Big Data

sreejata@leadsift.com

Agenda

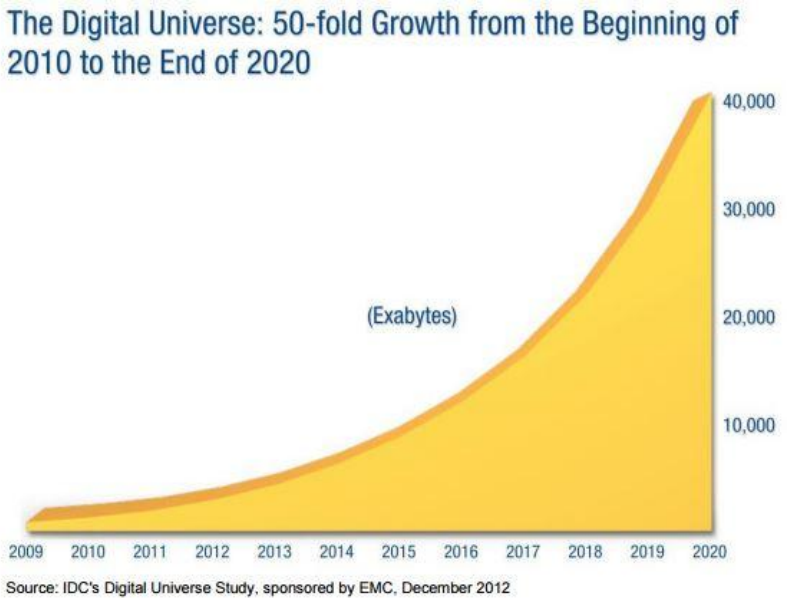
- What is BIG data?
- Why do I care?
- What are some sources?
- How can I collect it?
- Unstructured data collection
- Twitter data collection
- RSS data collection
- Databases
- Exercise (Find trending topics on Twitter and top 5 news items about it)

What is Big Data?

- Extremely large datasets (but no qualifying cut-offs!)
- Can be used to learn patterns, predict outcomes quantitatively
- Requires skills to make use of
- Can be text, numeric, images, mixed media
- Public vs Private (Examples?)

Why do I care?

- Because you're spending a lot of money on this degree!
- Information Explosion
- If there's data, then there's money to be made: but HOW? (Examples?)
- This is the bread and butter of all analysts.
- Some of the coolest tasks and applications in the modern world
- Also used for benevolent things like predicting disease outbreak (name some others?)



Sources?

- Social Media
- GPS data
- Bank transactional data
- Website visits and clicks
- … name others?

Collecting “Big Data”: Things to Consider

- Public vs Private
  - Is Facebook data public or private? (Depends on the data in question)
  - Blogs, Forums, Places you have to login to post/view
    Grocery store/bank transactions, DNA at 23&me
- Paid vs Free (Hybrid?)
  - Data as a Service
  - Twitter data can be downloaded limitedly or bought
  - Clearbit, Census Statistics Canada data, Bus routes and timing data, your GPS data
- Terms of Services: How can you use the data?
- Rate Limiting
- API vs RSS vs Database
- Visualization (Not my forte)

Unstructured Data

- Think about how GSP signals differ from a new article
- A News article differs from a tweet
- An Instagram caption differs from a tweet

Can not plan on collection or processing.

Training of a different model doesn’t work

Need to look at the data

API (Example: Twitter)

- Go to: https://apps.twitter.com/ while logged into your Twitter a/c
- Create a new app (note: you need your phone number)
- Generate your own access tokens
- Write a test script to ensure you're logged in
- Use Python-Twitter (or any other library you prefer)
- Try to fetch the last 200 mentions of Saint Mary's University

RSS

- Install and use any library (for example feedparser)
- Find your favourite news paper’s RSS feed link
- Write a quick script to find today’s top 5 stories

Read from an Index or Database

- 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/

We will use this in an assignment!

Tutorial

Find the top 10 trending topics on Twitter & search CBC or any new of your choice to find top stories.

The goal is to get you familiar with reading from an API/start playing with collecting data from different sources and playing with it.

If there’s any other Big Data problem that you fancy, please feel free to work on it.

Twitter Code Sample

```
import twitter # Load the Twitter Library

api = twitter.Api(consumer_key='sbAnI5uPhos56fux65hxXwOE2',
consumer_secret='oy3C8JLvk7vMihtIzaXdAAd4cK2k28lBDhBcGQlq882Uuj2X4O',
access_token_key='20185148-y7BudlZah0DM05tnqKLNmPzX4lWNoL5wFVGe9Zlaz',
access_token_secret='RVabWxPeqSmjwxGDxoanQnKUAleXaGclNVA92F5HTESe8')

# Please use your own access tokens from the Twitter app you set up

print(api.VerifyCredentials()) # Make sure you can authenticate

statuses = api.GetUserTimeline(screen_name='antiphobe') #Get a users posts

results = api.GetSearch(raw_query="q=smu%20&result_type=recent&count=200") #Search latest 200 SMU posts
```

RSS Code Sample

```
import feedparser
d = feedparser.parse('http://www.reddit.com/r/python/.rss')
print len(d['entries']) #How many articles in feed?
entry = d.entries[0] #first article
print entry.keys() #what are the dictionary keys?
#['updated', 'updated_parsed', 'links', 'author', 'tags', 'summary', 'content', 'guidislink', 'title_detail',
'href', 'link', 'authors', 'title', 'author_detail', 'id']

#Go over the feed

for post in d.entries:
      print post.title + ": " + post.link
```