

Prepared by Dr. Wesam Ahmed

IT393
Business Intelligence

Lecture 8



Predictive Analytics : Text, Web Analytics

Lecture 8

IT393

Business Intelligence

Prepared by
Dr. Wesam Ahmed

01

Learning Objectives:

- 1.1 Describe text analytics and understand the need for text mining
- 1.2 Differentiate among text analytics, text mining, and data mining
- 1.3 Understand the different application areas for text mining
- 1.4 Python libraries for data mining and text mining

02

Text mining

Unstructured data (versus structured data). Structured data is usually organized into records with simple data values (categorical, ordinal, and continuous variables) and stored in databases. In contrast, unstructured data does not have a predetermined format and is stored in the form of textual documents. In essence, the structured data is for the computers to process while the unstructured data is for humans to process and understand. In the recent years, the unstructured data generated over the *Internet of Things* (Web, sensor networks, radio-frequency identification [RFID]-enabled supply chain systems, surveillance networks, etc.) is increasing at an exponential pace, and there is no indication of it slowing down. This changing nature of data is forcing organizations to make text and Web analytics a critical part of their business intelligence/analytics infrastructure. The information age that we are living in is characterized by the rapid growth in the amount of data and information collected, stored, and made available in electronic format.

03

Text mining

A vast majority of **business data are stored in text documents that are virtually unstructured**. The unstructured data is doubling in size every 18 months. Because knowledge is power in today's business world , businesses that effectively and efficiently tap into their text data sources will have the necessary knowledge to make better decisions. **This is where the need for text analytics and text mining fits into the big picture of today's businesses.**

Even though the overarching goal for **both text analytics and text mining** is to **turn unstructured textual data into actionable information** through the application of natural language processing (NLP) and analytics, their definitions are somewhat different, at least to some experts in the field.

04

Text mining

According to them, **text analytics** is a broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms), as well as information extraction, data mining, and Web mining, whereas **text mining** is primarily focused on **discovering new and useful knowledge from the textual data sources**. Figure 8.1 illustrates the relationships between text analytics and text mining along with other related application areas. Based on this definition of text analytics and text mining, one could simply formulate the difference between the two as follows:

Text Analytics = Information Retrieval + Information Extraction + Data Mining + Web Mining,

or simply

Text Analytics = Information Retrieval + Text Mining

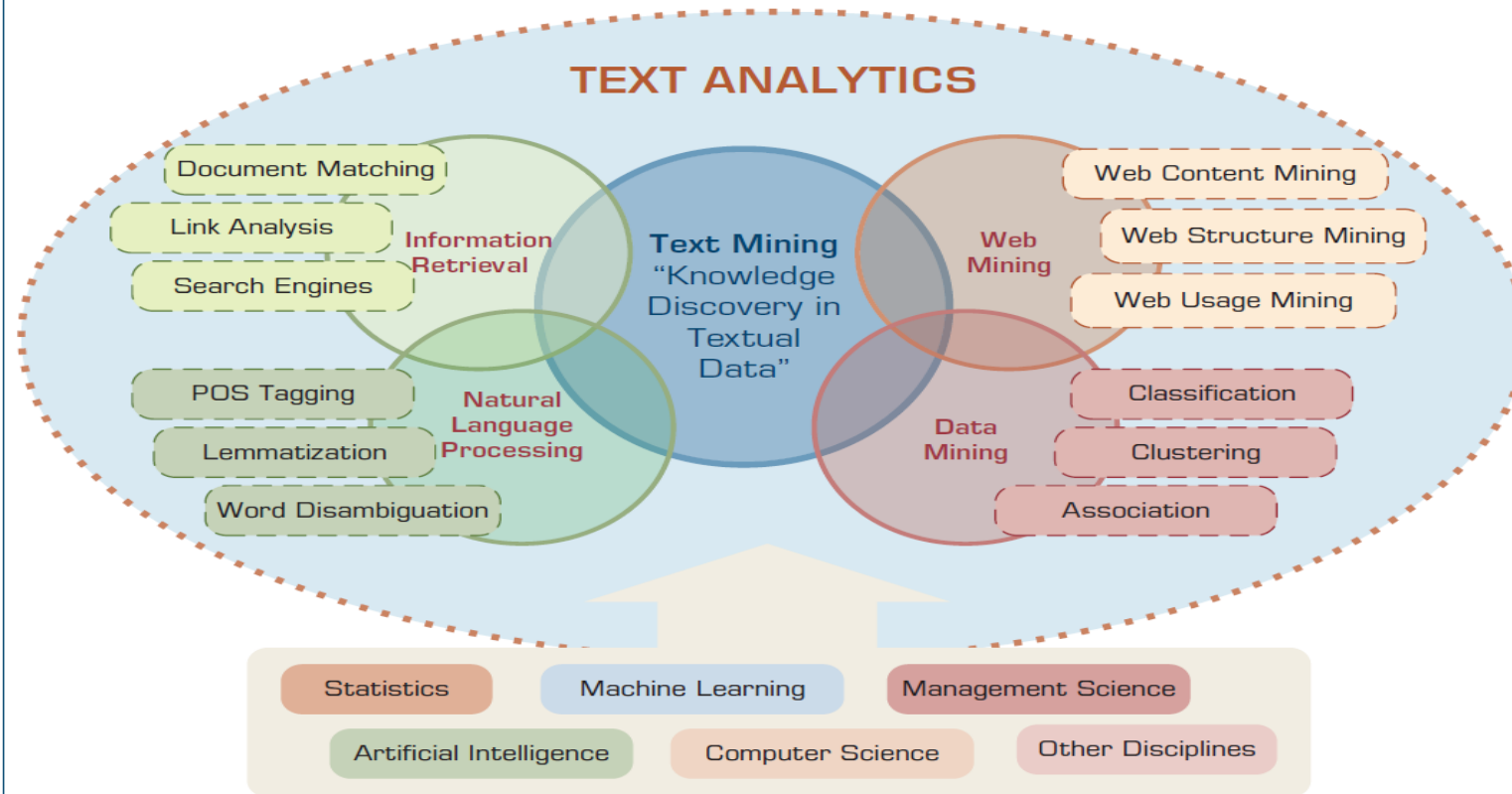


Figure 8.1
Text Analytics,
Related Application
Areas, and
Enabling
Disciplines

06

Text mining

Compared to text mining, **text analytics is a relatively new term**. With the recent emphasis on analytics, as has been the case in many **other related technical application areas (e.g., consumer analytics, complete analytics, visual analytics, social analytics)**, the field of text has also wanted to get on the analytics bandwagon. Although the term **text analytics** is more commonly used in a business application context, **text mining** is frequently used in academic research circles. **Text mining (also known as text data mining or knowledge discovery in textual databases)** is the semiautomated process of extracting patterns (useful information and knowledge) from large amounts of unstructured data sources. **Text mining is the same as data mining in that it has the same purpose and uses the same processes**, but with text mining **the input to the process is a collection of unstructured (or less structured) data files** such as Word documents, PDF files, text excerpts, XML files, and so on.

07

Text mining

In essence, text mining can be thought of as a process (with two main steps) that starts with imposing structure on the text-based data sources followed by extracting relevant information and knowledge from this structured text-based data using data mining techniques and tools. The benefits of text mining are obvious in the areas where very large amounts of textual data are being generated, such as law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), and marketing (customer comments).

08

Text mining

Following are among the most popular application areas of text mining:

- **Information extraction.** Identification of key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching.
- **Topic tracking.** Based on a user profile and documents that a user views, text mining can predict other documents of interest to the user.
- **Summarization.** Summarizing a document to save time on the part of the reader.
- **Question answering.** Finding the best answer to a given question through knowledge-driven pattern matching.
- **Clustering.** Grouping similar documents without having a predefined set of categories.

09

Text mining

Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics. It studies the problem of “understanding” the natural human language, with the view of converting depictions of human language (such as textual documents) into more formal representations (in the form of numeric and symbolic data) that are easier for computer programs to manipulate.

The goal of NLP is to move beyond syntax-driven text manipulation (which is often called “word counting”) to a true understanding and processing of natural language that considers grammatical and semantic constraints as well as the context. The following are just a few of the challenges commonly associated with the implementation of NLP: **Part-of-speech tagging**. It is difficult to mark up terms in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, or adverbs) because the part of speech depends not only on the definition of the term but also on the context within which it is used.

10

Text mining

The correct role is determined by *context*, not the dictionary

- "The **race** cars" → "race" is an adjective.
- "I **race** cars" → "race" is a verb.
- "I watched the **race**" → "race" is a noun.

Part-of-speech (POS)
tagging example

Text segmentation. Some written languages, such as Chinese, Japanese, and Thai, do not have single-word boundaries. In these instances, the text-parsing task requires the identification of word boundaries, which is often a difficult task.

Word sense disambiguation. Many words have more than one meaning. Selecting the meaning that makes the most sense can only be accomplished by taking into account the context within which the word is used.

11

Text mining

One prominent area where **the benefits of NLP and WordNet** are already being harvested is in customer relationship management (CRM). Broadly speaking, **the goal of CRM is to maximize customer value by better understanding and effectively responding to their actual and perceived needs**. An important area of CRM, where NLP is making a significant impact, is sentiment analysis. **Sentiment analysis** is a technique used to detect favorable and unfavorable opinions toward specific products and services using a large number of textual data sources (customer feedback in the form of Web postings).

NLP has successfully been applied to a variety of domains for a wide range of tasks via computer programs to automatically process natural human language that previously could only be done by humans. Following are among the most popular of these tasks:

- **Machine translation.** The automatic translation of one human language to another.

12

Text mining

- **Foreign language reading.** A computer program that assists a nonnative language speaker to read a foreign language with correct pronunciation and accents on different parts of the words.
- **Foreign language writing.** A computer program that assists a nonnative language user in writing in a foreign language.
- **Speech recognition.** Converts spoken words to machine-readable input. Given a sound clip of a person speaking, the system produces a text dictation.

NLP enables the **extraction of features from unstructured text** so that a wide variety of **data mining techniques can be used to extract knowledge** (novel and useful patterns and relationships) from it. In that sense, **text mining is a combination of NLP and data mining.**

13

Text mining

Text Mining Applications:

Marketing Applications : Text mining can be used to increase cross-selling and up-selling by analyzing the unstructured data generated by call centers. Text generated by **call center notes as well as transcriptions of voice conversations with customers** can be analyzed by text mining algorithms to extract novel, actionable information about customers' perceptions toward a company's products and services. In addition, blogs, user reviews of products at independent Web sites, and discussion board postings are a gold mine of customer sentiments. This rich collection of information, once properly analyzed, **can be used to increase satisfaction and the overall lifetime value of the customer .**

14

Text mining

Biomedical Applications:

Experimental techniques such as DNA microarray analysis, serial analysis of gene expression (SAGE), and mass spectrometry proteomics, among others, are generating large amounts of data related to genes and proteins. As in any other experimental approach, it is necessary to analyze this vast amount of data in the context of previously known information about the biological entities under study. The literature is a particularly valuable source of information for experiment validation and interpretation. Therefore, the development of automated text mining tools to assist in such interpretation is one of the main challenges in current bioinformatics research. Knowing the location of a protein within a cell can help to elucidate its role in biological processes and to determine its potential as a drug target.

