УДК [811.11+159.955.5]:81-139

TEXT AND DATA MINING IN CORPUS-APPLIED TRANSLATION STUDIES

Yuliya Demyanchuk
Lviv State University of Life Safety

Text mining and Data mining in Orange software is mostly used for, the "raw data" analysis that can be transformed into structured data and can be used for executing tasks such as Classification, Hierarchical clustering, Section clustering, Word count clustering, Character count clustering. This allows to gain insights from a wide range of data sources, such as texts of official and business discourse, social media posts, and news articles.

Text mining and Data Analytics are related but they have provide different processes for extracting statistical info from textual data. Text mining involves the application of natural language processing and machine learning techniques to discover new information (patterns, trends, models) from large volumes of written sources [1].

Although, Text Analytics (Text mining) focuses on extracting meaningful information, sentiments, context and key words from text often using statistical and linguistic methods. Mostly, extracted statistical information is involved to the process of Data Analytics (Data mining). In corpus and applied linguistics Data mining means extracting and discovering language patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems [2].

In my research I used *Cortical.io* for Data mining that offers free Natural Language Processing APIs for text processing tasks to find key words in certain clusters and to define semantic fingerprint of the term combinations (*full-scale invasion, humanitarian crisis, free-trade*). Using Data mining in uncovering hidden patterns and utilizing text analytics I defined sentiment in corpus. Both Text and Data mining play crucial role in transforming collection of documents into valuable knowledge, with Text mining exploring language patterns and data analytics providing interpretative statistical information.

Text mining same as Data mining is widely used in corpus and applied linguistics, in various fields such as natural language processing, information retrieval, and corpus-driven approach of daily newspapers analysis (The Guardian, The Washington Post, The New York Times). It has become an essential tool for linguistics to extract standard statistical measures and defined variables (words count, characters count, N-grams count, average word length) from collection of documents and to make data-driven decisions.

Based on this, to investigate the effect of corpus-driven approach in the new direction of Corpus-applied translation studies I reviewed 461 text documents from the British daily newspaper *The Guardian* for the period from 2023 to 2024. The findings show three clusters *C1* – full-scale invasion, *C2* – humanitarian crisis, *C3* – free-trade. I reveled a difference in Word, Character count and Section clustering. The data resources were used to determine p-value, degree of freedom, t-statistic. A range of procedures consisting widgets of Orange software were adopted to find out the distributions of attribute values (C1, C2, C3).

Thus, the results of the analysis served as a reference for the deep study. Once categorized into attribute values clusters were than singled out to discover the new data and language anomalies.

References

1. Ramzan T., Muhammad K., Shaeela A., Fakeeha F. (2016). Text Mining: Techniques, Applications and Issues. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.071153.

2. Degaetano-Ortlieb S., Fankhauser P., Kermes H., Lapshinova-Koltunski E., Ordan N., Teich E. (2014). Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers.