Project Marang-Rang

DSIDE Project Report

Phase 1: Data Exploration

12-07-2019

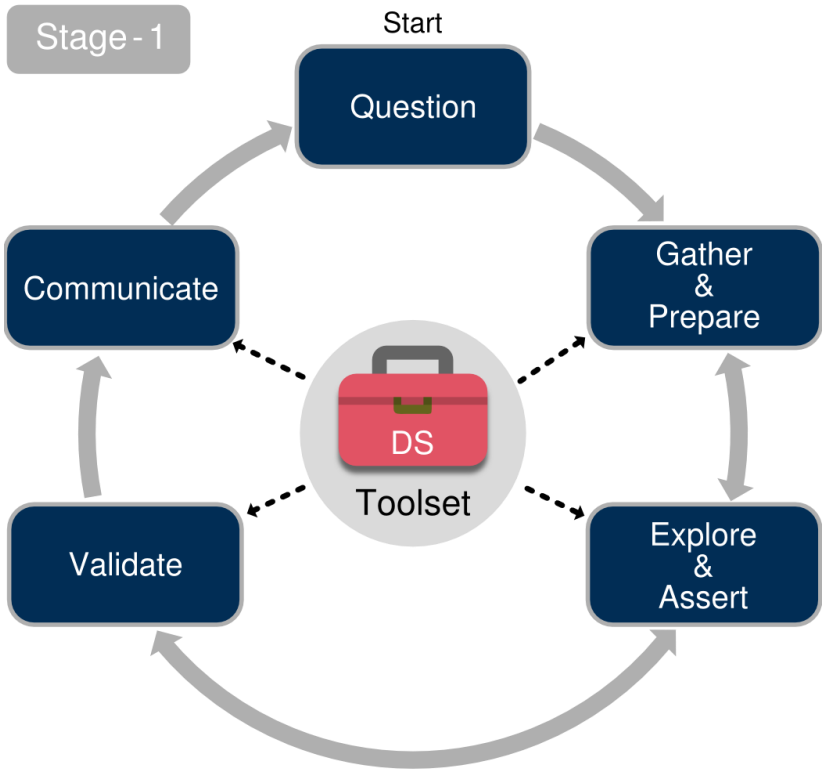
Version 1.0

Table of Contents

1. Executive Summary

The report shows the proposed design of a web-based system used to detect power and network outages. The data used to train and improve the system is gathered using open-source tools such as Nmap and Twitter. Nmap and IPinfo is used to obtain the IP addresses as well as the relevant information of each IP address such as the location of the device, whether device is online or offline, Province and so on. Tweepy API is used to extract data from Twitter using relevant keywords. Then the location of the tweet is obtained. Bar-graphs and geographical maps are plotted to analyse the data and obtain trends. After validation of the data, the proposed design and solution to solve the problem is to combine the two main tools, namely Twitter and Nmap. Nmap provides the initial phase detection of the problem, by observing the ratio of offline and online IP addresses sudden changes are detected and the locations tagged. Then to investigate the cause and correctness of the detection Tweepy is used to analyse the tweets (by searching for key words) around the area with the suspected problem. Ipyleaflet tool is used to plot the Geo-location of the data, showing the density of the offline devices, mark the tweets that reports more power or network outage in a region.

2. The DSIDE Data Analytics process



Network outage refers to the period where the devices are unable to transmit and receive data, it occurs due to link failure, power outage, network configurations and etc. Due to the occurring power and network problems in the world, web-based systems for automatic detection and incident warning has become a necessity for response task teams and formulating preventive measures. Therefore it would appear necessary to apply a method to solve power and network problems in South Africa using data analysis.

To gather and prepare data for this project open-source tools and software are are used such as Nmap for network scanning and Twitter API to search a specific keyword, such as “load shedding, power outage, network problems”, to detect power and network outages in a specific area.