15

Text mining process

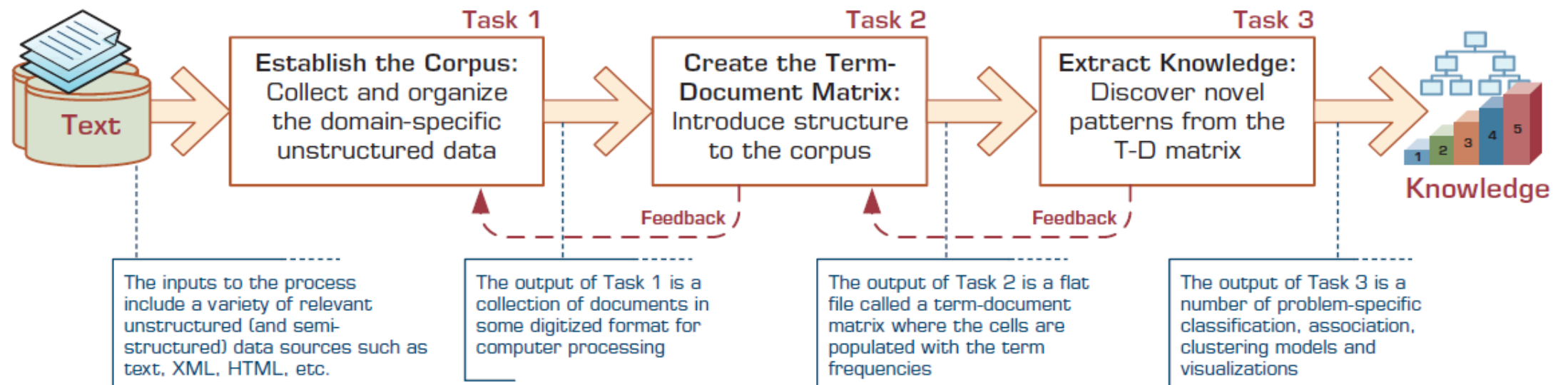


Fig 8.2 The Three-Step/Task Text Mining Process

16

Text mining process

The primary purpose of text mining (within the context of knowledge discovery) is to process unstructured (textual) data (along with structured data, if relevant to the problem being addressed and available) to **extract meaningful and actionable patterns for better decision making**. The text mining process can be **broken down into three consecutive tasks, each of which has specific inputs to generate certain outputs**.

Task 1: Establish the Corpus

The main purpose of the first task activity is to collect all the documents related to the context (domain of interest) being studied. This collection may include **textual documents, XML files, e-mails, Web pages, and short notes**. In addition to the readily available textual data, **voice recordings may also be transcribed using speech-recognition algorithms and made a part of the text collection**.

17

Text mining process

Once collected, the text documents are transformed and organized in a manner such that they are all in the same representational form (e.g., ASCII text files) for computer processing. The organization of the documents can be as simple as a collection of digitized text excerpts stored in a file folder or it can be a list of links to a collection of Web pages in a specific domain.

Task 2: Create the Term–Document Matrix

In this task, the digitized and organized documents (the corpus) are used to create the term–document matrix (TDM). In the TDM, rows represent the documents and columns represent the terms. The relationships between the terms and documents are characterized by indices (i.e., a relational measure that can be as simple as the number of occurrences of the term in respective documents).

Terms Documents	Investment Risk	Project Management	Software Engineering	Development	SAP	...
Document 1	1			1		
Document 2		1				
Document 3			3		1	
Document 4		1				
Document 5			2	1		
Document 6	1			1		
...						

A Simple Term–
Document Matrix

19

Text mining process

The goal is to convert the list of organized documents (the corpus) into a TDM where the cells are filled with the most appropriate indices. However, are all terms important when characterizing documents? Obviously, the answer is “no.” Some terms, such as articles, auxiliary verbs, and terms used in almost all the documents in the corpus, have no differentiating power and, therefore, should be excluded from the indexing process. This list of terms, commonly called *stop terms or stop words*, is specific to the domain of study and should be identified by the domain experts.

Another filtration that should take place to accurately create the indices is *stemming*, which refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of a verb are identified and indexed as the same word. For example, *stemming* will ensure that *modeling* and *modeled* will be recognized as the word *model*.

20

Text mining process

The first generation of the TDM includes all the unique terms identified in the corpus (as its columns), excluding the ones in the stop term list; all the documents (as its rows); and the occurrence count of each term for each document (as its cell values). If, as is commonly the case, the corpus includes a rather large number of documents, then there is a very good chance that the TDM will have a very large number of terms. Processing such a large matrix might be time-consuming and, more important, might lead to extraction of inaccurate patterns. At this point, one has to decide the following: (1) What is the best representation of the indices? and (2) How can we reduce the dimensionality of this matrix to a manageable size?

Reducing the Dimensionality of the Matrix Because the TDM is often very large and rather sparse (most of the cells filled with zeros. Several options are available for managing the matrix size:

21

Text mining process

Singular value decomposition (SVD), which is closely related to principal components analysis, reduces the overall dimensionality of the input matrix (number of input documents by number of extracted terms) to a lower-dimensional space, where each consecutive dimension represents the largest degree of variability (between words and documents) possible.

