Practical Business Python

Lecture 8: Text Analytics

Iegor Vyshnevskyi

Woosong University

November 9, 2023

Agenda

1. Intro to Text Mining
2. Text Mining: Process
3. Text Mining: Techniques
4. Text Mining: Methods
5. Text Mining: Areas & Applications
6. Text Mining: Practical use
7. In-class Assignment

1. Intro to Text Mining

Overview

- According to some reports in 2018, around ***2.5 quintillion bytes*** of data are created every day and it's going to increase every year.
- The data we create includes videos, audios, images, ***text*** and many more.
- Data is key to businesses and proper utilization of those data adds value to the organization.
- Data is collected from the customer through social media, emails, text messages and many more on a day to day basis.
- As most of the data such as email, customer feedback, text messages consists of text data -> ***text mining*** is becoming extremely important for *identifying important patterns and trends within text*.

Text as Data

- Textual data provides a means of understanding all human behavior through a data-driven, analytical approach.

Text as Data (cont.)

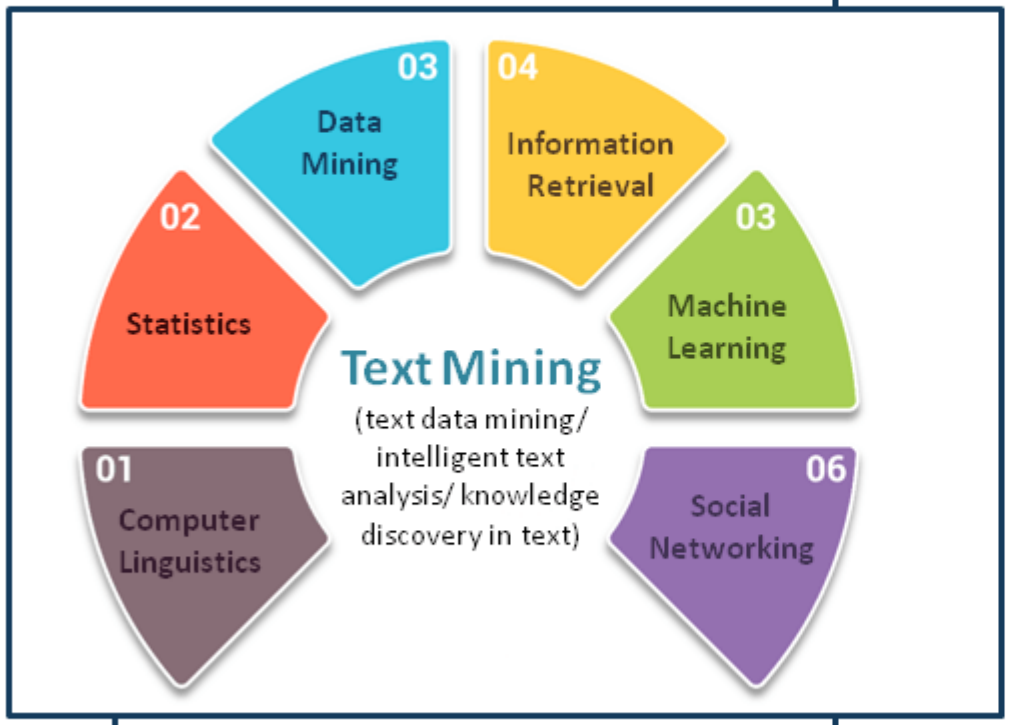
- There are *many reasons* why text has business value:
  - *Big Text*: there is more textual data than numerical data.
  - *Text is versatile*. Nuances and behavioral expressions are not conveyed with numbers, so analyzing text allows us to explore these aspects of human interaction.
  - *Text contains emotive content*. This has led to the ubiquity of "Sentiment analysis".
  - Text contains *opinions and connections*.
  - Numbers aggregate; *text disaggregates*. Text allows us to drill down into underlying behavior when understanding human interaction.

