Week 4: Data Structure & Sources

Day 1: Tuesday 14th January 2020

Process of measuring and the mathematics of its scales

Lesson 1: Scale of Measurements

During the module on intro to Data Science, one of the motivation for the profession was highlighted to be the datafication of every part of human interaction.

> *"Datafication is a technological trend turning many aspects of our life into data which is subsequently transferred into information realised as a new form of value" - Wikipedia*

To achieve datafication, the theory of measurement needs to be understood. Measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. S.S. Stevens an American psychologist suggests that there are broadly four major types of scale in which measurements can best be represented. In the table below, he presents these scale types, basic empirical operations that can performed, mathematical group structure and permissible statistics for each.

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics
Nominal	Determination of equality	Permutation group $x' = f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
Ordinal	Determination of greater or less	Isotonic group $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentile
Interval	Determination of equality of intervals or differences	General linear group $x' = a\,x + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
Ratio	Determination of equality of ratios	Similarity group $x' = a\,x$	Coefficient of variation

- Nominal Scale: This represents the assignment of numerals, words or letters as labels. A typical example is the use of numbers $[1,2,3,\ldots 11]$ to identify footballers or using the terms $[goalkeeper, defender, midfielder, striker]$ to class members of a football team based on position. The statistic relevance to nominal scale is to count the number football players who play in the midfield position or any other position. Based on this statistic we can analyse the class with the high number as well as look at the distribution of class among classes
- Ordinal Scale: This represents the assignment of numerals, words or letters as label based on the rule to preserve rank-ordering. A typical example is the use of $[Jss1, Jss2, Jss3, Sss1, Sss2, Sss3]$ to identify student in a secondary school education system. The statistical relevance of ordinal scale includes the

count as in Nominal scale as well as using median as a measure of central tendency of the data. Note that the assumption of linearity in ordinal scaling can be misleading as in some cases the rank order may not have a progressive meaning.

- Interval Scale: This represents the quantitative representation based on the rule of relationships without the knowledge of a true zero point. A typical example is the use of numbers $[1,2,3,...100]$ to identify the age (number of years) of a human in which case 0 years has not true meaning. The statistical relevance of interval scale includes those of nominal scale and ordinal scale as well as using mean and other numerical as a measure of central tendency of the data. There an obvious linearity as regards its rank-order as would be seen that someone with an age of 50 years would have obvious characteristics than someone with age 10 years.
- Ratio Scale: This represent the quantitive representation based on the rule of relationships with the knowledge that an absolute zero is always implied. This is very common in physical science with typical examples in the use of numbers$[-5.3, -3.4555, 0.123, 7.989]$ to identify the length of objects. The statistical relevance of ratio scale includes those of nominal, ordinal and interval scales as well as transformation to some derived magnitude in the sense that that they are mathematical functions of certain fundamental magnitudes.

Again, measurement is the assignment of numerals to things so as to represent facts or conventions about them. It is important to note that no measurement can be assumed to be precise or accurate as it is only as good as the empirical operations that underlines its representation. Therefore, it is best advised that no scale of measurement is void of the grounds of bias, low precision, restricted generality and human errors.

STUTERN

Lesson 2: Data Structure

Following our understanding on the different form in which data can be represented (nominal, ordinal, interval and ratio), we now look at how data can be best organised to enable efficient analysis. Taken from the field of mathematics, data structure are largely classified as primitive or abstract types. Some examples of each type is presented in the table below

Types	Structure examples
Primitive	Boolean, Characters, Integers and Floating-point
Abstract	Array, List and Dictionary

