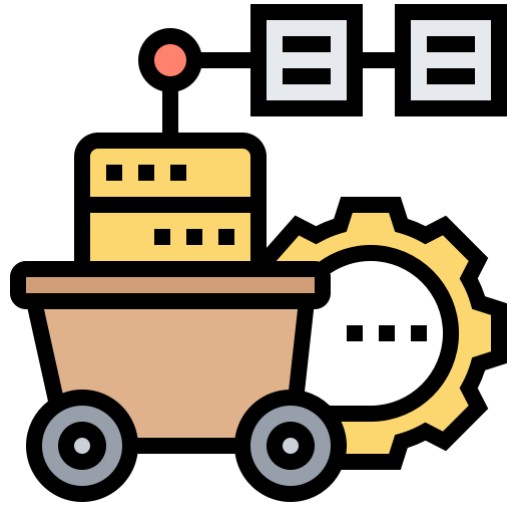
Text Data Mining
For Business Decisions



Module 1B

Introduction to Text Data Mining

What is Text Data Mining?

Agenda

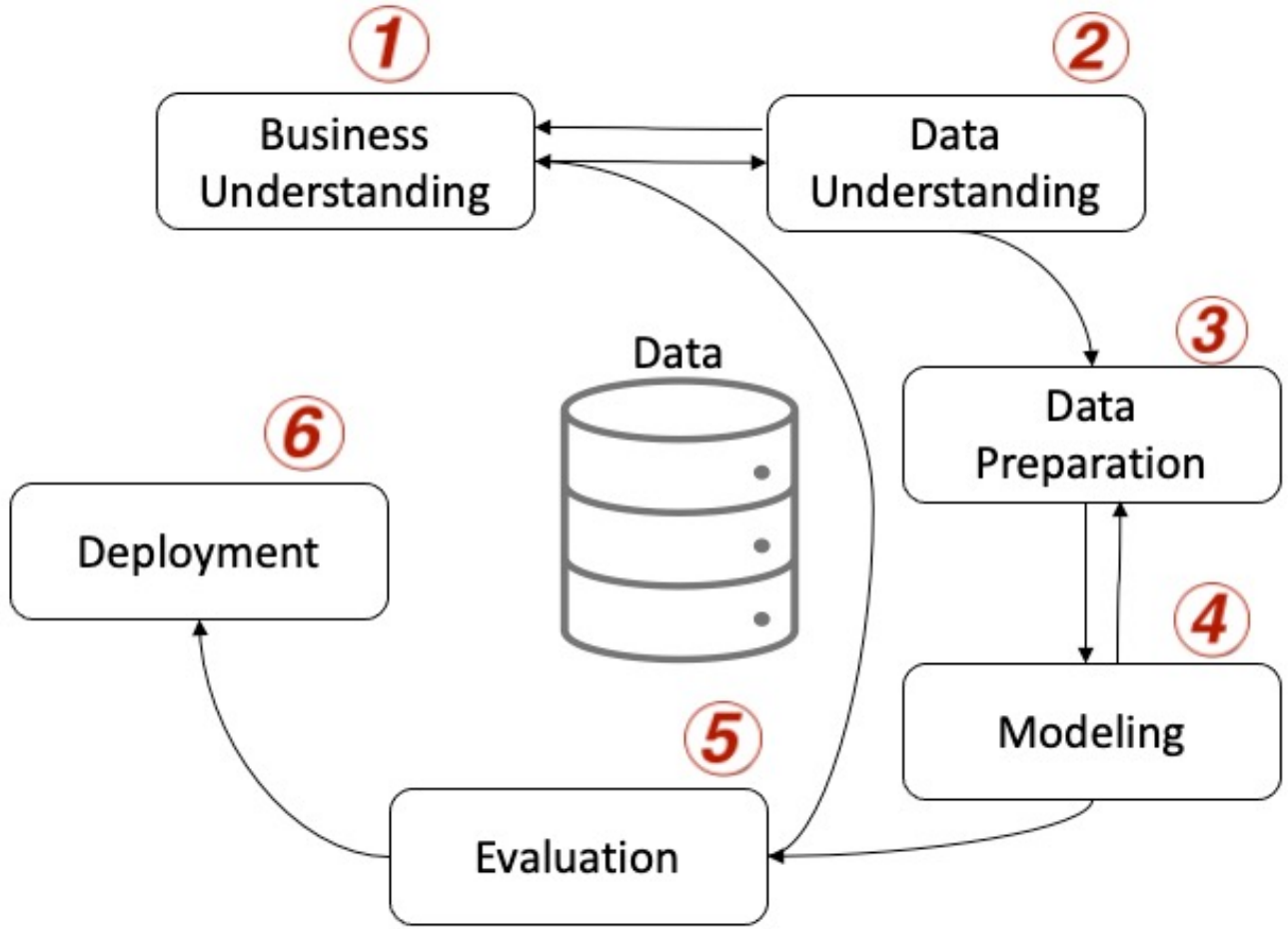
- What is Data Mining?
- What is Text Mining?
- What are some applications of Text Data Mining?
- Assignments

DATA MINING
WHAT IS IT?