Definition: Text-Mining

- ***Text mining*** is the large-scale, automated processing of plain text language in digital form to extract data that is converted into useful quantitative or qualitative information.
- Text mining is automated on big data that is not amenable to human processing within reasonable time frames. It entails extracting data that is converted into information of many types.
  - Simple: Text mining may be simple as key word searches and counts.
  - Complicated: It may require language parsing and complex rules for information extraction.
- Involves structured text, such as the information in forms and some kinds of web pages.
- May be applied to unstructured text is a much harder endeavor.
- Text mining is also aimed at unearthing unseen relationships in unstructured text as in meta analyses of research papers, see Van Noorden 2012.

Definition: Text-Mining

Importance of text mining.



"Text Mining is a valuable resource in social networking and blogging, customer relations management, tracking public opinion and text filtering."

2. Text Mining: Process

TM Process

The *process of text mining* combines several techniques that enable us to deduce the information from the unstructured data. The general process in text mining are:

1. The first step involved in text mining is *collection of data* i.e. text. The data can be collected from different sources such as websites, emails, social media, blogs and others. All the available data are gathered together in the initial step.
2. After the collection of data, *text pre-processing* is carried out. This is the main step in text mining and requires lots of time and effort. The collected data may be structured, semi-structured and unstructured. In text pre-processing all the available data is cleansed to create structured data. These steps consist of methods such as text cleanup, tokenization, filtering, stemming, lemmatization, linguistic processing, part of speech recognition and word sense disambiguation.

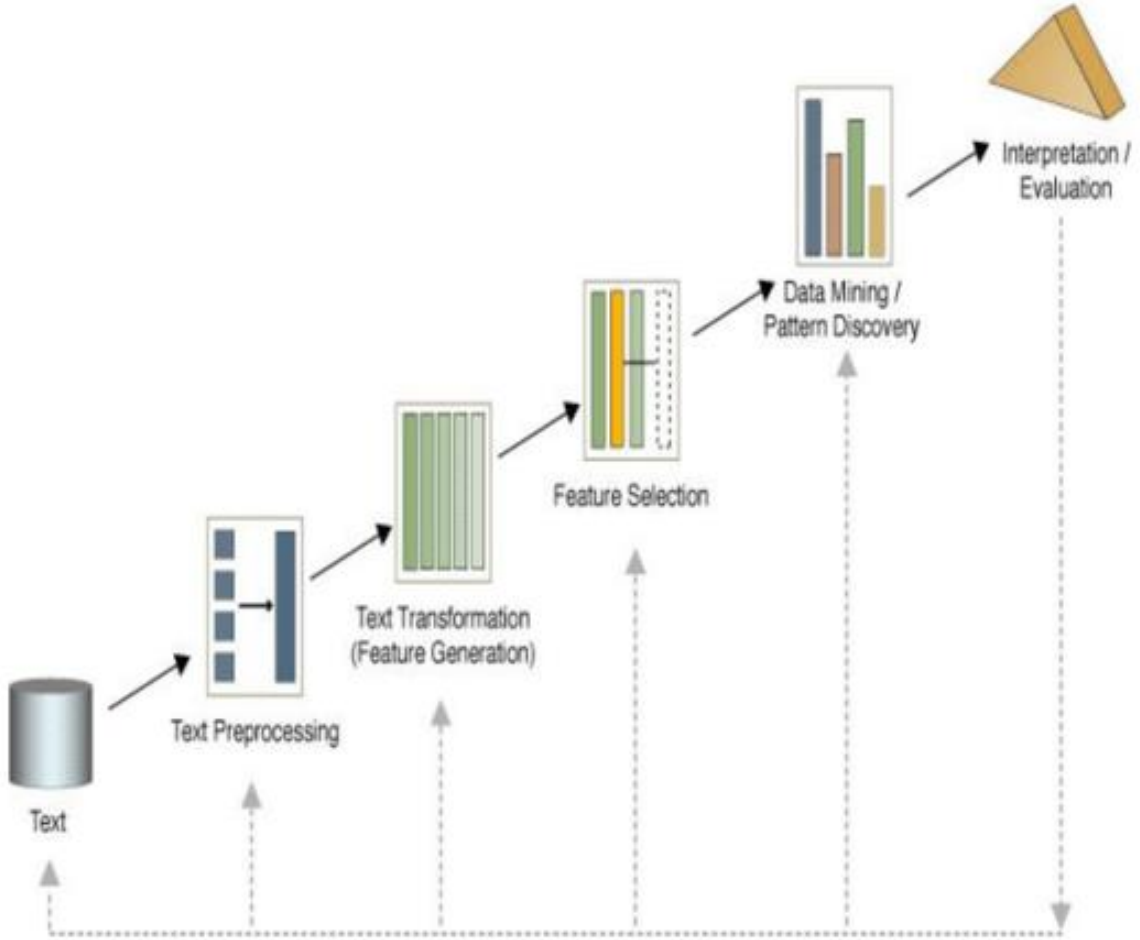
TM Process (cont.)