Task 3: Extract the Knowledge

Using the well-structured TDM, and potentially augmented with other structured data elements, novel patterns are extracted in the context of the specific problem being addressed. The main categories of knowledge extraction methods are **classification, clustering, association, and trend analysis.**

22

Text mining process

Sentiment Analysis:

Sentiment is a difficult word to define. It is often linked to or confused with other terms like *belief, view, opinion, and conviction*. Sentiment suggests a *settled opinion reflective of one's feelings*. Sentiment has some unique properties that set *it apart from other concepts that we may want to identify in text*. Often we want to categorize text by topic, which may involve dealing with whole taxonomies of topics. Sentiment classification, on the other hand, usually deals with *two classes (positive versus negative), a range of polarity (e.g., star ratings for movies), or even a range in strength of opinion*. These classes span many topics, users, and documents. Although dealing with only a few classes may seem like an easier task than standard text analysis, this is far from the truth.

23

Text mining process

As a field of research, sentiment analysis is closely related to **computational linguistics, NLP, and text mining**.

Sentiment analysis has many names. It's often referred to as **opinion mining, subjectivity analysis**, and appraisal extraction, with some connections to affective computing (computer recognition and expression of emotion).

Sentiment that appears in text comes in two flavors: **explicit**, where the subjective sentence directly expresses an opinion ("It's a wonderful day"), and **implicit**, where the text implies an opinion ("The handle breaks too easily").

Most of the earlier work done in sentiment analysis focused on the first kind of sentiment because it is easier to analyze. Current trends are to implement analytical methods to consider both implicit and explicit sentiments.

24

Text mining process

Sentiment Analysis



Positive



Negative



Neutral



Review



Sentiment Analyzer



Sentiment Analysis Process

25

Text mining process

Sentiment Analysis Applications:

Improving services. Companies such as Uber rely on sentiment analysis when they monitor social media to gather information pertaining to whether users are pleased with the newest version of their app.

Movie reviews. Review websites such as Rotten Tomatoes employ sentiment analysis to analyze movie reviews. Rotten Tomatoes performs sentiment analysis on huge stores of subjective opinion data. By using five values: negative, somewhat negative, neutral, somewhat positive, and positive, labels can be created that can classify all of the movie review phrases currently in the database. However, challenges occur when sentiments such as ambiguous language, brevity, sentence negation, or sarcasm are present in a review, all of which pose challenges for natural language processing tools that are designed to focus on the words rather than the intentions behind them.

26

Text mining process

Politics. Political Scientists rely on sentiment analysis to determine how announcements are received by the public. In 2012, the Obama administration applied sentiment analysis to evaluate policy announcements. In addition, this form of analytics can be used to study the number of negative mentions about candidates in various news and media sources.

Employee satisfaction. Businesses perform best when their employees are happy. Applying sentiment analysis can help an organization learn how engaged its employees are, how productive they have been, and ultimately how best to reduce turnover and ensure that workers are satisfied. By applying artificial intelligence to Slack messages, Glassdoor reviews, and employee surveys, it's possible to subjectively evaluate employee feedback and to discover the most common concerns and points of discussion.

27

Text mining process

HOW DOES SENTIMENT ANALYSIS WORK?



28

Text mining process

How Does NLP Work in Sentiment Analysis?

- **Text Preprocessing:** **Sentiment Analysis** begins with the preprocessing of raw text data. This involves cleaning and preparing the text.
- **Feature Extraction :** Once the text is preprocessed, **Sentiment Analysis** techniques are used to extract relevant features from the text.
- **Sentiment Classification:** **Sentiment Analysis** models, often machine learning-based, are trained on labeled data to classify text into sentiment categories.
- **Sentiment Scoring :** Sentiment analysis tools assign a sentiment score or polarity to the text. Scores can range from -1 (very negative) to 1 (very positive), with 0 typically indicating neutrality.

