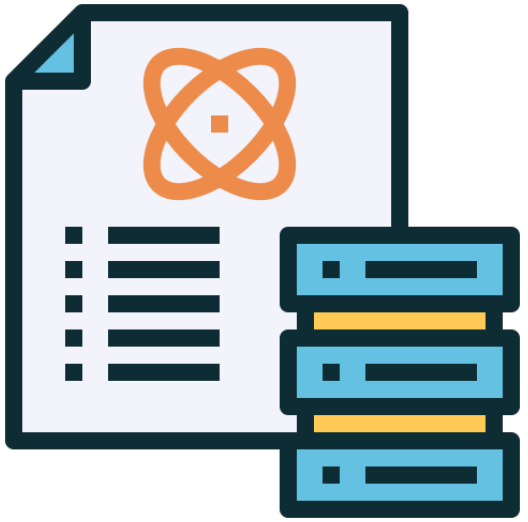
Text Data Mining
For Business Decisions

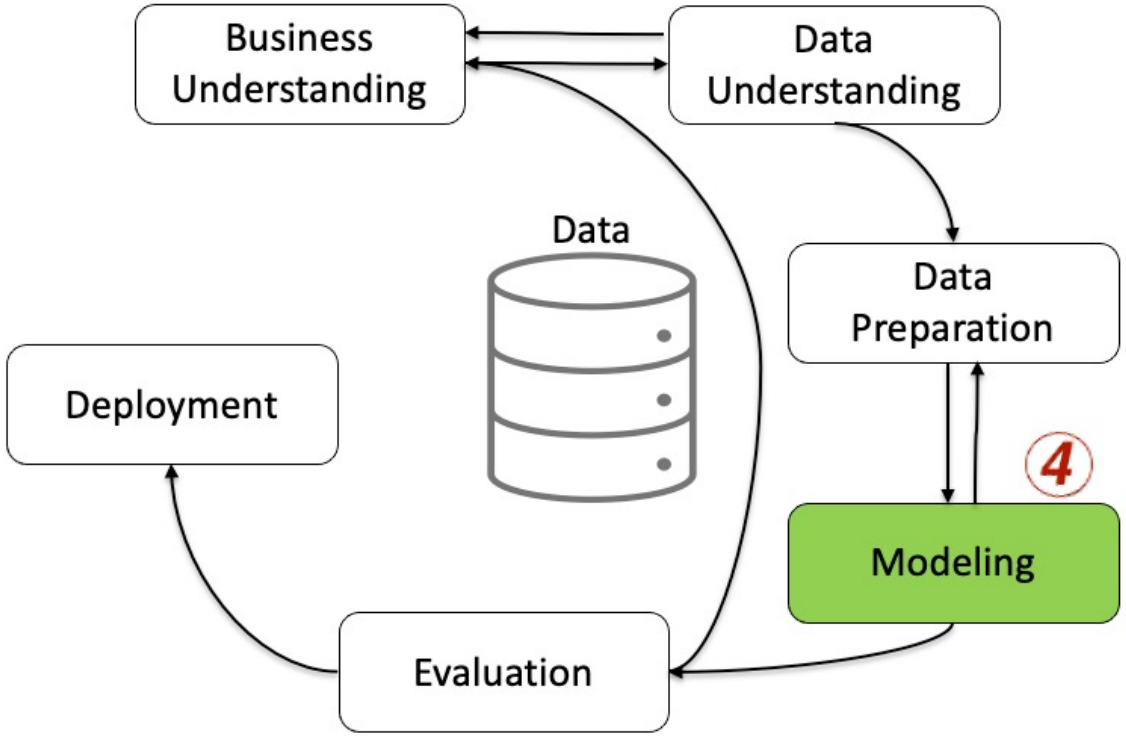