Let us review these data structure examples in a little more details

- Boolean: this is a data type that has one of two possible values intended to represent the two truth values of logic based Boolean algebra. Usual representation are $[true, false]$, $[1,0]$ or $[yes, no]$. Equality operators $==$, $!=$, Relational operators $>$, $<$, $\leq$, $\geq$ and logical operators $\neg$, $\vee$, $\wedge$ can be used on boolean data.
- Character: this is a data type that corresponds to symbol such as alphabets, or syllabary in the written form of natural language. They are mostly presented as sequential read-only objects called **Strings** $[hello, goodmorning, stay, bye]$**.** Only equality operators are usually performed on Character data
- Integers: this is a data type that represents some range of mathematical whole numbers that signed $(-,0,+)$ or unsigned $(0,+)$. Usual representation are 8 bit that ranges from $[-128,...,127]$, 16 bits that ranges from $[-(2^{15}), \ldots, (2^{15})-1]$, 32 bits that ranges from $[-(2^{31}), \ldots, (2^{31})-1]$ and 64 bits that ranges from $[-(2^{63}), \ldots, (2^{63})-1]$. Any mathematical operator such as equality, relational, logical and so on can be performed on integer data.
- Floating-point: this is a data type that depicts formulaic representation of real numbers as approximation to support the trade-off between range and precision. They are usually subject to rounding errors because they cannot represent base-10 numbers exactly.
- Array: this is a collection of primitive data structure stored at contiguous memory location that represents the storage of multiple values of the same data type together. It's most simplest form is a one-directional array such a mathematical vector $[1,4,9,16,-25,36,-49]$ or collection of strings $[bread, eggs, tea, coffee, butter]$
- List: this is a collection of primitive data structure and array but not stored at contiguous memory location. Examples of list includes $[3,5,7,[7,5,3],9,11]$
- Dictionary: this is a general-purpose data structure for storing a group of objects. A dictionary has a set of keys and each key has a single associated value. When presented with a key, the dictionary will return the associated value. Example of a dictionary is a collection names:age $\{'Detra': 17, Nova': 84, Charlie': 22, Henry': 75, Roxanne': 92, Elsa': 29\}$

Lesson 3: Data Storage

Data storage in the era of computing is the technology consisting of computer components and recording media that are used to record and store digital data.

There are a number of different approaches available for facilitating rapid data access including flat files such as .csv, traditional relational database, NoSQL database, dat warehouse and distributed processing such as Hadoop.

- Flat files (.CSV): A CSV is a comma-separated values file, which allows data to be saved in a tabular format. CSVs look like a garden-variety spreadsheet but with a .csv extension. CSV files can be used with most any spreadsheet program, such as Microsoft Excel or Google Spreadsheets. They differ from other spreadsheet file types because you can only have a single sheet in a file, they can not save cell, column, or row. Also, you cannot not save formulas in this format.
- Traditional Relational Database Management System (RDMS): A relational database is a collection of tables, each of which is assigned a unique name that uses Structured Query Language (SQL) statements to query and maintain. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.
- NoSQL Database: NoSQL databases are schema agnostic and provide the flexibility needed to store and manipulate large volumes of unstructured and semi-structured data. Users don't need to know what types of data will be stored during set-up, and the system can accommodate changes in data types and schema. Designed to distribute data across different nodes, NoSQL databases are generally more horizontally scalable and fault-tolerant.
- Data Warehouse: A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. A data warehouse is usually modelled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount). A data cube provides a multidimensional view of data and allows the pre-computation and fast access of summarised data.
- Distributed Processing (Hadoop): Apache Hadoop is an example of distributed data infrastructures that leverage clusters to store and process massive amounts of data, and what enables the Data Lake architecture. A cluster can be thought of as a single computer, but can dramatically improve the performance, availability, and scalability over a single, more powerful machine, and at a lower cost by using commodity hardware

Resources:

- On the theory of Scales of Measurement by S.S. Stevens, 1974 as part of the Journal of Science
- https://en.wikipedia.org/wiki/Data_structure
- https://en.wikipedia.org/wiki/List_of_data_structures
- https://towardsdatascience.com/everything-a-data-scientist-should-know-about-data-management-6877788c6a42
- Data Mining Concepts and Techniques by Morgan Kaufmann 3rd Edition (Chapter 1: Introduction to Data Mining and Database Technologies)

TASK: UNDERSTAND THE BASIC OF DATA MEASUREMENT, STRUCTURE & STORAGE

- Make a list of human and physical interaction that can measure. For the purpose of this module the list should not be more than 10 and should be a combination of nominal, ordinal, interval and rank scaled.
- Create an array to collect some data values for each of the interactions identified above. Note that there is some correlation between measurement and structure of your data
- Extend your work so far to create a dataframe (if your using R) or dictionary (if you are using Python) so that each row represents a tuple of 10 values that corresponds with the interactions identified in step 1.
- Save your table to a flat file stored in course directory (SGA07_DATASCI) as data.csv
- Go Beyond: Explore a database technology such as SQL or NoSQL to save your data in the cloud. Look at IBM, AWS or Google Cloud for some technology to get started with.

