

Text Data Mining For Business Decisions

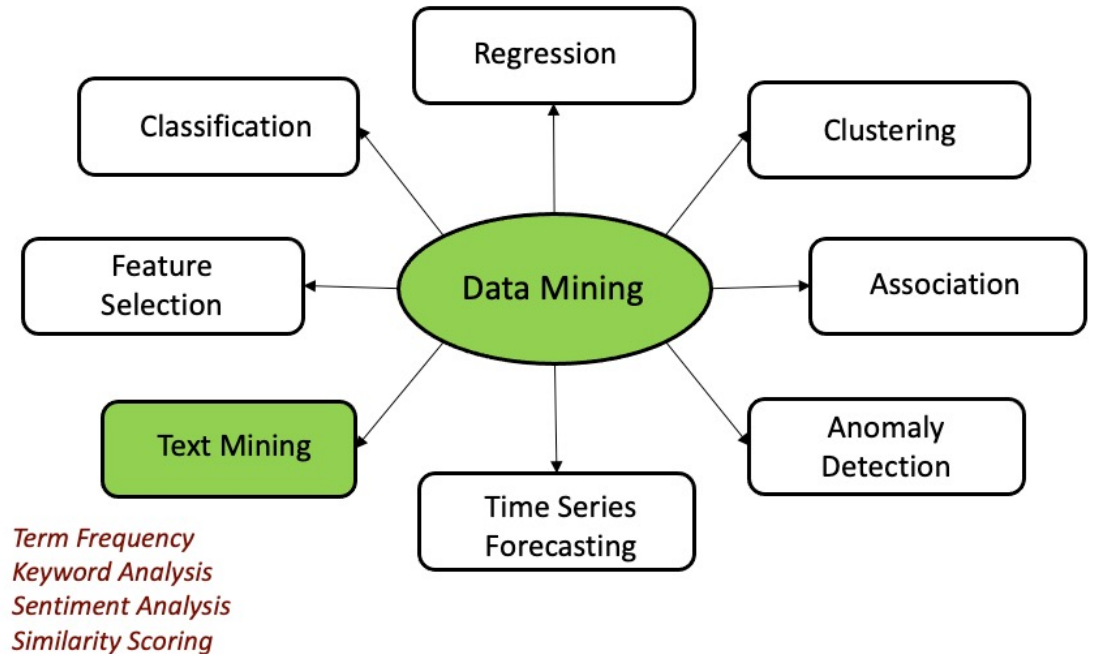
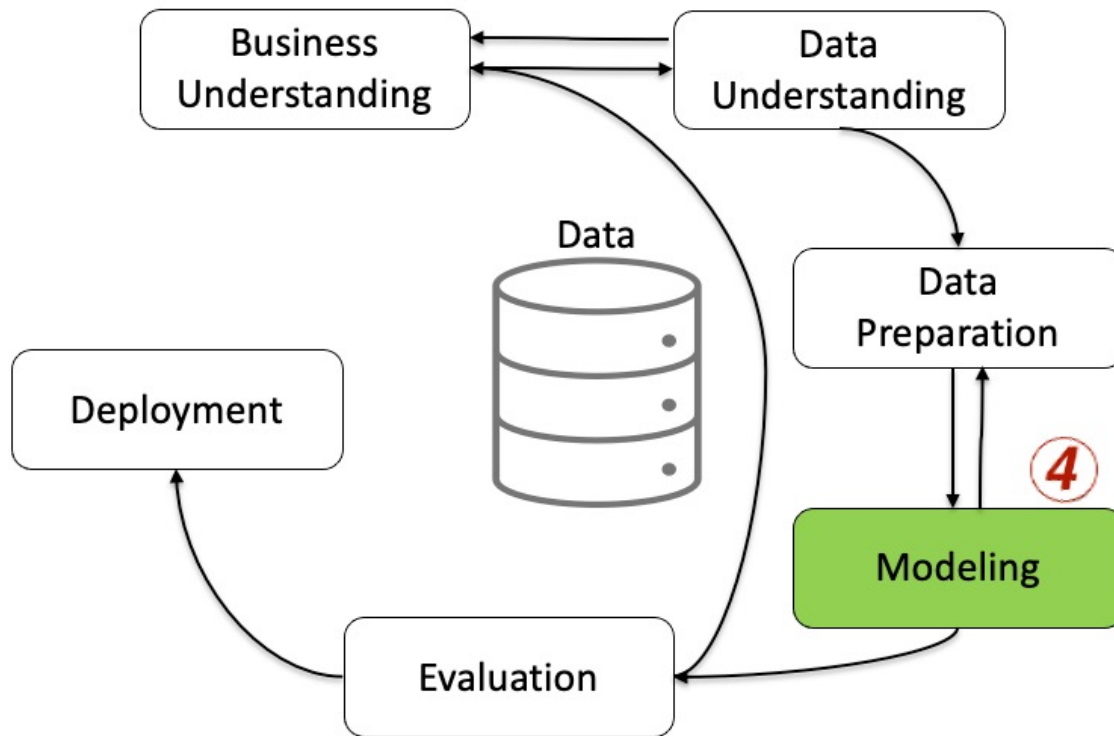
Module 5

