Introduction to Text and Web Mining

Text and Web Mining (H6751)

WKW School of Communication and Information,
NTU

What is Text Mining?

cna

Singapore Edition | Watch TV | CNA938 Live | Sign in | All Sections

Business

Predicting the next US recession

A protracted trade war between China and the United States, the world's largest economies, and a deteriorating global growth outlook has left investors apprehensive about the end to the longest expansion in American history.

Source: https://www.channelnewsasia.com/news/business/predicting-the-next-us-recession-11806122

News Headline

NEW YORK: A protracted trade **war** between **China** and the **United States**, the **world's largest economies**, and a deteriorating global growth outlook has left investors apprehensive about the end to the **longest expansion in American history**. The **recent rise in U.S.-China trade war tensions** has brought **forward the next U.S. recession**, **according to a majority of economists polled by Reuters** who now expect the Federal Reserve to cut rates again in September and once more next year. Trade tensions have pulled corporate confidence and global growth to multi-year lows and **U.S. President Donald Trump's announcement of more tariffs** have raised downside risks significantly, **Morgan Stanley analysts said in a recent note**.

How to handle high volume of text?

Business

After another cut in Singapore's GDP forecast, what could happen next? Experts weigh in

Asia

Family confirms body found in Malaysia rainforest is missing Irish teenager

Singapore

Kayak belonging to missing Singaporeans found as search operations enter fifth day in Malaysia

Asia

Scoot offers full refund, rebooking options for those flying between Hong Kong and Singapore

Lifestyle

Anxious? Depressed? Indigestion? Experts say kimchi or yoghurt can help

Asia

China hits back at UN rights boss over Hong Kong remarks

https://corenlp.run/

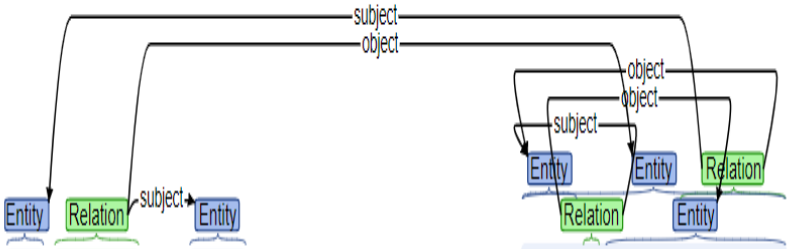
Named Entity Recognition:

1 NEW YORK (STATE_OR_PROVINCE) : A protracted trade war (CAUSE_OF_DEATH) between China (COUNTRY) and the United States (COUNTRY), the world 's largest economies , and a deteriorating global growth outlook has left investors apprehensive about the end to the longest expansion in American (NATIONALITY) history .

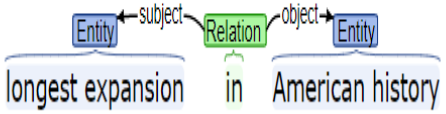
2 The recent rise in U.S.-China trade war (CAUSE_OF_DEATH) tensions has brought forward the next U.S. (COUNTRY) recession , according to a majority of economists polled by Reuters (ORGANIZATION) who now (PRESENT_REF, DATE) expect the Federal Reserve (ORGANIZATION) to cut rates again in September (DATE, 2019-09) and once (PAST_REF, DATE) more next year (DATE, 2020) .

3 Trade tensions have pulled corporate confidence and global growth to multi-year lows and U.S. (COUNTRY) President (TITLE) Donald Trump (PERSON) 's announcement of more tariffs have raised downside risks significantly , Morgan Stanley (ORGANIZATION) analysts said in a recent note .

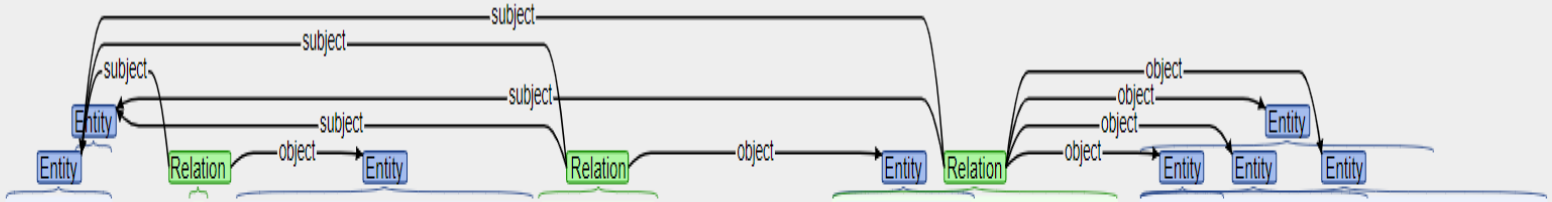
Open IE:



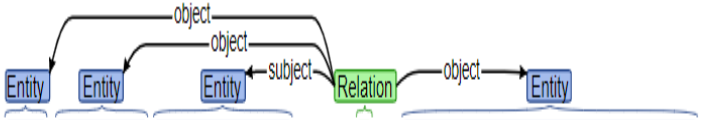
1 NEW YORK : A protracted trade war between China and the United States , the world 's largest economies , and a deteriorating global growth outlook has left investors



apprehensive about the end to the longest expansion in American history .



2 The recent rise in U.S.-China trade war tensions has brought forward the next U.S. recession , according to a majority of economists polled by Reuters who now expect the Federal Reserve to cut rates again in September and once more next year .



3 Trade tensions have pulled corporate confidence and global growth to multi-year lows and U.S. President Donald Trump 's announcement of more tariffs have raised

downside risks significantly , Morgan Stanley analysts said in a recent note .

What is Text Mining?

- Is finding **interesting regularities** in large **textual** dataset.
  - Where **interesting** means non-trivial, hidden, previously unknown and potentially useful.
  - E.g., extract **relations** between drugs and diseases.
  - E.g., stress is associated with migraines; stress can lead to loss of magnesium -> magnesium deficiency may cause migraine headache.
- Is finding semantic and abstract information from the surface form of textual data.
  - E.g., predict sentiments towards products
- The International Data Corporation estimated that approximately **80%** of the data in an organization is **text based**.
- Text Mining is also called **Text Analytics**.

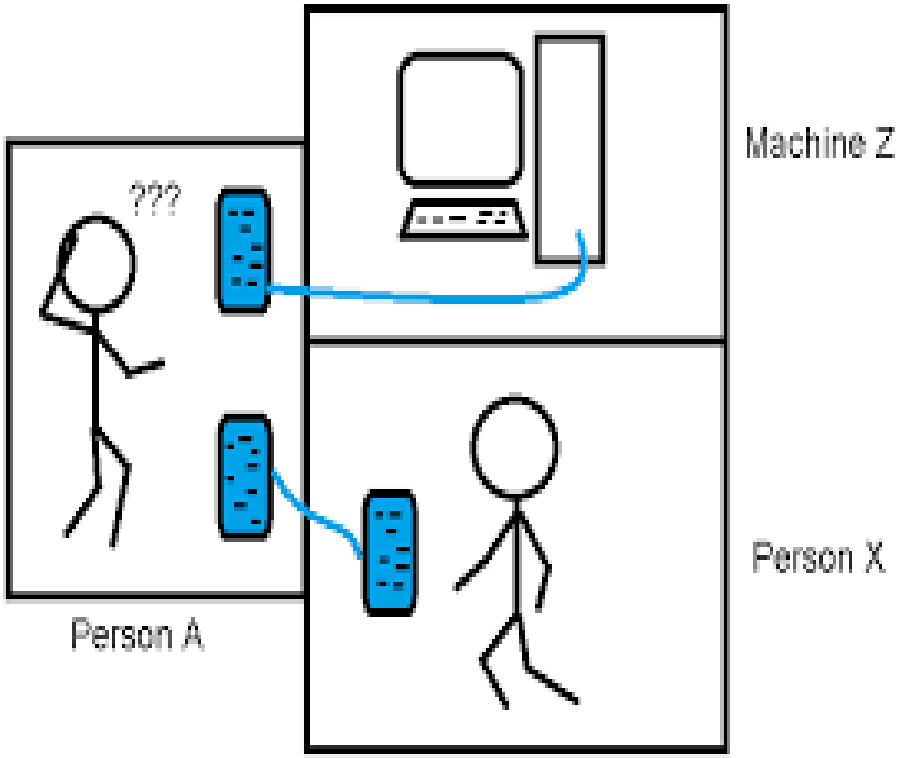
A bit of History

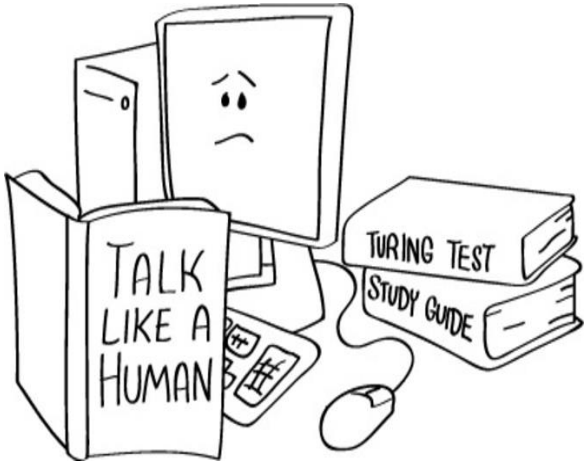
- Alan Turing (1912 – 1954)
  - Helped to break Enigma codes
  - "Father of AI" – Turing Test

Enigma machine

- **Can machines think?** (Comprehend Text)
  - The computer passes the test if a human interrogator, after posing some written questions, cannot tell if whether the written responses come from a person or not (Alan Turing, 1950)

Turing Test


???
Machine Z
Person X
Person A

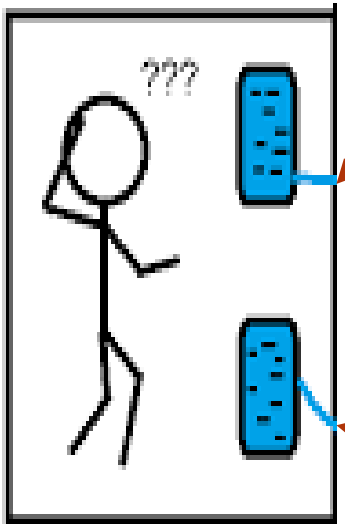

TALK LIKE A HUMAN
TURING TEST
STUDY GUIDE

Let's try the Turing Test

Machine or Human?

Question

What is your favorite subject in school?



A

My friends and teachers say I am totally stupid and untalented, so I like only music lessons, because music teacher tells me only that I'm tone-deaf and at least doesn't call me stupid.

B

My favorite subject is Science because I love doing experiments.

On 7 June 2014 a Turing test competition, organised by Huma Shah and Kevin Warwick to mark the 60th anniversary of Turing's death, was held at the Royal Society London and was won by the Russian chatter bot Eugene Goostman. The bot, during a series of five-minute-long text conversations, convinced 33% of the contest's judges that it was human

Morning Mix

A computer just passed the Turing Test in landmark trial

By Terrence McCoy June 9, 2014

Can machines think?

In 1950, famed London scientist Alan Turing, considered one of the fathers of artificial intelligence, published a paper that put forth that very question. But as quickly he asked the question, he called it "absurd." The idea of thinking was too difficult to define. Instead, he devised a separate way to quantify mechanical "thinking."

Alan Turing from archive of papers relating to the development of computing at the National Physical Laboratory between the late 1940s and the early 1970s. (Science Museum, London/SSPL)

"I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words," he wrote in the

- **Can machines think?**
  - The computer passes the test if a human interrogator, after posing some written questions, cannot tell if whether the written responses come from a person or not (Alan Turing, 1950)

- **Requires**
  - Natural language processing
  - Knowledge representation
  - Automated reasoning
  - Machine learning

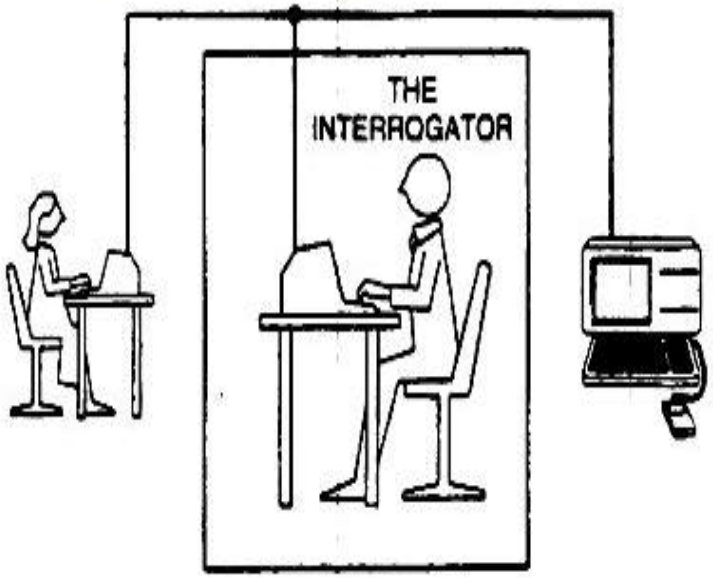
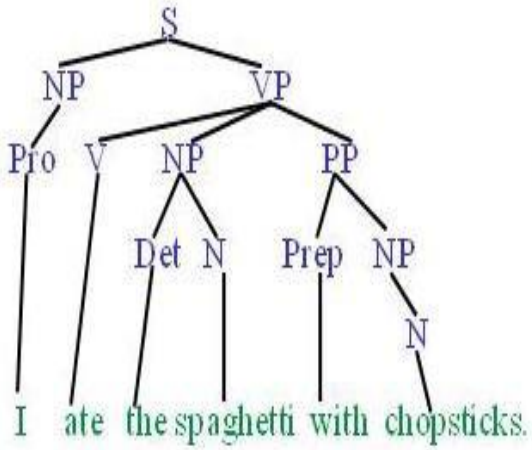


Figure 1.1 The Turing test.