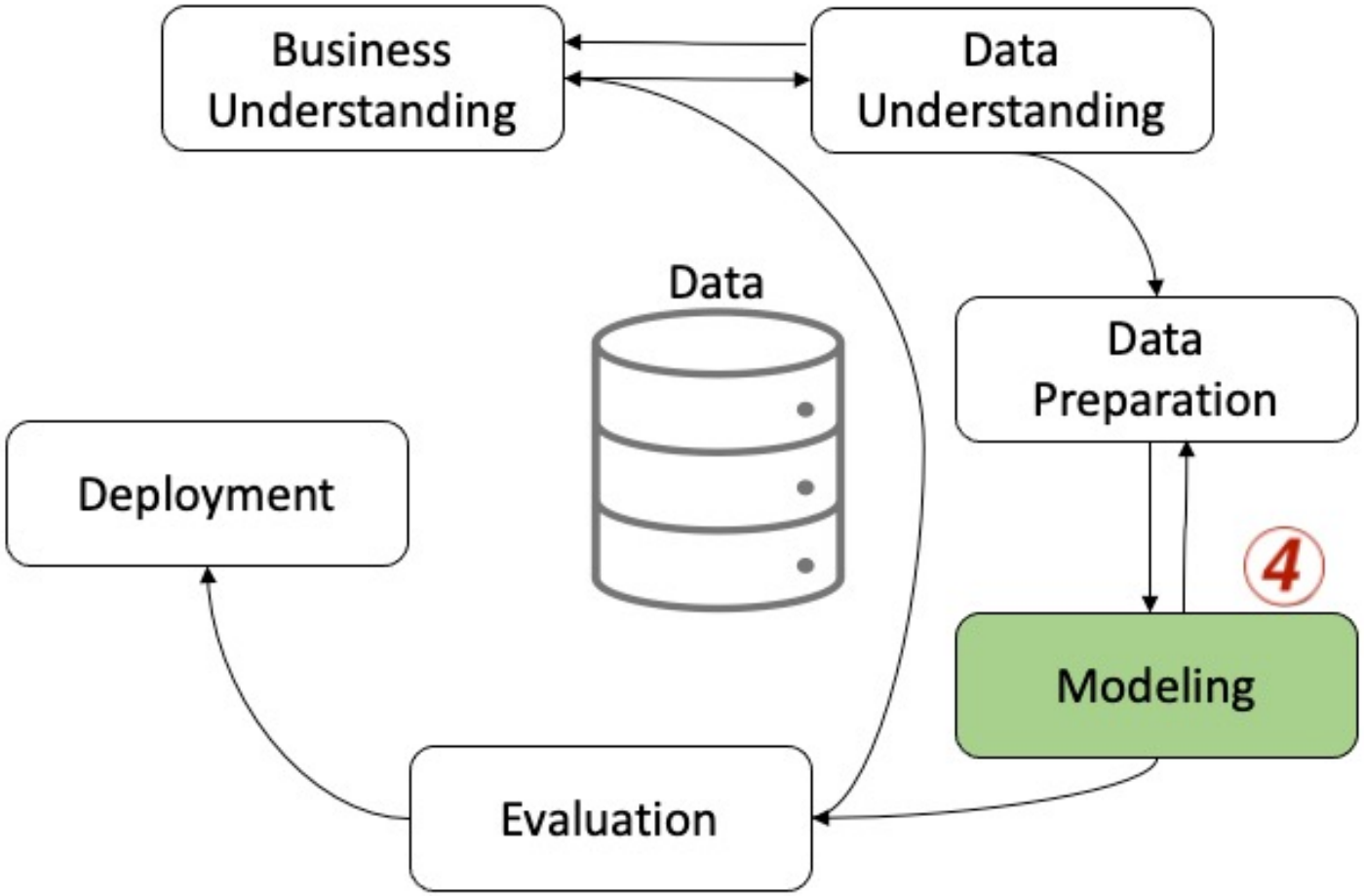
What is Data Mining?

- Data mining, in simple terms, is *finding useful patterns in the data*.
- Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or relationships in the data to make important decisions

