7 were found a number of 16 instances with a total of 31%. 

---

**Fig. 2.** The output of the application – view of information about restaurants from each cluster.

**Fig. 3.** The output of the application – view of restaurant from each cluster under the form of images.
Fig. 4. Capture during the run of the program – bringing data from the Internet using JSoup

Fig. 5: The resulted clusters and their centers
IV. CONCLUSIONS

The aim of the project described and implemented was to realize a clustering of the restaurants in order to suggest a restaurant representative for each cluster identified. The number of generated clusters was seven. This number has been specified as an input parameter of the k-means clustering algorithm, along with the number of seeds that was 500.

Based on the input data that the user provides meaning the paths to the locations that will save the files required for running the application (the downloaded data storage, the one saving the processing stems, one containing the words stop and the parameters for the clustering algorithm, one in which scores Tfidf are saved and the one in which it is made the instances assignment to clusters), the application through its modules, performs the clustering of the 51 restaurants given as incoming instances in 7 clusters.

The five modules of the design are relatively independent of each other, but the output of one module is the input for the next module at a lower level: for example, documents downloaded from the Internet are input to the algorithm of extracting stems, stems resulting from this algorithm are input into the process of calculating Tfidf, and the result of this process under the form of calculated scores is input for the k-means clustering algorithm that generates clusters and assigns instances. Based on this assignment, the display module suggests a restaurant for each cluster generated. This modularization of the application ensures its scalability and allows a relatively easy expansion of the project.

Through this chain of modules, the system is able to provide a justification for how the numbers of clusters are generated. This justification is visible during program execution via dynamic creation of output files. It can be tracked the creation of these files in real time.

In conclusion, we can say that the assignment of data instances into clusters using (and hence the suggestion process of restaurants) k-means algorithm depends on many factors such as:

- Clustering algorithm used (in this case it is about Simple KMeans);
- The number of clusters that the algorithm generates;
- The maximum number of iterations;
- Calculation of the distance function (in this case it is the Euclidean distance);
- True or False value of the order of instances keeping (boolean instance preserver order);
- The functioning of the algorithm (in this case using a set of training).

REFERENCES


Iulia-Alexandra Iancu is a student at Technical University of Cluj-Napoca in Computers Science and Information Technology area. She activates in the fields of human computer interaction and simulation techniques using high performance computing.

Eugenia Iancu is an associate professor at Stefan cel Mare University of Suceava. She is PhD in Computer Science since 2011. She is with the Informatics Department at the University “Ştefan cel Mare” of Suceava, Economic Sciences and Public Administration.