2.1 Project Question/Problem Statement/Research Question/ Project Objectives

1.1.1

Due to economical crisis, power and network problems are one of the prevalent societal problems experienced by the country today. Since this is an inevitable problem, it is convenient to develop a model which will detect power and network outages around South Africa which will notify the community about the current outages. Analysing the data from the power and network users can assist to predict the next coming power or network outage. The project will help the community to prepare themselves for the next outage

1.1.2

2.2 Gather and Prepare

TASK	TASK DESCRIPTION	TOOLS USED AND DESCRIPTION	CHALLENGES	HOW CHALLENGES WERE OVERCOME
Data collection (Network)	**This task is to collect the raw data that is used for analysis and exploration.**	• **Nmap: This is an open-source tool used to scan all IP addresses associated in a network range or region**	**The terminal is used to fully utilize this tool and optimize. This proved to be a major problem.** **Furthermore, storage of the data in a relevant file format.** **Lastly, using Nmap to collect the data takes time.**	**Online free tutorials and guides.** **A python code was made to solve the problems and automate the process.** **Unfortunately, this is an inherent problem Nmap contains and could not be overcame but worked around with by choosing a small area to scan which is the Gauteng Province.**
Data collection (Twitter data)	**This task involves obtaining data from twitter using relevant keywords.**	• **Tweepy: This is a twitter API used to source data from twitter.**	**Obtaining authentication keys from twitter.**	**Changing from researcher to student account.**
Data extraction	**This task involves extracting the relevant data that is useful for the scope of this project from the raw data obtained by Nmap.**	• **IPinfo: This is an open-source tool used to extract all the information associated with an IP address such as location, organization, province, region and so on.**	**A trade-off must always be made when obtaining accurate information from open-source tools. Also, IPinfo only allows a limited extraction of accurate data in a day. This proved to be a hurdle at this phase.**	**To solve this multiple accounts were created and the extraction was spanned over a number of days. The extracted data is stored in a database CSV file format.**

2.3 Explore and Assert

TASK	TASK DESCRIPTION	TOOLS USED AND DESCRIPTION	CHALLENGES	HOW CHALLENGES WERE OVERCOME
Data exploration (Network data)	**This task involves exploring the data obtained from the gathering and collecting stages. This task is analyzing the variables in the database and preparing for visualization.**	• **Pandas and Numpy: These libraries help with data analysis.**	**Using the full capabilities of the packages.**	**On-line sources such as text-based and video tutorials.**
Data exploration (Tweeter data)	**This task involves exploring data obtained .**	• **Pandas: This is a data analysis tool used for data preparation, for the Python programming language**	**Geo-coordinates not provided along tweets**	**Location provided by the user is used instead of Geo-coordinates.**
Data cleaning	**This Task involves removing all the tweets with no location**	• **Pandas: This is a data analysis tool used for data preparation, for the Python programming language**	**User provided location non-existent.**	**Dropped tweets with non-existent locations.**
Display tweet locations on map. Obtain approximate tweet Geo-coordinates from locations provided	**This task involves obtaining approximate latitude and longitude values from locations provided by use**	• **Pandas: This tool is used to import a csv file with cities and respective Geo-coordinates.**	**csv file not exhaustive of all South African places**	**Dropped tweets with non-existent locations.**
Data visualization (Twitter data)	**Display tweet locations on map using Ipyleaflet**	• **Ipyleaflet: This is a tool for visualizing the witter locations on a map.**	**Posed no challenges. Visualization was only limited by the data provide**	**Can overcome this in future by using accurate data.**
Data visualization (Network)	**This task involves plotting the data, observing trends and obtaining insight.**	• **Plotly and Matplotlib: These tools help visualization of data by allowing the easy-use of plotting functions such as plotting bar-graphs and histograms.** • **Ipyleaflet: This tool helps in plotting geographical data.**	**Plotly is not user-friendly and does not have all the capabilities programmers need. Consequently, Plotly has limitations in plotting geographical data that is not in the United States of America.** **Due to the open-source nature of the tool, the map contained in the database is not regularly updated so that causes inaccuracies.**	**Utilizing online information and guides.**
Data analysis and creating a model	**This task involves the creation of a model to solve the Problem Statement of the project. This task uses the insight and knowledge gained from the visualization stage. The initial analysis approach was to make the two databases and**	• **Plotly,Matplotlib and Ipyleaflet: This tools help in visualization as described in the stage above.**	**This approach and analysis yielded large knowledge gaps and incorrect results. This is because the tools used to obtain the data which is Nmap and Twitter do not contain all the relevant data needed to**	**Consultation with the mentor. New ideas and a new approach was suggested as shown in the validation section below.**

