# CSE 435: Data Mining

## Text Mining and Text Preprocessing; Web Mining; Introduction to Big Data Analytics

**Md Atikuzzaman**
Lecturer
Department of Computer Science & Engineering
atik@cse.green.edu.bd

GREEN UNIVERSITY OF BANGLADESH

# Outline

## Text Mining

- Text Mining Concepts
- Text Preprocessing
- Stemming and Lemmatization
- N-Grams
- Bag-of-Words
- TF-IDF
- Text Mining Tasks

## Web Mining

- Web Mining Concepts
- Web Content Mining
- Web Structure Mining
- PageRank
- Web Usage Mining

## Big Data Analytics

- Big Data and 5Vs
- Hadoop Architecture
- HDFS and YARN
- MapReduce
- Apache Spark
- Hadoop vs. Spark

## Lecture Flow

**Raw Text $\rightarrow$ Web Data $\rightarrow$ Distributed Big-Data Analytics**

# What is Text Mining?

## Definition

**Text Mining** is the process of extracting useful information, patterns, relationships, and knowledge from unstructured or semi-structured textual data.

**Common Text Sources**

- Emails and messages
- Social media posts
- News articles
- Customer reviews
- Research papers
- Medical reports
- Web pages

**Applications**

- Sentiment analysis
- Spam detection
- Document classification
- Information retrieval
- Topic discovery
- Text clustering
- Recommendation

## Main Challenge

Raw text is not directly suitable for most machine-learning algorithms. It must first be cleaned and transformed into numerical features.

# Text Mining Pipeline


## Input

"The product is extremely good and easy to use."

**Possible Processing**

- Tokenization
- Stop-word removal
- TF-IDF representation
- Classification

## Output

**Sentiment: Positive**

$P(\text{Positive}) = 0.94$

**Goal**

Transform human-readable language into machine-processable information.

# Text Preprocessing

## Original Text

"Data Mining is VERY useful! It helps us discover useful patterns."

**1. Lowercasing**

Data Mining $\rightarrow$ data mining

**2. Tokenization**

[data, mining, is, very, useful, . . .]

**3. Punctuation Removal**

! . , ? :

**4. Stop-Word Removal**

{is, the, a, an, of, to, and}

**5. Normalized Result**

[data, mining, useful, discover, patterns]

## Important

Preprocessing is task-dependent. Removing **"not"**, for example, can completely change the meaning of a sentiment.

# Stemming, Lemmatization and N-Grams

## Stemming

Uses rule-based reduction.

$$\text{playing} \rightarrow \text{play}$$
$$\text{played} \rightarrow \text{play}$$
$$\text{studies} \rightarrow \text{studi}$$

- Fast and simple
- Reduces vocabulary
- May produce invalid words

## Lemmatization

Finds the linguistic base form.

$$\text{cars} \rightarrow \text{car}$$
$$\text{running} \rightarrow \text{run}$$
$$\text{better} \rightarrow \text{good}$$

- Linguistically meaningful
- Uses language knowledge
- Usually more accurate

## N-Grams: Capturing Local Word Order

For “data mining techniques”:

Unigrams: [data], [mining], [techniques]
Bigrams: [data mining], [mining techniques]
Trigram: [data mining techniques]

# Bag-of-Words Representation

Consider:

$D_1$ : "data mining useful"
$D_2$ : "data science useful"
$D_3$ : "mining discovers patterns"

Vocabulary:

$$V = \{\text{data}, \text{mining}, \text{science}, \text{useful}, \text{discovers}, \text{patterns}\}$$

## Bag-of-Words

Each document is represented by the occurrence count of each term in the vocabulary.

| | *data* | *mining* | *science* | *useful* | *discovers* | *patterns* |
|---|---|---|---|---|---|---|
| $D_1$ | 1 | 1 | 0 | 1 | 0 | 0 |
| $D_2$ | 1 | 0 | 1 | 1 | 0 | 0 |
| $D_3$ | 0 | 1 | 0 | 0 | 1 | 1 |

**Limitation:** Bag-of-Words normally ignores word order, grammar, and contextual meaning.

# TF-IDF Representation

## Purpose

TF-IDF gives greater importance to terms that are frequent in a particular document but relatively uncommon across the collection.