3. After the pre-processing of data, various techniques are used to *analyse the data*. The analysis must be carried out on structured data as it gives efficient results. The common methods used in these steps are information extraction, information retrieval, categorization, clustering, visualization and summarization.

4. Finally the obtained *results are evaluated and stored* for future reference. In this way the data obtained from different sources are used to get the meaningful patterns using text mining.

TM Process (cont.)

- Text preprocessing
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Simple counting
  - Statistics
- Text/Data Mining
  - Classification-Supervised learning
  - Clustering-Unsupervised learning
- Analyzing results



3. Text Mining: Techniques

TM Techniques

Text mining techniques are used to discover the insights from structured text/data. These text mining techniques use different tools, methods and applications for their execution. The various text mining techniques are:

1. ***Information Extraction:*** It is one of the most famous text mining techniques. This technique focuses on identifying the extraction of entities, attributes and their relationships from the available textual data. Whatever information is extracted from the data is then stored in a database for future access and retrieval. The efficiency and relevancy of the results are evaluated using precision and recall processes.
2. ***Information Retrieval***: Information Retrieval is the process of extracting relevant and associated patterns based on a specific set of words or phrases. In this text mining technique, information retrieval systems make use of different algorithms to track and monitor user behaviors and discover relevant data accordingly.

TM Techniques (cont.)

3. Categorization: Categorization is one of the most popular supervised learning methods. In this technique, normal language texts are assigned to a predefined set of classes or topics depending upon their content. In this technique, the text documents are gathered, processed and analysed to find the right topics or indexes for each document. Naive Bayesian classifier, Decision tree, Nearest Neighbour classifier and Support Vendor Machines are commonly used to categorize the texts.

4. Clustering: Clustering is one of the most popular unsupervised learning methods in which the data points that are neither classified nor labeled. In this technique the group of text documents which have similar contents are divided into a cluster. The K-means clustering algorithm is one of the most used clustering techniques in which the available data are divided into different clusters by using mean values.

TM Techniques (cont.)

5. Visualization: Visualization is used to simplify and enhance the discovery of useful information with visual cues. It uses visual cues such as text flags to indicate individual documents or document categories and colours to indicate the density of a category, entity, phrase, etc. It is used in placing the large sources of textual data in visual hierarchy.

6. Summarization: Summarization is used to reduce the length of the document and summarize the document's details in brief. Summarization determines the most important points in a lengthy document and replaces the entire set of documents with new important points quickly and efficiently. Summarization involves three steps i.e. pre-processing, processing, and development. Pre-processing step involves building a structured representation of the text whereas different algorithms are used to get a summary of text in processing and finally the development step is where the final text summary is obtained.

4. Text Mining: Methods

TM Methods

There are lots of methods developed to solve the text mining problem which is relevant information retrieval according to the user's requirements. According to information retrieval there are four main methods used in text mining:

1. ***Term Based Method (TBM):*** Term refers to a word with semantic meaning. In term based methods a document is analyzed on the basis of the term and has the advantage of efficient computational performance as well as mature theories from their weighting. Term based method faces enormous challenges in case of polysemy and synonymy. Polysemy means a word hasmultiple meanings and synonymy is multiple words with the same meaning. The semantic meaning of many extracted terms is uncertain and does not provide full information for answering what the user wants.
2. ***Phrase Based Method (PBM):*** Phrase based method may have advantages over term based method as it carries more semantics like information and is less ambiguous. In phrase based methods the document is analyzed on a phrase basis as it is more discriminative than individual terms.

TM Methods (cont.)

