ISSN 1512-1232

Text classification using NLTK

Gela Besiashvili, Papuna Qarchava

Iv. Javakhishvili Tbilisi State University, The Faculty of Exact and Natural Sciences

Abstract

In this work, the task was to check how well-known technologies and tools for the English language work in the case of the Georgian language. Namely, using the example of text classification (NLP tasks).

The 20 newsgroups dataset (free dataset) was used as processing data. Basically, this dataset is used in the process of testing new algorithms for solving the problem of classifying text data. The dataset is presented in English and there is no translation into Georgian. For the current work, this dataset was translated into Georgian using an online translator from Google. The translated text was represented by two different coding methods (ASCII and UTF-8). Processing text in English and Georgian with ASCII encoding produces a good result, but processing text in Georgian with UTF-8 encoding takes longer to obtain the same result.

The results obtained allow us to say that for processing Georgian text using NLTK tools, it is better to present ASCII-encoded text.

Keywords*: Text classification, machine learning, NLTK*

Introduction

Practical work in Natural Language Processing typically uses large bodies of linguistic data or corpora. A text corpus is a large body of text. Many corpora are designed to contain a careful balance of material in one or more genres. For example, in English the language, there exist many different corpuses, such as are:

1. Gutenberg Corpus - contains some 25,000 free electronic books [1];
2. Brown Corpus - the first million-word electronic corpus of English, created in 1961 at Brown University. This corpus contains text from 500 sources, and the sources have been categorized by genre, such as news, editorial, and so on [2];
3. Reuters Corpus - contains 1,800,370 news documents totaling 1.8 million words. The documents have been classified into 90 topics [3]; and etc.

The process of searching for information is not the result of only one type of operation. Its success and relevance depend on the adequacy and completeness of the search cycle. One of the important places in this cycle is the classification stage.

Classification is a search process, the purpose of which is to automatically assign classes to documents from a predefined set. Classification uses the same models as information retrieval. However, there is a different approach - classification models are used for preliminary recognition of the appropriate class for documents. For this, the set of documents is divided into learning and testing sets before preprocessing. Initially, the system requires documents to be learned to create classifiers based on them. And then, based on the set of test documents, they predict the class of their belonging and classify them:

1. Natural Language Toolkit (NLTK) [4];
2. Stanford Core NLP [5];
3. SpaCy [6].

ISSN 1512-1232

Related works

Numerous studies have been conducted and many tools have been developed in the direction of processing the Georgian language. There was created a several Georgian corpuses, some of them are:

- Georgian language National Corpus [7]
- Ilya State University (ISU) created Georgian language Corpus in 2009-2016 [8].
- KaWaC - Kartvelian Web as a Corpus the Georgian language Corpus [9]. and other

Despite numerous studies, the task of processing the Georgian text is still unresolved. This is related to the peculiarities and complexity of the grammar of the Georgian language. The current research aims to study the question of how far it is possible to solve the task of classifying Georgian text using the existing tools, in particular, using the NLTK tool.

Text classification

There are mainly two types of text classification systems:

- Rule-based text classification - Rule-based techniques use a set of manually constructed language rules to categorize text into categories or groups. These rules tell the system to classify text into a particular category based on the content of a text by using semantically relevant textual elements.
- Machine learning-based text classification - Machine learning-based text classification is a supervised machine learning problem. It learns the mapping of input data (raw text) with the. This is similar to non-text classification problems where we train a supervised classification algorithm on a tabular dataset to predict a class, with the exception that in text classification, the input data is raw text instead of numeric features.

In the work there was used machine-learning supervised technique.

As it is known text classification involves some steps, such are:

1. Data gathering
2. Cleaning the data
3. Feature Extraction, etc.

The tools and various methods defined in the NLTK are used in the researches.

By using the NLTK tools the world has developed many applications for data classification in Arabic [10], Vietnam [11], and so on languages.

In the current work for cleaning the data, feature extraction, and so on, for both textual English and Georgian data used NLTK tools.

Experiments

The 20NewsGroup dataset [12] was taken as test data for the research. This dataset is a collection of newsgroup documents. The collection of 20 information groups has become a popular dataset for experiments in text applications of machine learning techniques such as text classification and text clustering. This data set contains about 20,000 documents, which are presented as separate files for each group.

Since the mentioned dataset is in English and there is no Georgian version of it, we translated it using the online translator (google.translate.com) resource developed by the Google organization.

The resulting translation was saved in two encodings: ASCII and UTF-8. Additionally, the data were divided into two groups with a ratio of 70/30 (study/test).

ISSN 1512-1232

English texts were processed taking into account the standard methods and approaches of the NLTK tool. As for Georgian text processing, there are no clearly defined components for Georgian text processing, such as stop-words, stemming, lemmatization, tagging, etc. There were manually developed appropriate components specific for the current work.

Georgian text was processed in the paper in the case of ASCII and UTF-8 encoded texts in parallel with English text processing. The methods and tools used in the processing of English-language texts were used in the analyzing of text data presented in ASCII encoding. Such an approach gave us results close to those obtained when processing English-language texts. As for the evaluating of Georgian text data represented by UTF-8 encoding, in this case, we could not achieve the desired results using existing methods, and it was necessary to develop additional approaches and tools. Also, data processing time increased, which leads to inappropriate consumption of mannan resources.)

Conclusion

As work has shown, processing English language text and Georgian text represented by ASCII encoding gave us good results, while processing text represented by UTF-8 encoding, the classification will need more preprocessing to get the desired result (matching of classification results).

The work carried out allows us to think that for the processing of Georgian texts by the existing NLTK tool, it is better to present text in ASCII encoding. In this case, according to the authors, it will be necessary to improve only the so-called abundance of stop-words for Georgian-language texts and as for NLTK (NLP technologies) tools, they can be used without any changes.

Of course, this is not the final result in this regard, considering that the existing approaches and tasks have significantly changed due to the introduction and use of artificial intelligence technologies. Therefore, additional studies are needed.

References

1. http://www.gutenberg.org/
2. http://icame.uib.no/brown/bcm-los.html.
3. https://trec.nist.gov/data/reuters/reuters.html
4. https://www.nltk.org/
5. https://stanfordnlp.github.io/CoreNLP/
6. https://spacy.io/usage/v3
7. http://gnc.gov.ge/gnc/page
8. http://corpora.iliauni.edu.ge
9. https://www.sketchengine.eu/kawac-georgian-corpus/
10. https://ieeexplore.ieee.org/abstract/document/8931818
11. http://export.arxiv.org/abs/1801.01331
12. http://qwone.com/~jason/20Newsgroups/.

Article received: 2023-03-03