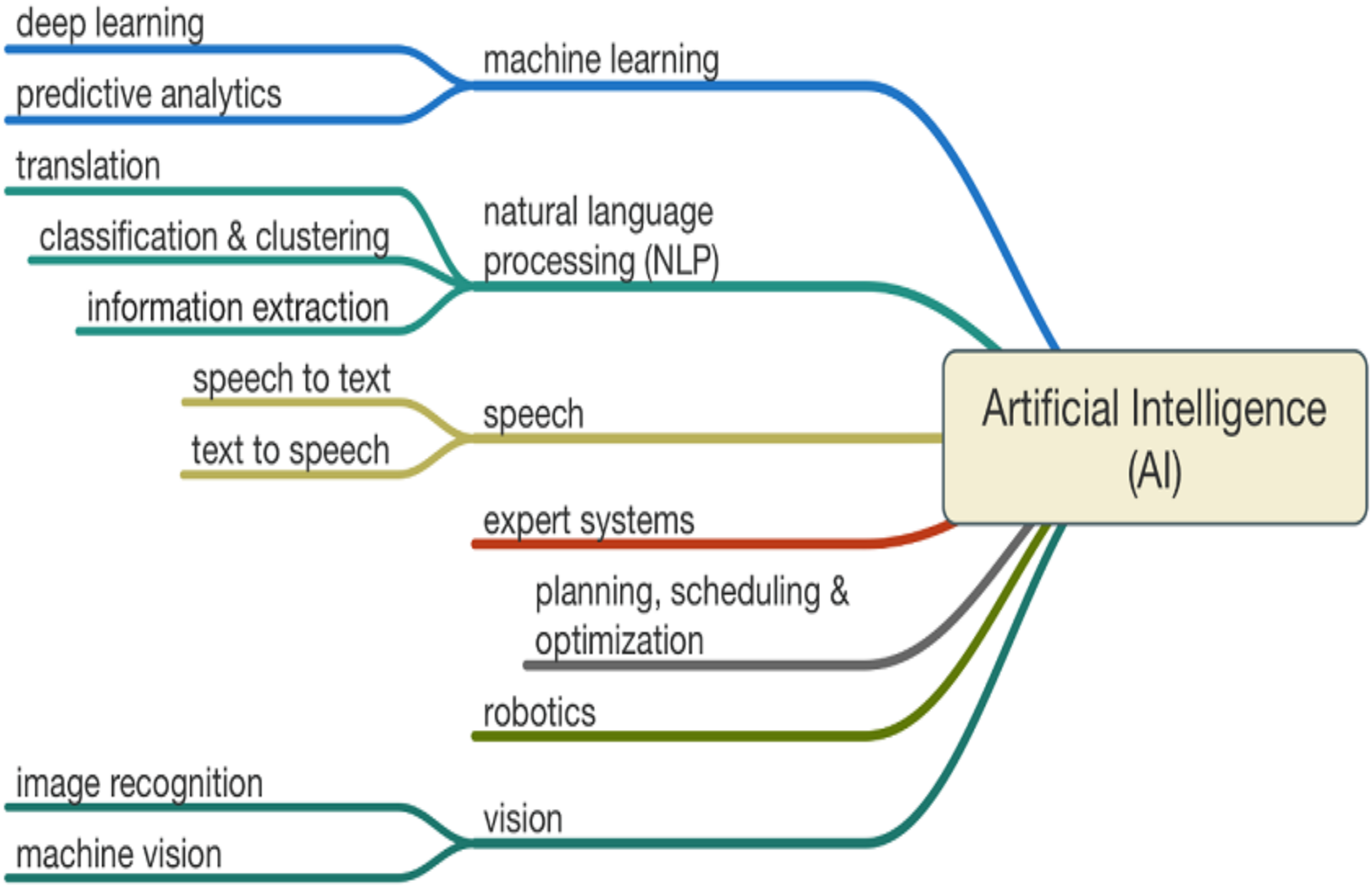
Which areas are related to Text Mining?

- **Data Mining**
  - Structured Data Analysis – numerical/categorical data, leverage on statistics/algorithms for discovery of unknown patterns
- **Machine Learning**
  - Prediction – focus on reproducing results from training data
- **Natural Language Processing**
  - Computational Linguistics – relate mathematical concepts to human languages
- **Information Retrieval**
  - Search & full-text indexing
- **Knowledge Representation and Reasoning**

  (e.g., *born-in(Albert Einstein, Ulm Germany)*)

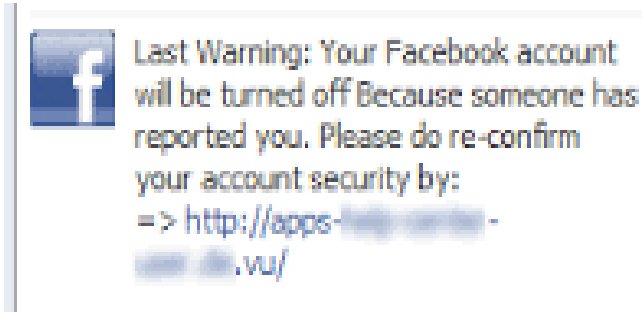


Text Mining in AI



Applications of Text Mining?

