Text Data Mining

Lecture 9.1

Objectives

- **Applications of Data mining in Text analysis**
- **Basics of text mining- what and why**
- **Areas**
- Key tasks
- Approaches
- Applications

Text Data Mining

- **Text data mining** can be described as the process of extracting essential data from standard language text
- All the data that we generate via text messages, documents, emails, files are written in common language text.
- Text mining is primarily used to draw **useful insights or patterns** from natural language data
- Text mining will **“turn text into numbers”**. Such as predictive data mining projects, the application of unsupervised learning methods.
- **Transforming data into information that machines can understand**

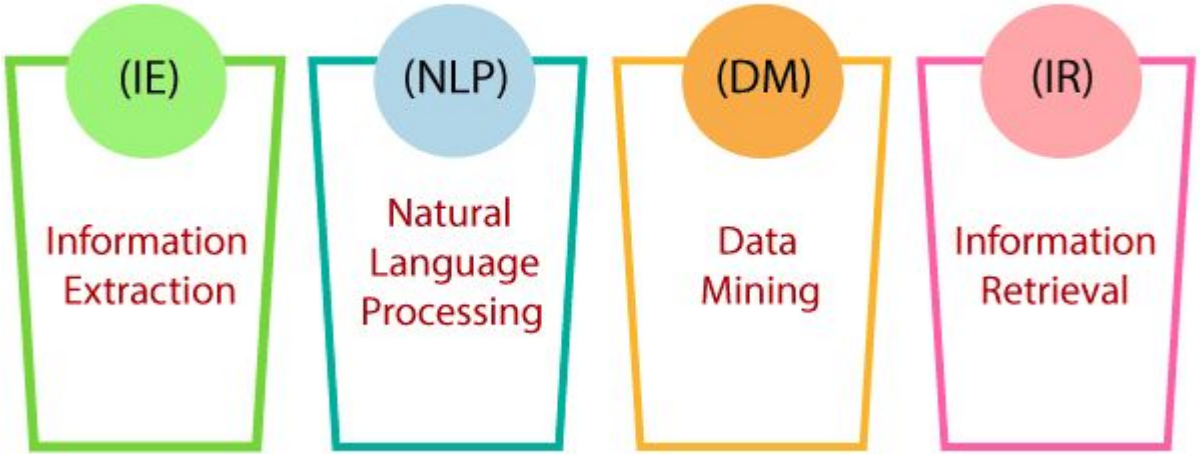
Text Data Mining -Why

- Text databases consist of huge collection of documents.
- They collect these information from several sources such as news articles, books, digital libraries, e-mail messages, web pages, etc.
- Due to increase in the amount of information, the text databases are growing rapidly. In many of the text databases, the data is unstructured or semi-structured.
- Manual processing of these data collection for useful information is not practical
- When **text mining and machine learning** are combined, automated text analysis becomes possible.

Text Mining vs Text Analytics

- **Text mining** combines notions of statistics, linguistics, and machine learning to create models that learn from training data and can predict results on new information based on their previous experience.
- Text analytics, on the other hand, uses results from analyses performed by text mining models, to create graphs and all kinds of data visualizations.

Text Mining: Areas



Methods in Text Mining

- Basic Methods
- Advanced Methods

Word Frequency Analysis

- Identify the **most recurrent terms** or concepts in a set of data
- Keyword identification
- Important in Information retrieval/web search
- Customer review analysis, social media chat analysis, etc.
- For instance, if the words *expensive, overpriced and overrated* frequently appear on your customer reviews, it may indicate you need to adjust your prices
- Toxic words in social media posts

Co-occurrence Analysis

- Co-occurrence/collocation
- A sequence of words that **commonly appear near each other**
- Important in Language Modeling (assigning probability to sentence)
- Help to describe the context better
- Word prediction, automatic completion, etc

Concordance

- Concordance is an alphabetical list of primary words used by an author.
- Concordance analysis is used to recognize the particular context or instance in which a word or set of words appears.
- The same word can be used in many different contexts. Analyzing the concordance of a word can help understand its exact meaning based on context.
- Example:

  The ***bank*** will not be accepting cash on Saturdays

  The river overflowed the ***bank***...

Text Classification

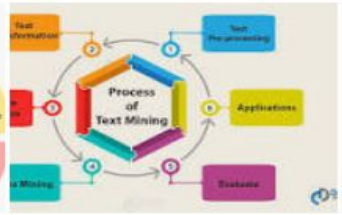
- Process of assigning **categories (tags) to unstructured text data**.
- Topic analysis: helps you understand the main themes or subjects of a text
- Sentiment Analysis: identify the sentiment (positive or negative) of the text
- Language Detection: classify text based on the language
- Spam filter: Classifying text into spam or not
- News classification- classify text into certain news groups (politics, economy, technology,...)

Text Extraction

text mining

All Books Images News Shopping More Settings Tools

About 51,00,00,000 results (0.48 seconds)



Text mining, also known as **text** analysis, is the process of transforming unstructured **text** data into meaningful and actionable information. **Text mining** utilizes different AI technologies to automatically process data and generate valuable insights, enabling companies to make data-driven decisions.

monkeylearn.com › text-mining

Text Mining: The Beginner's Guide - MonkeyLearn

About Featured Snippets Feedback

Text mining

Text mining, also referred to
similar to text analytics, is th
high-quality information from
discovery by computer of ne
information, by automatically
from different written resourc

Text Extraction

- Extracts specific pieces of data from a text, like keywords, entity names, addresses, emails, etc.
- By using text extraction, companies can avoid all the hassle of sorting through their data manually to pull out key information.
- Keyword extraction
- Question-answer extraction
- Named entity extraction
-

Clustering

- Seeks to identify intrinsic structures in textual information and organize them into relevant subgroups or ‘clusters’ for further analysis
- Form meaningful clusters from the unlabeled textual data without having any prior information on them.
- Cluster analysis is a standard text mining tool that assists in data distribution or acts as a pre-processing step for other text mining algorithms running on detected clusters.

Summarisation

- Automatically generating a compressed version of a specific text that holds valuable information for the end-user
- Browse through multiple text sources to craft summaries of texts containing a considerable proportion of information in a concise format, keeping the overall meaning and intent of the original documents essentially the same
- Example : News summarisation