3. Concept Based Method (CBM): In this method terms are analyzed on sentences and document level. Concept based methods can effectively discriminate between non important terms and meaningful terms which describe the meaning of the sentences. This model normally relies upon natural language processing techniques. Feature selection is used to optimize the representation and to remove the noise and ambiguity in the document.

4. Pattern Taxonomy Method (PTM): In this method documents are analyzed in terms of pattern basis. Taxonomy refers to the process of finding the root words. Patterns can be structured into a taxonomy by using is-a relationship and can be discovered using techniques like rule mining, frequent item set mining, sequential pattern mining and closed pattern mining. This method refines discovered patterns in text documents and has efficient performance than that of concept based and term based methods.

5. Text Mining: Areas & Applications

TM Areas

1. Information Retrieval (IR): Information retrieval is regarded as an extension to document retrieval where the documents that are returned are processed to condense or extract the particular information sought by the user. Thus document retrieval could be followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage using techniques. IR systems help to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very complex algorithms to large document collections, IR can speed up the analysis significantly by reducing the number of documents for analysis.

2. Data Mining (DM): Data mining can be simply described as looking for patterns in data. It can be more fully characterized as the extraction of hidden, previously unknown, and useful information from data. Data mining tools can predict behaviors and future trends, allowing businesses to make positive, knowledge based decisions. Data mining tools can answer business questions that have traditionally been too time consuming to resolve. They search databases for hidden and unknown patterns, finding critical information that experts may miss because it lies outside their expectations. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

TM Areas (cont.)

3. Natural Language Processing (NLP): NLP is one of the oldest, widest and most challenging problems in the field of artificial intelligence. It is the study of human language so that computers can understand natural languages as humans do. NLP research pursues the vague question of how we understand the meaning of a sentence or a document. The role of NLP in text mining is to deliver the system in the information extraction phase as an input.

4. Information Extraction (IE): Information Extraction is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases, this activity includes processing human language texts by means of natural language processing (NLP). The recent activities in multimedia document processing like automatic annotation and mining information out of images/audio/video could be seen as information extraction and the best practical and live example of IE is Google Search Engine. It involves defining the general form of the information that we are interested in as one or more templates, which are used to guide the extraction process. IE systems greatly depend on the data generated by NLP systems.

TM Applications

Text mining has improved user experiences and business decisions. Most of the companies are using text mining tools to add value to their organization and products. Some application areas of text mining includes:

1. ***Risk Management:*** One of the main causes of business failure is due to lack of proper or insufficient risk analysis. Integrating risk management software powered by text mining technologies canhelp businesses to stay updated with all current trends in the market and boost their abilities to cover up the potential risks.
2. ***Customer Care Service:*** When the business system is integrated with text analytical tools, feedback systems, chatbots, online reviews, support tickets and social media profiles, it enables us to improve the customer experience with speed. Text mining and sentiment analysis can provide mechanisms for us to prioritize key points for our customer, allowing us to respond to urgent issues in real-time and helps to increase the customer satisfaction.

