Text Mining

Abstract

(Lessions will be in English.) Lessons are divided between CS and DS + TTC students; the aim is to study and build search engines and recommendation systems. The first part is a presentations of Text Mining (an AI branch) with techniques of *Information Retrieval*, *Information Filtering*, *Text Classification* and *Summarization*, all using *open source* software. DS and TTC students will have labs using R and Python, CS students will use Java (Lucerne). The exam consists in a written test plus a project (done in groups of a maximum of 3 people), that can be followed by an oral test at will. On e-learning suggested books will be published.

Contents

1 Introduction and History

AI (Artificial Intelligence) was born in early '900: Machine Learning is only a small branch, started in '70s. Computational Linguistic and Statistics techniques are used as well. An Agent is considered *intelligent* if can understand the semantics of its Ambient, in that case texts written by humans in a *natural language*. To understand a text, the Agent must have a previous knowledge of the matter. Text Mining can be supported by Machine Learning but its techniques are more general and is built to emulate human behaviour. Other used techniques are DMs (Distributional Meanings) algorithms over time.

In 2004, Ian Witten (Weka project leader) published a paper titled "Text Mining" in which he reports that the first workshop was started in summer 1999: texts were processed in an automatic way to extract useful information.

Examples of Text Mining are *Sentiment* and *Opinion Analysis*; another example is *Text Summarization* (the ability to write a *snippet*, a short description of a text). *Reccomender Systems* are largely used by web-shopping platforms: they have revolutionized shopping and adverts suggesting useful information to users (positive filtering) or avoiding contents (negative filtering), analysing text content. Users are profiled by *Avatar*, their digital representations.

Text Analytics is used in healthcare to reduce the number of *fake news* on the web.

Knowledge Graphs are used to go behind traditional Text Mining techniques and have a better understanding of the text and improving the model.

The particularity of Text Mining is that a information in the text are not hidden but well written, and humans can reach it with no difficulty (but for text length); is largely used to analyze in a semi-automatic way lot of unstructured documents for decision making.

There are a lot of challenges in Text Mining:

- Text annotation;
- Dealing with large *corpora* and *streams*;
- Organizing semi- and un-structured data;
- Dealing with ambiguities on many levels (lexical, syntactic, semantic and pragmatic).

Information Retrieval aim is to find things on the Web, and has its roots in 70s.