H6751 Text and Web Mining

Zhao Rui

Course Webpage

Course Instructors

Zhao Rui
(Instructor)

rui.zhao@ntu.edu.sg

Chen Zhenghua
(Instructor)

zhenghua.chen@ntu.edu.sg

Course web: https://h6751.github.io/

Goals of this Course

Learn how to analyse unstructured text data

- Principles and concepts of text and web mining

- Various text mining techniques
  - Pre-processing, text categorization, document clustering, information extraction

- Practical text mining applications
  - Spam detection, sentiment analysis, knowledge graph

Course Assessment

- Class Participation (5%)

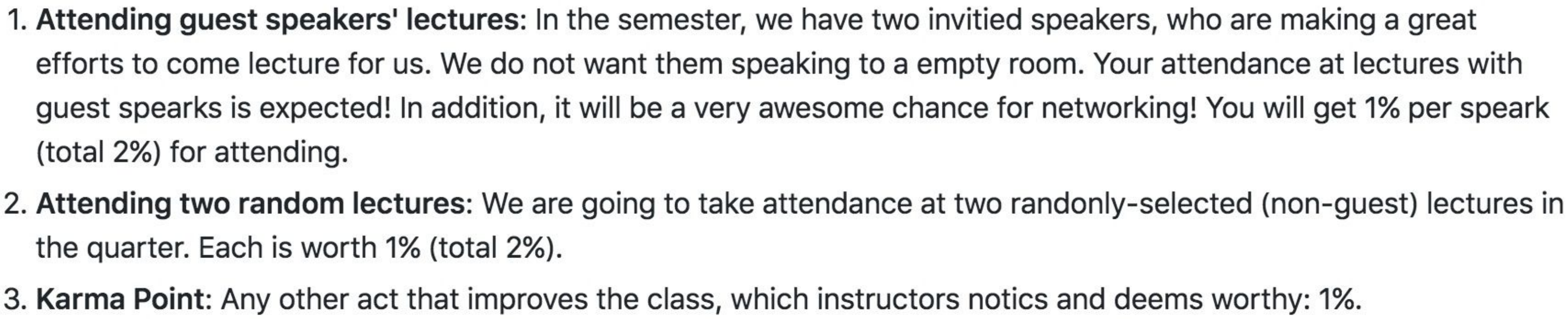
- Assignments:
  - A 90-minutes in-class assignment (15%)

- Group Project (30%)
  - Project Proposal (5%)
  - Final Presentation (10%)
  - Final Report (15%)

- Final Exam(50%)

Course Participation

- Class Participation (5%)

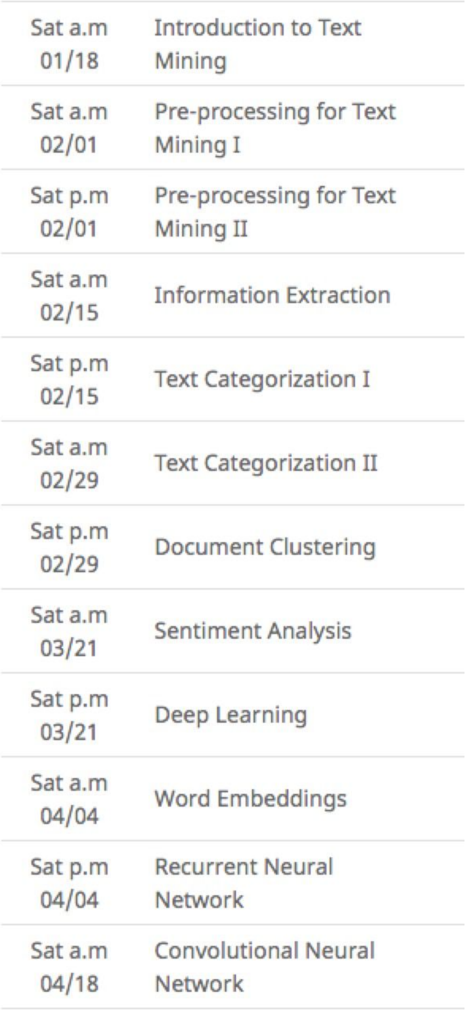


1. **Attending guest speakers' lectures:** In the semester, we have two invitied speakers, who are making a great efforts to come lecture for us. We do not want them speaking to a empty room. Your attendance at lectures with guest spearks is expected! In addition, it will be a very awesome chance for networking! You will get 1% per speark (total 2%) for attending.
2. **Attending two random lectures:** We are going to take attendance at two randonly-selected (non-guest) lectures in the quarter. Each is worth 1% (total 2%).
3. **Karma Point:** Any other act that improves the class, which instructors notics and deems worthy: 1%.

Assignments

- For in-class Assignments, it will be code-based exam. Open-book and Open-Internet.

- Details will be updated before the release of these assignments.

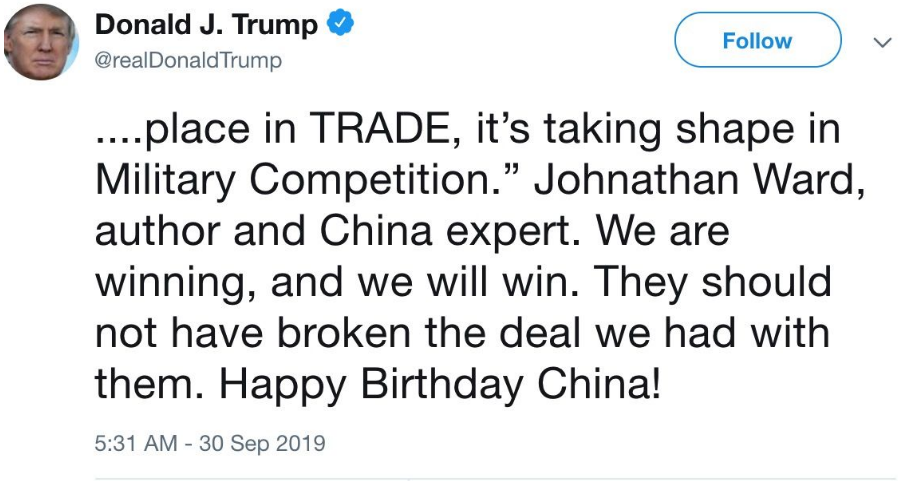


Sat a.m 01/18	Introduction to Text Mining	ZR
Sat a.m 02/01	Pre-processing for Text Mining I	ZR
Sat p.m 02/01	Pre-processing for Text Mining II	ZR
Sat a.m 02/15	Information Extraction	ZR
Sat p.m 02/15	Text Categorization I	CZH
Sat a.m 02/29	Text Categorization II	CZH
Sat p.m 02/29	Document Clustering	CZH
Sat a.m 03/21	Sentiment Analysis	CZH
Sat p.m 03/21	Deep Learning	ZR
Sat a.m 04/04	Word Embeddings	ZR
Sat p.m 04/04	Recurrent Neural Network	ZR
Sat a.m 04/18	Convolutional Neural Network	ZR

Let us Start

Twitter in Chief

- The President not only Make America Great Again, but also Twitter
- He tweets 4178 per year and 11 to 12 per day



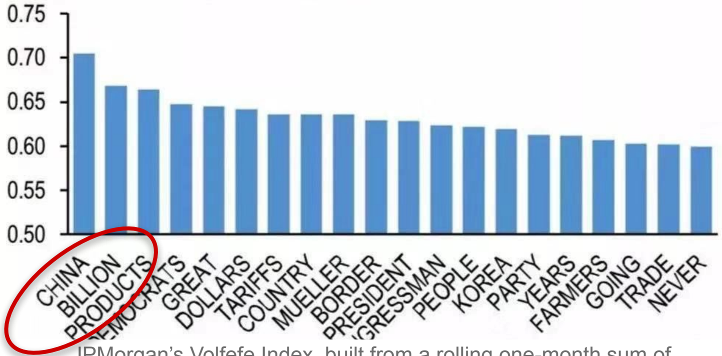
Donald J. Trump
@realDonaldTrump

....place in TRADE, it’s taking shape in Military Competition.” Johnathan Ward, author and China expert. We are winning, and we will win. They should not have broken the deal we had with them. Happy Birthday China!