Module 11

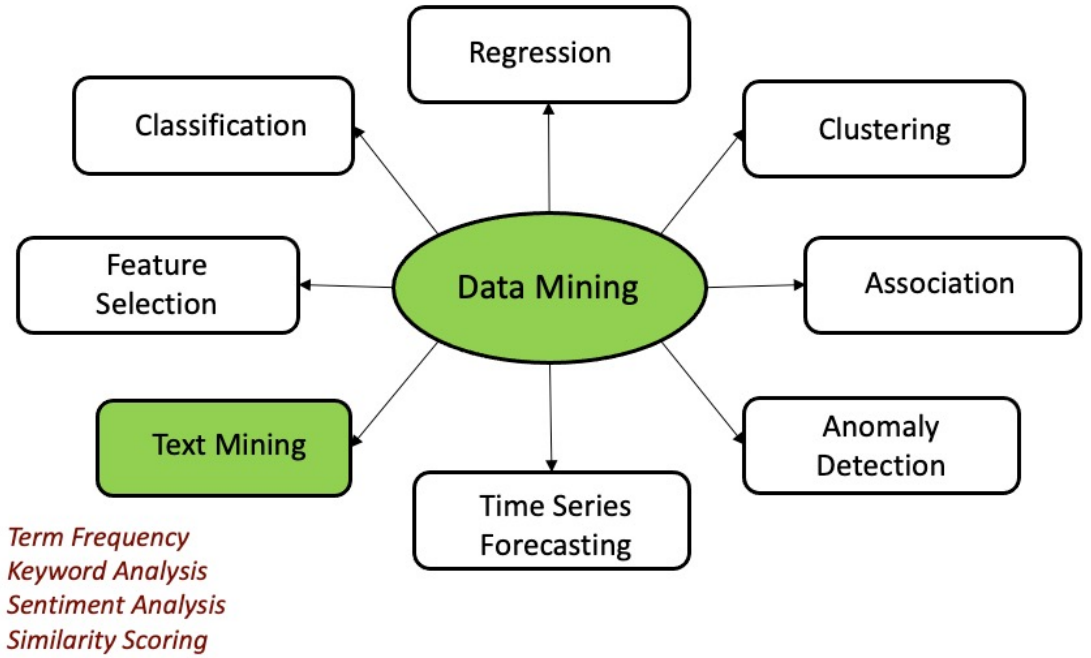
Topic Recognition

How do we extract the topics in a document?