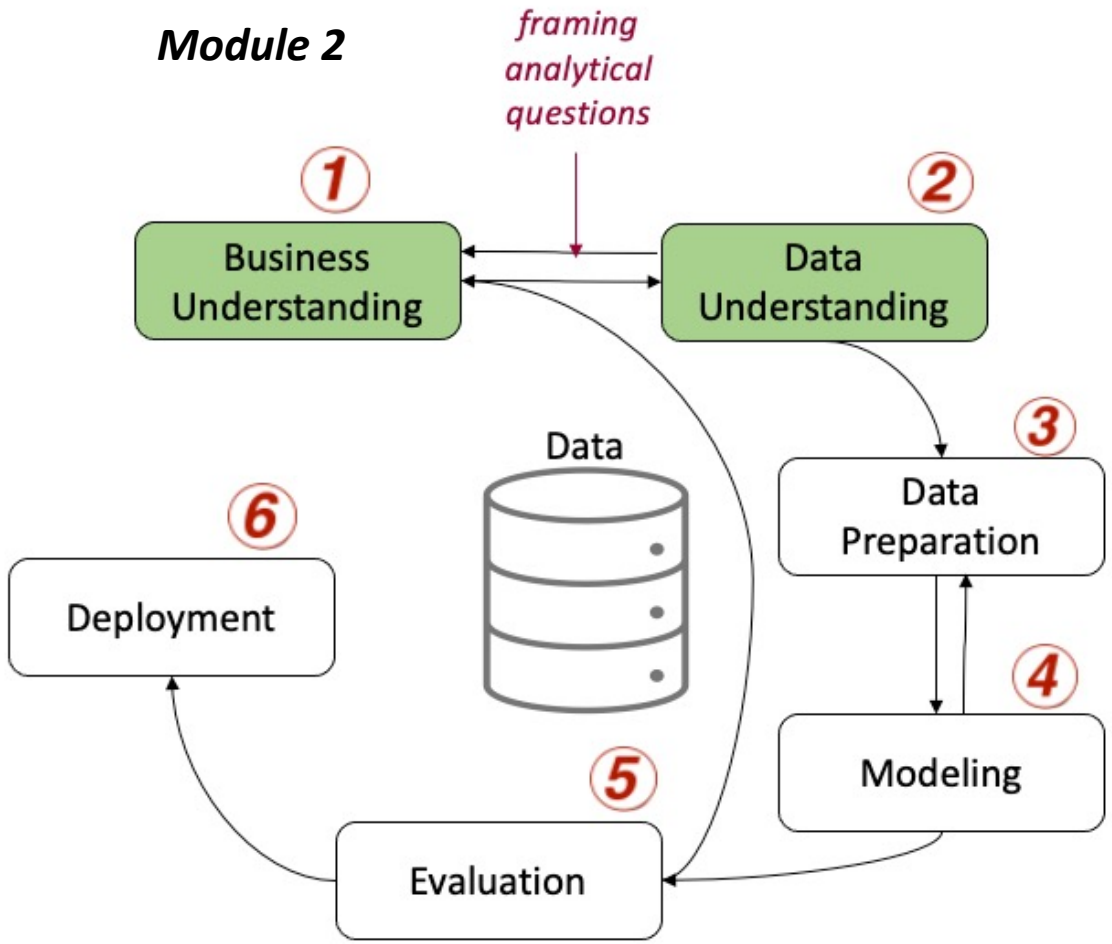
The Data Mining Process – CRISP-DM

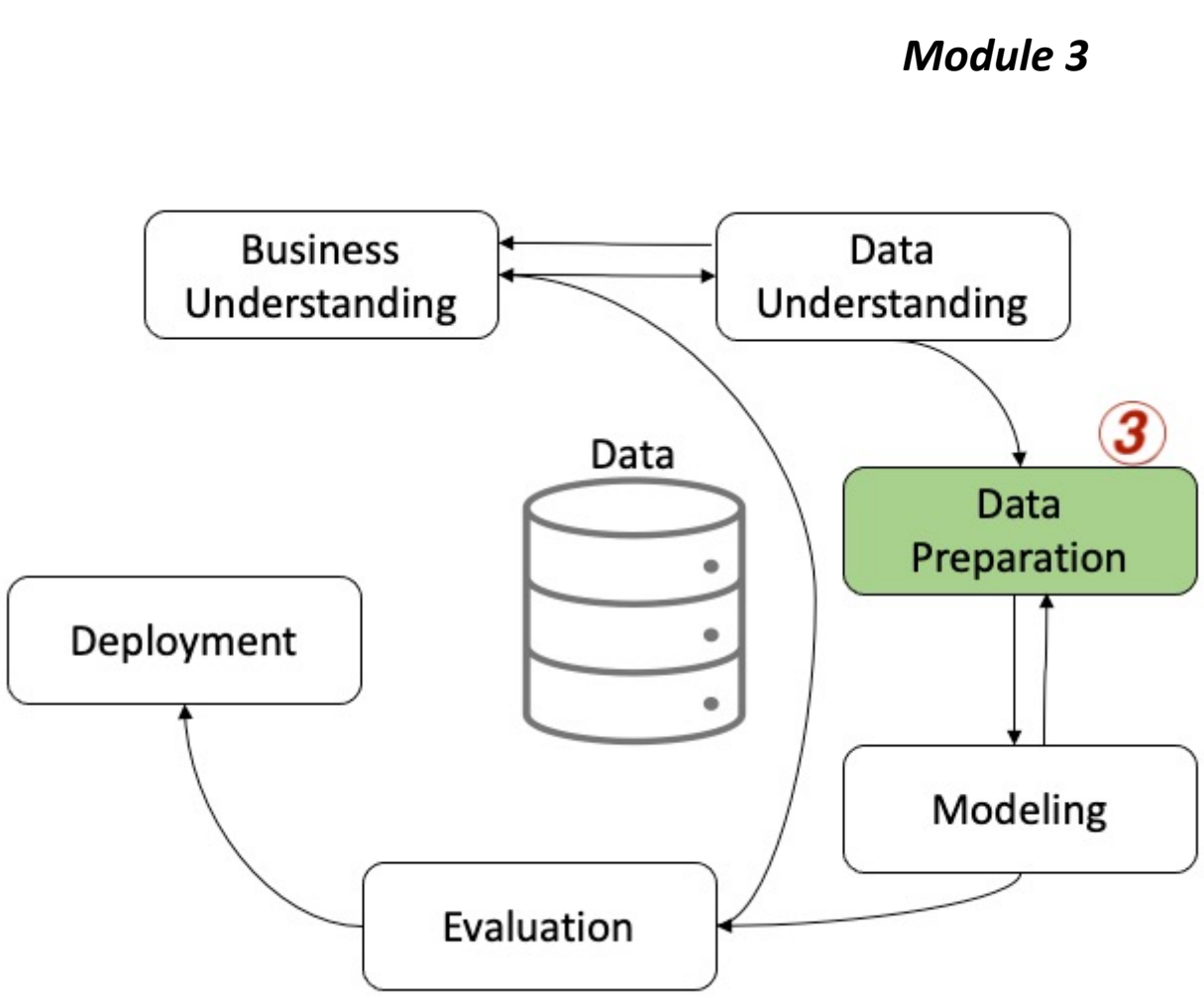