5:31 AM - 30 Sep 2019

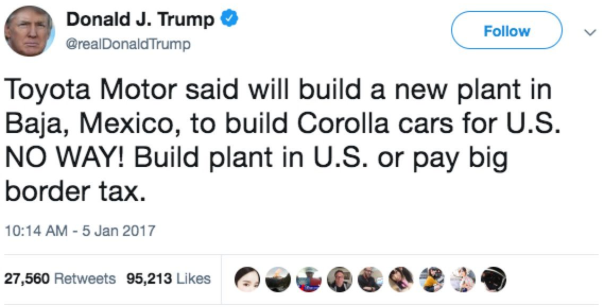
Volfefe Index

- Quantify the market impact of Trump's tweets
- Supervised learning and Natural Language Processing techniques are used to spot "market-moving" tweets
- Volfefe Index can explain moves in implied volatility



JPMorgan's Volfefe Index, built from a rolling one-month sum of inferred probability that each tweet is market moving (Source: Bloomberg)

What is More

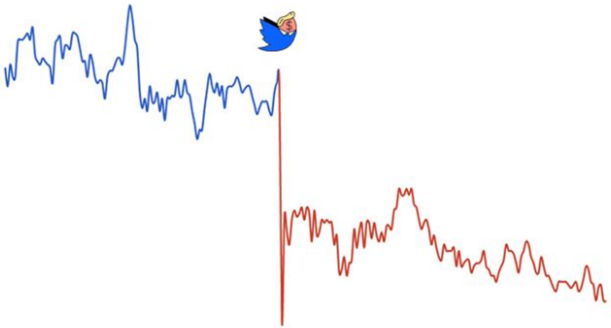


Donald J. Trump
@realDonaldTrump

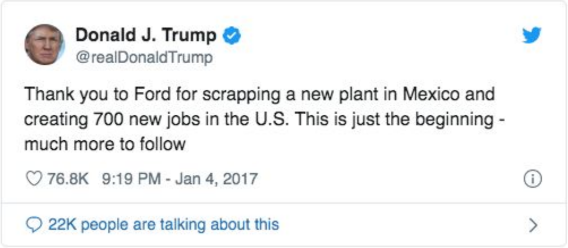
Toyota Motor said will build a new plant in Baja, Mexico, to build Corolla cars for U.S. NO WAY! Build plant in U.S. or pay big border tax.

10:14 AM - 5 Jan 2017

27,560 Retweets 95,213 Likes



Toyota's NYSE:TM stock price on January 5th 2017

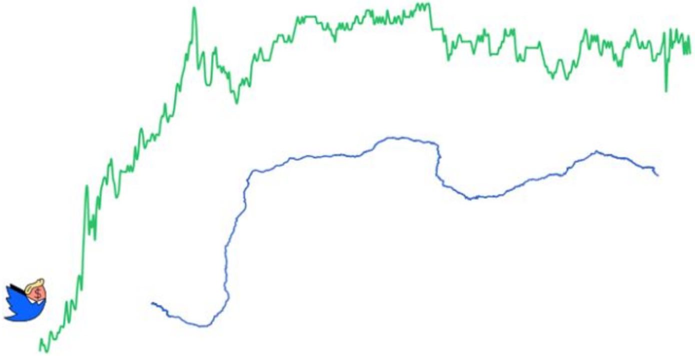


Donald J. Trump
@realDonaldTrump

Thank you to Ford for scrapping a new plant in Mexico and creating 700 new jobs in the U.S. This is just the beginning - much more to follow

♡ 76.8K 9:19 PM - Jan 4, 2017

22K people are talking about this

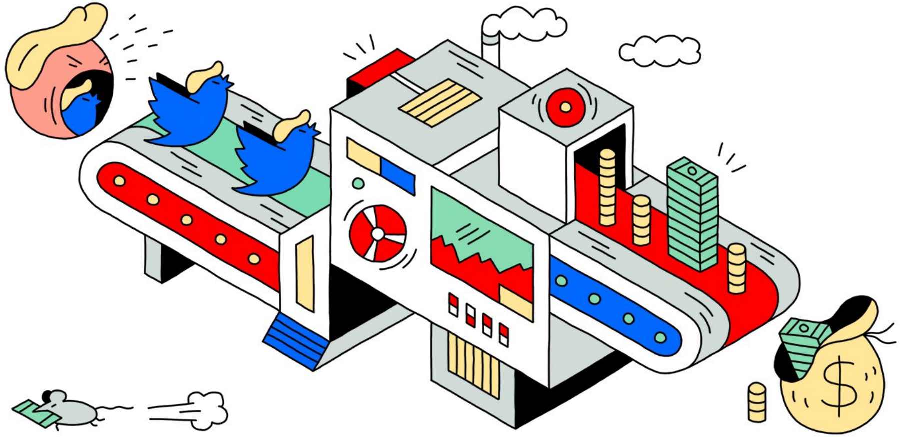


Ford's NYSE:F stock price on January 4th 2017 (Rio Grande for scale)

Source:
https://medium.com/@maxbraun/this-machine-turns-trump-tweets-into-planned-parenthood-donations-4ece8301e722#.yovbh4qc1/

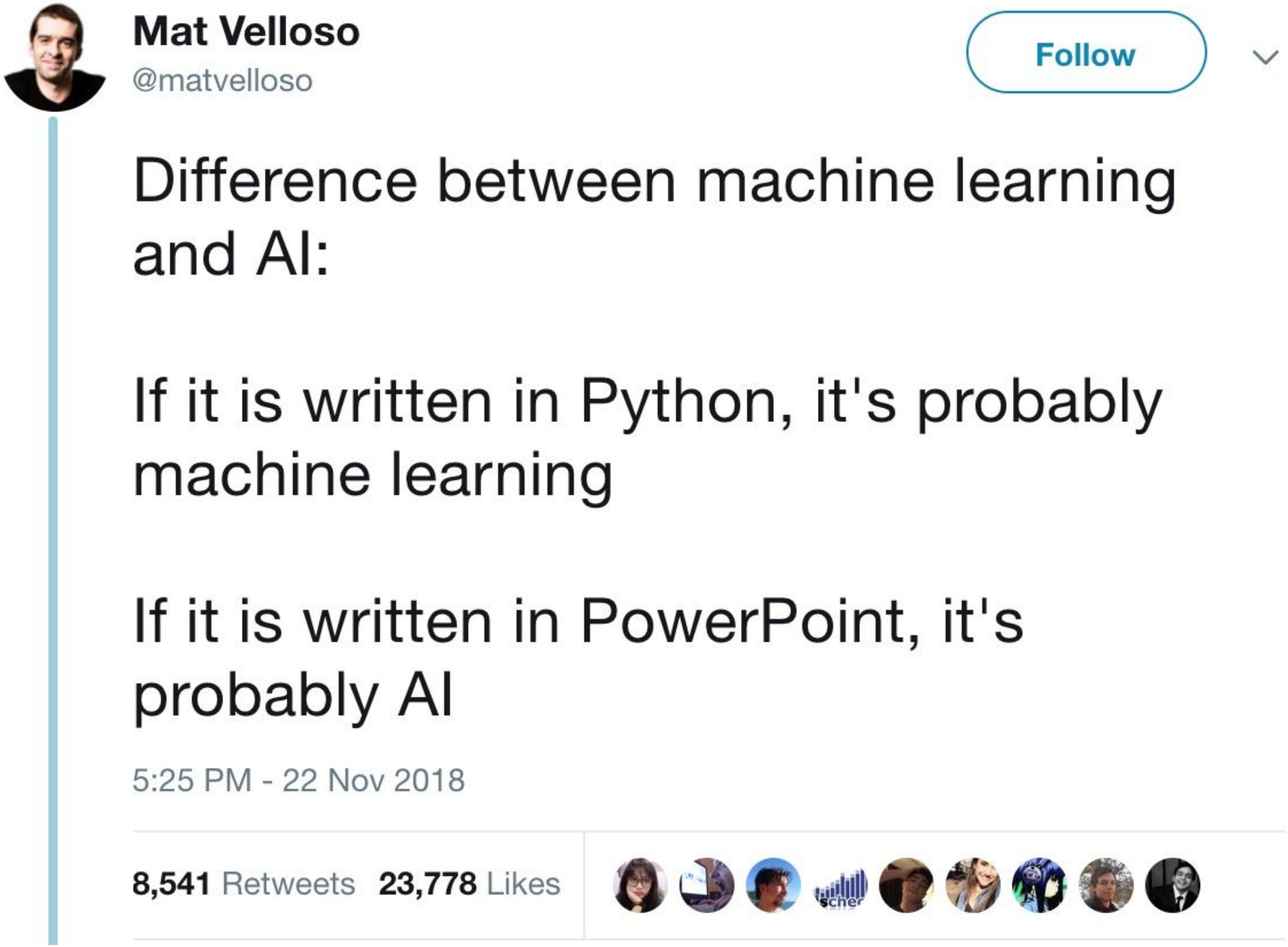
Trump2Money

1 Open your laptop and write some code
2. Monitor Trump’s twitter feed
3. Analyze the twitter
    If it mentions of any publicly traded stocks
    and compute its sentiment
a. Long it if the sentiment is positive
b. Short it if the sentiment is negative