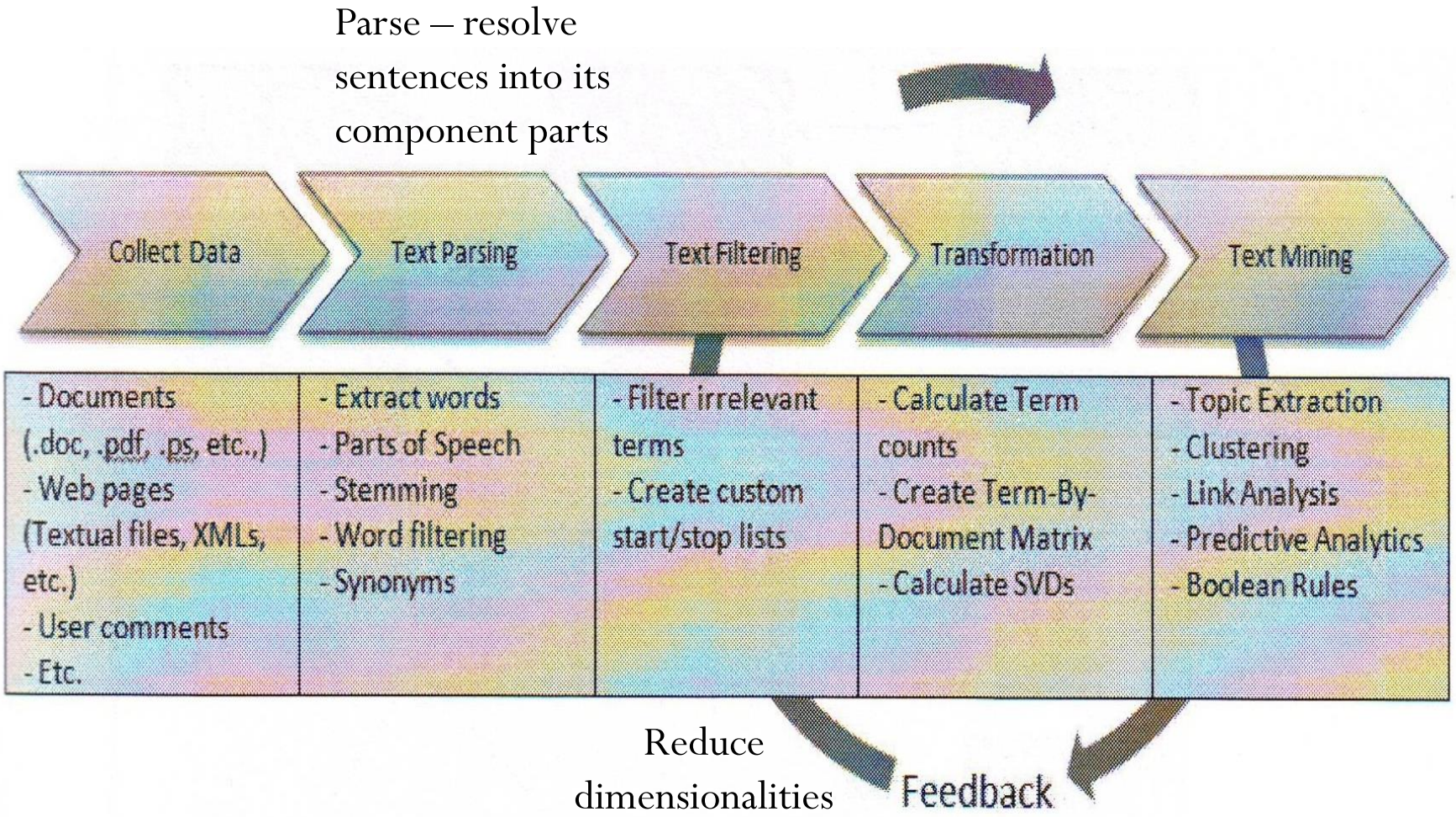
- Discovering trends in textual data in the business environment
- Sentiment analysis
- Crime prevention and fraud detection using social media data



- Improving patient outcomes and providing better care in hospitals
  - Watson, a Q&A system, used in hospitals.
- Mining biomedical literature to discover new drugs in pharmaceutical industry.
  - Also discover side effects and adverse drug reactions (ADRs)

Text Mining Process Flow

- A typical text mining project involves 5 steps.



Why Text is Tough? (M. Hearst 97)

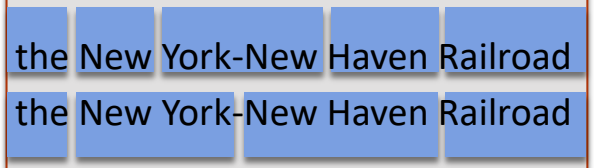
- **Many ways** to represent similar concepts
  - Synonyms – words with **same meaning** (e.g., space ship, flying saucer, and UFO)
  - Polysemy – word with **multiple meanings** (e.g., you were right, make a right turn, human right)
  - Dependent on **context**
- **"Countless" combinations** of subtle, abstract **relationships** among concepts.
  - E.g., relationships between drugs and diseases (concepts of treating diseases)
- **High dimensionality**
  - Tens or hundreds of thousands of features

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues



idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

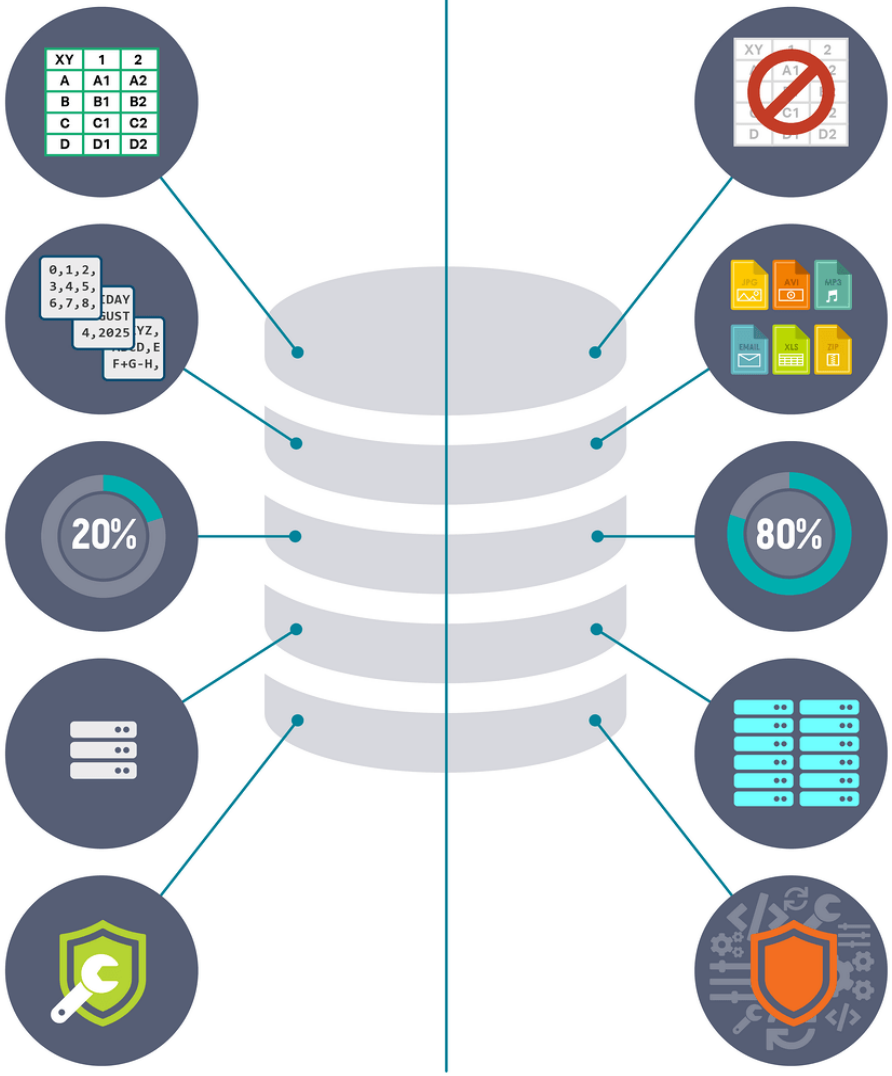
Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …
Let It Be was recorded …