Next

- Approaches
- Applications

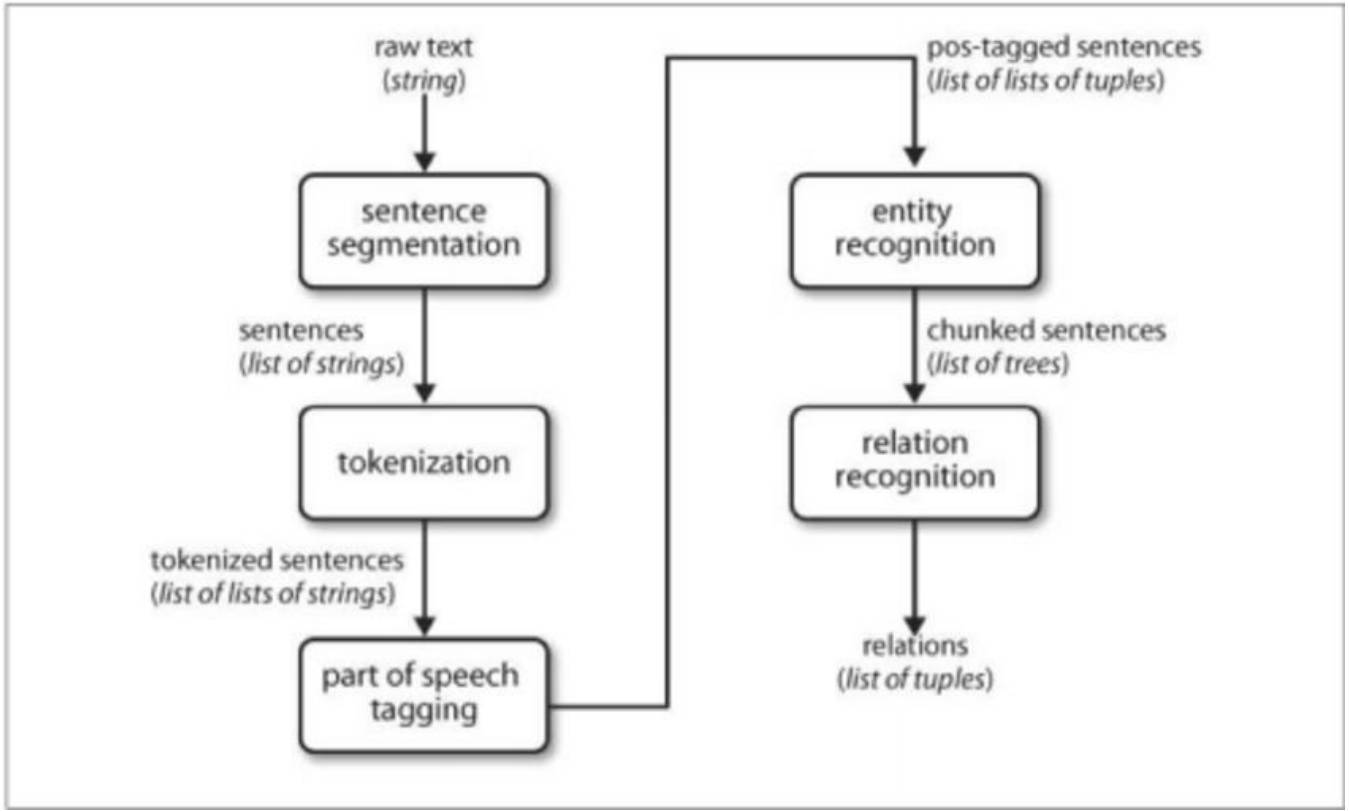
Text Data Mining- 2

Lecture 9.2

Recap

- Text mining- concepts
- Areas
- Methods

Basic Text Mining Pipeline

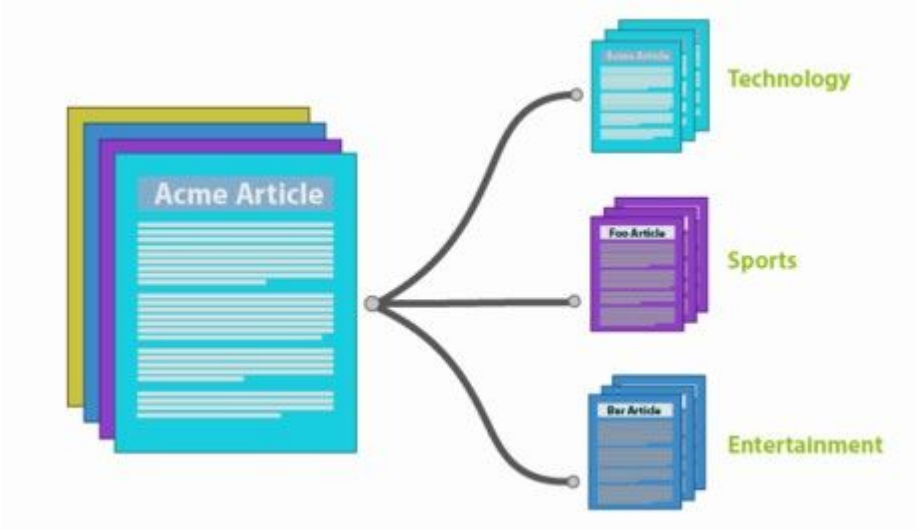


Text Mining

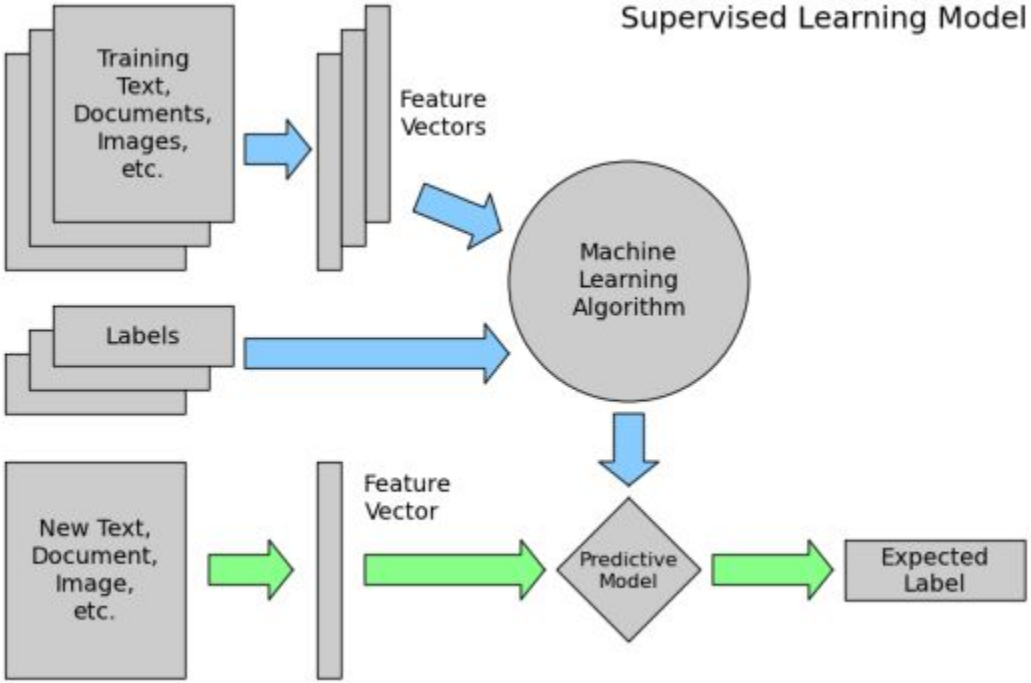
- Text Mining (analysis) can be performed in different levels
  - Document
  - Paragraph level
  - Sentence level
  - Word level
  - Character level

Document Classification

- Document classification is the act of labeling documents into categories according to their content
- Depending on the classification algorithm or strategy used, the classifier might also provide a confidence measure to indicate how confident it is that the classification label is correct.



Document Classification- Model



Loading the dataset

- The tagged dataset (in NLP aka Corpus)
- The dataset **needs to be large enough** to have an adequate number of documents in each class
- The dataset also **needs to be of a high enough quality** in terms of how distinct the documents in the different categories are from each other