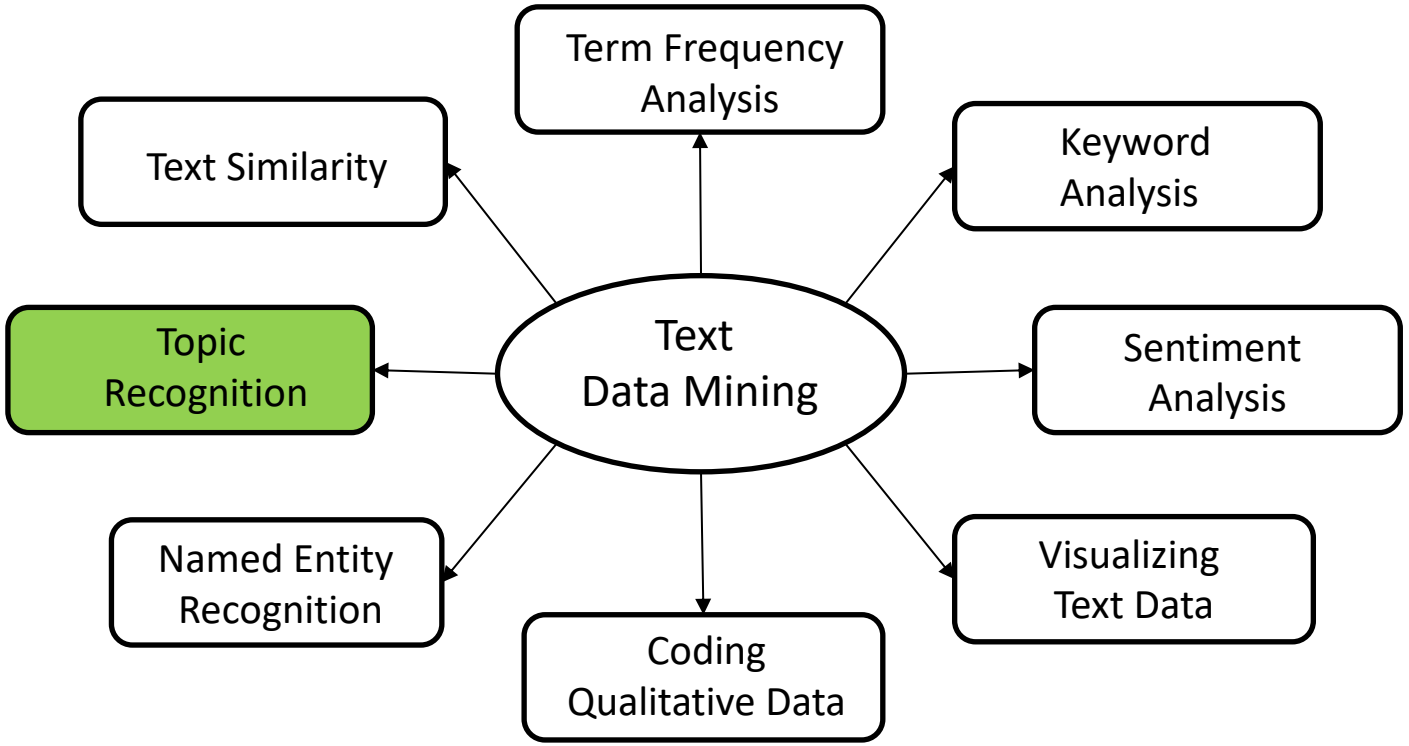


Data Mining- Continuing with Model-Making





Topic Extraction in Text Documents



What is Topic Extraction?

- Now, rather than extract only proper names, locations, and temporal elements, we also want to recognize whole topics in the texts.
- We simplistically did this work by term frequency analysis, keyword analysis and qualitative data coding.
- We now analyze the text with sophisticated machine learning algorithms to extract meaning through latent keywords and then relating those words to topics.
  - It helps us classify texts by comparing them to each other and grouping them by emerging topics.

What is Information Retrieval?

- *Information retrieval* (IR) is a field of study dealing with the representation, storage, organization of, and access to documents.
  - The documents may be books, reports, pictures, videos, or webpages.
  - The point of an IR system is to provide the analysis with easy access to documents containing the desired information.
  - An excellent example of an information retrieval system is the Google search engine.
- The difference between an information retrieval system and a data retrieval system like an RDBMS is that the IR system input is unstructured.
  - In contrast, data retrieval (a database management system or DBMS) deals with structured data with well-defined data constructs.
- For data extraction, or querying, extracting data from an RDBMS produces precise results or no results when there is no match, whereas querying an IR system produces multiple ranked results with partial matching.

What is Topic Retrieval?

- We identify and extract topics from unstructured text in this session - the unstructured text parts of documents that are written in prose.
- Most textual documents are characterized by three kinds of information about the document:
  - (1) its metadata,
    - very interesting, but the metadata elements may be collected as traditional numerical, categorical, and time variables, which can be analyzed with traditional means
  - (2) its formatting,
    - not too interesting to us
  - (3) its content.
    - The most interesting
  - Example - movies

What is Topic Retrieval?

- We identify and extract topics from unstructured text in this session - the unstructured text parts of documents that are written in prose.
- Most textual documents are characterized by three kinds of information about the document:
  - (1) its metadata,
    - very interesting, but the metadata elements may be collected as traditional numerical, categorical, and time variables, which can be analyzed with traditional means
  - (2) its formatting,
    - not too interesting to us
  - (3) its content.
    - The most interesting
  - Example - movies

What is a topic?

- Topic modeling is a method for finding and tracing clusters of words (called *topics*) in large bodies of texts. Topic modeling is very popular with digital humanities scholars, partly because it offers some meaningful improvements to simple word-frequency counts and partly because of the availability of some relatively easy-to-use tools for topic modeling.
- It can eb used as a text classification system