Structured Data vs Unstructured Data

Structured Data	Unstructured Data
Can be displayed in rows, columns and relational databases	Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings	Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 20% of enterprise data *(Gartner)*	Estimated 80% of enterprise data *(Gartner)*
Requires less storage	Requires more storage
Easier to manage and protect with legacy solutions	More difficult to manage and protect with legacy solutions



Structured or Unstructured Data?

- The **text** is usually a collection of **unstructured documents** with no special requirements for composing the documents.
- In **data mining** applications, the data must be prepared in a very special way (e.g., a spreadsheet format) before any learning methods can be applied.
  - Two types of information are expected: (a) ordered numerical and (b) categorical.

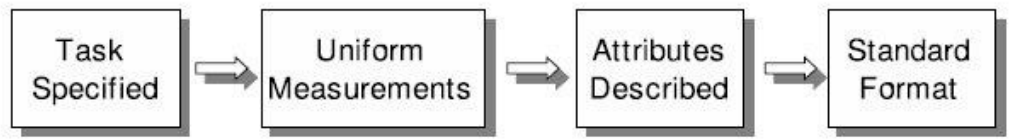


Fig. 1.1 Structured data in standard format

Fig. 1.2 A spreadsheet example of medical data

Gender	Systolic BP	Weight	Disease Code
M	175	65	3
F	141	72	1
...	...	...	...
F	160	59	2

Is Text Different from Numbers?

- One of the main themes supporting text mining is **the transformation of text into numerical data**.
- Although the initial presentation is document format, the data move into a classical data-mining encoding, a spreadsheet format.
- The unstructured data become **structured**.
- Each row represents a document and each column a word.

Company	Income	Job	Overseas
0	1	0	1
1	0	1	1
1	1	1	0
0	0	0	1

Fig. 1.3 A binary spreadsheet of words in documents

Is Text Different from Numbers? (cont.)

- The matrix is **sparse**.
  - An individual document will use only a tiny subset of the potential set of words in **a dictionary**, which is the total set of unique words in the collection.
  - Text mining methods mostly concentrate on **positive matches**, not worrying whether other words are absent from a document.
  - For text, **missing values** are a nonissue: words are either present or absent from a document.

DocID	Apple	Bear	Durian	…	…	…	…	…	Zoo	Animal?
1	0	1	0	0	0	0	0	0	1	1
2	1	0	1	0	0	0	0	0	0	0
…										

What types of problems can be solved?

- **Document Classification**
  - Given a sample of documents and **correct answers (text categories)** for each document, the objective is to find the correct answers for new documents.
  - The spreadsheet model **with one column corresponding to the correct answer** is the universal classification model for data, and the transformed text data can readily be combined with standard numerical data mining data.
  - The application is almost always **binary classification** because a document can usually appear in multiple folders.
  - E.g., automatically forwarding e-mail to the appropriate company department (Y/N) or detecting spam email (Y/N).

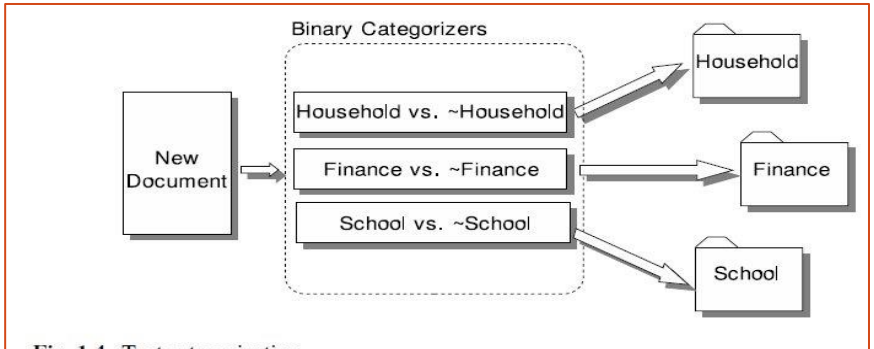


Fig. 1.4 Text categorization

What types of problems can be solved?

- **Document Clustering**
  - For document classification (or text categorization), the objective is to place new documents into the **predefined categories**.
    - E.g., spam detection and news articles categorization
  - Clustering is used when we have a collection of **documents with no known structure**.
    - E.g., Email complaints by users are clustered, and can learn about the categories and types of complaints.

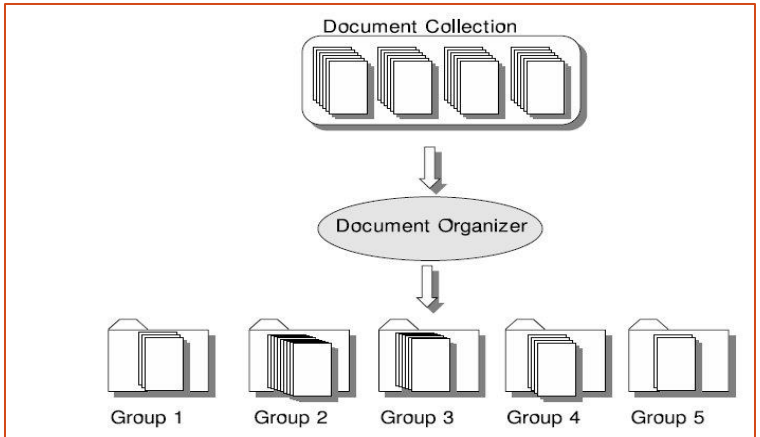


Fig. 1.6 Organizing documents into groups

  - Because there are many ways to cluster documents (group documents with similar features), it is not quite as powerful as assigning answers (i.e., known correct labels) to documents.