Where do we apply text data mining techniques?

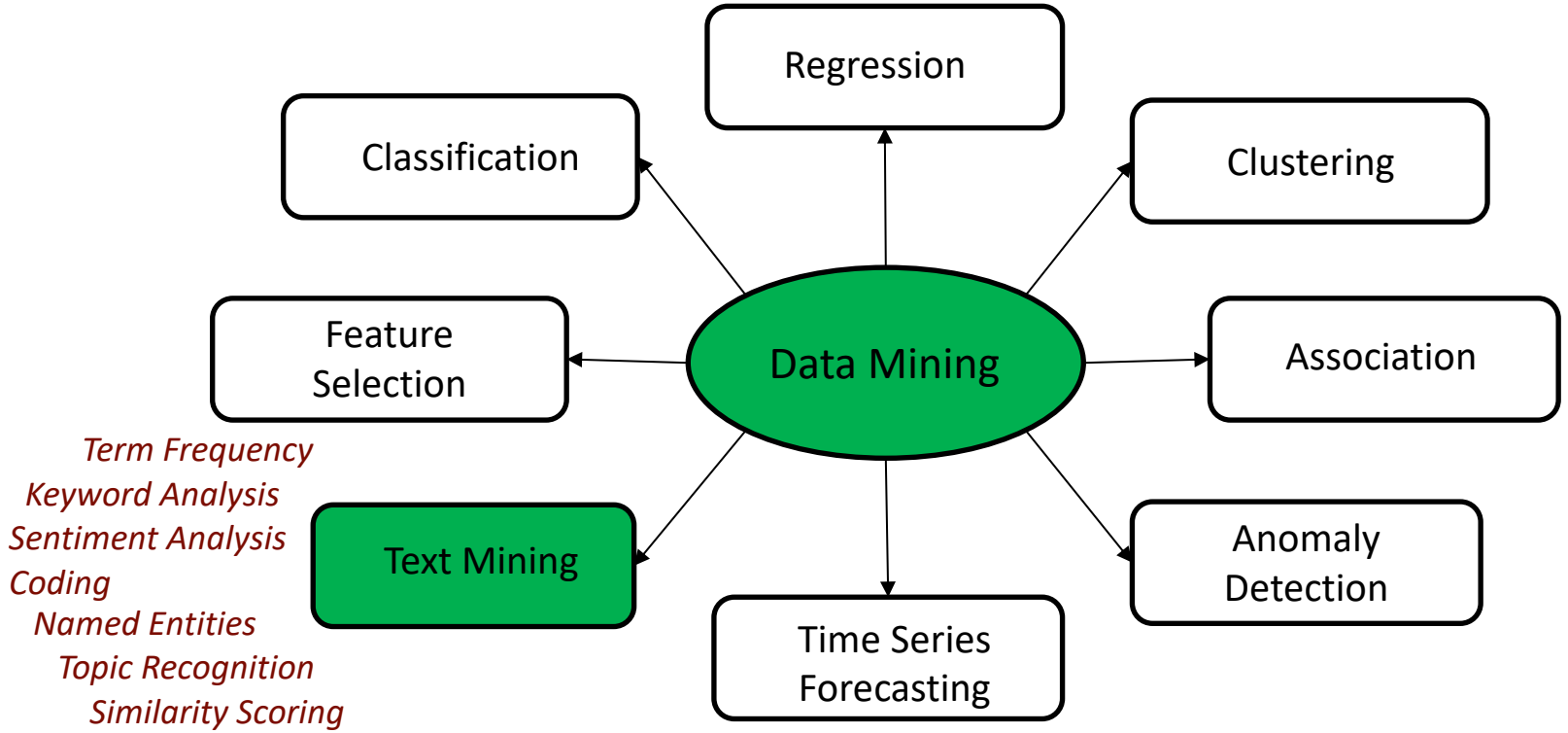


What do we have to do to get ready for text data mining?