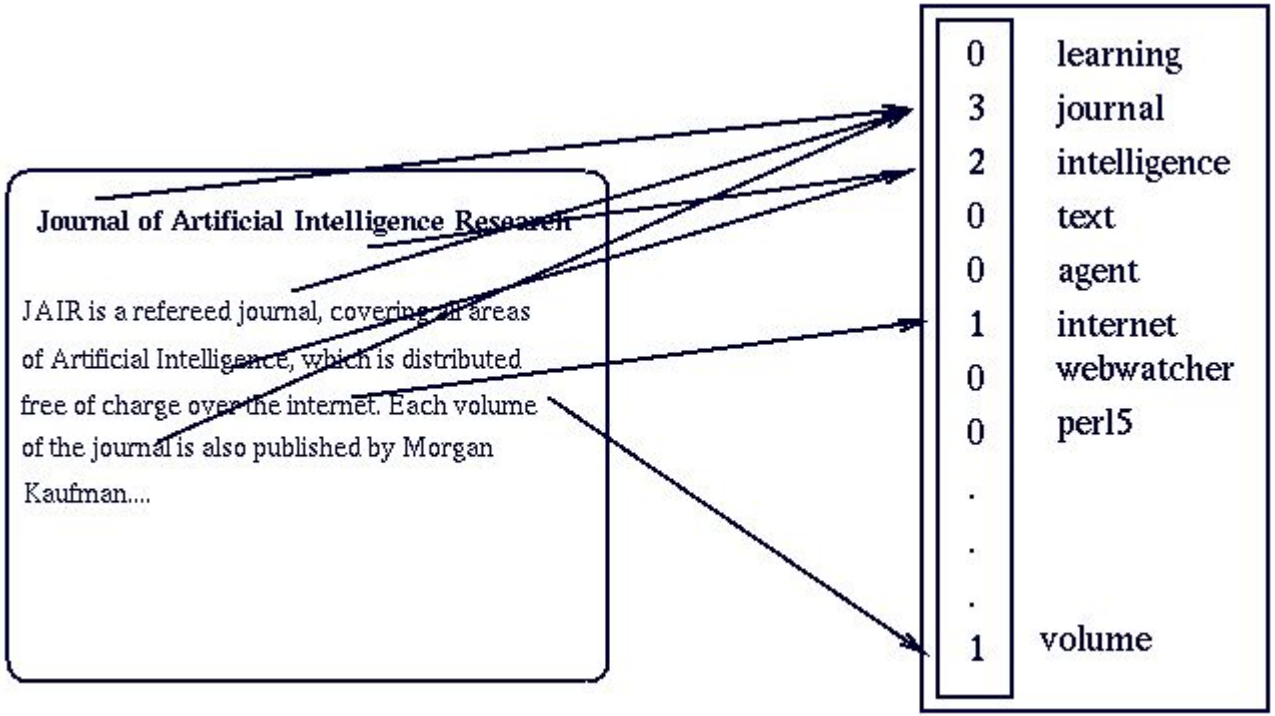
Preprocessing

- Text may contain numbers, special characters, and unwanted spaces.
- Depending upon the problem we face, we may or may not need to remove these special characters and numbers from text.
- Stop word removal
- NLP steps like stemming, lemmatization

Feature Engineering

- Next important step in transforming text documents into numerical vectors
- Generally, a document is represented by words in it
- **Vocabulary:** collection of distinct terms in the corpus
- The most common way to deal with documents is first to transform them into sparse numeric vectors and then deal with them with linear algebra operations
- This representation is referred to also as **“Bag-Of-Words”** or **“Vector-Space-Model”**

Bag-of-word Model



Bag-of-word Model

- In the bag-of-words representation each word is represented as a separate variable having numeric weight (importance)
- The most popular weighting schema (word embedding) is normalized word frequency TF-IDF

$$tfidf(w) = tf . \log(\frac{N}{df(w)})$$

Tf(w) – term frequency (number of word occurrences in a document)
Df(w) – document frequency (number of documents containing the word)
N – number of all documents

- The word is more important if it appears several times in a target document
- The word is more important if it appears in less documents

Bag-of-word Model

TRUMP MAKES BID FOR CONTROL OF RESORTS Casino owner and real estate Donald Trump has offered to acquire all Class B common shares of Resorts International Inc, a spokesman for Trump said. The estate of late Resorts chairman James M. Crosby owns 340,783 of the 752,297 Class B shares. Resorts also has about 6,432,000 Class A common shares outstanding. Each Class B share has 100 times the voting power of a Class A share, giving the Class B stock about 93 pct of Resorts' voting power.

Original text

[RESORTS:0.624] [CLASS:0.487] [TRUMP:0.367] [VOTING:0.171]
[ESTATE:0.166] [POWER:0.134] [CROSBY:0.134] [CASINO:0.119]
[DEVELOPER:0.118] [SHARES:0.117] [OWNER:0.102]
[DONALD:0.097] [COMMON:0.093] [GIVING:0.081] [OWNS:0.080]
[MAKES:0.078] [TIMES:0.075] [SHARE:0.072] [JAMES:0.070]
[REAL:0.068] [CONTROL:0.065] [ACQUIRE:0.064]
[OFFERED:0.063] [BID:0.063] [LATE:0.062] [OUTSTANDING:0.056]
[SPOKESMAN:0.049] [CHAIRMAN:0.049] [INTERNATIONAL:0.041]
[STOCK:0.035] [YORK:0.035] [PCT:0.022] [MARCH:0.011]

Bag-of-Words representation (high dimensional sparse vector)

Classification Algorithm

Popular algorithms for text categorization:

- Support Vector Machines
- Logistic Regression
- Perceptron algorithm
- Naive Bayesian classifier
- Winnow algorithm
- Nearest Neighbour

Evaluation of Model

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

- Accuracy = TP+TN/TP+FP+FN+TN
- Precision = TP/TP+FP
- Recall = TP/TP+FN
- F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Summary

Text Mining

Concept

Approaches

Example: Document classification

Assignment

Text/Document Clustering: Similarity based approach

- View from this class: https://www.coursera.org/lecture/text-mining/4-5-text-clustering-similarity-based-approaches-PsyKR
- A brief review on text/document clustering.
- Objective (purpose), approaches, applications, evaluation measures, challenges, and references
- You can submit as a presentation (max 15 slides) or a document with (max 1200 words) pages.
- **Due Date: 23rd October 2020, 23:59**