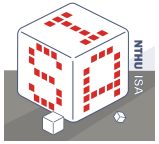


Data Mining Lab

Fall 2018

https://goo.gl/d6YTje

About us

Email: ellfae@gmail.com
Website: elvissaravia.com
Github: omarsar
Twitter: @omarsar0

Email: evan800112@gmail.com
Github: EvanYu800112

Email: gn01697933@gmail.com
Github: thisray

Objectives

Table of Contents

Requirements for this lab

- Learn how to use Jupyter notebooks to explore and visualize data
- Create Github account
- Find the working repository here: https://github.com/omarsar/data_mining_lab
- Fork it (*Complete the take home exercises and upload changes*) (How-to)

Assignment:

- Published on Github as a Jupyter Notebook (*Please learn Git as soon as possible*)
- **Due date: October 22, 2017**
- More details provided later

Data in a bygone era

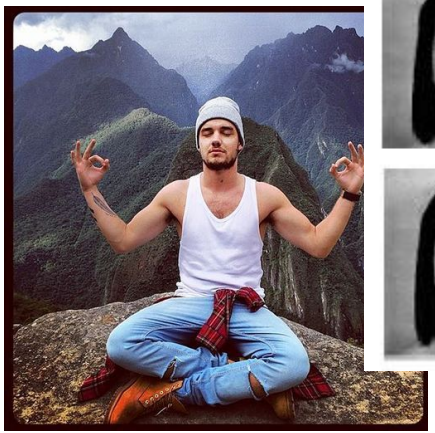
We used to share information through carvings and peculiar symbols

Observation: Very difficult to interpret any meaning; Any other observations?

Data Now

Now we share text messages, photos, and videos -- also known as Big Data era

Observation: Very difficult to interpret and gather knowledge from it; any other observations?



Data: Past vs Present

We all want to share and analyze data; we are all hungry for knowledge.

Problem: Very difficult to mine knowledge from it

Reasons: Data format / Data Scarcity / ?

Solutions: Advanced Computing / Fast Algorithms / Big Data / Dynamic Visualization Tools

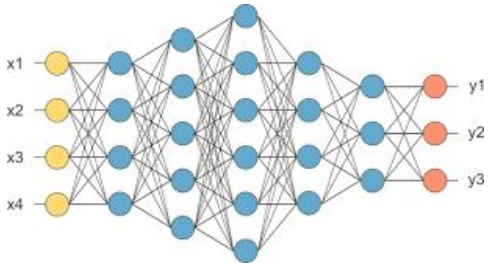
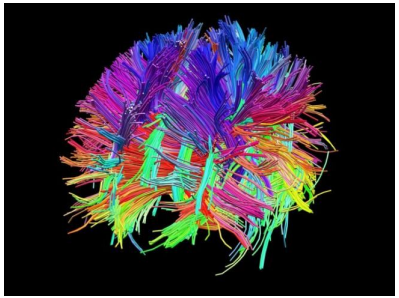


