Data Mining & Classification

sreejata@leadsift.com

Agenda

- What is data mining: The different kinds of data
- Why data mining?
- What makes unstructured text mining so interesting?
- Steps: Cleaning, transforming, collecting training data, selecting features, building a model, testing.
- Different databases: Elastic Search
- Exercise (Build a data mining system)

What is data mining & why do we do it?

- Examining extremely large datasets to uncover “new information”
- Because - so much is hidden in the huge volumes of data
- Again.. Because a lot of money is to be made
    - How?

Different kinds of data to be mined

- Numeric
- Text
- Images
- Mixed Media
- Records and Tables
- Transactional data
- Transient or time variant data
- Maps
- Genetic sequences
- Videos...

But what makes our job interesting (and well paid) is mining “unstructured data”.

Steps for mining data

- Data Gathering
- Cleaning
- Transforming
- Enriching & Integrating from different sources to make sense (Why?)
- Storage decisions (What are our options here?)
- Creating data cubes, indexes etc: Modelling
- Visualization (Not my forte!)

Initial steps: Realistic Timeline!

- Gathering: This is usually the fun and creative part of the process: finding all kinds of legal data sources to solve your problem (And sometimes grey areas, but you didn’t hear it from me)
- Cleaning: Boring! Look at the data and do it iteratively to take away the noise such as stopwords, spam, anything that you don’t need
- Transforming: No one gets this right in the first round, we decide on a schema and keep bloating it up and then change it every quarter for the first year
- Enrichment and Integration: Again, a fun part, unless the data is too sparse. Big data is good, sparse data is the most frustrating thing EVER! (Why?)

Storage of Big data (Carry over from last class)

- Options such as flat files, (json, csv), Relational databases
- NoSQL (MongoDB etc) and Key-Value Databases (BerkleyDB/TokyoCabinet etc)
- Very popular distributed real time indexing & searching: Elastic Search
- Install & run ElasticSearch & play around with it

  https://www.elastic.co/guide/en/elasticsearch/reference/current/install-elasticsearch.html

- Install Python library (pip install elasticsearch)
- Make sure it’s running from your browser (bin/elasticsearch)
- Pushing some data in & query it
- Try pushing in a LOT of data and testing with the speed of queries

Sample Code: https://elasticsearch-py.readthedocs.io/en/master/

Major “ways” to mine data

- Patterns Recognition
- Classifications
- Association Rule Mining
- Clustering
- Anomaly or Outlier detection
- Regression
- Prediction

Applications of Data Mining

Anywhere you need to ***convert data into knowledge***

- Finding unique users to market to (clustering)
- Understanding medical records
- Regression analysis for stock predictions, market predictions
- Analysing web behaviours
- Classifications for spam, sentiment, any “category”
- Detection for fraud, intrusion
- Data warehousing to understand users, predict
- What are some others?

Go over some applications - You tell me how can we solve these problems?

Some common Data Mining problems

- How can we drive up sales at the supermarket?
  - What are the things that consumers frequently buy together?
  - What do they not “plan to buy” but sometimes do anyway?
  - Which things do people usually always need to buy so we can place them farthest from each other?
- How can we find out more about a person using their Facebook/Social Media?
  - When are they active?
  - Where are their friends from?
  - What kind of topics do they always click on?

Some common Data Mining problems

- Fraud Detection
    - Steps? Things to consider?
- Detecting Cancer
    - Steps?

Tutorial

- Here's the Canadian datasets
- Use clustering or association rule mining to find interesting data points
- Put the data in ElasticSearch and use it's inbuilt functions to find "knowledge"

If you want some real life data/problems to play with here's one!

We are trying to find companies that look like each other. Aka, if we sold to company X, we can also sell to company Y, Z. Can you create a system where given company X, you can suggest some others that are exactly like it?

You can crawl it from Owler or other sources (bloomberg, angellist, crunchbase). Think of other datasets that you can use. If you want, I can give you a sample set of companies to play around with. Happy to discuss your solutions.

ElasticSearch Sample Code

```
pip install elasticsearch // cd elasticsearch<version#> // bin/elasticsearch
```

localhost:9200 → To ensure it’s running

```
from elasticsearch import Elasticsearch
es = Elasticsearch()
firstitem = {‘name’: “Sreejata”, “doc”: “Just a sample json/dict”}
res = es.index(index="test-index", doc_type=test-schema, id=1, body=firstitem) #Use loop to push more data
es.indices.refresh(index="test-index") #recerate the index
res = es.search(index=”test-index", body={"query": {"match_all": {}}}) #search
print res['hits']['total']
for doc in res['hits']['hits']: #loop over results
        print("%s) %s" % (doc['_id'], doc['_source']['name']))

res = es.search(index="test-index", doc_type="test-schema", body={"query": {"match": {"name": "Sreejata"}}})
```