What types of problems can be solved?

- **Information Retrieval**
  - Instead of a few words used in a search engine, **a complete document** is presented as a set of clues.
  - The input document is then matched to all stored documents, retrieving the best-matched documents.
  - A basic concept for IR is **measuring similarity**: a comparison is made between two documents, measuring how similar the documents are.

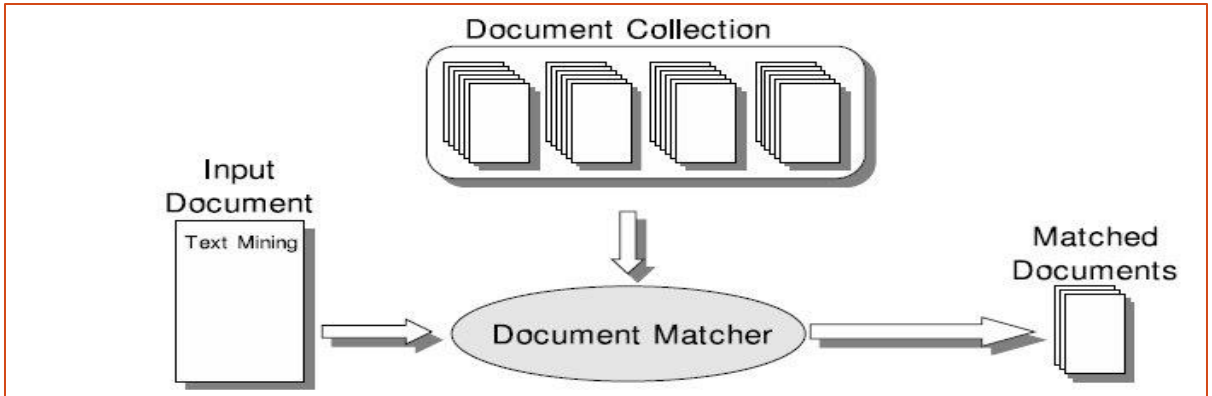


Fig. 1.5 Retrieving matched documents

What types of problems can be solved?

- An example of Document Clustering: Consider the **comments** made by **patients** about the best thing they liked about the hospital.

1. ***Friendliness*** *of the* ***doctor*** *and* ***staff***
2. ***Service*** *at the eye* ***clinic*** *was* ***fast***.
3. *The* ***doctor*** *and other people were very, very* ***friendly***.
4. *Waiting time has been excellent and* ***staff*** *has been very* ***helpful***.
5. *The way the* ***treatment*** *was done.*
6. *No hassles in* ***scheduling*** *an appointment.*
7. *Speed of the* ***service***.
8. *The way I was treated and my* ***results***.

Table 1.2: Clustering Results from Text Mining

Cluster No.	Comment	Key Words
1	1, 3, 4	doctor, staff, friendly, helpful
2	5, 6, 8	treatment, results, time, schedule
3	2, 7	service, clinic, fast

What types of problems can be solved?

- **Information Extraction**
  - One of the objectives is to take an unstructured document and **automatically fill in the values of a spreadsheet**.
  - In a spreadsheet, the columns are not just words but can be **higher-level concepts** that are found by the information extraction process.
    - E.g., people, organizations, places, addresses, dates, times, etc.

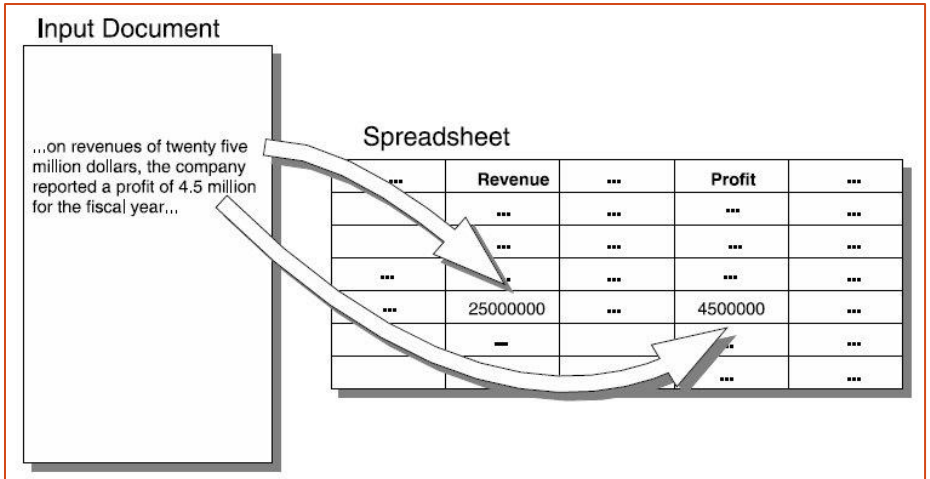
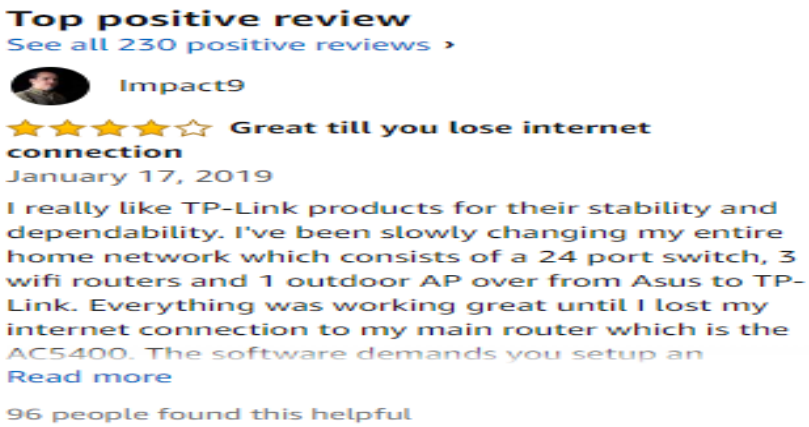
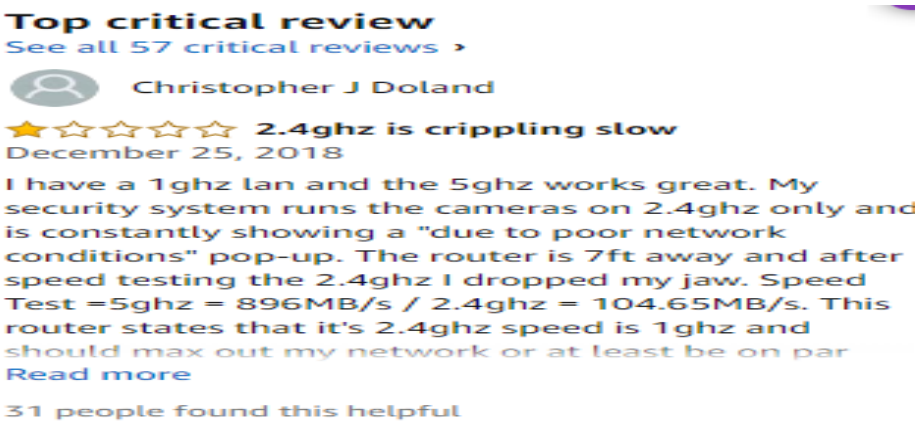


