HW3 Answer 8

Apr 14, 2015

Text analysis problem (14 points)

Suppose that you work for a large health insurance company that is trying to update its educational programs for people that suffer from chronic kidney disease. The company has started crawling discussion forums at social media sites such as KidneySpaces.org to find clues about what information kidney disease patients need.

a. Describe how you would use text analysis to discover 10-20 of the main topics that are discussed, and their relative frequencies. Your topic descriptions must be something that senior management will understand easily. **[7 points]**

Answer:

There are at least two different answers for this question.

One could use topic models, for example, LDA, to discover the main topics. LDA has the advantage that a single document can be about several topics (e.g., sports and finance), which (perhaps) allows it to form better topics than other topic-finding methods. For each document, topic modeling will provide the distribution of topics associated with the document (e.g., 60% sports and 40% finance for a document about the business models of major sports teams), so it is possible to determine the relative frequency of different topics in the corpus. The disadvantage is that the topics are described as language models, which will be difficult for senior management to understand. Instead, you can manually inspect and name each topic, and then show senior management your names and example documents from each topic.

One could also use clustering, for example, k-means, or hierarchical clustering, to identify the main groupings of documents. Assume that each cluster represents a topic. The cluster sizes indicate the relative frequency of each topic. You can manually inspect and name each cluster, and then show senior management your names and example documents from each cluster.

You could also categorize the documents into MeSH categories, then display to senior management the most frequent MeSH codes and their descriptions from the MeSH hierarchy. This wasn't an answer that I expected, and it has some weaknesses (e.g., trying to map social media content onto the categories of medical researchers), but it is an acceptable answer.

Grading Notes:

- 5 points for the general approach, 1 points for relative frequencies, and 1 points for topic descriptions.
- -1 points for irrelevant or wrong information (e.g., NB).
- Half credit for frequency analysis answers if they produced 10-20 topics. The top 10-20 bigrams doesn't cut it.

- I didn't accept descriptions that consist of terms and frequencies. I don't believe that senior management will find that meaningful. However, I did accept word clouds.

b. Senior management liked your initial analysis. Now they want more information. Describe how you would provide a more detailed analysis of the issues or concepts associated with each topic, and their relative frequency or importance. **[7 points]**

Answer:
There are several possible answers to this question.

Frequency analysis would work here. Each word, phrase, or entity can be treated as a unique concept. The importance of a concept to a topic can be determined by its frequency in documents about the topic (although this is a little weak), or by the PMI or phi-square between the concept and the topic. This provides a lot of detail.

If you used topic modeling for the answer above, you could do hierarchical topic modeling to provide more detail. For example, you could apply LDA to the entire corpus to get the high-level topics, and then apply LDA again to the documents that are associated with each individual topic to obtain more fine-grained topics. The importance of each fine-grained topic is determined as described above for the high-level topics.

Likewise, if you used clustering for the answer above, you could do hierarchical clustering to provide more detail.

Grading notes:
- 5 points for the general approach, 2 points for relative frequencies or importance.
- -1 points for irrelevant or wrong information (e.g., NB).
- People interpreted "more detailed analysis" in many different ways. I was liberal about accepting them, as long as they made sense.
- Sentiment analysis towards topics was -2, because I asked for a more detailed analysis of the issues or concepts associated with each topic. Sentiment analysis is more information about each topic, but doesn't address the issues or concepts associated with the topic.
- -2 for answers that talk about the location of a topic in a document.