Source: https://github.com/maxbbraun/trump2cash

What is Machine Learning



Mat Velloso
@matvelloso

Difference between machine learning and AI:

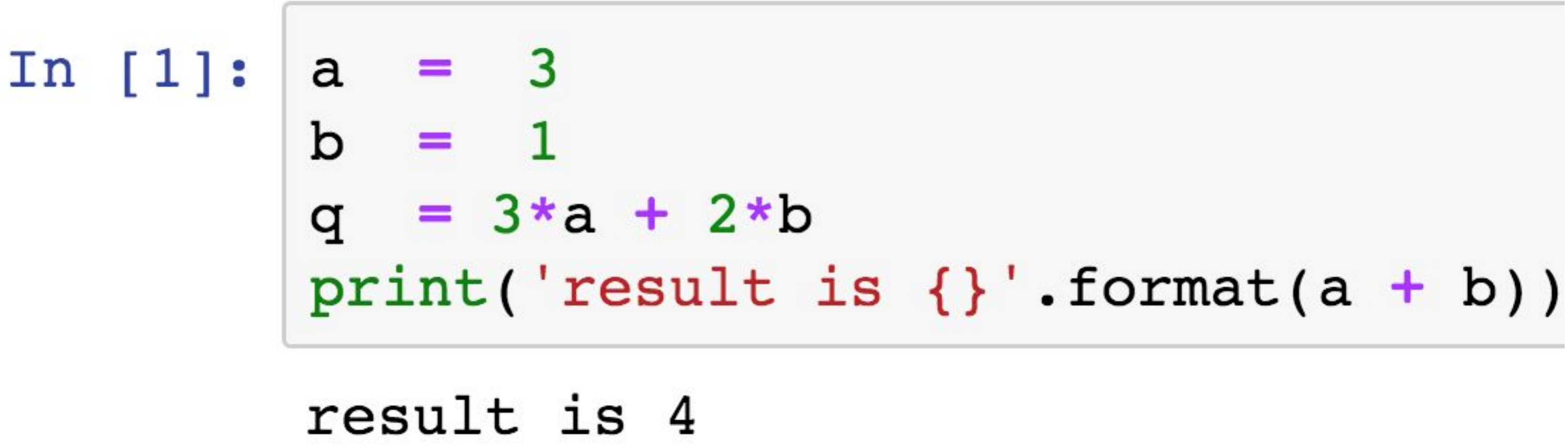
If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

5:25 PM - 22 Nov 2018

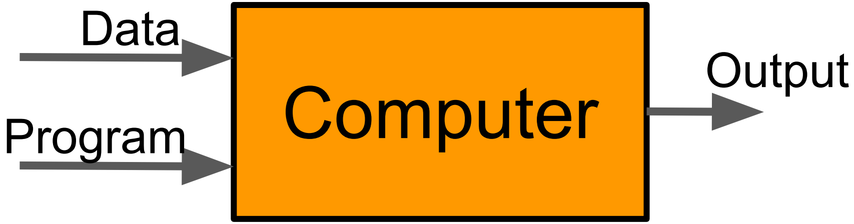
8,541 Retweets 23,778 Likes

Python Programming



```
In [1]: a   =  3
        b   =  1
        q   = 3*a + 2*b
        print('result is {}'.format(a + b))

        result is 4
```



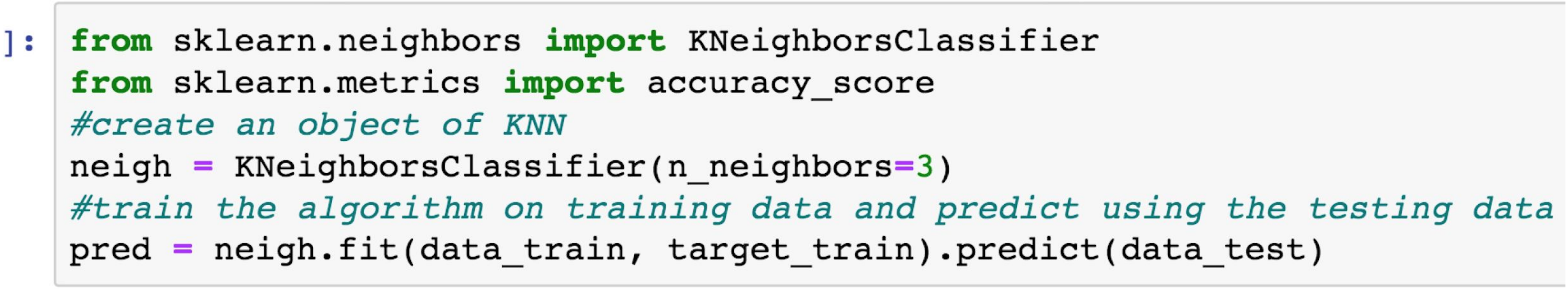
Data

Program

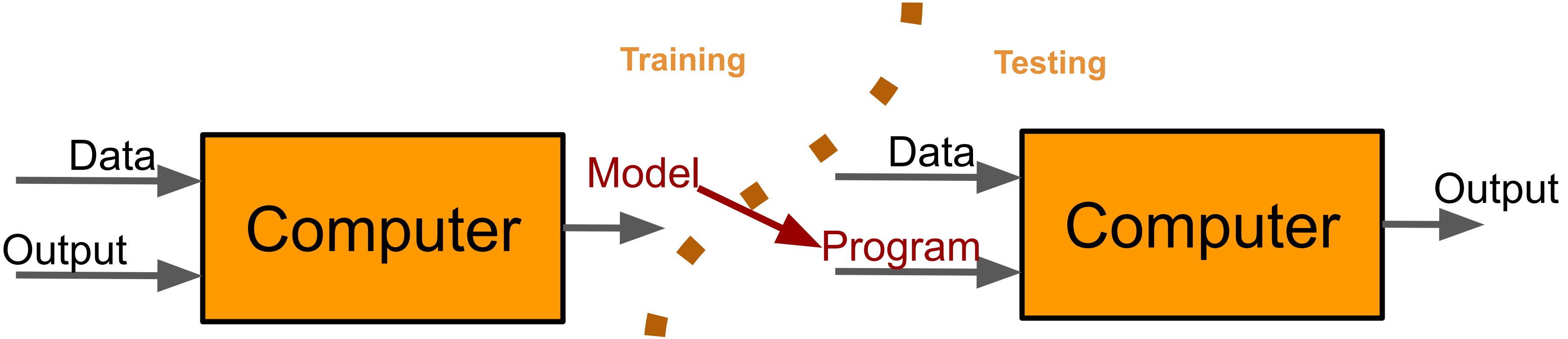
Computer

Output

Machine Learning



```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
#create an object of KNN
neigh = KNeighborsClassifier(n_neighbors=3)
#train the algorithm on training data and predict using the testing data
pred = neigh.fit(data_train, target_train).predict(data_test)
```



Definition of Machine Learning

"A computer program is said to learn from experience **E**
with respect to some class of tasks **T**
and performance measure **P**,
if its performance at tasks in **T**, as measured by **P**,
improves with experience **E**"