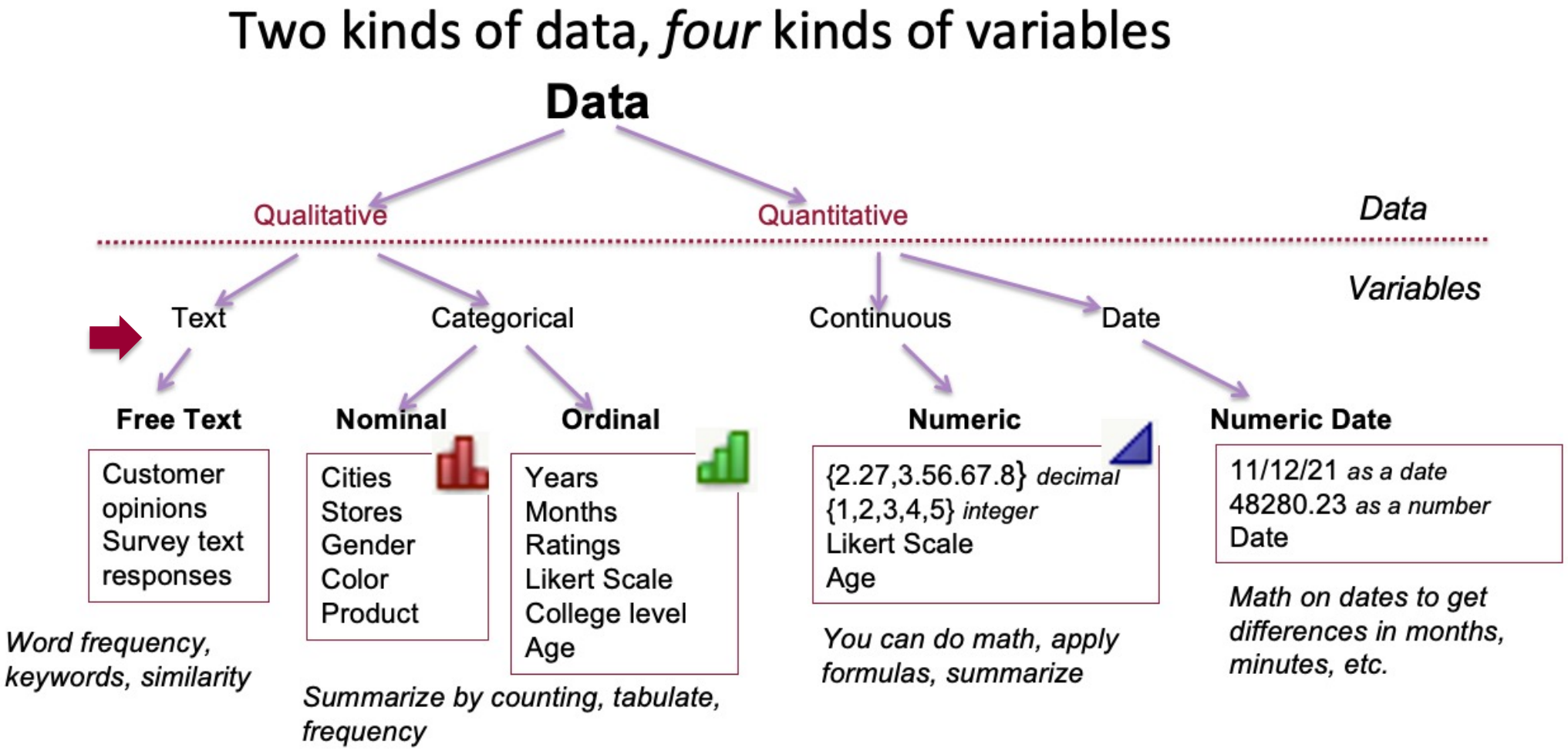
Data Mining Tasks and Text Mining



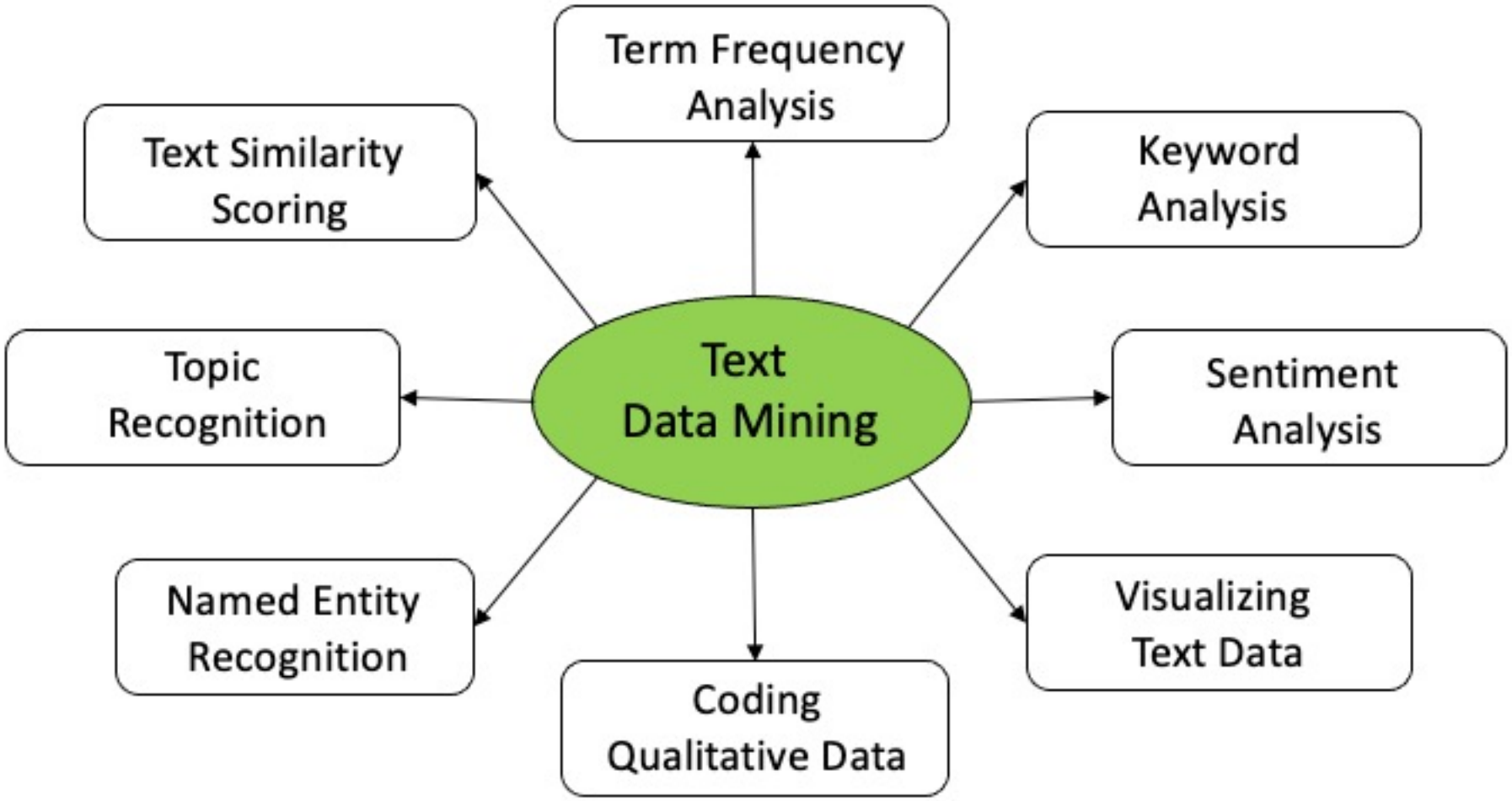
What Data Mining Is Not

- Descriptive statistics
- Exploratory visualization
- Dimensional slicing
- Hypothesis testing
- Queries

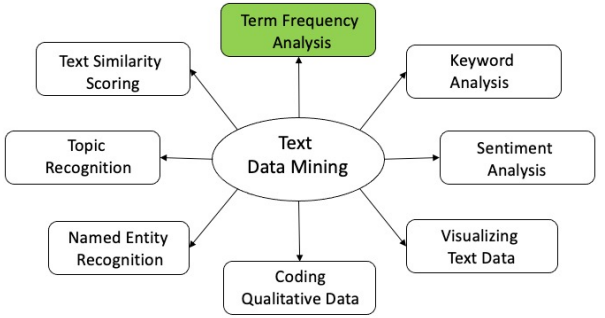
Where does text data fit in as a variable?



What is Text Data Mining?

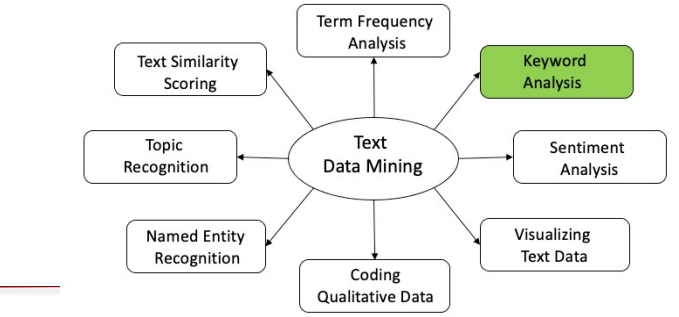


Term Frequency Analysis



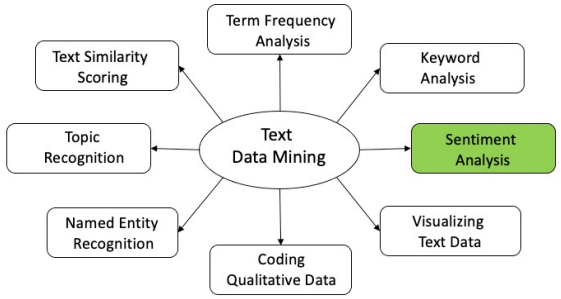
- Term frequency (TF) means **how often a term occurs in a document**. Term frequency is commonly used in Text Mining, Machine Learning, and Information Retrieval tasks.
- For example: Term frequency is the number of times a given term or query appears within a search index. Term frequency is a key component for determining the relevance of a given document for a particular query, and is an essential piece of the widely used TF-IDF relevancy algorithm.

