A primer on data- and text mining for content analysis

In the big data era, the importance of the ability to analyze varying forms of data is becoming ever more important. This is especially true for textual data, which is often claimed to account for around 80% of all data used for decision making. This presentation will introduce the audience to commonly used text mining approaches for content analysis of large amounts of textual data. Firstly, the special nature of text will be discussed and the relation between data and text mining will be presented. Then, commonly used preprocessing steps from the realm of Natural Language Processing (NLP) will be presented and motivated. The focus will be on preparing data for a word frequency-based approach (Bag of Words), but other preprocessing approaches will also be touched upon. After introducing the bag of words approach, the presentation will illustrate how common data mining operations can be applied to perform content based analysis, including document classification and sentiment analysis, document clustering and topic modeling, and word association analysis. Operations will be illustrated using examples form the literature.