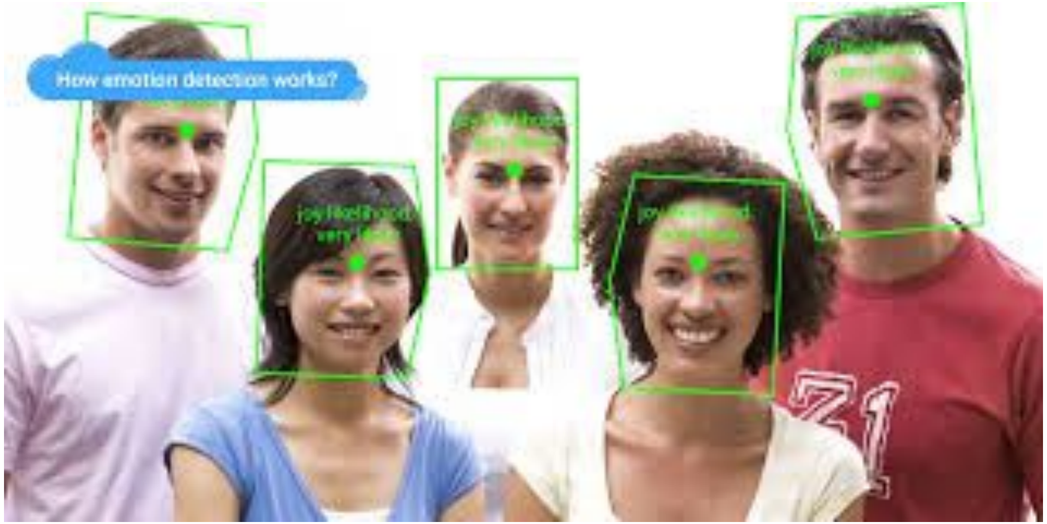
Tom Mitchell

T, **P**, **E** are three basic elements to define a complete machine learning tasks

AlphaGo

T : Play Go Games
P : Win rates of all matches
E : Match Experiences with many go players or itself

Face Recognition



T : **Identify or verify human faces**
P : **Accuracy that human faces are detected**
E : **Dataset of labelled human faces**

More E

- For machine learning algorithms, E is **data**.

- When **data is text (unstructured data type)**, we then have text mining.

Text mining is not only limited to machine learning approaches, since we can also hand-craft rules (old days).

Text Mining

What is Text Mining

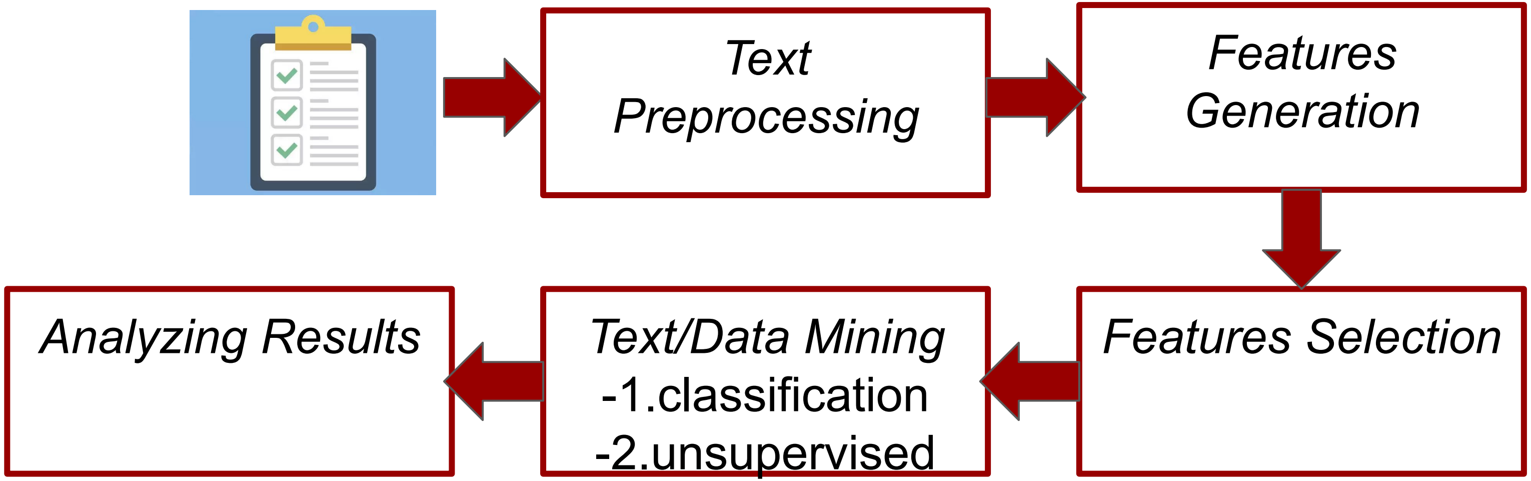
- Is finding **interesting regularities** in large **textual** dataset.
    - Where **interesting** means non-trivial, hidden, previously unknown and potentially useful.
    - E.g., extract **relations** between all of the entities.
    - E.g., **NTU** is in **Singapore**.
- Is finding semantic and abstract information from the surface form of text data:
    - E.g., predict sentiment 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**.

Which Topics are related to Text Mining

- Data Mining
- Machine Learning
- **Natural Language Processing**
  - Computational Linguistics
- **Information Retrieval**
  - Search & full-text indexing
- **Knowledge Management**
  - Knowledge Representation and Reasoning
  - Used in Question & Answering Systems

Text Mining Process Flow

- A typical text mining project involves 5 steps



Text Preprocessing

Features Generation

Features Selection

Text/Data Mining
-1.classification
-2.unsupervised

Analyzing Results

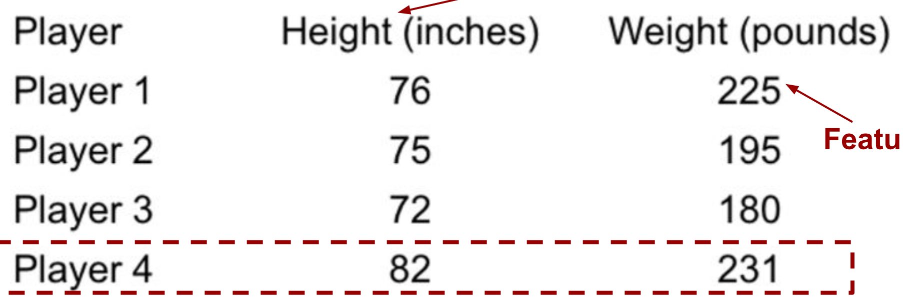
Unstructured Data: Text

Structured Data

- Structured Data
  - Machine learning/predictive algorithms need fixed-length vectors as inputs
  - Structured data is easily to be handled/prepared by our humans
  - Can be represented by columns and rows.
  - Each row is a data sample. Each column is attribute/feature.
- A toy task: predict the position of the basketball player

Structured Data for Toy Example

- Structured: just like the excel file or csv



Features

Labels

Player	Height (inches)	Weight (pounds)	Position
Player 1	76	225	C
Player 2	75	195	PG
Player 3	72	180	SF
Player 4	82	231	PF

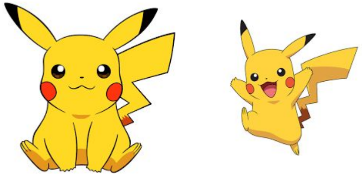
Feature Values

Data Sample

Unstructured

- The original data can not be stored in an “table”
- More abstract, more fuzzy, and more high-dimensionality