Keyword Analysis



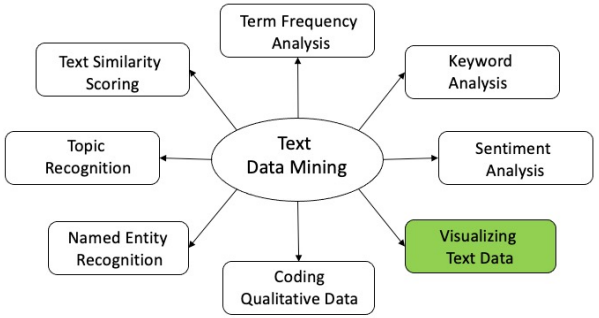
- Keyword analysis is the process of analyzing the keywords or search phrases that bring visitors to your website through organic and paid search. As such, keyword analysis is the starting point and cornerstone of search marketing campaigns.
- By understanding what queries qualified visitors to your website type into search engines, search marketers can better customize their content and landing pages to drive more traffic and increase conversion rates. For this reason, keyword analysis is an important skill for both SEO and PPC experts.

Sentiment Analysis



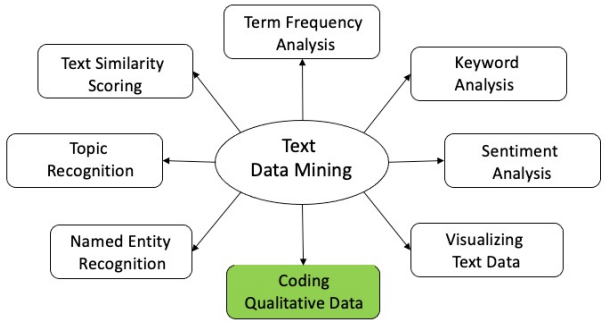
- Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.
- Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Visualizing Text Data



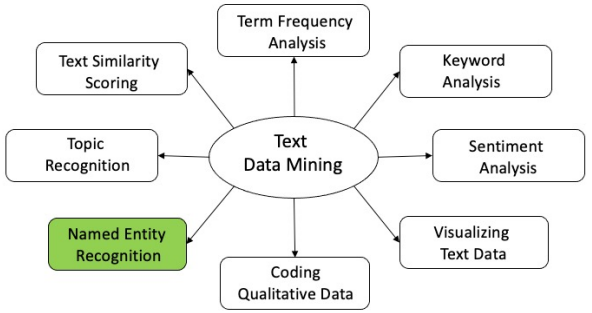
- Text visualization is the **technique of using graphs, charts, or word clouds** to showcase written data in a visual manner. This provides quick insight into the most relevant keywords in a text, summarizes content, and reveals trends and patterns across documents.

Coding Qualitative Data



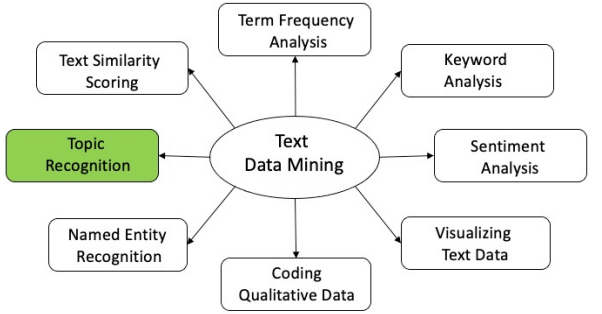
- Coding is the process of labeling and organizing your qualitative data to identify different themes and the relationships between them.
- When coding customer feedback, you assign labels to words or phrases that represent important (and recurring) themes in each response. These labels can be words, phrases, or numbers; we recommend using words or short phrases, since they're easier to remember, skim, and organize.
- Coding qualitative research to find common themes and concepts is part of thematic analysis, which is part of qualitative data analysis. Thematic analysis extracts themes from text by analyzing the word and sentence structure.

