Using NLP Python Tools in Methods of Content Analysis

Maria Razno [0000-0003-3356-5027], Nina Khairova [0000-0002-9826-0286]
National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

mari.razno@gmail.com, nina_khajrova@yahoo.com

Abstract. This article describes the relevance of the content analysis task applied on Twitter text data using various NLP methods, that can be implemented on Python programming language. It also includes the concept of different NLP methods and algorithms, its main varieties and the most popular Python packages and libraries for working with text data. The content analysis algorithms based on the text processing is introduced in this study. It shows how to use NLP methods in practice.

Keywords: Content analysis, Python, Text classification, Text processing NLP, NLTK, Twitter

Social networks have become widespread in recent years, primarily as a tool for communication, exchange of ideas and information. Very often, such network reflection affects the response of people to a particular event. Social networks are increasingly being used as a source of information, including information related to global world events. The idea of using NLP Python tools in methods of content analysis is extremely actual due to the fact, that the sharing of structural and content data potentially allows social networks to be used to solve a wide range of tasks, including identification of trends in the modern world. In the last decade, online social networks where social interaction is carried out through web technologies played a main role in society. Over 70% of Internet users use online social networks. The social network Twitter has more than 2 billion users nowadays.

Content analysis is a method of obtaining reliable and valid conclusions from texts. Content analysis allows you to systematically and as objectively as possible investigate large arrays of documents and identify implicit information - the purpose of the author, the representation of the addressee of the message, etc. A social network is any social interaction that can be represented by many social units and the relationship between them. Natural language processing (NLP) is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

Today quite popular method of content analysis is "text-mining". It can be considered, on the one hand, as a set of techniques for identifying non-trivial trends in

COLINS'2020, Volume II: Workshop. Lviv, Ukraine, April 23-24, 2020, ISSN 2523-4013
http://colins.in.ua, online

the textual data that may interest the researcher (including content analyst), and on the other - as an interdisciplinary field of research covering information processing techniques, machine learning training, neural networks, statistical classification, database work, etc. The main elements of Text Mining are: text classification, clustering, construction of semantic networks, feature extraction, summarization, question answering, thematic indexing, keyword searching etc. The Python libraries NLTK and SpaCy has all the necessary algorithms for managing these methods.

Named entity recognition (NER) is a sub-task of extracting information that attempts to find and classify named entities in unstructured text into predefined categories, such as names of people, organizations, places, times, numbers, monetary values, percentages, etc. NER is used in many fields in Natural Language Processing (NLP), and it can help answering many real-world questions such as: which companies were mentioned in the news article; were specified products mentioned in complaints or reviews; does the tweet contain the name of a person; does the tweet contain this person's location. The most popular RIS platforms are: GATE, OpenNLP and SpaCy.

The next approach of content analysis is a co-referents resolution. Its main task is to identify all the entities of the text and to establish for each group of the noun the text to which the essence of the given noun refers. It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction. The basic complexity of co-referents resolution stems from the fundamental problem of linguistic polysemy. There is a NeuralCoref library and Coreference Resolution in SpaCy with Neural Networks in Python to accomplish this.

In addition we will use a sentiment analysis as it is one of the most important aspects of content analysis of social networks, because it allows you to determine the opinion of a group of actors about a particular event. If given text, which is a subjective expression, then assuming that the expression has a single object, it helps reveal the emotional connotation of the text as one of two polarities: positive or negative. The main Python libraries for this task are: NLTK, SpaCy, TextBlob, Stanford CoreNLP and Gensim.

An approach called dependency grammar is to relate individual words. A binary asymmetric link is established between the word pairs, which indicates the main word and the dependent one. The dependency tree is represented as a labeled oriented graph in which nodes are lexical units and the marked arcs represent the relationship of the dependencies between the main and dependent words. It has a potential usefulness of bilexical relations in disambiguation and gains in efficiency that result from the more constrained parsing problem for these representations. There are the following Python libraries for this purpose: NLTK, Stanford Parser, SpaCy.

In this study we use a dataset of messages from users in the Twitter social network. The collection has about 120 million tweets with relevant judgments about over 500 events that have been going on during 2012-2016 [3]. This dataset was created by scientists at Glasgow University to tackle Event Detection problem from the text data, in this social network. The dataset contains only User ID pairs and Tweet ID pairs.

The dataset covers a number of interesting and significant events, including Hurricane Sandy, the US Presidential Election and others. To retrieve text data from tweets, it is required to use the Python wrapper, such as the twitter-dataset-collector library. The data is got in json format.

To summarize, in the course of our research, we can say that Python is a wonderful programming language, which provides a lot of great libraries for creating good models for content analysis using libraries that contain the most popular natural language processing methods. As the result of compiling our content analysis model, the user gets some statistics with graphs, scores and rates on three different topics: sport, politics and natural disasters. These results will show entities, key words, short summarization and related words on each topic, sentiment scores of tweets.

References

1. Rubtsova, Y.: Constructing a corpus for sentiment classification training. SOFTWARE SYSTEMS 1(109), 72-78 (2015).
2. A list of Twitter datasets and related resources. https://github.com/shaypal5/awesome-twitter-data. Last accessed 10 April 2020.
3. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, 110-120 (2018).