Survey of text mining and analysis toolkits

Introduction

What is text mining? From my own perspective, text mining is a technology that uses NLP (natural language processing) to transform the unstructured text into structured data, which is more suitable for analysis and easier for computer to understand.

I found an article "Text mining for Bioinformatics using Biomedical Literature" Biomedical literature is a powerful source for nowadays biomedical knowledge. Why we need this biomedical literature? Because it is important that when a research group working on a specific topic, they need to know if other groups of people have done some research on the same topic or not. Searching for this specific topic manually requires a lot of efforts, also sounds impossible. And under this circumstance, we need text mining. In the article I found, it says that "Text mining aims at using IE methods to process text documents. " (IE stands for information extraction) , the main challenges text mining facing is that since it deals with unstructured text, it is hard to develop algorithms which can be used on unstructured text to obtain structured information. As it mentioned in the article "Biomedical literature is particularly challenging to text mining algorithms for several reasons. The writing style differs from other types of literature since it is more formal and complex. Furthermore, different types of documents have different styles, depending on whether the document is a journal paper, patent or clinical report [2]. Finally, there are a wide variety of terms that can be used, referring to genes, species, procedures, and techniques and, within each specific term, it is also common to have multiple spellings, abbreviations and database identifiers." It means that biomedical literature is hard to develop text mining algorithms for 3 main reasons:

1. Writing style is more formal and complex than other types of literature
2. It has many different types of literature.
3. A lot of terminology.

How to overcome these challenges makes biomedical text mining more interesting. And I am going to presents some ideas shows in the paper I found about text mining.

NLP

I think it is necessary to talk about NLP when we are introducing text mining. Although we have learned a lot about it during our class, I just want to combine them with the information I got from this paper and make a summery. NLP, also known as Natural Language Processing, it is used to determine sentences structure or sentiment, which can be used for text mining. The following concepts in NLP are used in text mining: Token: a sequence of characters with some meaning, Part-of-speech (POS): the lexical category of each token, Lemma and stem: the base form of a word, Sentence splitting: the NLP task consisting of identifying the sentence boundaries of a text, and Entity: a segment of text with relevance to a specific domain.

What is text mining tasks?

As I mentioned above, basically it is used to extracted useful information from unstructured text. Below are some specific tasks discussed in the paper that text mining needs to do:

1. Topic modeling: the classification of documents according to their topics or themes.
2. Named Entity Recognition (NER): consists of identifying entities that are mentioned in the text
3. Normalization: consists of matching each entity to an identifier belonging to a knowledge base that unequivocally represents its concept.
4. Relationship Extraction (RE): the identification of entities that participate in a relationship described in the text.

The paper also shows some approaches to accomplish the tasks above, which are:

1. Classic approaches: approaches based on statistics that can be calculated on a large corpus of documents
2. Rule-based methods: consist of defining a set of rules to extract the desired information.
3. Machine learning (ML) algorithms: are used for automatically learning various tasks.
4. Distant supervision (DS): a learning process which heuristically assigns labels to the data according to the information provided by a knowledge base.

Text mining toolkits

Above are just some basic idea about text mining, how can we deal with some specific content like biomedical literature. Well according to the paper, those basic text mining tools can be used as a starting point for specialized approach. As it mentioned in the article, "These general tools can be adapted to specific domains, either by using models trained with biomedical datasets or by developing pre- and post-processing rules developed for this type of text." We can develop many text mining toolkits based on this framework. As it mentioned in the article, Stanford CoreNLP which can be used in biomedical text mining to pre-process the data. Below are some different kinds of toolkits I searched on websites:

"NLTK , another NLP toolkit, was implemented as a Python library. This toolkit provides interfaces to various NLP resources, such as WordNet, tokenizers, stopwords lists, and datasets from community challenges. It is often used by developers who are getting started in text mining, due to its well-designed API, and to the availability of various online tutorials for this toolkit. More recently, another Python-based toolkit was released, spaCy3 , which is more focused on computational performance, using state-of-the-art algorithms."

"ClearTK is a text mining toolkit based on machine learning and the Apache Unstructured Information Management Architecture (UIMA). This framework provides interfaces to several machine learning libraries and feature extractors"

"GATE [30] is one of the few text mining toolkits which has features specially designed for biomedical text mining. This toolkit provides plugins for bioinformatics resources such as

Linked Life Data and other ontologies, and specialized biomedical NLP tools. Furthermore, a graphical user interface is available to visualize and edit the data and system architecture."

Conclusion

What I am showing in this tech review is some basic idea of text mining, and by showing this specific idea about Biomedical Text Mining, I want to demonstrate that text mining toolkits can evolve, which means that we can develop special tools for specific problems based on the characteristics. Text mining and data mining are popular these days, and in order to do data mining, we need to have a good text mining tool. Also, since I am interested in artificial intelligence. I think it will be useful to include deep learning in text mining. I believe in the future there will be more different kinds of text mining toolkits.