Named Entity Recognition



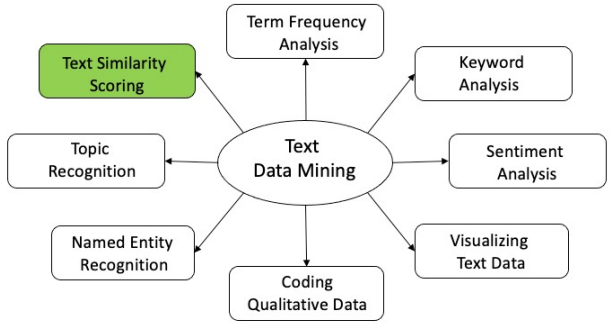
- Named-entity recognition (NER) (also known as (named) entity identification, entity chunking, and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

Topic Recognition



- Topic analysis is a Natural Language Processing (NLP) technique that allows us to automatically extract meaning from text by identifying recurrent themes or topics.
- A topic model is a type of statistical model for discovering the abstract “topics” that occur in a collection of documents.
- Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts.
- Topic models provide a simple way to analyze large volumes of unlabeled text. A “topic” consists of a cluster of words that frequently occur together.

Text Similarity Scoring



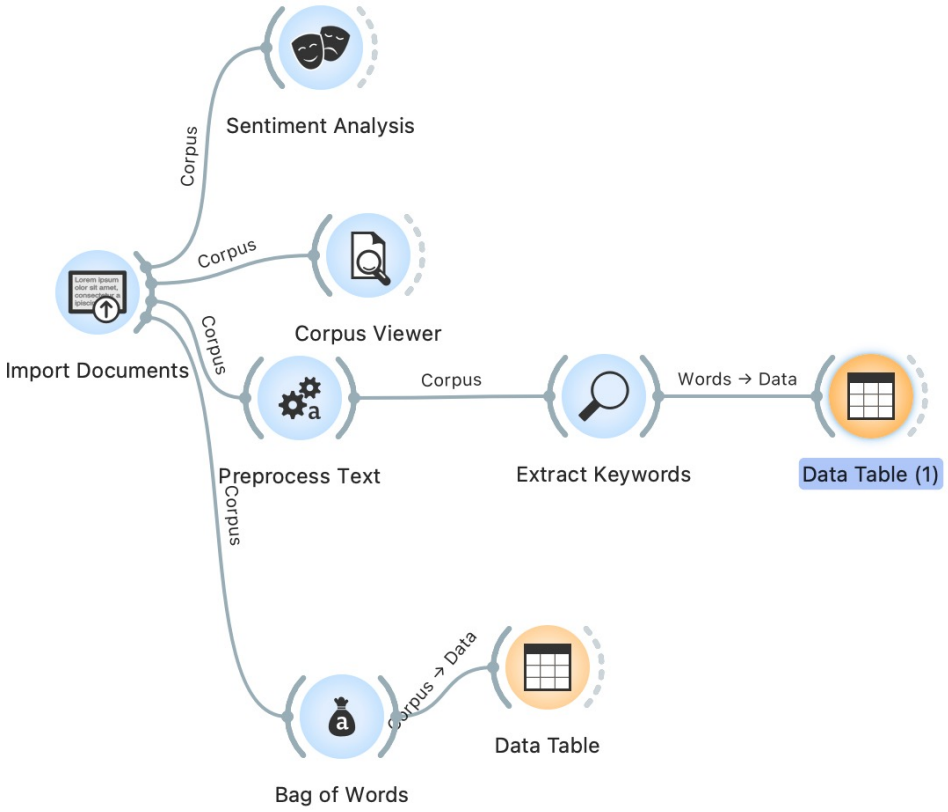
- What is text similarity?
- Text similarity helps us determine how 'close' two pieces of text are both in surface closeness [lexical similarity] and meaning [semantic similarity].

Visual Programming

- Visual programming is a type of programming language that lets humans describe processes using illustration.
- Whereas a typical text-based programming language makes the programmer think like a computer, a visual programming language lets the programmer describe the process in terms that make sense to humans.

Visual Programming vs. Command Line Code

Python Visual Programming



Python Code

Solver.py

```python
import math


class Solver:

    def demo(self, a, b, c):
        d = b ** 2 - 4 * a * c
        if d > 0:
            disc: float = math.sqrt(d)
            root1 = (-b + disc) / (2 * a)
            root2 = (-b - disc) / (2 * a)
            return root1, root2
        elif d == 0:
            return -b / (2 * a)
        else:
            return "This equation has no roots"


if __name__ == '__main__':
    solver = Solver()

    while True:
        a = int(input("a: "))
        b = int(input("b: "))
        c = int(input("c: "))
        result = solver.demo(a, b, c)
        print(result)
```

FORTINO Global Education

Transformers

- A transformer is a deep learning model that adopts the mechanism of attention, differentially weighing the significance of each part of the input data. It is used primarily in the field of natural language processing (NLP) and in computer vision (CV).

What are some text analytic applications?

- Word frequency analysis of web pages to determine keywords
- Word/phrase location on text
- Sentiment analysis
- Analysis of a “corpus” - a set of static texts:
  - Contracts
  - Emails
- Analysis of streaming text
  - Twitter feeds

Techniques

Results

Disciplines

FIGURE I.1

Text mining draws upon contributions of many text analytical components and knowledge input from many external disciplines (shown in the 'bar with upward arrow' at the bottom), which result in directional decisions affecting external results (shown by the arrow point to the right arrow at the top).