Fig. 1.7 Extracting information from a document

What types of problems can be solved?

- **Sentiment Analysis**
  - The rapid growth of ***user-generated content***, called ***social media*** - Weblogs, Discussion Boards, User and Critic Review Web sites, Twitter, Facebook, etc.
    - **Online shoppers** are influenced by product reviews and are willing to pay more for products highly rated by other consumers.

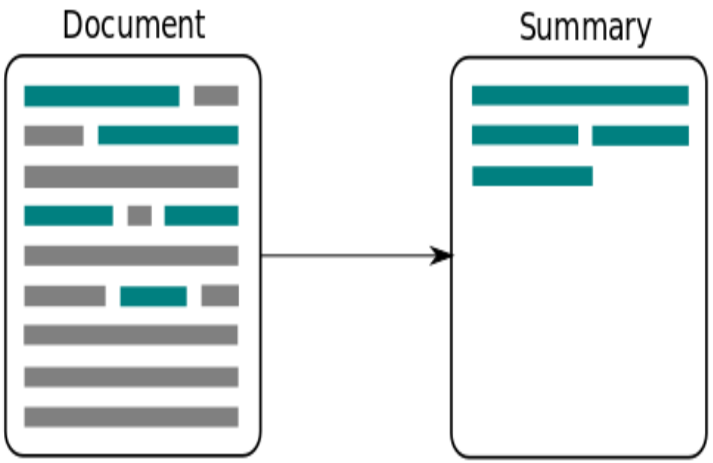




  - ***Opinion mining*** or ***sentiment analysis***
    - A type of subjectivity analysis which analyzes sentiment in a given textual unit with the objective of understanding the ***sentiment polarities (i.e. positive, negative, or neutral)*** of the opinions toward various aspects of a subject.

What types of problems can be solved?

- **Text Summarization**
  - **Task**: the task is to produce shorter, summary version of an original document.
  - Two main approaches to the problem:
    - **Selection based**
      - Output consists from topmost (frequency-based) text units (sentences).
    - **Knowledge rich**–performing semantic analysis, representing the meaning and generating the text satisfying length restriction
      - Latent Dirichlet Allocation



What types of problems can be solved?

- **Emerging Directions**
  - Handling big (text) data
    - Challenges posed by the three Vs: **Variety, Velocity, and Volume**
    - Unstructured data will occupy 90% of the data.
    - *Apache Hadoop* is a framework for storage and large-scale processing big data on clusters of machines.
    - Kubernetes – containers orchestration
  - Voice mining
    - In Call centers, each voice call can be analyzed to predict the customer's likelihood to cancel or close the account.
      - call length, emotion, stress detection, number of transfers, etc.
  - Real-time text analytics
    - Need to address data that is streaming continuously on social media, such as Twitter.
    - E.g., Governments predict medical epidemics, terrorist attacks, etc.
    - E.g., Companies analyze their customers' negative comments about their brand or products.

Natural Language Processing Technology

making good progress

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC
Einstein met with UN officials in Princeton		

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Topics Covered

- Preprocessing for Text Mining
- Text Classification
- Text Clustering
- Information Extraction
- Opinion Mining and Sentiment Analysis
- Text Classification using Deep Learning
- Text Mining Tools
  - Python, nltk, scikit-learn, Keras, etc.

Python Programming

```
from nltk.tokenize import word_tokenize

input_str = "NLTK is a leading platform for building Python programs to work with
tokens = word_tokenize(input_str)
print (tokens)
```

```
['NLTK', 'is', 'a', 'leading', 'platform', 'for', 'building', 'Python', 'progr
ams', 'to', 'work', 'with', 'human', 'language', 'data', '.']
```

```
from nltk.tokenize import TreebankWordTokenizer
s = '''Good muffins cost $3.88\nin New York. Please buy me\ntwo of them.\nThanks.'
print(TreebankWordTokenizer().tokenize(s))
```

```
['Good', 'muffins', 'cost', '$', '3.88', 'in', 'New', 'York.', 'Please', 'bu
y', 'me', 'two', 'of', 'them.', 'Thanks', '.']
```

Assessment Components

- Class Participation (class interactions and attendance) – 10%
- Coursework (individual and group assignments) – 40%
  - Group assignment – team of 3 to 4 (max 12 groups)
  - Team members will receive same marks
  - Assessment criteria (methodology and innovative ideas) – identify and solve text mining related problem
    - Case study
    - Data collection
    - Data pre-processing
    - Analysis and modelling
    - Observation and evaluation of results
    - Innovation/Challenging problem
- Final Examinations (3 hours, close book) – 50%