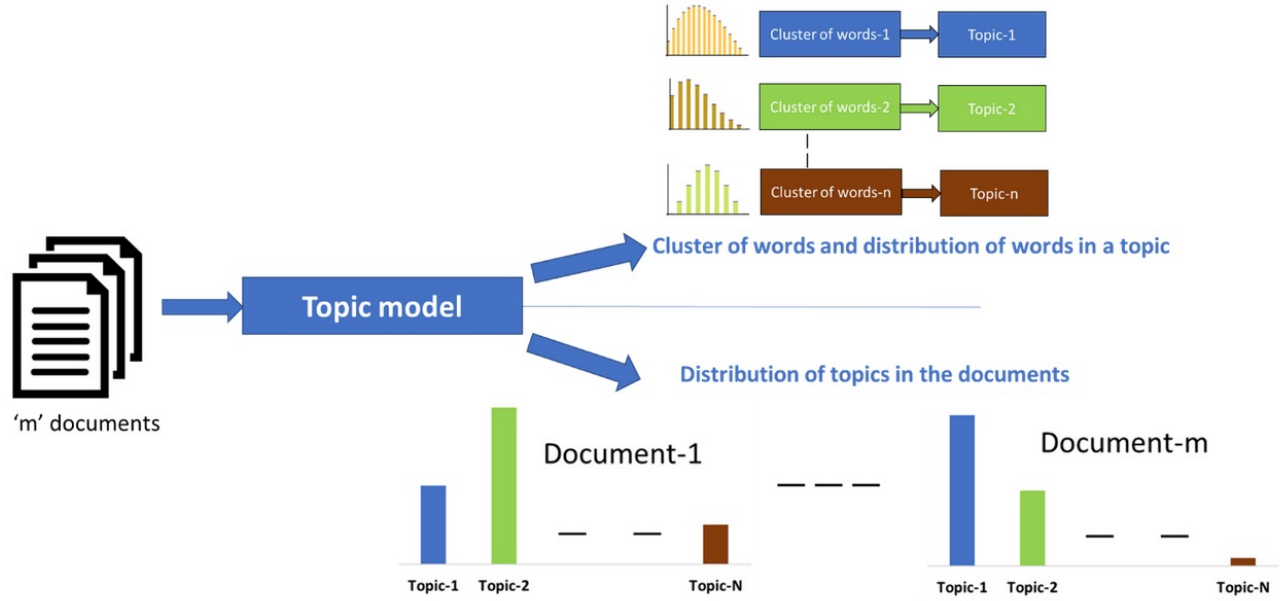
Tools for Topic Extraction

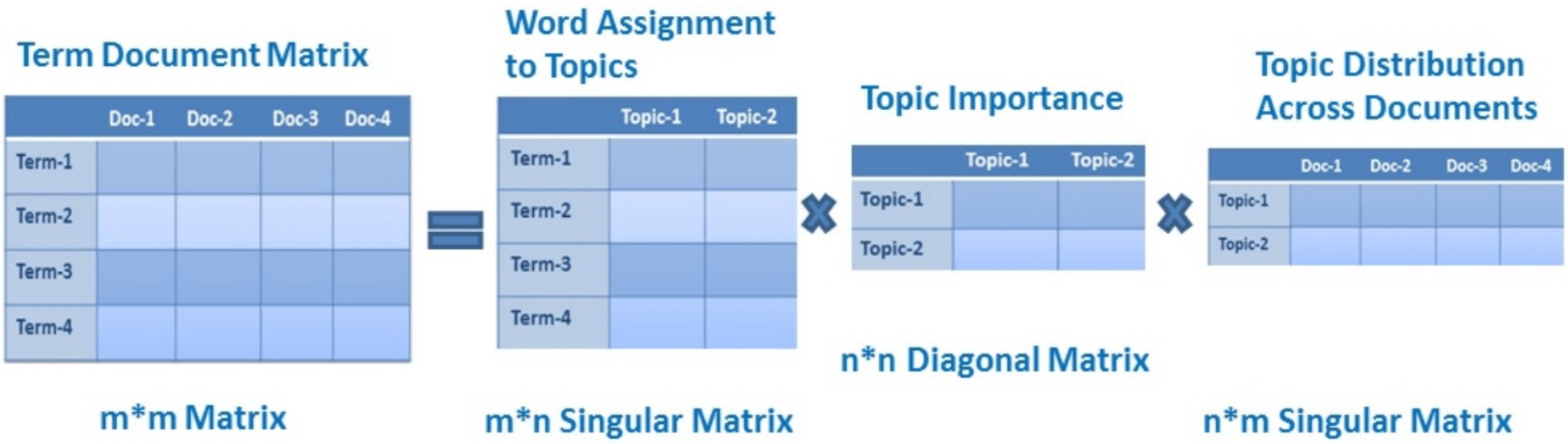
- Topic modeling is a method for finding and tracing clusters of words (called *topics*) in large bodies of texts.
- The most popular natural language processing tool for topic extraction is called MALLET, a package of Java code.
  - Topic Modeling Tool, which implements MALLET in a graphical user interface (GUI), meaning you can plug files in and receive output without entering a line of code. *(See Chapter 16 of the Text Analytics book)*
- We will use a reduced version of it embedded in Voyant, a tool called *Topics*
- Both use a machine learning algorithm called *Latent Semantic Analysis*

Assumptions of all topic modeling algorithms

- Each document consists of more than one topics, and
- Each topic consists of a collection of words.
- The semantics of our document is actually being governed by some hidden, or “latent,” variables that we are not observing directly after seeing the textual material
- The core idea of LSA is to take a matrix of documents and terms and try to decompose it into separate two matrices –
  - A document-topic matrix
  - A topic-term matrix.

Latent Semantic Analysis





FORTINO Global Education