Introduction to Text Analytics

CPE 393/623: Text Analytics

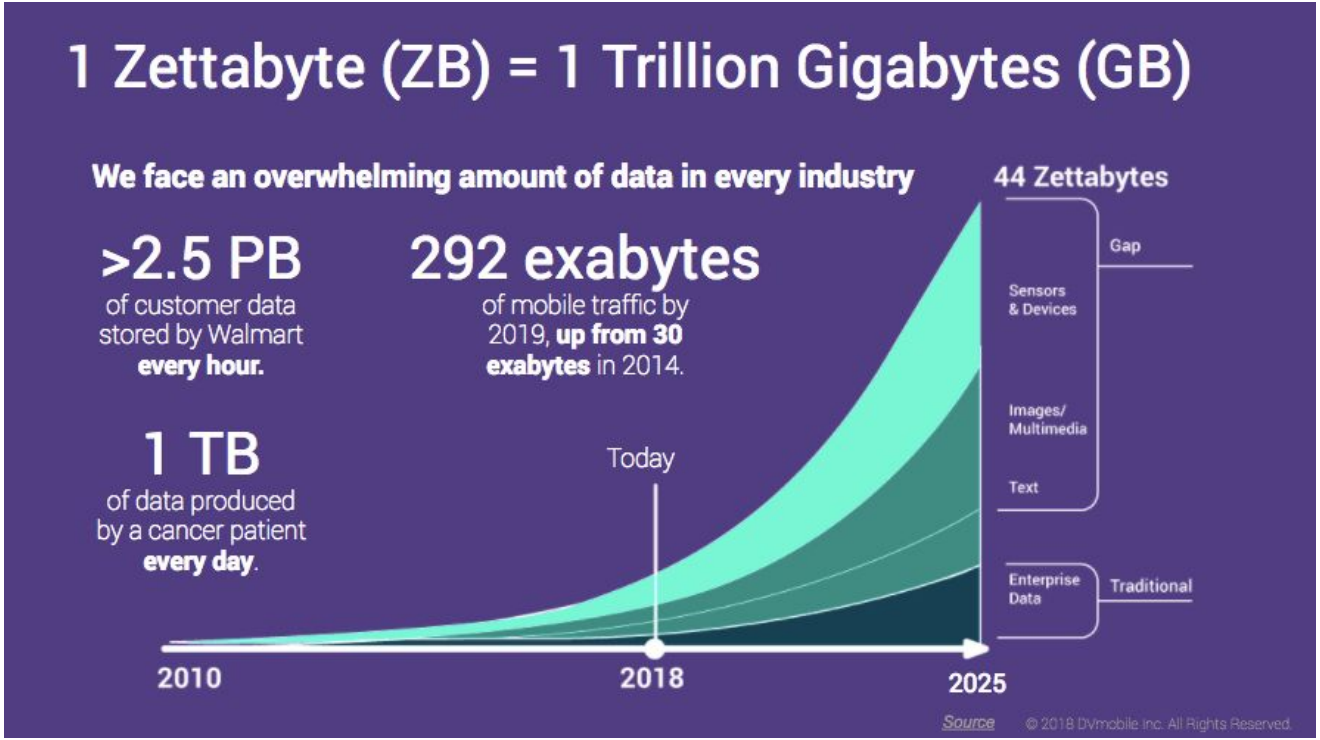
Dr. Sansiri Tarnpradab

Department of Computer Engineering
King Mongkut's University of Technology Thonburi

Outline

- Text data and their impact
  - Growth
  - Benefits
  - Applications

- This course
  - Overview
  - Syllabus



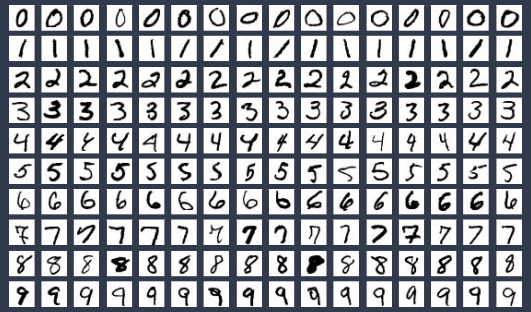
Ref: https://dcencompass.com.au/it-solutions-sydney/data-backup-solutions/