**Term Frequency**

$$TF(t, d) = \frac{\text{Frequency of } t \text{ in } d}{\text{Number of terms in } d}$$

Measures how frequently term $t$ occurs in document $d$.

**Inverse Document Frequency**

$$IDF(t) = \log\left(\frac{N}{df(t)}\right)$$

where:

- $N$: total number of documents
- $df(t)$: documents containing $t$

## TF-IDF Weight

$$\boxed{TFIDF(t, d) = TF(t, d) \times IDF(t)}$$

# Worked Example: TF-IDF

Documents:

$D_1$ : "data mining mining"
$D_2$ : "data science"
$D_3$ : "web mining"

Find the TF-IDF of **mining** in $D_1$.

## Step 1: TF

Mining occurs twice among three terms:

$$TF = \frac{2}{3} = 0.667$$

## Step 2: IDF

Mining occurs in 2 of 3 documents:

$$IDF = \log\left(\frac{3}{2}\right) \approx 0.405$$

## Step 3: TF-IDF

$$TFIDF = \frac{2}{3} \times 0.405 = \boxed{0.270}$$

**Interpretation:** the value represents the importance of *mining* in $D_1$ relative to the document collection.

# Common Text Mining Tasks

| Task | Purpose | Example |
|---|---|---|
| Text Classification | Assign predefined categories | Spam / non-spam email |
| Sentiment Analysis | Identify opinion or emotion | Positive / negative review |
| Text Clustering | Group similar documents | News article grouping |
| Topic Modeling | Discover hidden themes | Sports, politics, technology |
| Information Extraction | Extract entities and relationships | Person, company, location |
| Information Retrieval | Find relevant documents | Search engine results |

## Typical Classification Workflow

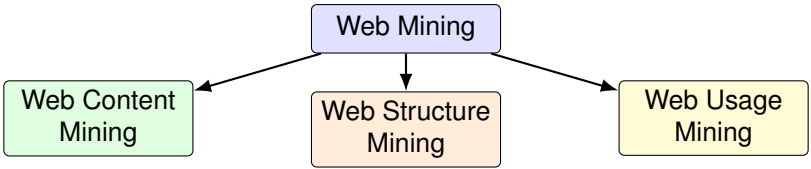
Document $\rightarrow$ Preprocessing $\rightarrow$ TF-IDF $\rightarrow$ Classifier $\rightarrow$ Label

**Common Models:** Naive Bayes, Logistic Regression, SVM, Neural Networks

# Introduction to Web Mining

## Definition

**Web Mining** applies data-mining techniques to discover useful patterns and knowledge from Web data.




**Web Data**

- Page text and HTML
- Images and multimedia
- Hyperlinks
- Search results

**Behavioral Data**

- Server logs
- User clicks
- Browsing sessions
- Online transactions

# Web Content Mining

## Definition

**Web Content Mining** extracts useful information from the actual contents of Web pages.

**Content Types**

- Text and HTML
- Tables
- Images
- Video and audio
- Product information
- Customer reviews

**Applications**

- Search engines
- Product comparison
- News aggregation
- Sentiment analysis
- Recommendation
- Information extraction

## Example

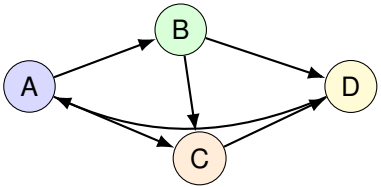
$$\text{E-commerce HTML Page} \rightarrow \{\text{Name, Price, Brand, Rating, Review}\}$$

The unstructured page is transformed into structured attributes that can be analyzed using data-mining algorithms.

# Web Structure Mining

## Definition

**Web Structure Mining** analyzes relationships among Web pages using hyperlinks.




Web pages are modeled as:

$$G = (V, E)$$

where:

$V =$ Web pages

$E =$ directed hyperlinks

It can identify:

- important pages;
- hubs and authorities;

# PageRank: Ranking Web Pages

## Concept

PageRank estimates the importance of a page using the quantity and quality of links pointing to it.

$$PR(A) = \frac{1-d}{N} + d \sum_{i \in In(A)} \frac{PR(i)}{L(i)}$$

**Variables**

- $PR(A)$: rank of page $A$
- $d$: damping factor
- $N$: total number of pages
- $In(A)$: pages linking to $A$
- $L(i)$: outgoing links from page $i$
- usually $d \approx 0.85$

