Data Processing

Tran Giang Son, tran-giang.son@usth.edu.vn

ICT Department, USTH

Review

Data Mining

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
  - The use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data.
  - The discovery of models for data

Data Models

- Different types of models
  - Models that explain the data (e.g., a single function)
  - Models that predict the future data instances.
  - Models that summarize the data
  - Models the extract the most prominent features of the data.

Why Data Mining?

- Huge amounts of complex data generated from multiple sources and interconnected
  - Scientific data: Weather, astronomy, physics, biological microarrays, genomics
  - Text collections: The Web, scientific articles, news, tweets, facebook postings.
  - Transaction data: Retail store records, credit card records
  - Behavioral data: Mobile phone data, query logs, browsing behavior, ad clicks
  - Networked data: The Web, Social Networks, IM networks, email network, biological networks.
- All these types of data can be combined in many ways
  - Facebook has a network, text, images, user behavior, ad transactions

Why Data Mining?

- Analyze this data to extract knowledge
    - Knowledge can be used for commercial or scientific purposes.
    - Our solutions should scale to the size of the data

Data Processing

Pipeline

- Mining is not the only step in the analysis process

Pipeline

- Preprocessing
  - Real data: noisy, incomplete and inconsistent
  - Cleaning: remove noise, make sense of the data
  - Techniques: Sampling, Dimensionality Reduction, Feature selection
  - A dirty, but important work
- Post-Processing: Make the data actionable and useful to the user
  - Statistical analysis of importance
  - Visualization

Data Quality

- Data is not clean
  - Noise and outliers
  - Missing values
  - Duplicate data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

Data Sampling

- Sampling: main technique for data selection
  - Preliminary investigation
  - Final data analysis.

Why Data Sampling?

- Difficult to **obtain** the entire set of data of interest
  - Example: What is the average height of a person in Hanoi?
- Difficult to **process** entire set of data of interest
  - Example: We have 1M documents. What fraction has at least 100 words in common?
    - Computing number of common words for all pairs requires 10^{12} comparisons
  - Example: What fraction of tweets in a year contain the word "Data"?
    - 300M tweets per day, if 100 characters on average, 86.5TB to store all tweets

Data Sampling

- **Representative** samples
  - Share approximately the same property (of interest) as the original set of data
  - Work almost as well as using the entire data sets
- Non-representative samples: bias
  - Sample from the university campus to compute the average height of a person in Hanoi?

Data Sampling

- Simple Random Sampling: Equal probability of selecting any particular item
- Sampling without replacement: Remove the selected item from the population
- Sampling with replacement: Do not remove the selected item from the population
  - Same object can be picked up more than once
  - Easier analytical computation of probabilities

Data Sampling

- 100 people
  - 51 women
  - 49 men
- Probability of picking two people that both are women
  - Sampling with replacement:

Data Sampling

- 100 people
  - 51 women
  - 49 men
- Probability of picking two people that both are women
  - Sampling with replacement: $P(W, W) = 0.51^2$
  - Sampling without replacement:

Data Sampling

- 100 people
  - 51 women
  - 49 men
- Probability of picking two people that both are women
  - Sampling with replacement: $P(W, W) = 0.51^2$
  - Sampling without replacement: $P(W, W) = 51/100 * 50/99$

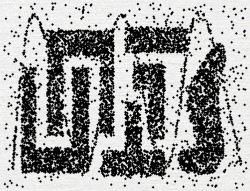
Data Sampling

- Stratified sampling
  - Split the data into several groups
  - Pick random samples from each group
    - Ensures that all groups are represented

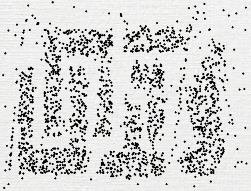
Data Sampling

- Example: Legitimate and fraudulent credit card transactions
  - 0.1% fraudulent
  - 99.9% legitimate
  - 1000 transactions at random?
    - 1 expected fraudulent transaction
  - Better: sample 1000 legitimate and 1000 fraudulent transactions

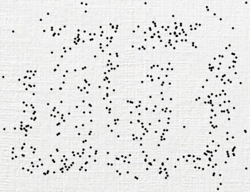
Sample Size



8000 Points

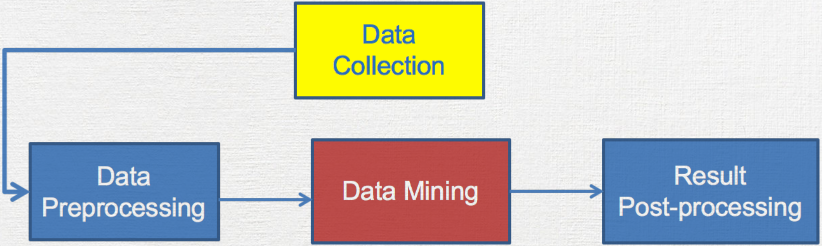


2000 Points



500 Points

Data Collection



- A lot of data online
  - Facebook, Twitter, Wikipedia, Web, etc. . .
- First step: collect the data
  - Customized crawlers: public APIs, HTML parsers
  - Additional cleaning/processing to parse out the useful parts
  - Respect of crawling ethics

Data Collection: Example

- Comment analysis from Yelp dataset
  - Yelp: reviews, businesses, users, tips, and check-in data
  - Data: Yelp dataset on Kaggle
  - Task: Find few terms that describe restaurants?