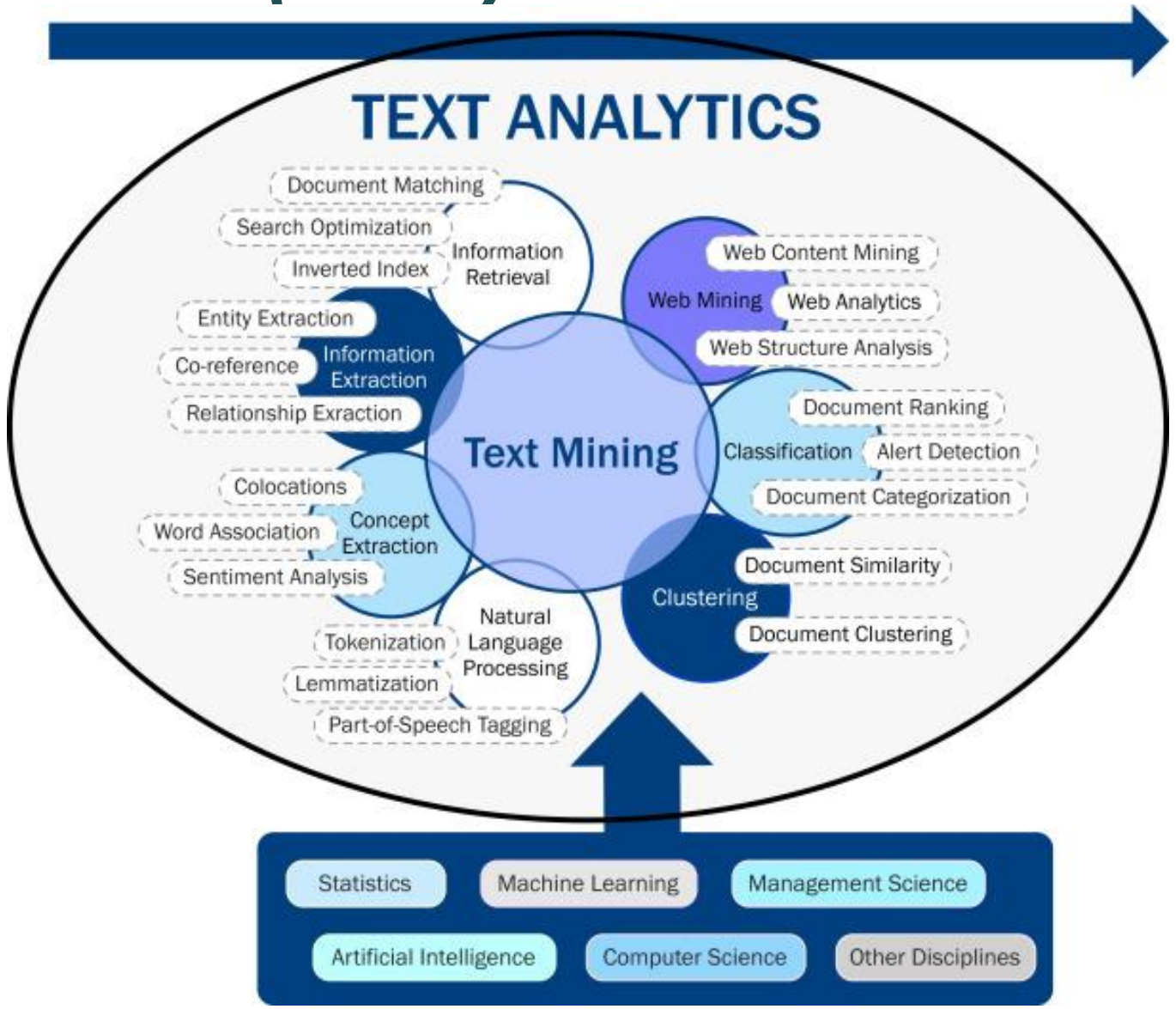
TM Applications (cont.)

3. Healthcare: One of the major applications of text mining is the healthcare sector as it provides valuable information to the researchers. Manual investigation of medical research can be very costly and time consuming. As text mining provides an automation method for extracting the valuable information from the medical literature, it is becoming extremely popular in the medical field as well.

4. Spam Filtering: Text mining is used to filter and exclude the emails from inboxes and thus improving the overall user experiences. With the help of this application it reduces the risk of cyber attacks to the end users.

5. Fraud Detection: By combining the outcomes of the text analysis with relevant structured data we are able to process the user profile and claims efficiently as well as to detect and prevent frauds.

TM Applications (cont.)