Keyword Analysis

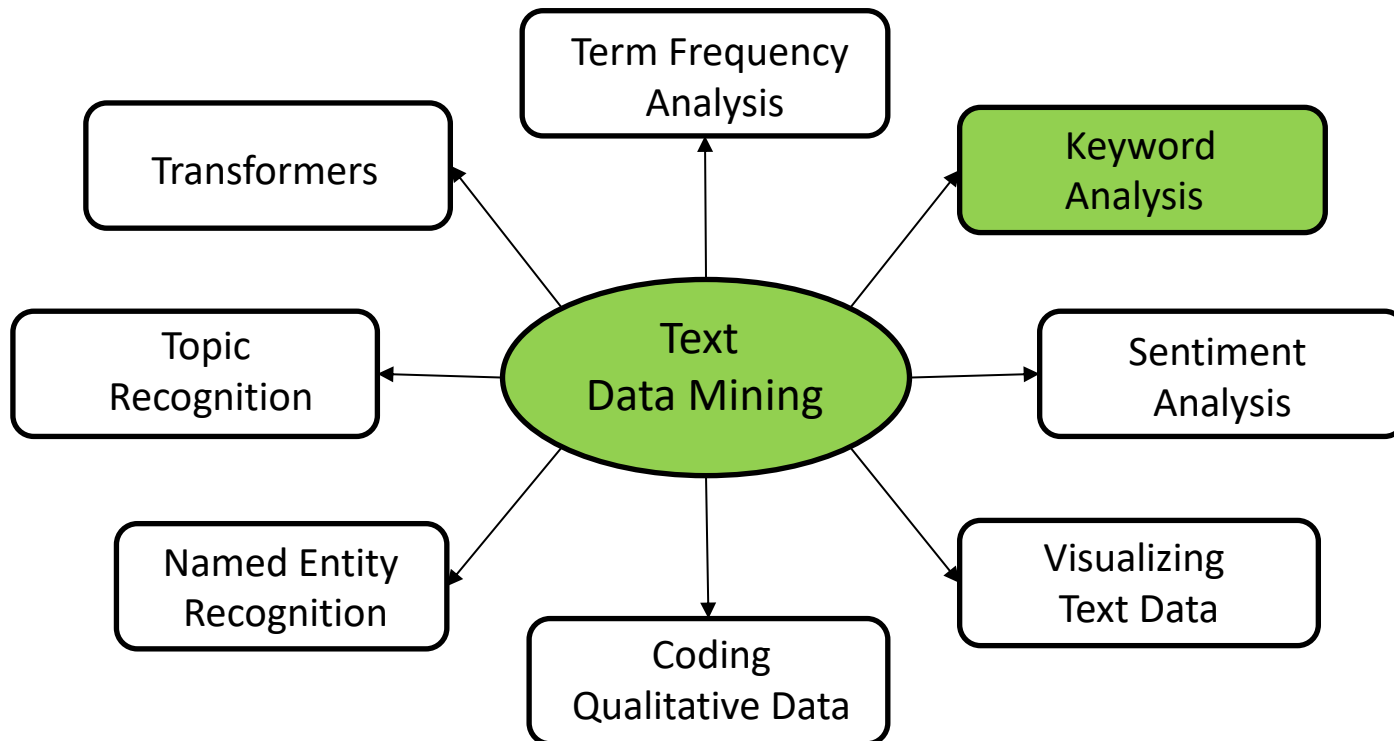
How do we discover and compare the frequency of keywords among documents?