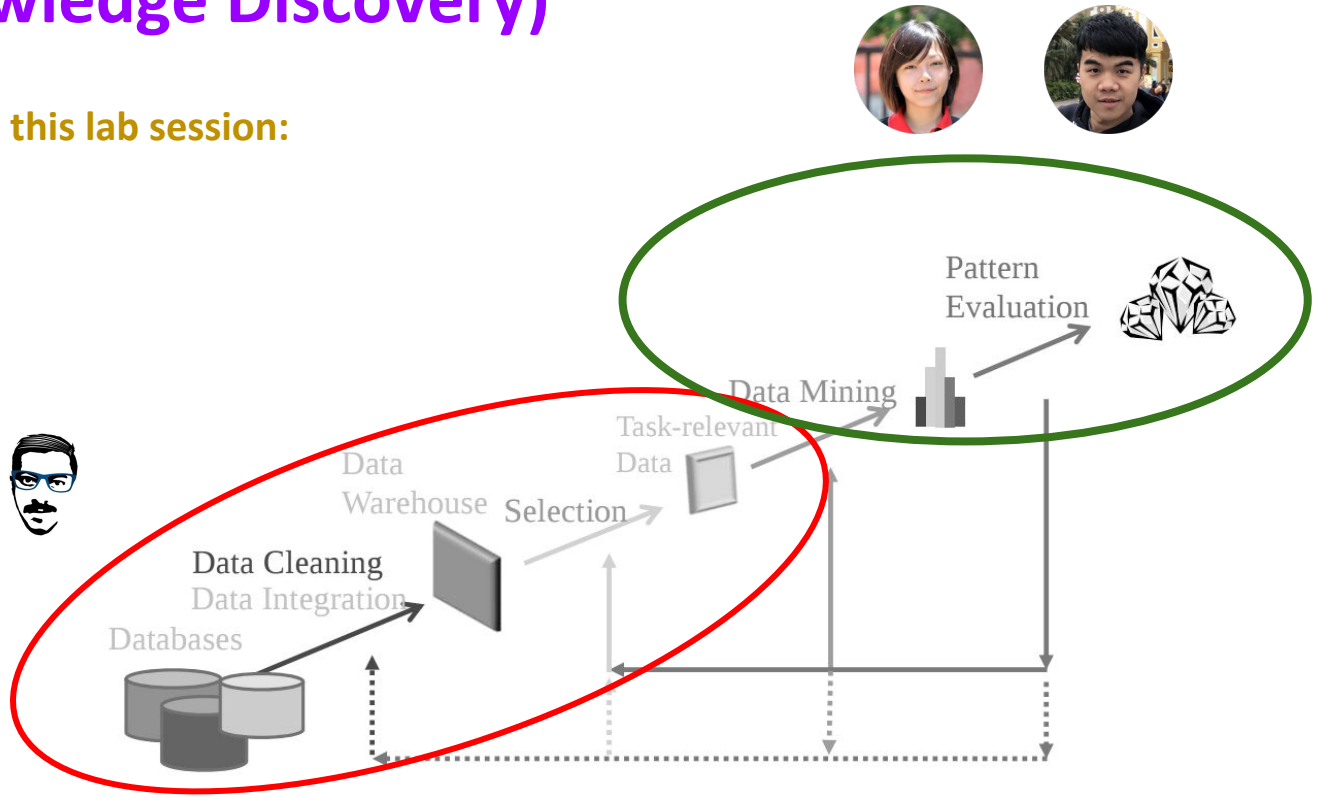
Diagram of Neural Network


Google's Tensor Processing Unit


The Human Connectome Project

KDD (Knowledge Discovery)

Our main focus in this lab session:



Data Sources

Worst Case: Web crawling data / Graph data / API data / Surveys

Best case: Already processed and formatted for us

Data usually comes in one of three kinds: *data matrix, document data, transaction-based data.*

Some popular sites for searching/finding open data:

- UCI Machine Learning Repository
- Kaggle
- Google Dataset Search

Our Dataset

Name: 20 Newsgroups

Type: Text (*unstructured*)

Characteristics:

- 20 newsgroups categories (1000 articles for each)
- Initially collected from multiple sources and compiled into one

```
alt.atheism
comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x
misc.forsale
rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey
sci.crypt
sci.electronics
sci.med
sci.space
soc.religion.christian
talk.politics.guns
talk.politics.mideast
talk.politics.misc
talk.religion.misc
```

Document Data

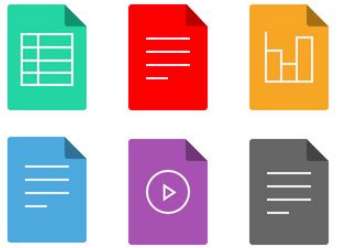
Our main focus is on *document data*

Why:

- We get to work with both ***structured*** and ***unstructured*** data
- Our ***final project*** will heavily focus on this type of dataset

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Where to store data?



Options for storing data:

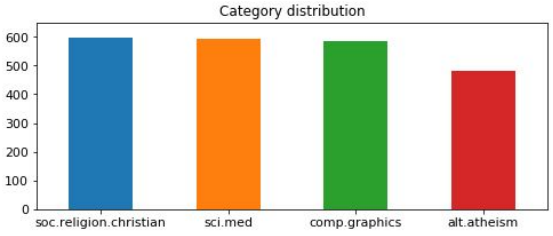
- Raw files (csv, txt, xml, json, etc.)
- Database
    - SQL / NoSQL / Document-based / Distributed
- Warehouse
- Elasticsearch (*fast for search*)
- **Pandas Dataframes - fast in-memory data processing (*we will use this*)**
- etc.