Day 2: Thursday 16th January 2020

Process of sourcing for information in the digital age

Lesson 1: Web Crawling

A lot of your job as a data scientist will be to find information (either in the form of raw or clean data) to help build your intuition about a given problem statement. Now that we have covered the fundamentals of measurement, different data structure and how to store/retrieve data from database, we will now briefly look at how to source for information (online and offline).

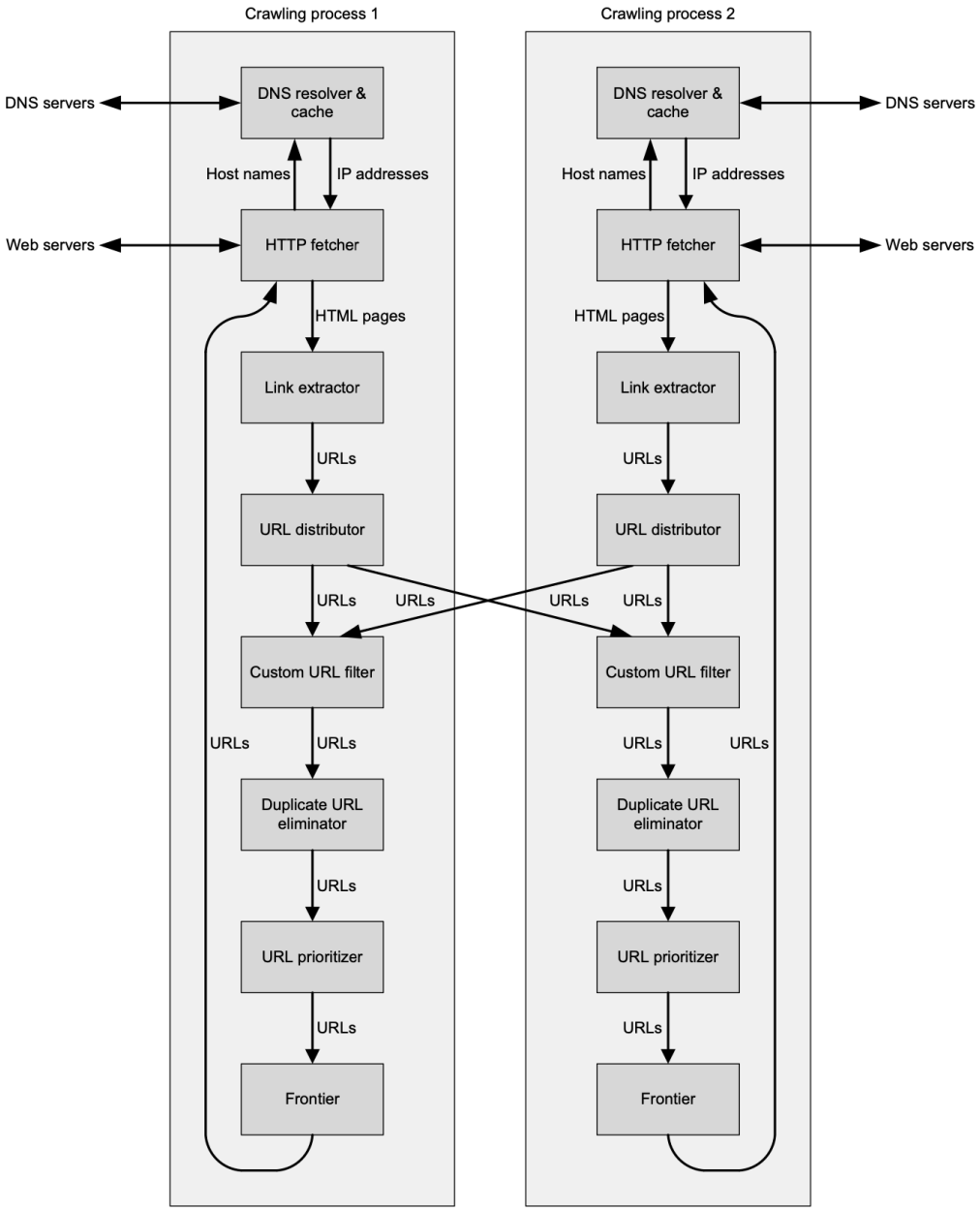
The most used approach to source for information on the internet is a web crawler. A web crawler is an agent (software program) designed to search a website(s) so as to scrape, clean and organise its data for efficient analysis. Most internet search engines (such as google, yahoo and bing) rely on the techniques of a web crawler so as to index keywords and phrases to help people find useful pages.