Data Mining- Continuing Model-Making



Keyword Analysis is one of the Basic Text Mining Semantic Algorithms



How do we count *important* words?

- Counting words is the most basic technique in text analysis.
 - The starting point for most investigations - *lexical analysis*
- To measure the frequency of *important* words
 - We tabulate their frequency of all terms
 - We *select* certain terms because they likely have some greater significance in the context of our work.
 - We stress interpretation -what is important to us as humans - *semantic analysis*
- We count the most *meaningful* words within a context to determine which are more important and which are less important.

Keyword Analysis – What is it?

- *Keyword analysis* is also known as keyword extraction or keyword detection.
- Keyword analysis is a text analysis technique that extracts the most frequent and most important words and expressions from text data.
- It helps summarize the content of texts and recognize the main topics discussed.
- It is most powerfully used to optimize how search engines find and index webpages from your company's website.
- It is an integral part of Search Engine Optimization (SEO).

Why do we count keywords for business purposes?

- It helps summarize the content of texts and recognize the main topics discussed.
 - It is most powerfully used to optimize how search engines find and index webpages from your company's website. It is an integral part of Search Engine Optimization (SEO).
- With keyword analysis, you can find keywords from all types of text data: documents, business reports, social media postings, online reviews, and news reports.
- Suppose you want to analyze your product's many online reviews on your favorite e-commerce site, like Amazon.
 - Keyword extraction helps you sift through all the text data comprising the reviews and obtain the most important and frequent words that best describe the reviews very quickly.
 - Then you can see what your customers are mentioning most often, saving you the work of examining all the reviews manually.
- Let's say you want to know what percentage of your customers are price sensitive. Looking at social media postings or product feedback data, you can do a keyword search across all postings and determine how frequent "price" as a keyword is.

The Process

- We can use the techniques we studied in the last session to compute the word frequencies in a document to find keywords.
- It is just a matter of determining how many of the most frequent words are deemed to be enough to fully characterize a document by its keywords.
- And which are important in the context of the analysis
- That decision is made by the analyst. The top five? The top 10? The top 25? And which ones?
- You determine that by inspection.

How to do it – Example A

- In this example, we try to match a resume to a set of job descriptions.
- Suppose you are looking for work and select a small number of jobs that seem interesting.
- Can you narrow the search and prioritize the assignments using keyword analysis?
- Essentially, you want to answer the question

Which jobs that am I interested in applying to match my resume most closely?

- We extract the most frequent words from a resume and use the most appropriate ones for a keyword search through the list of jobs.

Team Case Study 5 - Keywords

- Use a browser to go to the each of the websites.
- Your team should scrape all the program web pages and create text file for each one.
- Work with your team to decide three major keywords that you think people would use to find your program
- Using Voyant generate the TF analysis to for all the web page texts
- Decide with your teams which of the frequent words are keywords
- Generate a table for the frequency of each keyword in each program web page
- Create a bar graph of the frequency directions for each keyword by program
- Analyze the MASY program with respect to all the other programs. What do you see?
- What would you recommend that the program director do to increase the visibility of her program