UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

COMPUTER SCIENCE

MIT 805, Big Data

Semester Project Assignment Instructions

Stacey Baror

1 Introduction

Due to an increase in the world's population, technological advances, and the expansion of big corporations, many organisations face challenges when collecting and analysing large amounts of data daily. This assignment provides a broad overview and a hands-on approach to the main phases of dealing with big data: collecting, processing, reducing, and visualising the data. For the written part of your assignment, LaTeX is recommended, and the coding Github is required. There are free online version of LaTeX `https://www.overleaf.com/` and GitHub resources `https://github.com`. The semester assignment consists of two parts and should be completed in detail.

The parts are classified as follows:

1. Part 1 (Assignment 1): Data decision, collection & process: Collect and analyse the big dataset that you have chosen for you assignment.

2. Part 2 (Assignment 2): Map-reduction & Visualisation: The processed dataset is reduced to a sensible output, the visualise the dataset and then produced output to find useful relationships.

3. **For Part 1 should submit a report and part 2 a report and a 20 mins video demo that shows how your system works - that is, for processing, map reduce and visualisation aspects of your assignments output** - *Add a link to the video demo and written pdf in your Github.*

4. Ensure to submit your assignments in parts as at when due. Since the parts follow one another, it is important to thoroughly think about the decisions for your dataset in part one, as this will influence the output of your overall analysis in part 2.

5. **Hadoop Ecosystem & Github** Students are expected to show code implementation of their project using Github. Github ensures the proper grading of your contributions to all aspects of your project. When you set up Github for MIT 805, use your `tuks.co.za` email to afford you more 'private' and free resources and efficient processing. Invite me to your MIT805 Project ***Github - username:staceybaror***.

6. There are free and available resources to host and process your data and endeavour to explore them. To avoid long delays in your data processing, start with smaller datasets and intermediate increments.

7. **Plagiarism** -plagiarism is serious academic misconduct. It involves appropriating someone else's work and passing it off as one's work afterwards. Thus, you commit plagiarism when you present someone else's written or creative work (words, images, ideas, opinions, discoveries, artwork, music, recordings, computer-generated work, etc.) as your own. Only hand in your original work. Indicate precisely and accurately (Bibliography) when you have used information provided by someone else.

8. Referencing must be done following a recognised system. By referencing the resources, you should indicate whether you have downloaded information from any source.

   For more details, visit the library's website `http://www.library.up.ac.za/plagiarism/index.html`.

2 Semester Assignment Instruction

2.1 Part 1 - Collection and Process

The first part of any big data project is collecting the necessary data, which is later processed and analysed to make decisions. Since you do not have the time, budget, and resources to collect your data, you will use an existing dataset. A wide variety of free datasets are available online, and your task is to choose one that you will use for part 2 of this assignment. You can select any dataset, in the case you do not know where to start with data search, check here:

- `https://www.kaggle.com/`
- `http://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html`
- `https://www.mathworks.com/solutions/deep-learning/ai-signal-processing.html`
- `https://matlabacademy.mathworks.com/details/matlab-for-data-processing-and-visualiza mlvi`
- `https://github.com/caesar0301/awesome-public-datasets`
- `https://www.the-numbers.com/`

- And others

.

When choosing a dataset, you should consider the following:

- **Size**: Do not choose a dataset that is too large or too small. Some of these sets contain terabytes of data. You will have to process the dataset on your own computer, and if the dataset is too large, it will take a lot of time to process the entire set; if too small, it will skew your result as useful information and relationships in the data are essential for the other parts of the assignment.

  We recommend choosing a dataset between 5GB and 20GB, depending on the machine (use Google Colab to avert potential resource issues. See section 2.4)

- **Diversity**: Make sure there is enough Diversity in the dataset so that you can nicely reduce, summarise, and visualise the data later on.

- **Format**: You will find that the datasets have a wide variety of formats. We recommend using raw text, CSV, or SQL-based datasets. Technically any format should work, as long as you understand the format and can manipulate it through some code. Try to avoid datasets with binary entries (e.g., images, audio, and video) since it will increase your processing time and will be not helpful for the level of data analysis you will be doing in this module.

- **Document format:** Paper size A4, font 12pt, 1-column (No double column submission), PDF document with clear explanation of the data

2.2 Deliverable Part 1 (Assignment 1)

Once you have chosen a dataset, **write a report about the data. The report should contain some technical aspects of the dataset, such as the size, format, and age. Provide a brief overview of the data contained in the set and the reasons/procedures followed by the organisation that collected the data. Also, predict which relationships and correlations you expect to find in the data.** The majority of your report should focus on describing your dataset according to the V's of the collection and processing phases, variety, veracity, volume, and velocity. Note that since you are not working on a live database, the velocity might not be obvious or even determinable. Some datasets contain dates or timestamps that can guide you with the velocity. You may also incorporate other V's if applicable.