29

Python for data mining and text mining

This section provides a comprehensive examination of the most common functions used for data mining and text mining in the pandas and seaborn libraries:

- Pandas: <https://pandas.pydata.org/docs/>
- Seaborn: <https://seaborn.pydata.org/>

Functions like `read_csv()` and `to_csv()` are fundamental for loading and saving text data, often used in initial data ingestion and final output stages.

```
df = pd.read_csv('file.csv')
```

30

Python for data mining and text mining

df.describe(): After loading, quickly check scale, range, and possible outliers in numeric columns.

df.isna().sum: Determine which columns need imputation or row-dropping before modelling.

Boolean indexing (no function, but a pattern):

Give me only rows that satisfy business rules (e.g., revenue>0, age between 18-65).”

```
subset = df[(df['sepal_length'] > 5) &
            (df['species'] == 'versicolor')]
print(subset.head())
```

DataFrame.corr:

A **DataFrame** (square correlation matrix)

```
corr = df[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']].corr()
print(corr)
```

String Methods for Text Mining:

The pandas string methods, accessible via the `str accessor`, are essential for text preprocessing. For instance, `str.split()` is frequently used for tokenization, breaking text into individual words or tokens. Other common functions include `str.lower()` for case normalization, `str.replace()` for cleaning unwanted characters, and `str.contains()` for filtering based on patterns. These methods are critical for tasks such as removing punctuation, converting text to lowercase, and identifying specific patterns, which are foundational steps in text mining workflows.

The following table 1 summarizes the most common pandas functions for text mining.

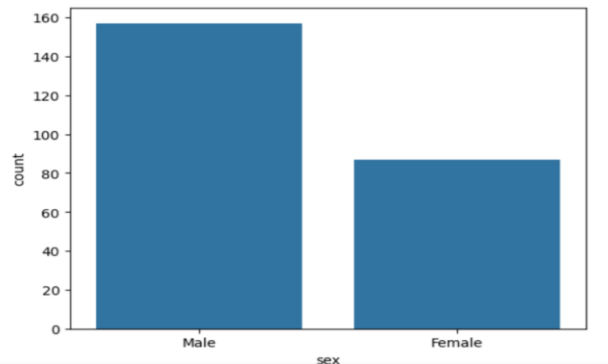
Python for data mining and text mining

Function	Description	Example Usage in Text Mining
<code>str.split()</code>	Splits strings into lists, used for tokenization	Breaking text into words for frequency analysis
<code>str.lower()</code>	Converts strings to lowercase for normalization	Standardizing text for consistent processing
<code>str.replace()</code>	Replaces patterns in strings, used for cleaning	Removing unwanted characters or replacing terms
<code>str.contains()</code>	Checks for patterns in strings, used for filtering	Identifying documents containing specific keywords
<code>groupby()</code>	Groups data for aggregation, useful for categorizing text data	Grouping by document or category for analysis
<code>value_counts()</code>	Computes frequency of unique values, key for word frequency analysis	Counting occurrences of words after preprocessing
<code>apply()</code>	Applies custom functions, used for advanced text processing	Applying sentiment analysis or custom tokenization

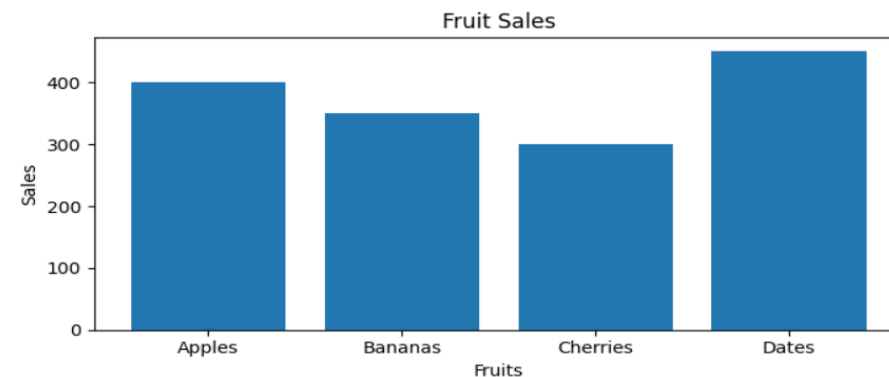
33

Python for data mining and text mining

- **Countplot()** is one of the most common functions for text mining, used to display the frequency of categorical data, such as word frequencies, n-grams, or named entity recognition results.
- **Barplot()** offers similar functionality but with more flexibility, allowing for comparisons across categories, such as comparing word frequencies across different documents or categories.