Growth of Data

1 Bit	Binary digit (0/1)	**1024 GB**	1 TB (**Terabyte**)
8 Bits	1 **Byte**	**1024 TB**	1 PB (**Petabyte**)
1024 Bytes	1 KB (**Kilobyte**)	**1024 PB**	1 EB (**Exabyte**)
1024 KB	1 MB (**Megabyte**)	**1024 EB**	1 ZB (**Zettabyte**)
1024 MB	1 GB (**Gigabyte**)	**1024 ZB**	1 YB (**Yottabyte**)

Units in Computer

Different Types of Data



Ref: https://en.wikipedia.org/wiki/MNIST_database

Ref: https://cv.gluon.ai/build/examples_datasets/imagenet.html

Traditional Data (Tabular)

Image

Audio

Video

Text

Multimedia

Sensor

Devices

IMDb Dataset - From 1888 to 2023
Ref: https://www.kaggle.com/datasets/komalkhetlani/imdb-dataset

Structured

vs

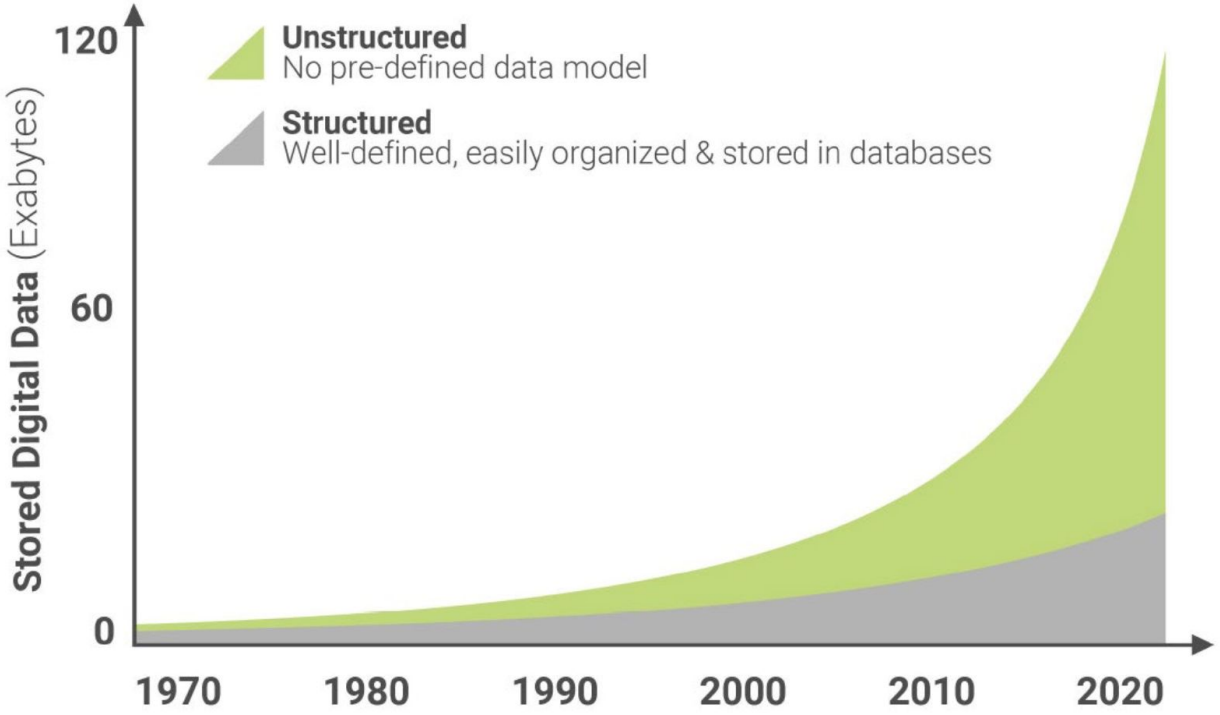
Unstructured

Structured:

- Follows a predefined/organized format
- Tabular data, .csv, spreadsheets

Unstructured:

- No predefined/organized format
- Images, Videos, Text, Posts



Growth of Structured vs Unstructured data

Text Data:

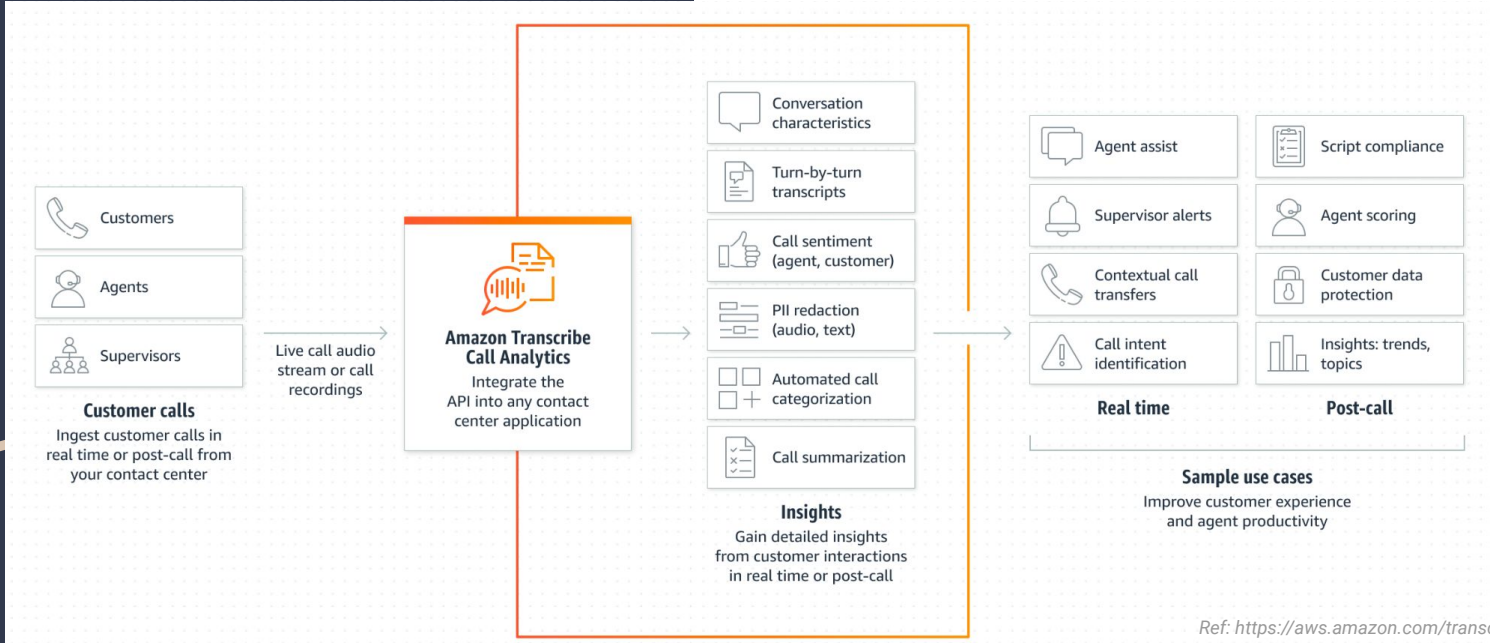
Applications

- Emails
- Chat logs
- Surveys
- Call center
- Voice transcriptions
- Meeting minutes
- Social media (Facebook, X, etc)
- Customer/Member transactions
- Reviews
- … and more

Gain Insights from

Call Center Notes

- What the customer wants
- How the customer feels (sentiment analysis)
- Q&A
- Categorize activities based on calls
  - Agent introduction
  - Account cancellation
  - Competitor mentioning



Ref: https://aws.amazon.com/transcribe/call-analytics/

Gain Insights from Online Reviews

- Sentiment Analysis
- Organic evaluation
- Understand customer emotions
- Make informed decisions
- Improve business growth

Ref: https://www.zendesk.com/th/blog/customer-reviews/#georedirect

Ref: https://hyken.com/customer-experience/guest-blog-how-much-does-negative-review-hurt-business/

Text **Mining**

Text **Analytics**

Same thing?