	methodologies used which is network scanning and tweets independent. Furthermore, each method was to obtain all the results independently and be able to solve the Problem Statement.		**correctly detect an outage they only show a perspective. In the instance of network scanning, Nmap and visualization tools show behavior and trends of a network outage however, it does not show the reason of the network outage which can be caused by a power outage or network failure. Similarly, this problem occurs from the twitter data. This creates large knowledge gaps which affect the final results.**	
Stream tweets from Twitter	**Display tweet texts as posted by the public on screen.**	• **Use StreamListener, a tool provided by Twitter library Tweepy, to stream and display tweets in Python.**	**Could not find function/method to stop and store streams.**	**Resorted to finding tweets manually using twitter.Cursor.**
Map points	**Display point locations from tweets obtained.**	• **Ipyleaflet:: This is an interactive map tool for visualizing point locations in Python.**	**Setting the initial zoom argument.**	**Set a zoom value that displays whole of South Africa.**

2.4 Validate

TASK	**TASK DESCRIPTION**	**TOOLS USED AND DESCRIPTION**	**CHALLENGES**	**HOW CHALLENGES WERE OVERCOME**
Mentor discussions and analysis re-evaluation	**This task involves discussing the methodologies and analysis approach with the mentor and improving the analysis. The previous section shows the analysis approach which is keeping the two methodologies independent. This approach is modified and changed.**	• **Pandas: This is a data analysis tool used for data preparation, for the Python programming language** • **Ipyleaflet: the tool for visualization the data**	**Large knowledge gaps from analysis which will cause inaccurate results.**	**A new approach is suggested. Contradictory to the initial approach, the analysis shows more insight if the two methodologies and tools are used together complimentary rather than independently. This solves the large knowledge gaps. For instance, on the scenario shown in section 2.3, the network scan by Nmap will show the trends and behavior of a network outage, however the cause of the network outage will be validated by the twitter data which should also show trends and data that suggest a power outage in that region. Thus, using the two tools together yields better results and closes the knowledge gaps by exploring a deeper level of abstraction.**

Confirming map positions to be true	**Making sure that a place corresponds correctly to its coordinates (latitude and longitude) without any misrepresentation.**	• **Google searched csv file containing correct geographical (longitude and latitude) positions corresponding to the place.**	**Csv file contained certain cities in South Africa and not all.**	**Used tweets only having the location contained in the csv positions file**
Further improvements	**This task will enable the users to know which areas affected by network outage and power outage in the map, and predict the future network outages and or power outages in the area based on the previous trends**	• **In progress**	**In progress**	**In progress**

2.5 Communicate

TASK	TASK DESCRIPTION	TOOLS USED AND DESCRIPTION	CHALLENGES	HOW CHALLENGES WERE OVERCOME
Data visualization (Network data)	**This task involves plotting maps, and bar graphs indicating the chart of online devices and online device.**	• **Ipyleaflet: the tool for plotting the geographical data to show the density of the offline devices in the areas.**	**Due to the open-source nature of the tool, the map contained in the database is not regularly updated so that causes inaccuracies.**	
Twitter data visualization	**This task involves displaying map coordinates markers on a map.**	• **Ipyleaflet: This is a visualization tool used to show markers on a map**	**ELK stack installation.**	**Resorted to using Ipyleaflet**
Data storage	**Store the twitter data collectively in a dataframe.**	• **Pandas, a data set tool for Python.**	**Storing the twitter data correctly in a data frame**	**Used Pandas to sort twitter data in a readable manner in a pandas data frame.**

2.

3. Final Project Outcome

3.1 Initial-stage project deliverables

- Detecting the regions in the Gauteng province that have online and offline devices. This helps to understand which area is affected mostly by network and power outages. Furthermore, this allows for analysis of the density of the network in a region this in turn helps in analysis of the related network problems such as connection and network congestion problems.

This is achieved and visualized by various plots of connectivity of devices in that region as well as a map that shows the geographical locations of those devices.

- Tweets were obtained from twitter using Python. Tweet locations were extracted from the tweets and used to find approximate geographical coordinates. These coordinates were then

used to plot markers on a map using Ipyleaflet. These markers indicate areas were there are possible power and network outages.

- One limitation regarding this method of search is that the tweets might have the keywords searched for, but might not necessarily be stating a power/network outage in that location. The tweet text should be analyzed for more accurate location extraction. Tweets were stored on the computer's storage, the python script had a higher run-time each time new tweets were added
- The final solution was not delivered due to mainly time constraints and also the lack of historical data. The lack of historical data prevents the accurate analysis of the data-set to develop accurate models.

