Log Analysis: Topic Modeling applications on fine-features data processing system

KeyWords: Data Processing, Topic Modeling, Coherence Score, Similarity, LDA, LSI, NMF, ProdLDA, BERTopic

An ever-increasing number of operations is performed nowadays with the help of computer systems. They are everywhere around us and each of them produces a huge quantity of log files for each operation.

In this context, the aim of this project is to study and create an architecture able to read and categorize the logs produced by computer systems based on logs content.

The outcomes of this project form the foundation of a subsequent task of anomaly detection, which is highly requested since lets companies to detect malicious attacks or failures and by which the focus is moved towards the most relevant events.

As previously mentioned, I deal with log files. They contain the sequential and chronological records of the operations done by computer systems. Since new systems are extremely complex, the interaction of their software components causes the registration of a big number of operations.

Accordingly, analyse and gather useful information from this huge amount of data is challenging and requires many efforts, because both field expertise and cutting-edge technologies are needed.

The starting point is the LogHub dataset repository, where several computer system log types are collected. Specifically, for this research the focus is on logs coming from “HDFS” and “Spark” distributed systems.

One of the issues of this study is the object manipulated: the log. Differently from common texts like books, logs are short texts with a lot of technical words and useless elements like path or web keys. For this reason, the manipulation and the cleaning phases are essential to provide useful information to models adopted.

So, the first part of the project is the process of making the data more suitable and it is made by several steps. They are principally based on NLP and Regular Expression crafting, which allow to obtain meaningful words from unstructured logs.

Afterward, the topic models are implemented. Concisely, they consist in annotating documents with thematic information and in grouping together documents that share similar contents.

About the chosen algorithms, the project is split into two branches. The first one investigates standard approaches, based on probabilistic or algebraical operations, where the models analysed are LDA, NMF and LSA.

The second branch, instead, follows a modern approach based on Neural Network and Feature Embedding. For this part the proposed algorithms are ProdLDA, Top2Vec/BERTopic and, by combining many of their approaches, my custom model called GEAC.

All models have been tested with the CV coherence score. Inside the world of coherence scores, the CV is the most meaningful to evaluate the goodness of topics clusters since it provides results close to the human ones, the ground truth for this task.

One issue with Topic Models is about the quality, since the models cited do not always guarantee that each topic cluster is independent from the others. For this reason, I perform an analysis about the similarity between topic clusters by exploiting the spatial representation of each word inside a topic through Feature Embedding. Thanks to this analysis, I can select a model that has a high CV metric score and that is also optimal for what the supervisor might infer regarding similarity.

Overall, at the end of this project, a generic log line can be labelled with an almost black box approach that provides optimal outputs both under human and machine perspectives.