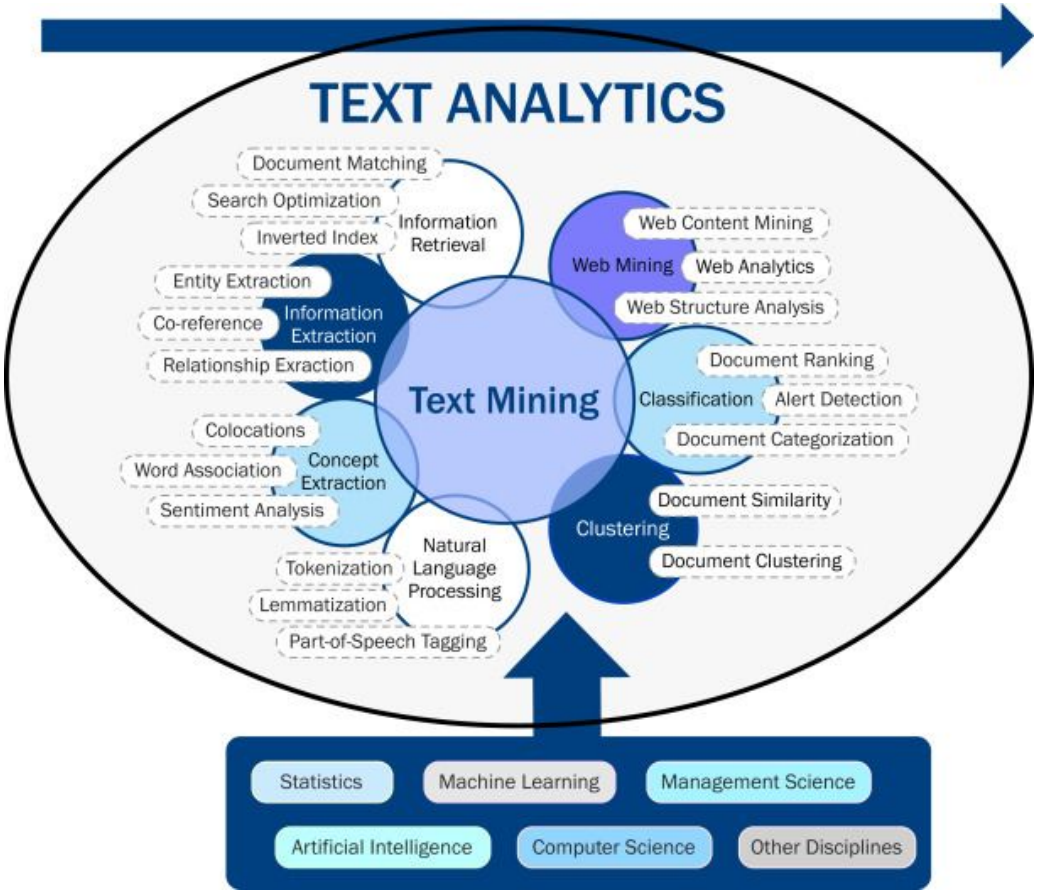
Text Mining vs Text Analytics

Text Mining

- Aka Text Data Mining
- Text Mining = Data Mining + Text Data
- Discover patterns, relationships, knowledge from textual data
- Extract valuable information
- Topic modeling, Clustering, etc