2.2.1 Deliverable Part 1 - date

- **Date issued: 28 July 2023**
- **Deliverable: Written report completed and process dataset**

- **Due date: Upload your report no later than Friday 18 August 2023 23:30 to the ClickUp module website and GitHub** *invite me to your MIT805 Project Github: Github username:staceybaror)*

2.3 Part 2 - Map-Reduction and Visualisation

For this part of the assignment, you should install Hadoop on a computer and familiarise yourself with the way the framework works. Based on your dataset, decide which interesting information could be extracted from it. Suppose you are working on a set collected by a commercial company. In that case, you may ask yourself which information the company might want to extract in order to give them a competitive advantage. Then write a small Java program for Hadoop that will extract and reduce the information of your dataset.

Hadoop is a free and open-source framework developed by Apache for processing big datasets. Hadoop is written in Java, and you will be able to run it under any operating system as long as you have Java installed. If you are not familiar with Hadoop, we suggest setting up a Ubuntu virtual machine since it will make the installation, compilation, and execution a lot easier. You can use any version of Hadoop, but we suggest using the latest stable version. There are many tutorials on the web explaining how to set up Hadoop. Decide on datasets that are most interesting. Determine which useful information can be extracted from the set. Code the MapReduce algorithm to extract and summarise your data. You may use separate algorithms to extract different narrow information or create a complex algorithm that combines a number of attributes to find previously unseen information. The latter approach is advised since it applies to Big Data in the industry. For instance, a grocery store might want to determine which other product a consumer might purchase if they already have product X in the basket (e.g.,: macaroni and cheese). This stands in contrast to traditional data reduction, such as simply calculating what the average consumer spends in the store. Afterwards, visualising your dataset in order to extract some meaningful information. One aim of visualisation is to discover information, relationships, and clusters that are often not discovered intuitively through writing MapReduce algorithms. Visualise the entire original dataset in order to extract new information that you did not know about when coding your MapReduce algorithm. There are many different visualisation tools. Some of these tools can be directly integrated into Hadoop to access your MapReduce code. However, integration is not expected of you for doing this assignment, do what works for you.

Here are a few suggestions for visualisation tools. Some have a unique data focus (e.g.: geographical datasets). We highly recommend Hue since it is free and cross-platform, can directly communicate with Hadoop, and can utilise Hadoops' HDFS, SQL databases, and CSV files. It is also relatively easy to install and use and has good online support. The commercial software listed here has demo versions which should be fine for this assignment. You may also use any other visualisation tool or directly create graphs from within Java, JavaScript, Python, or R.

2.4 Resources

- It is advisable to use Google Colabs for your implementations - `https://colab.research.google.com`
- Tensorflow should come in handy `https://www.tensorflow.org/`

2.4.1 Tutorial Big data Techniques

- `https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluste html`
- `https://www.crayondata.com/blog/top-big-data-technologies-used-store-analyse-data/`
- `https://www.sciencedirect.com/topics/computer-science/big-data-technology`
- m `http://spark.apache.org/docs/latest/quick-start.html`

2.4.2 Open-source or Free

- Hue: `http://gethue.com`
- Apache Zeppelin: `https://zeppelin.apache.org`
- Elastic Kibana: `https://www.elastic.co/products/kibana`

2.4.3 Commercial or Demo Versions

- Tableau: `https://www.tableau.com` – if you are interested in a free academic license, please let me know.
- Datameer: `https://www.datameer.com`
- Carto: `https://carto.com`
- ChartBlocks: `https://www.chartblocks.com`
- ZoomData: https://www.zoomdata.com

2.5 Expected deliverable for Part 2 (Assignment 2)

1. For the deliverable (Mapreduce), you should (1) push your MapReduce code to your GitHub - *(invite me to your MIT805 Project Github: Github username:staceybaror).* **(2) submit a short report briefly explaining your dataset, the approach you followed for the MapReduce algorithm, the reasons why you chose your specific algorithm/attributes, and a short discussion on the results you retrieved** from the MapReduce algorithm. The format should be similar to part 1.

2. For the deliverable (Visualisation), you should **Write a report with a discussion on the different ways you visualised your data. You should have at least two (preferable three of four) different ways of visualising subsets of your dataset. Include screenshots of your graphs and maps and discuss which interesting information you discovered through this process. The report should be about five pages, but you may add more if you have a lot of screenshots.**

3. **A 20 mins video showing your system's works - for both the Map reduce and visualisation output.**

2.5.1 Deliverable Part 2 - Due date: 06 October 2023; Time: 23:30

- **Date issued: 28 July 2023**.
- **Expected Deliverable:**
  - Written report and Hadoop code completed.
  - Submitted Report
  - A video project Demo uploaded video demo here `https://forms.gle/nLSePkKyWN1BDDUW9`
  - Code in Github - Your code should be on your GitHub (with a continuous and frequent push to your GitHub is a requirement)
  - Invite me to your GitHub for grading purposes

Proposed Exam date, sometimes between 25 Oct to 3rd November 2023.