3. 3.2 Future recommendations

- An improvement would be to obtain a much faster way to obtain the data such as scanning for IP addresses. This would help in getting a much larger database which helps in creating more accurate models as the is more data to analyze.
- Also, obtaining a system that has high computational and storage capabilities. This will allow for the designed system to be used in real-time.
- A full automation of the process and use of the system. This allows the system to be fully independent and free of human maintenance.
- To improve the robust nature of the system by designing other network-related issues such as network congestion and connection problems.
- Using a database such as MongoDB will significantly reduce run-time.

4. DSIDE Program feedback

- To improve the allocation of mentors and communication between teams and mentors.
- Entabeni lodge is very cold. It makes the stay very uncomfortable. Heaters provided were not effective.

5. Project Team

3.1

NAME	CONTACT DETAILS	UNIVERSITY and current degree	STATUS OF STUDY	CONTRIBUTION TO PROJECT
Sinawo Dlulisa	**Cell: 0734754872** **Email: sinawodlulisa00@gmail.com**	**University of the Witwatersrand, Bsc Electrical Engineering**	**In progress**	**Data collection:This task is to collect the raw data that is used for analysis and exploration.** **Data Exploration: This task involves exploring the data obtained from the gathering and collecting stages. This task is analyzing the variables in the database and preparing for visualization.** **Data extraction: This task involves extracting the relevant data that is useful for the scope of this project from the raw data obtained by Nmap.** **Data visualization: This task involves plotting the data,**

				observing trends and obtaining insight. **Data analysis and creating a model: This task involves the creation of a model to solve the Problem Statement of the project. This task uses the insight and knowledge gained from the visualization stage. The initial analysis approach was to make the two databases and methodologies used which is network scanning and tweets independent. Furthermore, each method was to obtain all the results independently and be able to solve the Problem Statement.**
Zibuyisile Magubane	**Cell: 0734573313** **Email: buyinkomose@gmail.com**	**University of Zululand, MSc Computer Science**	**In progress**	**Data collection:This task is to collect the raw data that is used for analysis and exploration.** **Data Exploration: This task involves exploring the data obtained from the gathering and collecting stages. This task is analyzing the variables in the database and preparing for visualization.** **Data extraction: This task involves extracting the relevant data that is useful for the scope of this project from the raw data obtained by Nmap.** **Data visualization: This task involves plotting the data, observing trends and obtaining insight.**
Nondumiso Khumalo	**Cell: 064 832 7172** **Email: ndumidurbs@gmail.com**	**University of Kwakiutl-Natal, B Sc Physics (honours)**	**In Progress**	**Data collection using Tweepy and data exploration.**
Siphesihle Gama	**Cell: 062 535 4074** **Email: gamahlanganani@gmail.com**	**University of Witwatersrand ,Bsc Mechanical Engineering**	**In progress**	**Data cleaning and visualization using pandas and Ipyleaflet.**

6. Appendix

6.1 Detailed description of the dataset(s) provided

- Data was obtained from twitter using different Tweepy endpoints. The username, location, timestamp and Geo-coordinates were the endpoints used. This data was saved in a pandas dataframe. This data was obtained in order to extract latitude and longitude coordinate points for each tweet. Tweets that had the keywords but did not have location attached to them were dropped. A CVS file containing a list of cities and their respective coordinate positions was also imported as a dataframe named cities using pandas. Each user location was checked whether it matched any of the cities imported. The latitude and longitude were extracted from the cities dataframe if there was a match. These latitude and longitude values were stored in a dataframe that was used to display map markers using Ipyleaflet.
- The data obtained by scanning the connected devices at a certain range using Nmap, IPinfo to obtain all the details of the connected devices such as IP address, location, Province and the name of the organization. Ipyleaflet was used to visualize and analyze the data.

6.2 Detailed description of your development environment

Software	**Function**
Python 3.7.3	Wrote code in language
Jupyter notebook	Ran and edited code on interface
Twitter API	Allowed us authenticated access to data on Twitter.
Tweepy	Provided tools for manipulating Twitter data.
Nmap	Scan the devices connected in the network and save the data
IPinfo	Extract the data obtained with Nmap and find the IP addresses of the online devices and offline device, the location of the devices and the name of the organization
Ipyleaflet	Data Visualization tool

Project can be found under the repository: git@gitlab.com:marang-rang-2019/network-outage.git