To build an efficient web crawler it must be . The table below give an overview of the most common web crawling techniques

Technique	Search Pattern	Benefits	Drawbacks
Breadth- First Crawling	Scans neighbor node from root level, if result not achieved then go to next level.	where the branches are small and resultant objective is identical.	When the branches or tree is very deep then goes into infinite.
Depth First Crawling	Scans from the root node and traverse next to its child leftmost node	the assurance of scanning of all node is achieved	Takes more time when the child node is large.
Targeted Crawling	Uses random (heuristics) crawling process	retrieves the greatest number of pages relating to a particular subject by using the minimum bandwidth.	Takes more time when specific topics are very large.
Page Rank Algorithm	works on the importance of the web pages. It calculates inlinks or backlinks to that page.	More accurate search result.	Difficult to manage and update page index repository.

Overall, efficient web crawling techniques follow an approach that either adopts a pull model where they will proactively scour the web for new or updated information, or they could try to establish a convention and a set of protocols enabling content providers to push content of interest to the aggregators. The figure below gives the high-level architecture of a distributed web-crawler.

STUTERN



TASK: WRITE A PYTHON PROGRAM TO CRAWL, PARSE AND STORE DATA FROM A WEBSITE

- Create a new file web-crawler.py in your SGA07_DATASCI directory
- Copy the code provided in the link https://github.com/akinlabiceo/SGA07_DATASCI.git/hotel-web-crawler.py
- Install the following packages: requests and BeautifulSoup4
- Run the code python3 web-crawler.py in your terminal

Lesson 2: Social Media

Technologies for Websites have continued to evolve from the initial era in which users simply acted as consumers of information on a static pages hosted on a server. In today's world, websites are designed with the Web 2.0 mindset to emphasise user-generated contents to dynamically build on page contents.

Social media are platforms (dynamic websites) that leverages on the technologies of Web 2.0 to curate and share information through virtual communities and networks. Most popular social media platforms. Include Facebook, Twitter, Instagram, YouTube, WhatsApp, Blogs, etc.

While, we will not dive into the psychology social media website, we will concentrate our effort on how to harness the dynamic power of social media to source for information.

TASK: WRITE A PYTHON PROGRAM TO CRAWL, PARSE AND STORE DATA FROM TWITTER

- Create a new file twitter-web-crawler.py in your SGA07_DATASCI directory
- Create a Twitter developer account
- Copy the code provided in the link https://github.com/akinlabiceo/SGA07_DATASCI.git/twitter-web-crawler.py
- Install the following packages: tweepy
- Run the code python3 web-crawler.py in your terminal

Lesson 3: Field Research Methods

Notably so, we live in an information age in which the world wide web provides a rich resource for information. As a data scientist, this plethora of data gives you power to effectively deliver on various projects. However, I believe that to provide relevant context to a successful delivery there is the need to augment data sourced online with data sourced through field research.