Data Operations

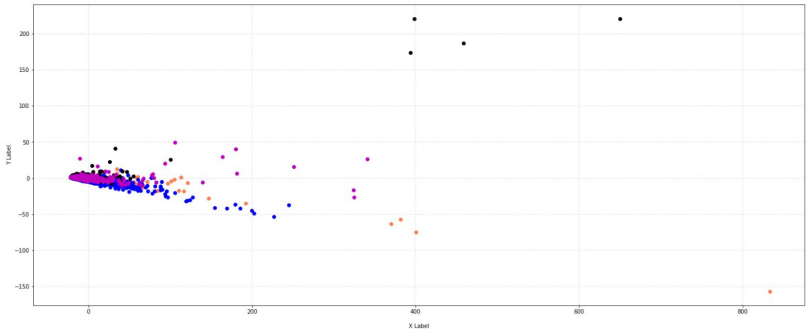
Snapshots of what we will do with the dataset

	text	category	category_name	unigrams	bin_category
0	From: sd345@city.ac.uk (Michael Collier) Subje...	1	comp.graphics	[From, :, sd345, @, city.ac.uk, (, Michael, Co...	[0, 1, 0, 0]
1	From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...	1	comp.graphics	[From, :, ani, @, ms.uky.edu, (, Aniruddha, B....	[0, 1, 0, 0]
2	From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...	3	soc.religion.christian	[From, :, djohnson, @, cs.ucsd.edu, (, Darin, ...	[0, 0, 0, 1]
3	From: s0612596@let.rug.nl (M.M. Zwart) Subject...	3	soc.religion.christian	[From, :, s0612596, @, let.rug.nl, (, M.M, ., ...	[0, 0, 0, 1]
4	From: stanly@grok11.columbiasc.ncr.com (stanly...	3	soc.religion.christian	[From, :, stanly, @, grok11.columbiasc.ncr.com...	[0, 0, 0, 1]
5	From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...	3	soc.religion.christian	[From, :, vbv, @, lor.eeap.cwru.edu, (, Virgil...	[0, 0, 0, 1]
6	From: jodfishe@silver.ucs.indiana.edu (joseph ...	3	soc.religion.christian	[From, :, jodfishe, @, silver.ucs.indiana.edu,...	[0, 0, 0, 1]
7	From: aldridge@netcom.com (Jacquelin Aldridge)...	2	sci.med	[From, :, aldridge, @, netcom.com, (, Jacqueli...	[0, 0, 1, 0]
8	From: geb@cs.pitt.edu (Gordon Banks) Subject: ...	2	sci.med	[From, :, geb, @, cs.pitt.edu, (, Gordon, Bank...	[0, 0, 1, 0]

Transformation and Binarization



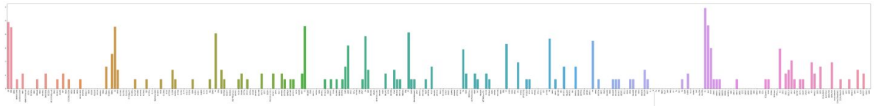
Data Distribution



PCA (Dimensionality Reduction)