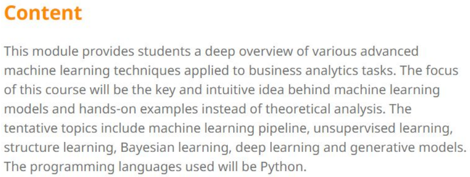
Images



Audio

Video

Text

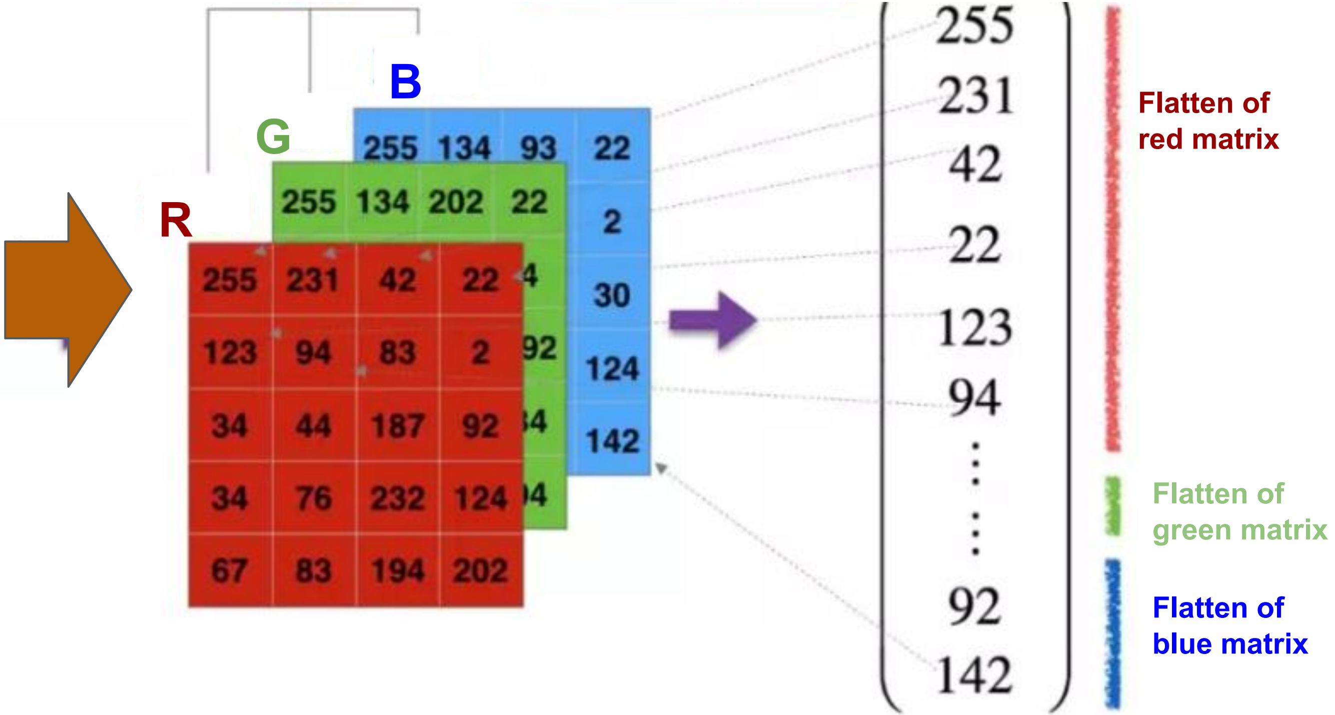


Content

This module provides students a deep overview of various advanced machine learning techniques applied to business analytics tasks. The focus of this course will be the key and intuitive idea behind machine learning models and hands-on examples instead of theoretical analysis. The tentative topics include machine learning pipeline, unsupervised learning, structure learning, Bayesian learning, deep learning and generative models. The programming languages used will be Python.

Environment around agent

For Images



255	231	42	22
123	94	83	2
34	44	187	92
34	76	232	124
67	83	194	202

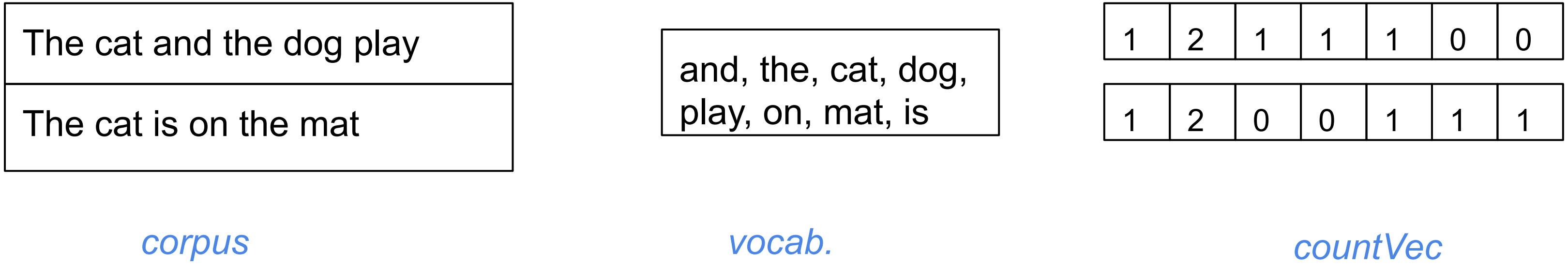
Flatten of red matrix

Flatten of green matrix

Flatten of blue matrix

For Text

- 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 (from unstructured to structured).
  - Each data is a vector
  - The length of the vector should be fixed
- Each row represents a document and each column a word.



The cat and the dog play
The cat is on the mat

and, the, cat, dog, play, on, mat, is

1	2	1	1	1	0	0
1	2	0	0	1	1	1

corpus

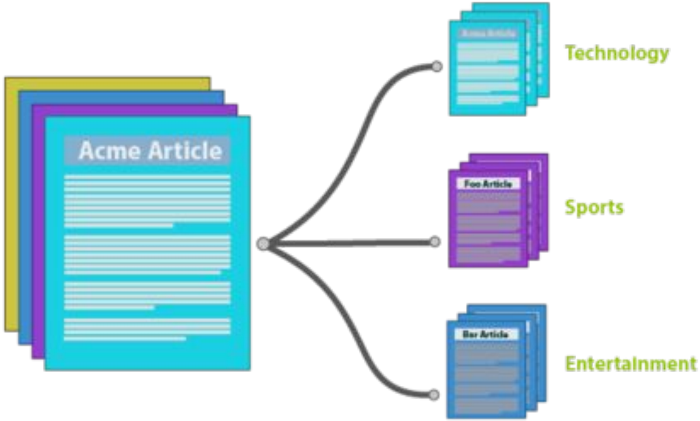
vocab.

countVec

Text Mining Applications

Applications

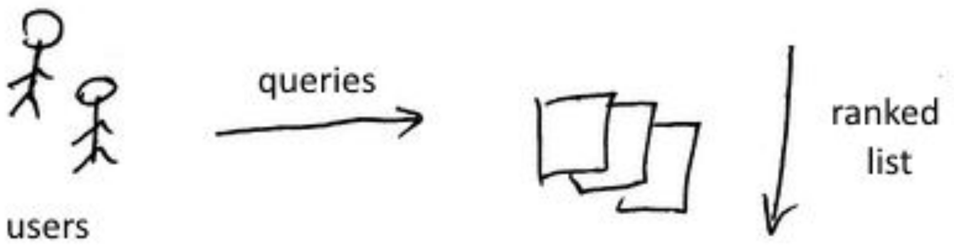
- **Document Classification**: given a sample of documents and correct answers (text categories) for each document, the objective is to find the correct answers for new documents.

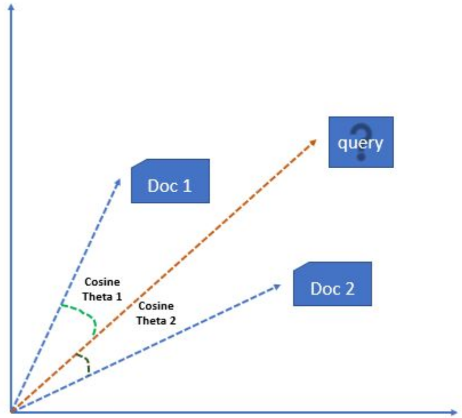


- Assign topic into each document/piece of text
- Email spam detection (binary classification) or new topic categorization (multiple classification)

Applications

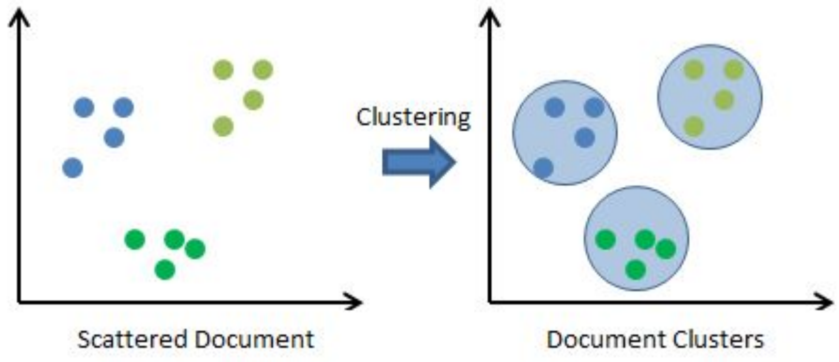
- **Information Retrieval** is the science of searching for documents or information in documents.
    - The input document is matched to all documents, retrieving the best-matched documents.
    - A basic concept for IR is **measuring similarity**: a comparison is made between two documents, measuring how similar the documents are.
    - Similarity can be computed after documents have been encoded as vectors



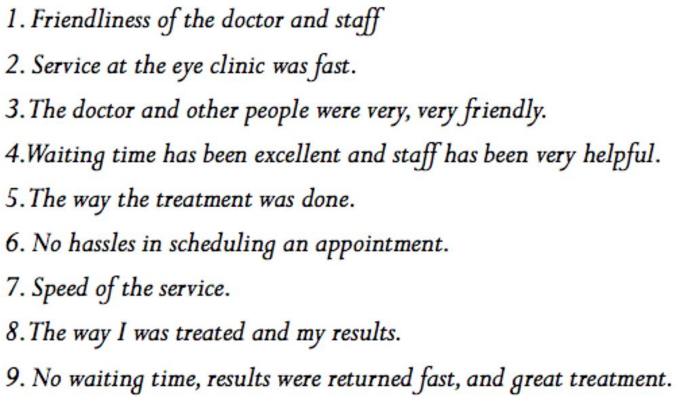


Applications

- **Document Clustering** is used when we have a collection of **documents with no known structure or no predefined categories**.
  - E.g., email complaints by users are clustered, and can learn about the categories and types of complaints.
- Because there are many ways to cluster documents, it is not quite as powerful as assigning answers(i.e., known correct labels) to documents.