TM Intersection

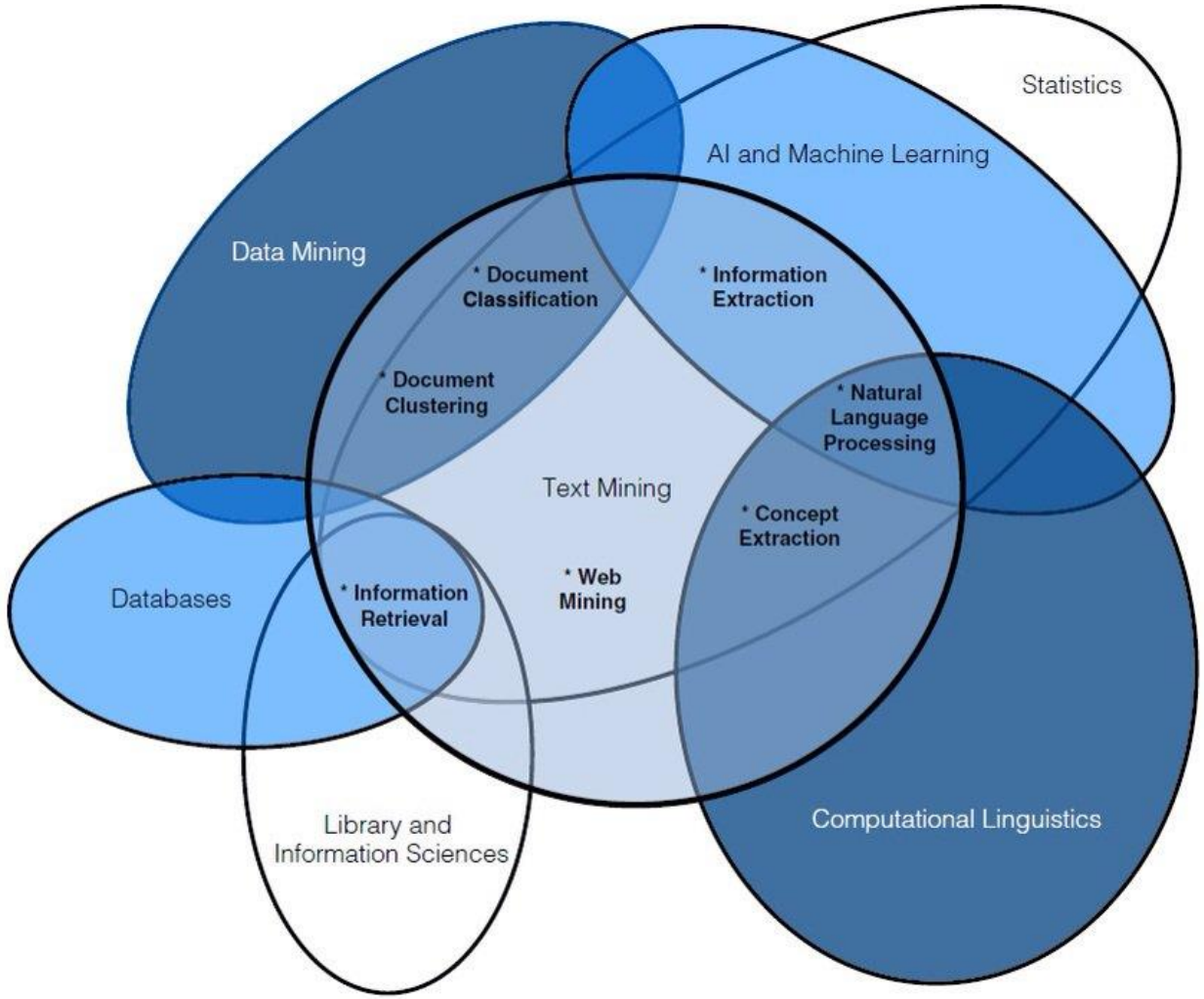


FIGURE 2.1
A Venn diagram of the intersection of text mining and six related fields (shown as ovals), such as data mining, statistics, and computational linguistics. The seven text mining practice areas exist at the major intersections of text mining with its six related fields.

6. Text Mining: Practical use

Please open the script

7. In-class Assignment

You will be working with ‘Twitter Data.csv’ file. Please load the file into Python:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   tweet_id                      14640 non-null  float64
 1   airline_sentiment             14640 non-null  object 
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object 
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object 
 6   airline_sentiment_gold        40 non-null     object 
 7   name                          14640 non-null  object 
 8   negativereason_gold           32 non-null     object 
 9   retweet_count                 14640 non-null  int64  
 10  text                          14640 non-null  object 
 11  tweet_coord                   1019 non-null   object 
 12  tweet_created                 14640 non-null  object 
 13  tweet_location                9907 non-null   object 
 14  user_timezone                 9820 non-null   object 
dtypes: float64(3), int64(1), object(11)
memory usage: 1.7+ MB
```

Acknowledgment

Used freely available online materials.