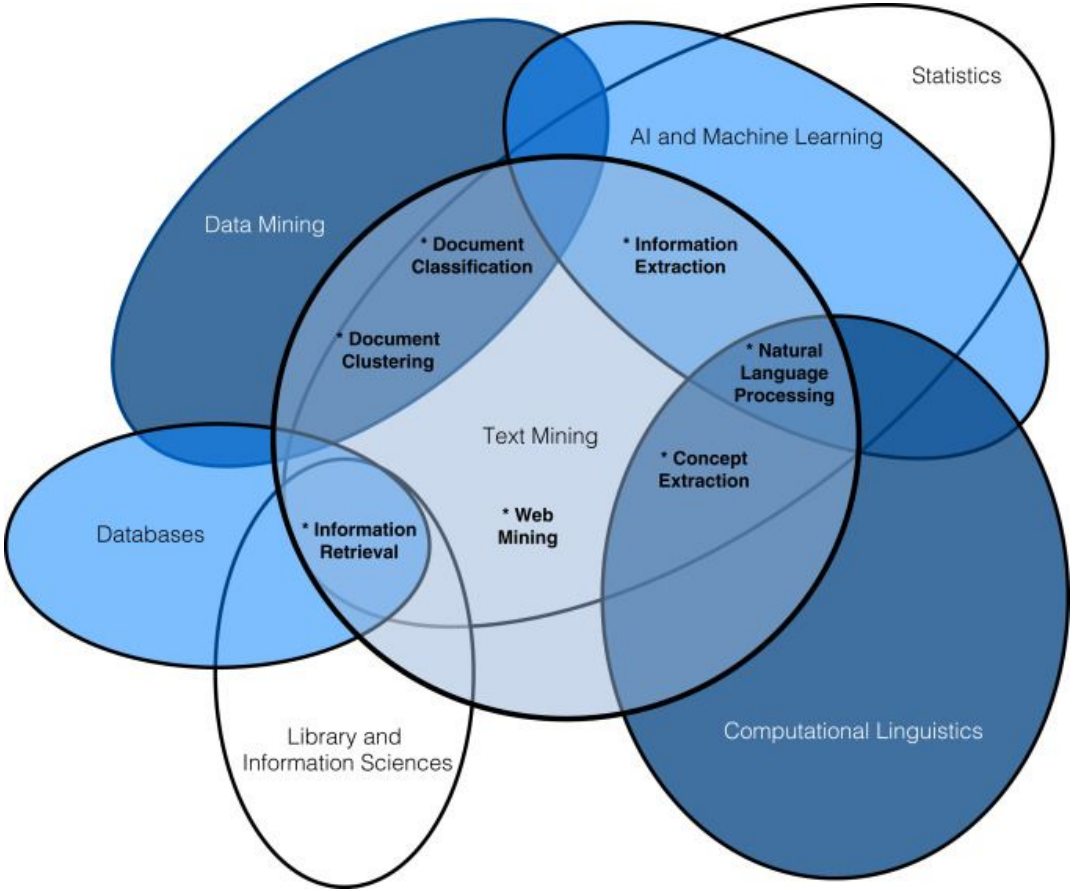
Text Analytics

- Broader term
- Keyword: Analytics
- Use results from analysis from Text Mining models to create data visualizations



Ref: Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, 2012

Text Analytics



Ref: [Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, 2012]

Text Mining

This Course

This Course:

Course Learning Outcomes (CLOs)

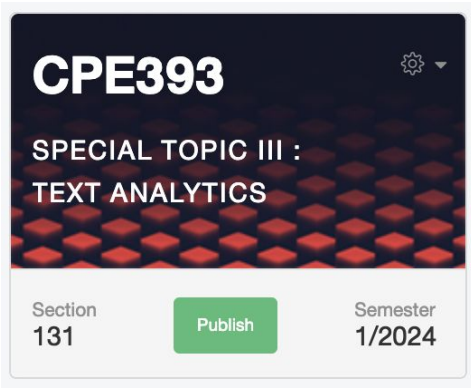
Students can...

- Use suitable methods to process and analyze the text data
- Understand and use text mining to address the requirements

This Course:

Platform & Grading

LEB2



Grading

Lab/Homework	20%
Quiz	20%
Midterm Exam	20%
Final Exam	20%
Project & Presentation	20%

This Course:

Project

- Weight of 20%
- Group Work (Number of members is TBA)
- To gain hands on experience – applying what have been studied in the project
- Select dataset of your choice
- Present your understanding, results, insights by the end of the semester
- Deliverables:
  - Proposal (1page)
  - Report
  - Presentation

This Course:

Policies

Honesty:

The Computer Engineering Department's honesty policy will be strictly enforced. Any assigned work including lab work, if copied with permission, all persons involved will receive a negative score equivalent to the full score of the assigned work for first violation; a second violation will result in F for the course.

Late:

Student are given up to one week to complete the labs/assignments. For each day late, 10% will be deducted.

Work submitted after 5 days past the original due date will not be accepted and receive a zero.

This Course:

Schedule

Up until Midterm

Week	Topics	Remarks
[1] 07/08	Introduction to Text Analytics	
[2] 14/08	Pattern Matching	
[3] 21/08	Textual Data Visualization	
[4] 28/08	Web Scraping	
[5] 04/09	Textual Data Preparation for Analytics	
	Exam I Period - NO CLASS	
[6] 18/09	Textual Feature Representation	
[7] 25/09	**Midterm Exam**	In-class

This Course:

Schedule

Post Midterm

Week	Topics	Remarks
[8] 02/10	Text Classification	Proposal due
[9] 09/10	Text Clustering & Topic Modeling	
[10] 16/10	Extractive Summarization	
	National Holiday - NO CLASS	
[11] 30/10	Abstractive Summarization	Progress report
[12] 06/11	Advanced Topic in Text Analytics	
[13] 13/11		TBA
[14] 20/11	Project Presentation	Project due
[15] 27/11	**Final Exam**	In-class

Q & A