1. Unstructured Data Analysis for UBH Knowledge Base

by Pinar Karagoz (Middle East Technical University, Turkey)

A knowledge base is a collection of structured and unstructured data to provide information about and to analyze a certain topic. In a knowledge base, in addition to structured data such as databases and and forms, a large amount of data is unstructured, mostly textual, in the form of documents, web sites, and social media messages. The knowlede based aimed in Cost Action18110, which is ***Underground Built Heritage (UBH) knowledge base***, will include a variety of unstructured data such as descriptions of underground built heritage, surveys conducted, information collected from social media and web resources. The automated analysis of textual data provides a considerable potential for extracting the information from the content and to facilate the use of and hence to increase the effectiveness of the knowledge base.

The most basic functionality that is commonly provided on knowledge base is ***keyword based search***. This functionality is familiar from popularly used web search engines. For knowledge bases, a similar keyword based searching software mechanism, yet generally in smaller scale, is installed. Another analysis type that can be complementary to keyword based search is ***automated classification of the documents***. Once a classification system is defined for UBH, including classes such as *Urban UBH*, *rural UBH* or *UBH related to social interactions*, then the documents in the UBH knowledge base can be automatically annotated with such class labels. This is considered as a computational task generally solved through ***Artificial Intelligence (AI) methods***. Automated classification of the documents can be augmented with similar analysis on images. Hence unstructured data of both text and image can be automatically classified and searched according to the class labels.

As another automated unstructured textual data analysis task, ***sentiment analysis*** (aka opinion mining) provides high potential for extracting valuable information for UBH on subjective textual content. Sentiment analysis aims to automatically detect the orientation of the subjective information from the text as ***positive*** and ***negative***. Within UBH knowledge base context, sentiment analysis can be applied on ***transcripts of the interviews*** and ***open-ended responses of survey***s. In addition to general subjective orientation, ***sentiment polarity for particular aspects*** of the topic mentioned in the text can be extracted as well. For instance, in a stackholder interview on a particular UBH asset, the opinion on economical aspect and safety aspect of the UBH could be discussed.

In addition to the overall sentiment orientation, the subjectivity value for economical and safety aspects can be automatically extracted, as well. For search engines, the trend and one of the research focus is to improve user experience for search through AI-supported interactive solutions. Hence, rather than keyword based search, a user directly type the question.

Such a user interface may enable UBH knowledge based user to drectly get answers for questions such as “In which countries are there social intection related UBH assets?”, instead of keyword based search. Furthermore, such automated question answering systems can be enhanced towards chatbots to answer a series of questions or to improve the information retrieval performance.