- An example of Document Clustering: consider the comments made by the patients about the best thing they liked about the hospital.
- Because there are many ways to cluster documents, it is not quite as powerful as assigning answers(i.e., known correct labels) to documents.



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.

9. No waiting time, results were returned fast, and great treatment.

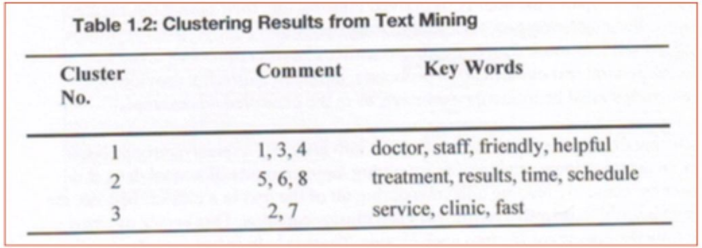


Table 1.2: Clustering Results from Text Mining

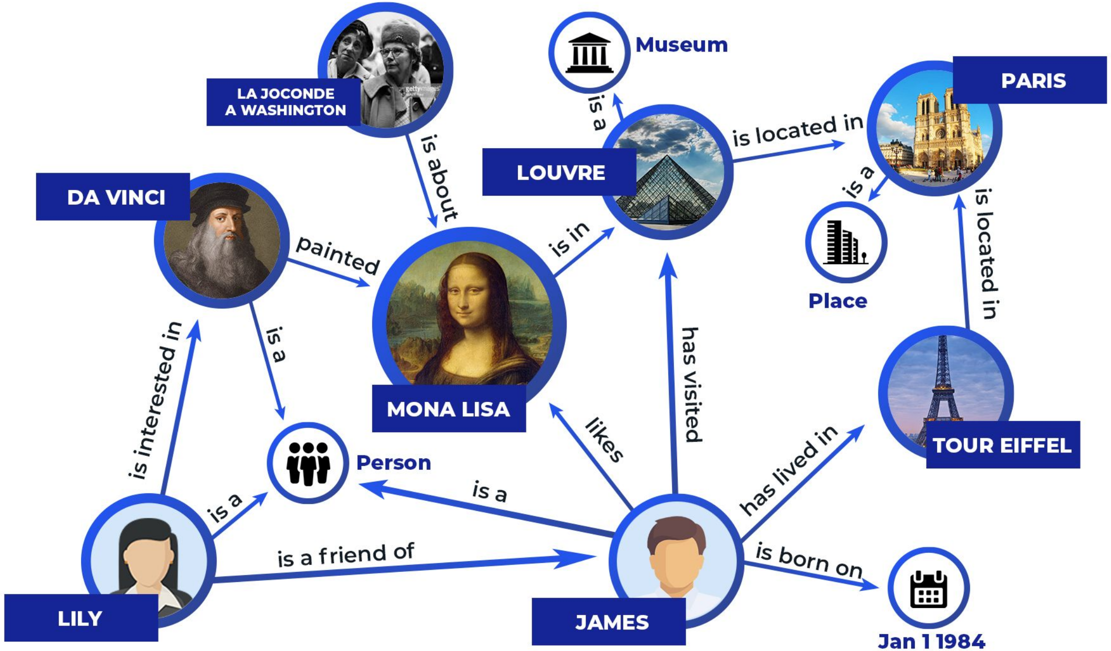
Cluster No.	Comment	Key Words
1	1, 3, 4	doctor, staff, friendly, helpful
2	5, 6, 8	treatment, results, time, schedule
3	2, 7	service, clinic, fast

Applications

- **Text Summarization**
  - Task: the task is to produce shorter, summary version of an original document.
  - Two main approaches to the problem:
    - Extraction-based: output consists from topmost text units
    - Abstraction-based: perform semantic analysis, representing the meaning and generating the text satisfying length restriction.

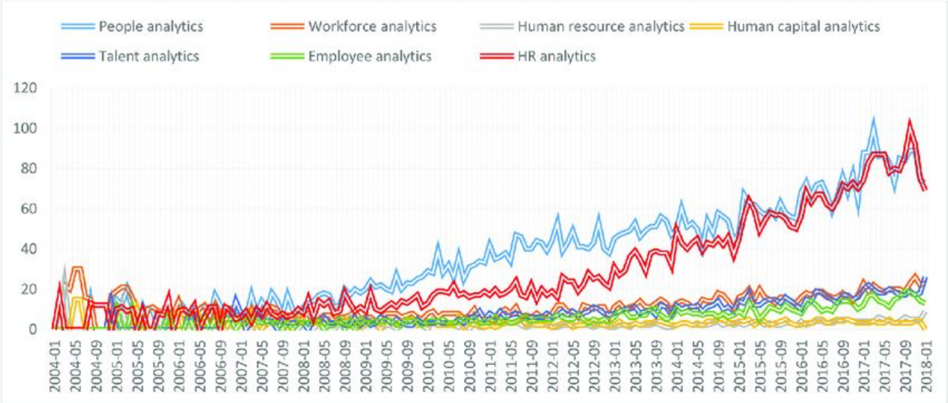
Applications

- **Knowledge Management**
    - Knowledge Graph: nodes(entities) and edge (relationship between entities)



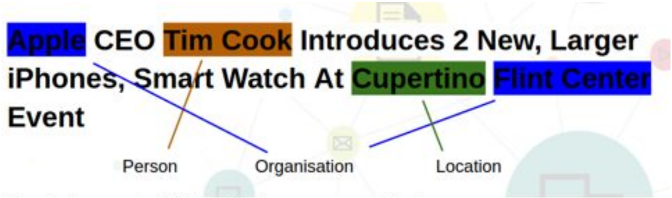
Applications

- **Trend Analysis**: Given a set of documents with a time stamp, text mining can be used to identify trends of different topics that exist in the text.
- Examples
  - Tracking the trends in research from scientific literature
  - Summarizing events from news articles.
- Google Trends provides a facility to identify the trends in various topics over a period of time.
  - Topic: Text Analytics



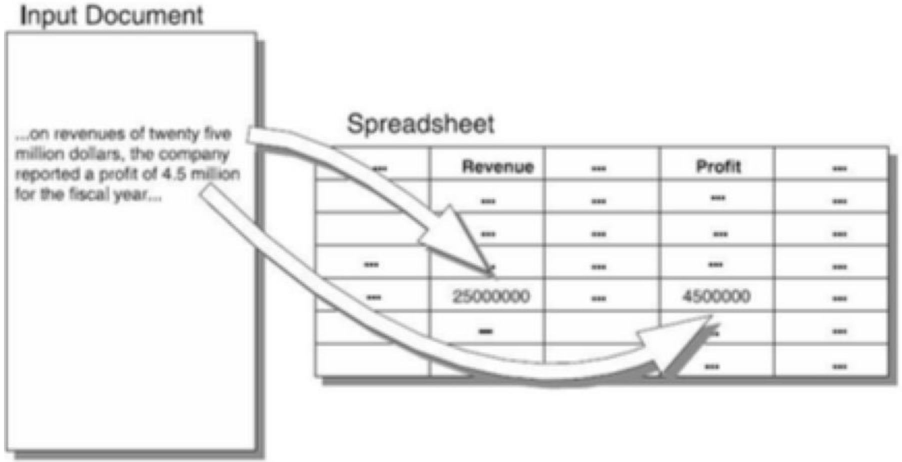
Applications

- **Information Extraction**
  - Take an unstructured document and automatically turn them into structured format
  - In the structured format, the columns are not just words but higher-level concepts that are found by the information extraction process.
    - E.g., people, organization, places, addresses, dates.