Data Collection: Example

- Download dataset
- Preprocessing
  - Remove punctuations, lower case, trim…
  - Break by space
  - Count occurence
  - Remove stop words: a, in, of, the, at…
  - Remove commonly used words in reviews: like, get…
- Statistically find words using TF-IDF

Data Collection: Example

I heard so many good things about this place so I was pretty juiced to try it. I'm from Cali and I heard Shake Shack is comparable to IN-N-OUT and I gotta say, Shake Shake wins hands down. Surprisingly, the line was short and we waited about 10 MIN. to order. I ordered a regular cheeseburger, fries and a black/white shake. So yummerz. I love the location too! It's in the middle of the city and the view is breathtaking. Definitely one of my favorite places to eat in NYC.

I'm from California and I must say, Shake Shack is better than IN-N-OUT, all day, err'day.

Would I pay $15+ for a burger here? No. But for the price point they are asking for, this is a definite bang for your buck (though for some, the opportunity cost of waiting in line might outweigh the cost savings) Thankfully, I came in before the lunch swarm descended and I ordered a shake shack (the special burger with the patty + fried cheese & portabella topping) and a coffee milk shake. The beef patty was very juicy and snugly packed within a soft potato roll. On the downside, I could do without the fried portabella-thingy, as the crispy taste conflicted with the juicy, tender burger. How does shake shack compare with in-and-out or 5-guys? I say a very close tie, and I think it comes down to personal affliations. On the shake side, true to its name, the shake was well churned and very thick and luscious. The coffee flavor added a tangy taste and complemented the vanilla shake well.

Data Collection: Example

- Normalized, Word-counted

the 27514	the 16710	the 16010	the 14241
and 14508	and 9139	and 9504	and 8237
i 13088	a 8583	i 7966	a 8182
a 12152	i 8415	to 6524	i 7001
to 10672	to 7003	a 6370	to 6727
of 8702	in 5363	it 5169	of 4874
ramen 8518	it 4606	of 5159	you 4515
was 8274	of 4365	is 4519	it 4308
is 6835	is 4340	sauce 4020	is 4016
it 6802	burger 432	in 3951	was 3791
in 6402	was 4070	this 3519	pastrami 3748
for 6145	for 3441	was 3453	in 3508
but 5254	but 3284	for 3327	for 3424
that 4540	shack 3278	you 3220	sandwich 2928
you 4366	shake 3172	that 2769	that 2728
with 4181	that 3005	but 2590	but 2715
pork 4115	you 2985	food 2497	on 2247
my 3841	my 2514	on 2350	this 2099
this 3487	line 2389	my 2311	my 2064
wait 3184	this 2242	cart 2236	with 2040
not 3016	fries 2240	chicken 2220	not 1655
we 2984	on 2204	with 2195	your 1622
at 2980	are 2142	rice 2049	so 1610
on 2922	with 2095	so 1825	have 1585

Data Collection: Example

- Non-stop-word highlighted

the 27514	the 16710	the 16010	the 14241
and 14508	and 9139	and 9504	and 8237
i 13088	a 8583	i 7966	a 8182
a 12152	i 8415	to 6524	i 7001
to 10672	to 7003	a 6370	to 6727
of 8702	in 5363	it 5169	of 4874
ramen 8518	it 4606	of 5159	you 4515
was 8274	of 4365	is 4519	it 4308
is 6835	is 4340	**sauce 4020**	is 4016
it 6802	**burger 432**	in 3951	was 3791
in 6402	was 4070	this 3519	**pastrami 3748**
for 6145	for 3441	was 3453	in 3508
but 5254	but 3284	for 3327	for 3424
that 4540	**shack 3278**	you 3220	**sandwich 2928**
you 4366	**shake 3172**	that 2769	that 2728
with 4181	that 3005	but 2590	but 2715
pork 4115	you 2985	food 2497	on 2247
my 3841	my 2514	on 2350	this 2099
this 3487	line 2389	my 2311	my 2064
wait 3184	this 2242	**cart 2236**	with 2040
not 3016	**fries 2240**	**chicken 2220**	not 1655
we 2984	on 2204	with 2195	your 1622
at 2980	are 2142	rice 2049	so 1610
on 2922	with 2095	so 1825	have 1585