Field Research is the application of both qualitative and quantitative methods of data collection that aims to observe, interact and understand people while they are in a natural environment. For a data scientist, it is the application of such techniques as observation, interviews and survey to gather information that helps challenge your assumptions as well as better understand people and context.

If you are to take every data science project as a field research, the starting point is usually to formulate a research question to make sure that all stakeholders involved in the project have a common research aim. This research question can be derived from a client's brief, from customer complaints, from a previous research or any other place. The research question are often broad and vague in the beginning, but narrow down to one or more specific questions throughout the data science process.

Next is to effectively plan the research - that is organise how you will approach the research question and arrive a plausible answer. The goal of planning is to help reflect very early on the resources that you may need as well as possible constraints. A good starting point is to take a look at what is already out there so you are "standing on the shoulder of giants".

Additionally, you have to select a sample size and context for your research. As was mentioned, at the early stage of your project the problem statement may be broad and vague. However, you can intuitively start to narrow down on the scope as you get better understanding of what is out there via sample selection. For instance, if I am to work on a project that applies data science techniques to a lending business, I can start to effectively narrow down my thinking by exploring just small lending businesses or mortgage houses.

Finally, you can then begin to collect your data. There are various methods for data collection some (web crawling and social media) of which we have just reviewed. Here are other qualitative approaches that can be used for data collection:

- Direct Observation: In this method, the data is collected via an observational method or subjects in a natural environment. In this method, the behaviour or outcome of situation is not interfered in any way by the researcher. The advantage of direct observation is that it offers contextual data on people, situations, interactions and the surroundings. This method of field research is widely used in a public setting or environment but not in a private environment as it raises an ethical dilemma.
- Qualitative Interviews: Qualitative interviews are close-ended questions that are asked directly to the research subjects. The qualitative interviews could be either informal and conversational, semi-structured, standardised and open-ended or a mix of all the above three. This provides a wealth of data to the researcher that they can sort through. This also helps collect relational data. This method of field research can use a mix of one-on-one interviews, focus groups and text analysis.
- Survey Questionnaires: A questionnaire is a research instrument that consists of a set of questions or other types of prompts that aims to collect information from a respondent. A research questionnaire is typically a mix of close-ended questions and open-ended questions. Open-ended, long-form questions

offer the respondent the ability to elaborate on their thoughts. Research questionnaires were developed in 1838 by the Statistical Society of London.

TASK: USE FIELD RESEARCH TO BUILD ON INTUITION FOR FINAL PROJECT

- Develop problem statement that encompasses your final project
- Develop a project plan with tasks, timeline and milestones for the project (constraint to at least 6 weeks)
- Compile a list of interview questions you would like to ask an expert in the domain of your project
- Develop a survey questionnaire (you can use survey monkey) that you share through your social media to get a general perspective for your project

Resources:

- https://www.webfx.com/blog/internet/what-is-a-web-crawler/
- https://computer.howstuffworks.com/internet/basics/search-engine1.htm
- https://www.octoparse.com/blog/web-crawling-how-to-build-a-crawler-to-extract-web-data
- Analysing Different Web Crawling Methods by Bhavin M. Jasani, International Journal of Computer Applications (0975 - 8887) Volume 107 - No 5, December 2014
- Foundations and Trend R in Information Retrieval Vol. 4, No. 3 (2010) 175–246 2010 C. Olston and M. Najork DOI: 10.1561/1500000017
- https://www.analyticsvidhya.com/blog/2019/10/web-scraping-hands-on-introduction-python/
- https://en.wikipedia.org/wiki/Social_media
- https://ourworldindata.org/rise-of-social-media
- http://www-public.imtbs-tsp.eu/~gibson/Teaching/Teaching-ReadingMaterial/OReilly07.pdf
- https://www.znetlive.com/blog/web-2-0/
- https://www.promptcloud.com/blog/scrape-twitter-data-using-python-r/
- https://www.questionpro.com/blog/field-research/
- This is Service Design Doing: Applying Service Design Thinking in the Real World by Marc Stickdorn, Markus Edgar Hormess, Adam Lawrence and Jakob Schneider: Chapter 5 Research