Apple CEO Tim Cook Introduces 2 New, Larger iPhones, Smart Watch At Cupertino Flint Center Event

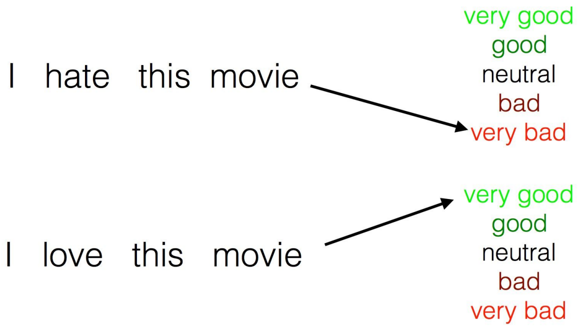
Figure 1: An example of NER application on an example text



...on revenues of twenty five million dollars, the company reported a profit of 4.5 million for the fiscal year...

Applications

- **Sentiment Analysis**
  - A type of subjective 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.
  - It is also called as opinion mining.
  - Importance of social media and online opinions
    - Online shoppers are influenced by product reviews and are willing to pay more for products highly rated by other consumers.
    - Users are more influenced by reviews of fellow consumers rather than those generated by professionals.



I hate this movie

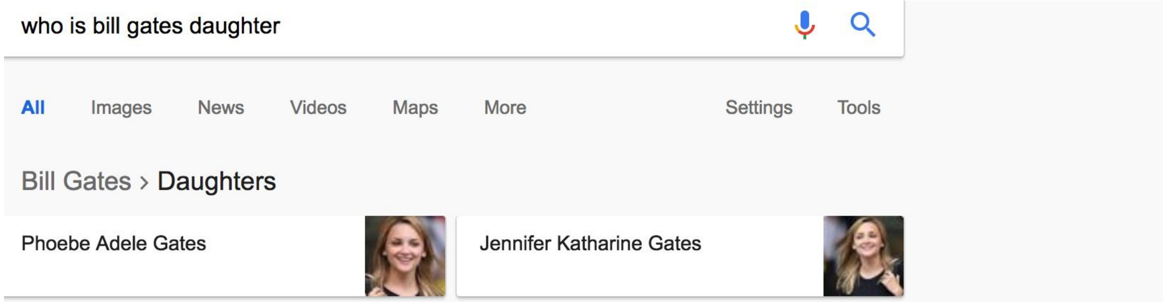
very good
good
neutral
bad
very bad

I love this movie

very good
good
neutral
bad
very bad

Applications

- **Question Answering**



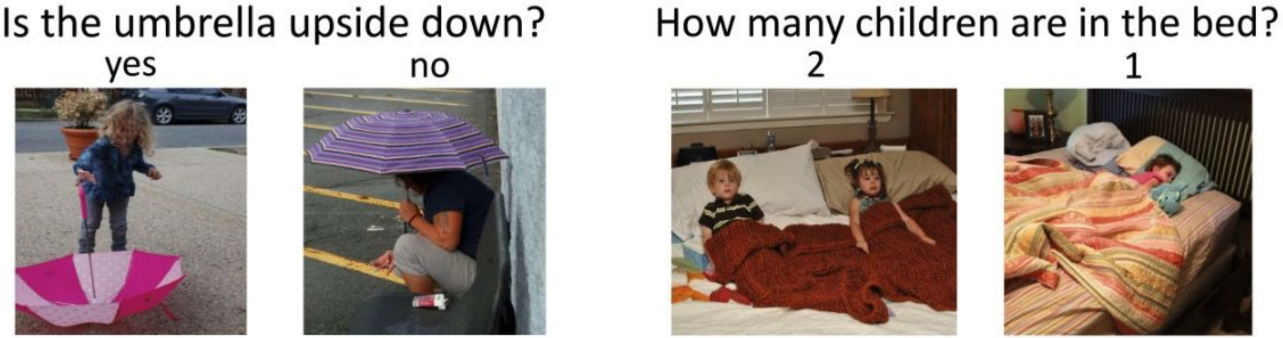
who is bill gates daughter

Bill Gates › Daughters

Phoebe Adele Gates

Jennifer Katharine Gates

- **Visual Question Answering**



Is the umbrella upside down?

yes

no

How many children are in the bed?

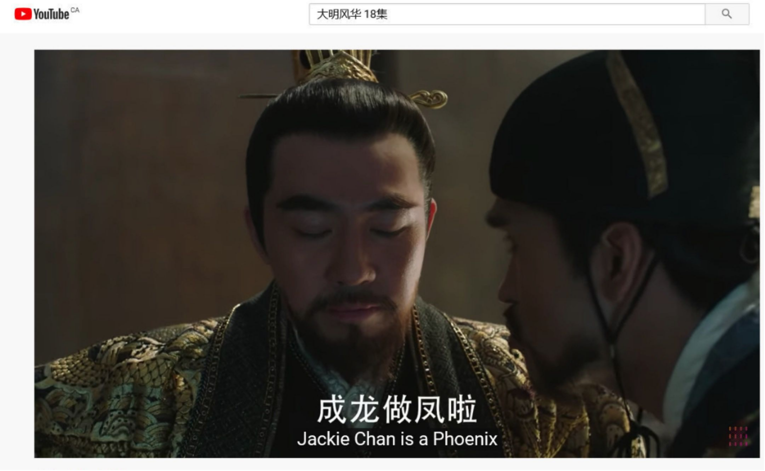
2

1

Why Text Mining is Tough?

- Many ways to represent similar concepts
  - E.g., space ship, flying saucer, and UFO
- “Countless” combinations of subtle, abstract relationships among concepts
  - E.g., relationship between drugs and diseases
- High dimensionality
  - Tens of hundreds of thousands of features
- Data Variation
  - We have ImageNet, while we do not have such huge labelled volume text data
- Ambiguity of Language
  - Word level: bank
  - Sentence level: I heard his cell phone in my office

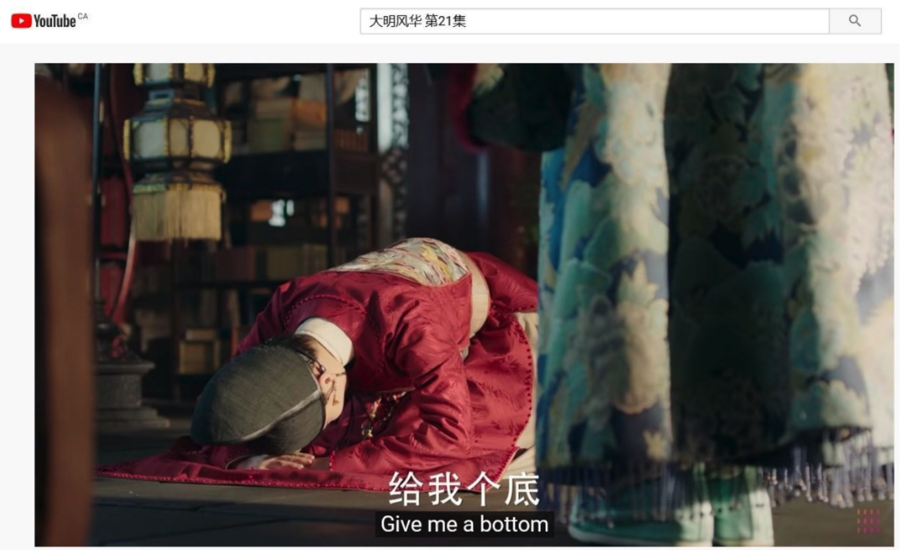
Text mining/NLP is really hard



大明风华 18集

成龙做凤啦

Jackie Chan is a Phoenix



大明风华 第21集

给我个底

Give me a bottom

Group Projects

Project Guidelines and Grading Policy

https://h6751.github.io/project/h6751_guidlines_grading.pdf

Project Hint 1

- Find a new problem which can be solved by text mining/machine learning models
    - **Generate a poem based on images**

sometimes alone
followed by vistas
you are a light seeker
and the light finds you

do colours really convey moods
colors answer feeling in man
shapes answer thought
motion answers will

in the celtic tongue
a glen is any dale
touched by the natural magic
of green shade

what is life
it is the flash of a firefly in the night
it is the breath of a buffalo in the wintertime
it is the little shadow which runs across the grass and
loses itself in the sunset

we are all falling this hand is falling too
all have this falling sickness none withstands
and still there's always one whose gentle hands
this universal falling can't fall through

come on down to my boat baby
come on down where we can play
come on down to my boat baby
come on down we'll sail away

the man bent over his guitar
a shearsman of sorts the day was green
they said 'you have a blue guitar
you do not play things as they are

my walls outside must have some flowers
my walls within must have some books
a house that's small a garden large
and in it leafy nooks