Data Collection: Example

- Stop-word removed

ramen 8572
pork 4152
wait 3195
good 2867
place 2361
noodles 2279
ippudo 2261
buns 2251
broth 2041
like 1902
just 1896
get 1641
time 1613
one 1460

burger 4340
shack 3291
shake 3221
line 2397
fries 2260
good 1920
burgers 1643
wait 1508
just 1412
cheese 1307
like 1204
food 1175
get 1162
place 1159

sauce 4023
food 2507
cart 2239
chicken 2238
rice 2052
hot 1835
white 1782
line 1755
good 1629
lamb 1422
halal 1343
just 1338
get 1332
one 1222
like 1096

pastrami 3782
sandwich 2934
place 1480
good 1341
get 1251
katz's 1223
just 1214
like 1207
meat 1168
one 1071
deli 984
best 965
go 961
ticket 955

Data Collection: Example

- Important words: unique to the document
    - differentiating compared to the rest
- Document Frequency $DF(w)$: fraction of documents that contain word w.

$$DF(w) = \frac{D(w)}{D} \tag{1}$$

- $D(w)$: number of documents containing word w
- D: total number of documents

Data Collection: Example

- Inverse Document Frequency $IDF(w)$:

$$IDF(w) = log(\frac{\mid D \mid}{DF(w)}) \tag{2}$$

- Maximum when unique to one document : $IDF(w) = log(D)$
- Minimum when the word is common to all documents: $IDF(w) = 0$

Data Collection: Example

- **VERY** important words: unique to the document
- $TF(w, d)$: term frequency of word w in document d
  - Number of times that the word appears in the document over total number of words in the document
  - Natural measure of **importance** of the word for the document
- $IDF(w)$: inverse document frequency
  - Natural measure of the uniqueness of the word w

$$TF - IDF(w, d) = TF(w, d) \times IDF(w) \tag{3}$$

Data Collection: Example

- Words with their $TF - IDF$ sorted:

ramen 3057.417	fries 806.085	lamb 985.655	pastrami 1931.942
akamaru 2353.241	custard 729.607	halal 686.038	katz's 1120.623
noodles 1579.682	shakes 628.473	53rd 375.685	rye 1004.289
broth 1414.713	shroom 515.779	gyro 305.809	corned 906.113
miso 1252.606	burger 457.264	pita 304.984	pickles 640.487
hirata 709.196	crinkle 398.34	cart 235.902	reuben 515.779
hakata 591.764	burgers 366.624	platter 139.459	matzo 430.583
shiromaru 587.115	madison 350.939	chicken/lamb 135.852	sally 428.110
noodle 581.844	shackburger 292.428	carts 120.27437415	harry 226.323
tonkotsu 529.595	'shroom 287.823	hilton 84.298	mustard 216.079
ippudo 504.527	portobello 239.806	lamb/chicken 82.893	cutter 209.535
buns 502.296	custards 211.837	yogurt 70.007	carnegie 198.655
ippudo's 453.609	concrete 195.169	52nd 67.596	katz 194.387
modern 394.839	bun 186.962	6th 60.793	knish 184.206
egg 367.368	milkshakes 174.996	4am 55.451	sandwiches 181.415
shoyu 352.295	concretes 165.786	yellow 54.447	brisket 131.945
chashu 347.690	portabello 163.483	tzatziki 52.959	fries 131.613
karaka 336.177	shack's 159.334	lettuce 51.323	salami 127.621
kakuni 276.310	patty 152.226	sammy's 50.65	knishes 124.339
ramens 262.494	ss 149.668	sw 50.566	delicatessen 117.482
bun 236.512	patties 148.068	platters 49.906	deli's 117.431
wasabi 232.366	cam 105.949	falafel 49.479	carver 115.129
dama 221.0481	milkshake 103.972	sober 49.221	brown's 109.441
brulee 201.179	lamps 99.011	moma 48.158	matzoh 108.222

Data Collection: Example

- Advantages
  - $TF - IDF$ of stopwords?

Data Collection: Example

- Advantages
  - $TF - IDF$ of stopwords?
  - No need remove stopwords
    - $IDF(w) = 0$

Data Collection: Example

- Advantages
  - $TF - IDF$ of stopwords?
  - No need remove stopwords
    - $IDF(w) = 0$
- Challenges
  $AAAAAAAAAAAAAA$
  $AAAAAAAAAAAAAAAAAAAAAAAAAAA$
  $AAAAAAAAAAAAAAAAAAAAAAAAAAA$ AAA

Practical work 1: Data preprocessing

- Implement the above preprocessing steps in Python
  - Name it «01.preprocessing.py»
- Push your code to corresponding forked Github repository