Count plot

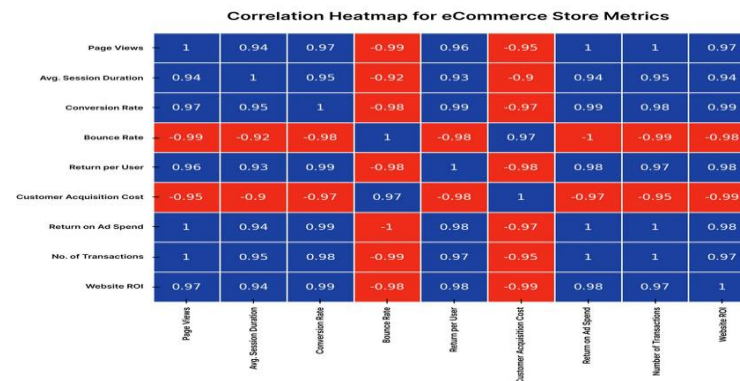


Bar plot

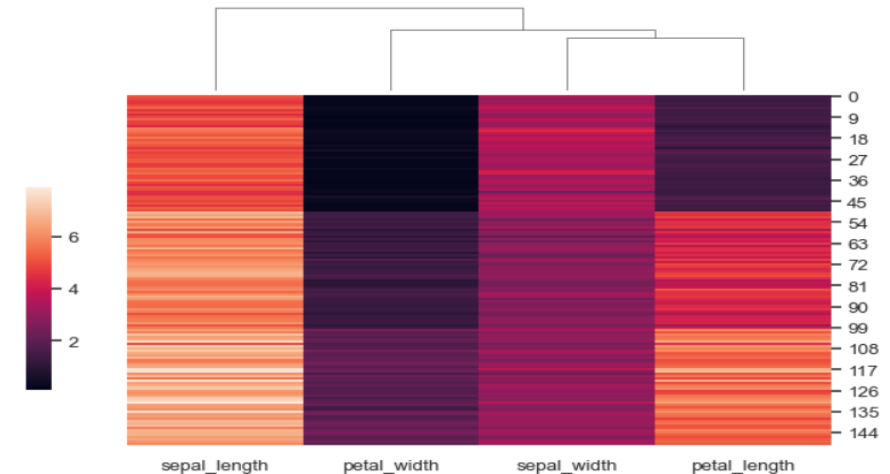
33

Python for data mining and text mining

- **Heatmap()** is frequently used to visualize co-occurrence matrices or correlation matrices between terms, aiding in understanding relationships in text data.
- **Clustermap()** is utilized for hierarchical clustering, which can visualize relationships between terms or documents, particularly useful in topic modeling.



Heatmap



Clustermap

Plot	Function	Typical call
Distribution	<code>sns.histplot</code>	<code>sns.histplot(df['age'])</code>
Count / bar	<code>sns.countplot</code>	<code>sns.countplot(x='label', data=df)</code>
Boxplot / outliers	<code>sns.boxplot</code>	<code>sns.boxplot(x='cat', y='val', data=df)</code>
Correlation heat-map	<code>sns.heatmap</code>	<code>sns.heatmap(df.corr(), annot=True)</code>
Scatter / relation	<code>sns.scatterplot</code>	<code>sns.scatterplot(x='x', y='y', hue='group', data=df)</code>
Pair-wise overview	<code>sns.pairplot</code>	<code>sns.pairplot(df[cols])</code>

36

Python for data mining and text mining

Function	Description	Example Usage in Text Mining
countplot()	Creates bar plots for frequency counts, key for word frequency visualization	Displaying frequencies of top words or n-grams
barplot()	Flexible bar plots for comparing frequencies across categories	Comparing word frequencies across different documents or categories
heatmap()	Visualizes matrices, such as co-occurrence or correlation, for term relationships	Showing term co-occurrence in a document set
clustermap()	Performs hierarchical clustering, useful for visualizing term or document clusters	Visualizing clusters in topic modeling results

37

Summary

1. Text mining.
2. Text mining process.
3. Python libraries for data mining and text mining.



Thank You