Chapter 6
Text Preprocessing: A Tool of Information Visualization and Digital Humanities

Piotr Malak
University of Wroclaw, Poland

ABSTRACT

Digital humanities and information visualization rely on huge sets of digital data. Those data are mostly delivered in the text form. Although computational linguistics provides a lot of valuable tools for text processing, the initial phase (text preprocessing) is very involved and time-consuming. The problems arise due to a human factor – they are not always errors; there is also inconsistency in forms, affecting data quality. In this chapter, the author describes and discusses the main issues that arise during the preprocessing phase of textual data gathering for InfoVis. Chosen examples of InfoVis applications are presented. Except for problems with raw, original data, solutions are also referred. Canonical approaches used in text preprocessing and common issues affecting the process and ways to prevent them are also presented. The quality of data from different sources is also discussed. The content of this chapter is a result of a few years of practical experience in natural language processing gained during realization of different projects and evaluation campaigns.

INTRODUCTION

Big Data is a term used to describe a vast storage and source of data. Those data can be extremely valuable if properly gathered, compared and analyzed. Back, in the Machine Learning (ML) era, it was obvious to use charts in order to approximate to human analytical perspective and abilities any huge data set, presented originally in tables. Visualizing data allows us to easily detect and track changes and trends hidden in data sets, which are too extensive to analyze them only on the basis of raw data. Pie-, bar- and many other charts, are still useful tools for presenting and analyzing data. They are permanently used for analyzing huge amounts of structured data. Such data have clearly and strictly defined categories with clear, strong borders between types and distinct meaning. Typically, in the case of structured data, those categories are presented as columns or rows headers in data tables.

DOI: 10.4018/978-1-5225-4990-1.ch006
Information visualization (InfoVis), however, deals not only with data of clear, properly defined structure. This analytic and reasoning tool is also, if not mainly, used for data of latent structure or of a structure without strictly defined categories. One can find it useful in market analyzes, flow control or processes simulations. There is, of course, much more possibilities of applying InfoVis to analyzing real life processes. The ideas of use are limited by our imagination only.

One should be aware that Information visualization is so powerful tool, for it is a final step of automatic data analysis processes, called Machine Learning. InfoVis approximates, scales and presents in a comprehend manner results of ML data processing. Machine learning approach can help us to find and analyze any distinct and reliable relations, whether they are obvious or hidden, between data stored in different forms, not necessary in tables. As such, ML, can be and is successfully used also in Digital Humanities.

Any of those comprehend, very informative and nice looking graphical data or information presentation starts from raw data. Data which can, but not necessary are, cleared, unambiguous, and structured. The very preliminary step towards reliable graphical data, information, relations or trends presentation is text preprocessing. For properly prepared textual data we can apply ML approach and finally Information Visualization.

The goal of this chapter is to present preprocessing techniques and their role in modern and novel technique which is Information Visualization. We will discuss canonical approaches used in text preprocessing, common issues affecting the process and ways to prevent them. As substantial part of data can be, and already is, gathered from Web resources author presents chosen channels of gathering date for Digital Humanities (DH) purposes. The quality of data from different sources is also discussed. The content of this chapter is a result of few years of practical experience in Natural Language Processing (NLP) gained during realization of different projects and evaluation campaigns.

BACKGROUND

Big Data components and sources, such as data bases, Web- and Web2.0 sites, social media etc. are ever-growing sources of digital data. Despite of improving data transferring speed over the Internet, which make it possible to present in real time sounds, pictures and video, textual data are still the most popular in scale of document types. And text documents themselves are still the richest source of data to analyze and visualize.

From the human comprehension perspective the reception of text is a complex system, that employs not only reading abilities but also other cognitive predispositions. We, the humans, do always comprehend text in the context, which is defined by our education, knowledge and experience. Thus we can clearly detect and understand metaphors, and social-, political- and any other context and relations included non-explicitly in the text. Thanks to the ability of reading the latent information, so called “reading between the lines”, a raw text itself can be very informative for us, while, in contrast, not for computer systems.

There are, of course, similarities in perception of texts by humans and by computers, but in the very core of the process there are substantial differences. Computational text analysis are surface processes, this is recognizing of graphical representations of sounds and words. In order to adjust any meaning to recognized signs and their sequences computer needs a background guidance information. In order to make computational text processes useful and reliable we, humans, need to provide for computers additional data, relations descriptions, meaning linkage, disambiguation rules, computational routines, etc.