Contact Information

Dr. Luke Tan Kien Weng

Email: luke.tan@ntu.edu.sg / luketan@hotmail.com

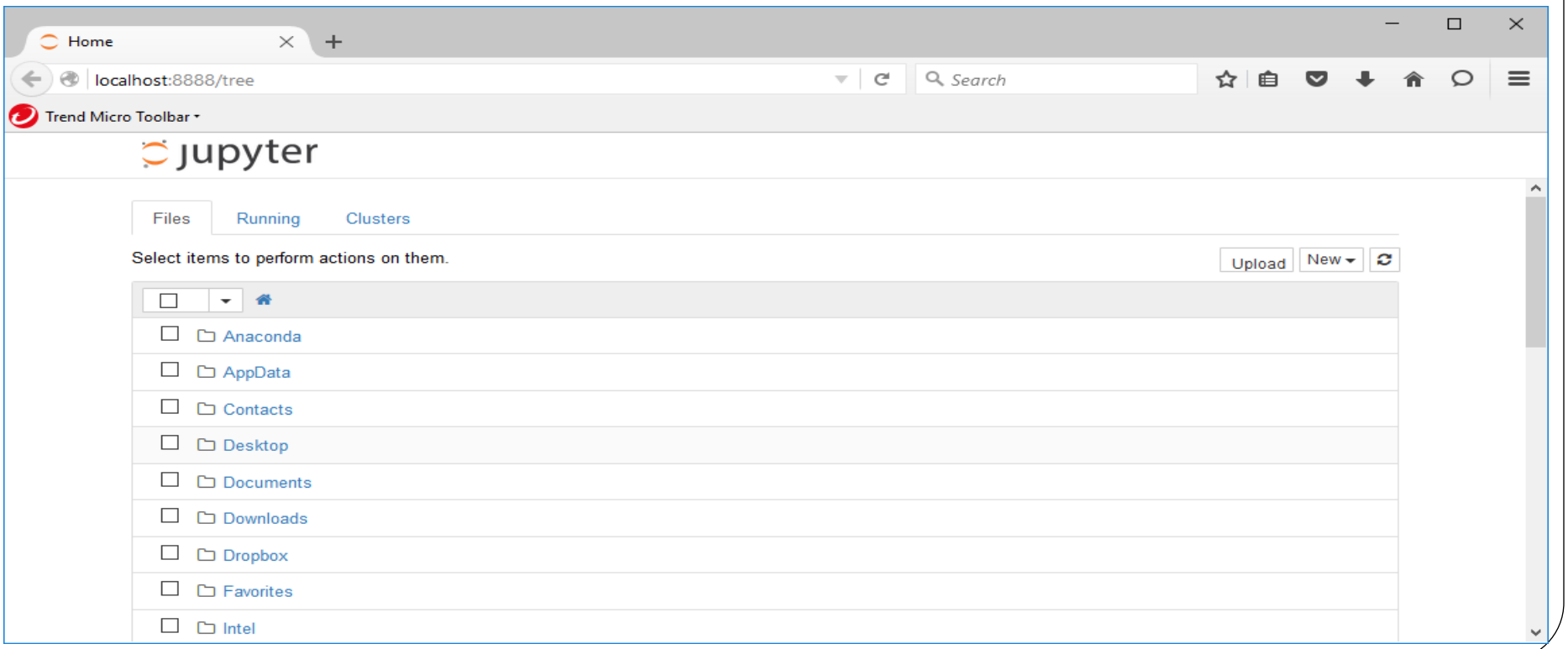
Referenced Materials

- Fundamentals of Predictive Text Mining, Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang, Springer.
  - Chapter 1
- Text-Mining Tutorial, Marko Grobelnik, Dunja Mladenic, J. Stefan Institute, Slovenia
- NLP Introduction, Dan Jurafsky and Christopher Manning, http://www.stanford.edu/~jurafsky/NLPCourseraSlides.html

Introduction to Python and Jupyter

- **1. Install Python tools on a Local Machine.**
- – Anaconda (https://www.continuum.io/downloads) has been installed on lab machines.
- Windows 64-bit Graphical Installer: Anaconda3-x.x.x-Windows-x86_64.exe

- **Launching *Jupyter Notebook App***
- The Jupyter Notebook App can be launched by clicking on the ***Jupyter Notebook*** icon on task bar (or type in "**Jupyter Notebook**" in Window Menu) or by typing in a terminal (*cmd* on Windows): *ipython notebook*
- This will launch a new browser window (or a new tab) showing the Notebook Dashboard, a sort of control panel that allows the selection of which notebook to open.



String Formatting

Let's say you have two strings:

```
>>>name = "Joel"
>>>job = "Programmer"
```

```
>>>title = name + " the " + job
>>>title
>"Joel the Programmer"
```

String Joining

Another nifty Pythonic trick is the **join()** method, which takes a list of strings and combines them into one string. Here's an example:

```
>>>availability = ["Monday", "Wednesday", "Friday", "Saturday"]
>>>result = " - ".join(availability)
>>>result
>'Monday - Wednesday - Friday - Saturday'
```

Boolean Values

Like in all other programming languages, comparison operators evaluate to a boolean result: either **True** or **False**. Here are all the comparison operators in Python:

```
>>>x = 10
>>>print(x == 10) # True
>>>print(x != 10) # False
>>>print(x <> 10) # False, same as != operator
>>>print(x > 5) # True
>>>print(x < 15) # True
>>>print(x >= 10) # True
>>>print(x <= 10) # True
```

The in Operator

If you just want to check if a value exists within an iterable object, like a list or a dictionary, then the quickest way is to use the **in** operator:

```
>>>availability = ["Monday", "Tuesday", "Friday"]
>>>request = "Saturday"
>>>if request in availability:
>>>    print("I'm available on that day!")
```

The is and not Operators

The **==**, **!=**, and **<>** operators above are used to compare the values of two variables. If you want to check if two variables point to the same exact object, then you'll need to use the **is** operator:

```
>>>a = [1,2,3]
>>>b = [1,2,3]
>>>c = a
>>>print(a == b) # True
>>>print(a is b) # False
>>>print(a is c) # True
```

You can negate a boolean value by preceding it with the **not** operator:

```
>>>a = [1,2,3]
>>>b = [1,2,3]
>>>if a is not b:
>>>    # Do something here
```

```
>>>x = False
>>>if not x:
>>>    # Do something here
```

Loops

The most basic type of loop in Python is the **while** loop, which keeps repeating as long as the conditional statement evaluates to True:

```
>>>i = 0
>>>while i < 10:
>>>     print(i)
>>>     i = i + 1
```

This could also be structured like so:

```
>>>i = 0
>>>while True:
>>>     print(i)
>>>     if i >= 10:
>>>     break
```

The **break** statement is used to immediately exit out of a loop. If you just want to skip the rest of the current loop and start the next iteration, you can use **continue**.

The For Loop

The more Pythonic approach is to use **for** loops. The for loop in Python is nothing like the for loop that you'd find in a C-related language like Java or C#. It's much closer in design to the **foreach** loops in those languages.

In short, the for loop iterates over an iterable object (like a list or dictionary) using the **in** operator:

```
>>>weekdays = ["Monday", "Tuesday", "Wednesday", "Thursday", "Friday"
>>>for day in weekdays:
>>>    print(day)
```

How to declare an empty dict:

```
>>>d = {}
```

How to assign a dict key to a value:

```
>>>d = {}
>>>d["one_key"] = 10
>>>d["two_key"] = 25
>>>d["another_key"] = "Whatever you want"
```

Reference https://docs.python.org/3/tutorial/