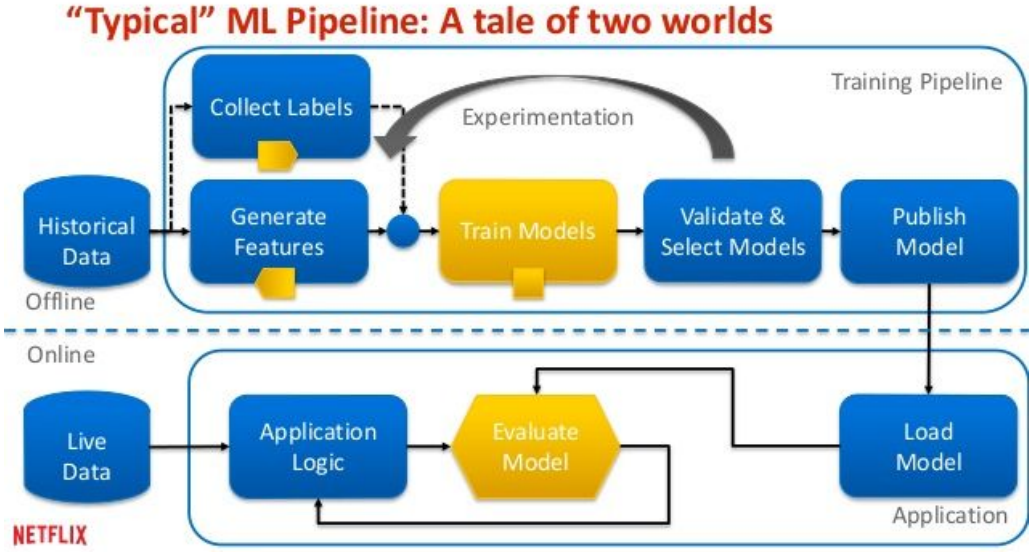
but now the psyche of thy being
still shyly doth essay her delicate wing
like to that airy nurseling of the sun
when first it breaketh through its dun

is it so small a thing
to have enjoy'd the sun
to have lived light in the spring
to have loved to have thought to have done

Source: https://github.com/researchmm/img2poem

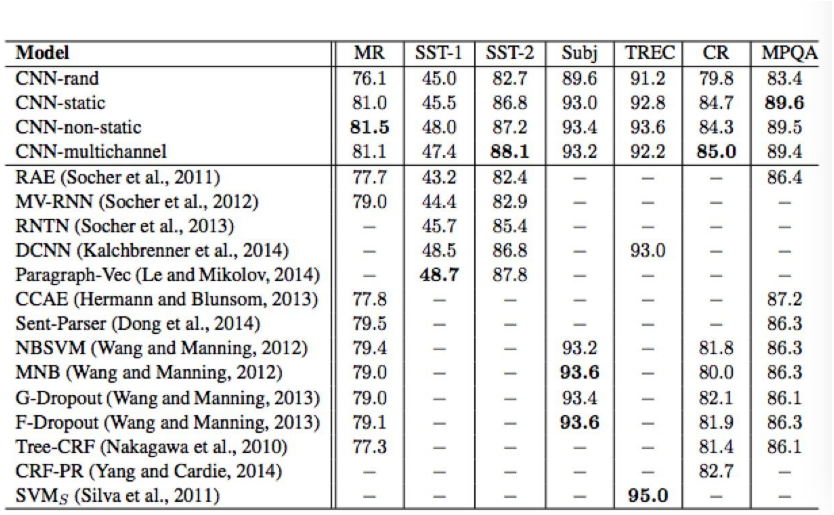
Project Hint 2

- Build a whole pipeline text mining system (real-life one)



Project Hints 3

- In-depth analysis of text mining/machine learning algorithms on one specific application
- **Try to explain the findings**

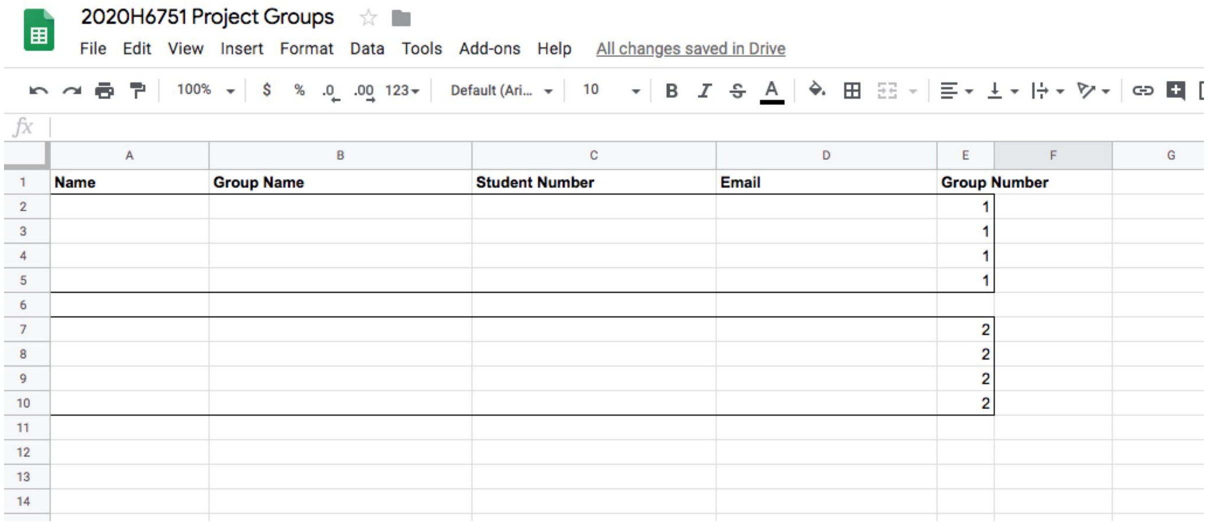


Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-non-static	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**	89.4
RAE (Socher et al., 2011)	77.7	43.2	82.4	–	–	–	86.4
MV-RNN (Socher et al., 2012)	79.0	44.4	82.9	–	–	–	–
RNTN (Socher et al., 2013)	–	45.7	85.4	–	–	–	–
DCNN (Kalchbrenner et al., 2014)	–	48.5	86.8	–	93.0	–	–
Paragraph-Vec (Le and Mikolov, 2014)	–	**48.7**	87.8	–	–	–	–
CCAE (Hermann and Blunsom, 2013)	77.8	–	–	–	–	–	87.2
Sent-Parser (Dong et al., 2014)	79.5	–	–	–	–	–	86.3
NBSVM (Wang and Manning, 2012)	79.4	–	–	93.2	–	81.8	86.3
MNB (Wang and Manning, 2012)	79.0	–	–	**93.6**	–	80.0	86.3
G-Dropout (Wang and Manning, 2013)	79.0	–	–	93.4	–	82.1	86.1
F-Dropout (Wang and Manning, 2013)	79.1	–	–	**93.6**	–	81.9	86.3
Tree-CRF (Nakagawa et al., 2010)	77.3	–	–	–	–	81.4	86.1
CRF-PR (Yang and Cardie, 2014)	–	–	–	–	–	82.7	–
SVM$_S$ (Silva et al., 2011)	–	–	–	–	**95.0**	–	–

From Yoon Kim

Take Action

- Form your group
- Find a cool team name
- The link has been shared in the course website



2020H6751 Project Groups

	A	B	C	D	E	F	G
1	Name	Group Name	Student Number	Email	Group Number		
2					1		
3					1		
4					1		
5					1		
6							
7					2		
8					2		
9					2		
10					2		
11							
12							
13							
14							