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0
	3	8	5	2	5	8	2	5	3	2

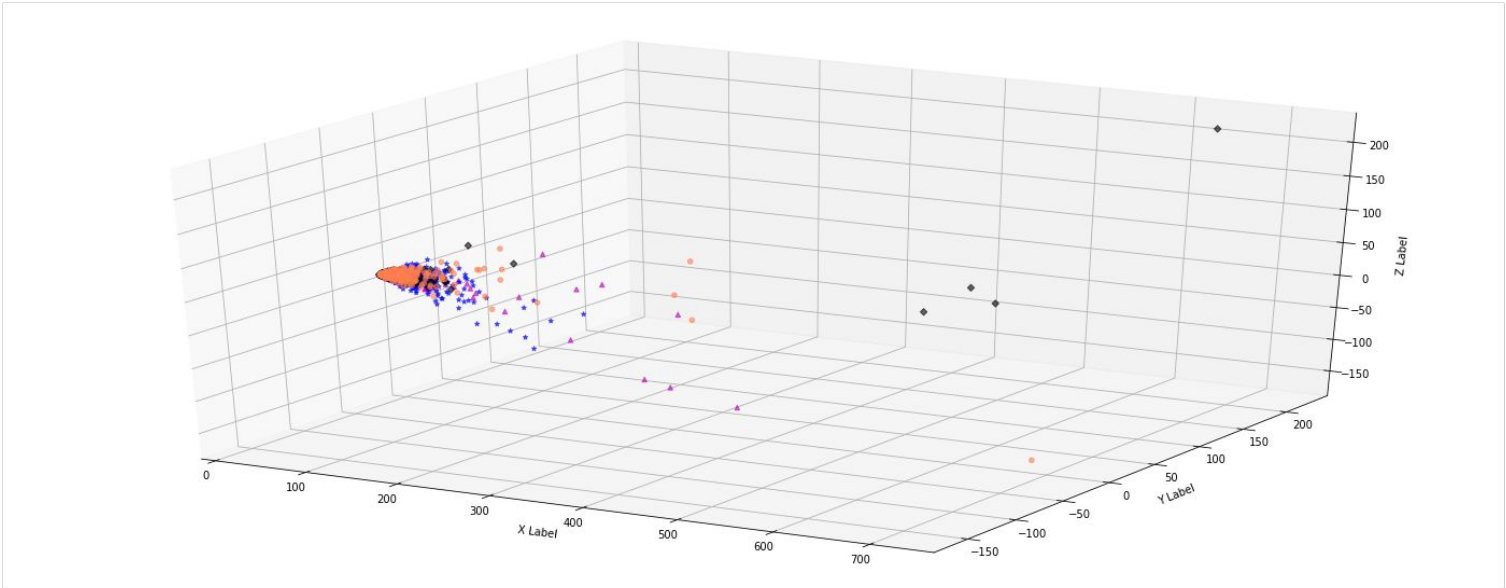
Matrix Operations (Pandas)



Term Frequency Distribution

Dimensionality Reduction (PCA)

We want to reduce the dimensionality of our dataset, especially if we are working on a dataset with high dimensionality, as it is very common in natural language information.



Data Exploration

The most important parts that will be covered for data exploration.

	text	category	category_name	unigrams	bin_category
0	From: sd345@city.ac.uk (Michael Collier) Subje...	1	comp.graphics	[From, :, sd345, @, city.ac.uk, (, Michael, Co...	[0, 1, 0, 0]
1	From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...	1	comp.graphics	[From, :, ani, @, ms.uky.edu, (, Aniruddha, B....	[0, 1, 0, 0]
2	From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...	3	soc.religion.christian	[From, :, djohnson, @, cs.ucsd.edu, (, Darin, ...	[0, 0, 0, 1]
3	From: s0612596@let.rug.nl (M.M. Zwart) Subject...	3	soc.religion.christian	[From, :, s0612596, @, let.rug.nl, (, M.M, ., ...	[0, 0, 0, 1]
4	From: stanly@grok11.columbiasc.ncr.com (stanly...	3	soc.religion.christian	[From, :, stanly, @, grok11.columbiasc.ncr.com...	[0, 0, 0, 1]
5	From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...	3	soc.religion.christian	[From, :, vbv, @, lor.eeap.cwru.edu, (, Virgil...	[0, 0, 0, 1]
6	From: jodfishe@silver.ucs.indiana.edu (joseph ...	3	soc.religion.christian	[From, :, jodfishe, @, silver.ucs.indiana.edu,...	[0, 0, 0, 1]
7	From: aldridge@netcom.com (Jacquelin Aldridge)...	2	sci.med	[From, :, aldridge, @, netcom.com, (, Jacqueli...	[0, 0, 1, 0]
8	From: geb@cs.pitt.edu (Gordon Banks) Subject: ...	2	sci.med	[From, :, geb, @, cs.pitt.edu, (, Gordon, Bank...	[0, 0, 1, 0]

Demo Time!

https://github.com/omarsar/data_mining_lab