## Main Idea

A link from an important page contributes more PageRank than a link from a less important

# Web Usage Mining

## Definition

**Web Usage Mining** discovers behavioral patterns from users' interactions with websites and Web applications.

**Example Web Log**

{User ID, IP, Time, URL, Device, Referrer, Action}


**Patterns**

- Frequently visited pages
- Navigation sequences
- Session duration

**Applications**

- Recommendation
- Personalization
- Website optimization

# Introduction to Big Data Analytics

## Definition

**Big Data** refers to datasets whose size, speed, or complexity make them difficult to efficiently process using conventional single-machine data-processing systems.

**The 5Vs**

1. **Volume**: amount of data
2. **Velocity**: generation speed
3. **Variety**: multiple formats
4. **Veracity**: quality and reliability
5. **Value**: useful insight

**Examples**

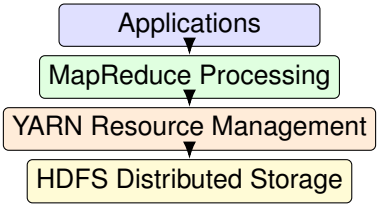
- Social media streams
- IoT sensor data
- Banking transactions
- E-commerce activities
- Video platforms

## Core Principle

Big Data platforms distribute **storage** and **computation** across multiple machines.

# Hadoop Architecture

Apache Hadoop supports distributed storage, resource management, and large-scale data processing.




**Data Locality**

Move computation **close to the data**.

**HDFS**

- Stores large files
- Splits files into blocks
- Distributes blocks across nodes
- Replicates blocks for fault tolerance

**YARN**

- Manages cluster resources
- Schedules applications
- Coordinates distributed jobs

# MapReduce: Distributed Word Count

**Input:** “data mining data”

## Map Phase

Generate key-value pairs:

$$\text{data} \rightarrow (\text{data}, 1)$$
$$\text{mining} \rightarrow (\text{mining}, 1)$$
$$\text{data} \rightarrow (\text{data}, 1)$$

## Reduce Phase

Combine identical keys:

$$(\text{data}, [1, 1]) \rightarrow (\text{data}, 2)$$

$$(\text{mining}, [1]) \rightarrow (\text{mining}, 1)$$

$$\text{Input} \rightarrow \text{Map} \rightarrow \text{Shuffle/Sort} \rightarrow \text{Reduce} \rightarrow \text{Output}$$

## Limitation

Traditional MapReduce frequently writes intermediate results to disk, making iterative processing relatively expensive.

# Apache Spark and Hadoop MapReduce

## Apache Spark

A distributed computing engine designed for fast, large-scale analytics.

**Key Features**

- In-memory processing
- Batch analytics
- Stream processing
- Iterative computation
- Spark SQL
- MLlib

Data $\rightarrow$ Transform $\rightarrow$ Action

Examples:
`map`, `filter`, `reduce`

| Feature | Hadoop MapReduce | Spark |
|---|---|---|
| Processing | Primarily disk-based | Supports in-memory reuse |
| Iteration | Repeated disk I/O | Efficient iterative processing |
| Batch | Strong | Strong |
| Streaming | Not primary model | Structured Streaming |
| Machine Learning | Additional ecosystem tools | MLlib |
| Typical Use | Large-scale batch jobs | Interactive, iterative, streaming |

## Key Takeaway

Hadoop emphasizes reliable distributed storage and batch processing, while Spark is designed for fast, reusable distributed computation.

# References



J. Han, J. Pei, and H. Tong,
*Data Mining: Concepts and Techniques*,
4th ed., Morgan Kaufmann, 2022.


C. D. Manning, P. Raghavan, and H. Schütze,
*Introduction to Information Retrieval*,
Cambridge University Press, 2008.


B. Liu,
*Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*,
2nd ed., Springer, 2011.


L. Page, S. Brin, R. Motwani, and T. Winograd,
"The PageRank Citation Ranking: Bringing Order to the Web,"
Stanford InfoLab, 1999.


T. White,
*Hadoop: The Definitive Guide*,
4th ed., O'Reilly Media, 2015.