INF1002 Programming Fundamentals Python Projects

(Confidential)

A/Prof. Daniel, Wang Zhengkui

Below are the project ideas for your teams to select. You are welcome to propose your own project ideas. The innovative and useful projects will receive bonus marks.

Project 1: Talent Recruitment Competence Study via Online JD Analytics

Problem statement:

There are many popular job platforms (such as LinkedIn, Indeed etc.) for companies to post their job openings. The objective of this project is to analyze the job descriptions from different areas. The job descriptions generally include the position levels (entry, associate, mid-senior, executive, director), country, company, job requirements or qualifications, post time etc. We are trying to study what are the skillsets that required by different positions.

Suggested datasets to be used: Students may develop a data crawler to crawl the Job description data from LinkedIn or Indeed.

Detailed tasks to perform:

- Define several job positions that you want to study. For example, to make your analysis meaningful, you may consider different SIT programmes as your study topic. For example, in ICT, there are multiple undergraduate degrees like information security, AAI, supply chain etc. You can search all the positions related to these programmes, and analyze what are the skillsets required in jobs related to each programme.
- Develop a data crawler to crawl the job descriptions. This means that we collect data that is accessible to the general public. For each job description, the information needs to crawl includes position levels (entry, associate, mid-senior, executive, director), job location, company, job descriptions(e.g. requirements or qualifications), posting time, the JD URL crawled.
- To manage your data, you can put the job descriptions in each position in one excel.
- Develop a data cleaner to clean the data and make sure the data is ready for analysis. The main analysis is try to identify what are the key skills/competencies that the JD requires. This can be done by perform the topic modeling, key words extraction or entity recognition. To clean the data, below tasks may be needed:
    - Remove the non-English JDs
    - Try to extract the qualifications/requirements / required skills sections which may potentially have the qualifications information to analyze.
    - Other data cleaning tasks, like removing the stop-words etc.
- Store all your cleaned data into another excel
- Generate all the skills that required in each position/area.

Project 2: A Comprehensive List of Latest CVE Vulnerabilities Abused by Ransomware Gangs

Problem statement:
Ransomware is one of the most severe cyber threats for business and operations. The ransomware attacks grew by 105% globally in 2021 alone. Governments worldwide saw a 1,885% increase in ransomware attacks, and the health care industry faced a 755% increase in these attacks in 2021. One reason is that the ransomware gangs may exploit the vulnerabilities published in the Common Vulnerabilities and Exposures (CVE) website. Hence the project will search and list the latest vulnerabilities with CVE numbers abused by the ransomware gangs from 2015 to 2022, and also list out the most popular ransomware vulnerabilities.

Possible datasets to use:
(1) Researchers compile list of vulnerabilities abused by ransomware gangs, https://www.bleepingcomputer.com/news/security/researchers-compile-list-of-vulnerabilities-abused-by-ransomware-gangs/
(2) Latest Ransomware CVEs – Vulnerabilities Abused by Ransomware Actors, https://www.socinvestigation.com/latest-ransomware-cves-vulnerabilities-abused-by-ransomware-actors/
(3) https://www.cve.org/ to find the latest ransomware vulnerabilities, especially in 2022
(4) Ransomware gangs are exploiting CVE-2022-26134 RCE in Atlassian Confluence servers, https://securityaffairs.co/wordpress/132186/cyber-crime/ransomware-gangs-cve-2022-26134-rce-atlassian-confluence.html
(5) https://www.kroll.com/en/insights/publications/cyber/q2-2022-threat-landscape-ransomware-healthcare-hit

Expected outcomes:

- A csv file with all vulnerabilities with CVE numbers abused by the ransomware gangs from 2015 to 2022
- A csv file with the most popular ransomware vulnerabilities with CVE numbers
- Other useful information from the datasets.

Project 3: Healthcare workers' emotions and factors during COVID19 pandemic

Problem statement:

The COVID-19 Pandemic has disrupted the lives of millions worldwide and brought challenges and turmoil to the mental and psychological well-being of healthcare workers. Studies about the past pandemic situation have also proven that healthcare workers experience emotional issues and stress, requiring support from organizations and peers. With high levels of stress, lack of rest and burnout, healthcare workers' well-being is negatively affected. Many healthcare workers have been working longer hours than usual in high-pressure environments and faced challenges of low staffing levels. The current pandemic situation has also shown an urgent need for more healthcare workers to play a part in the public health crisis. The Singapore government has even recalled individuals with healthcare backgrounds to join the industry to cope with the COVID-19 effect. This also includes the redeployment of Singapore Airlines (SIA) cabin crew members as patient care officers.

Recent studies in Singapore are focusing more on the general effects of COVID-19 on the mental and psychological well-being of healthcare workers. However, there seems to be a research gap in the factors that contribute to various emotions of healthcare workers during COVID-19. By doing so, we may shed light on designing the effective strategies to improve the well-being of healthcare workers in Singapore.

Suggested datasets to be used:
Students may develop a data crawler to crawl the profile and tweets data from Twitter.

Detailed tasks to perform:

- Develop a data crawler to identify the profiles of healthcare workers (e.g., including physicians, nurses, emergency medical personnel, dental professionals and students, medical and nursing students, laboratory technicians, pharmacists, hospital volunteers, and administrative staff) in Twitter, and then crawl the relevant data in COVID-19 period.
- Conduct sentiment analysis to identify various emotions and group the data based on categories of emotions
- Conduct topic modeling to identify the topics mentioned when healthcare workers had various emotions.

Project 4: Public responses towards healthcare workers during COVID19 pandemic

Problem statement:
During the COVID19 pandemic, the healthcare industry and hospitals have become one of the big concerns of the people. Therefore, people have got a lot of discussions about them online.

At the same time, human resource management research has shown that healthcare workers are greatly motivated by public appreciation but greatly demotivated by public ostracism. Specifically, healthcare workers reported that they felt the tensions of others in public places just because they were healthcare workers and might be exposed to viruses. All these have been proven to decrease their work engagement and well-being. On the other hand, public encouragement and appreciation, in terms of psychological and material resources given to healthcare workers, have been demonstrated to increase their work engagement and well-being.

However, no research has revealed what is happening online and thus there is no policy or guidance of people's online narrative discourse about healthcare workers. This might endanger healthcare workers' work engagement and well-being, thus might endanger the healthcare system during this public health crisis. This project has the potential to uncover the current emotions of public discussions of healthcare workers and the related topics. The results might shed light on our understanding of how to positively navigate online discourse about healthcare workers and thus help them the most.

Suggested datasets to be used:

Students may develop a data crawler to crawl tweets data from Twitter or other platforms.

Detailed tasks to perform:

- Develop a data crawler to identify the tweets with the hashtags of healthcare workers (e.g., including physicians, nurses, emergency medical personnel, dental professionals and students, medical and nursing students, laboratory technicians, pharmacists, hospital volunteers, and administrative staff) in Twitter, and then crawl the relevant data in COVID-19 period.
- Conduct sentiment analysis to identify various emotions and group the data based on categories of emotions
- Conduct topic modeling to identify the topics mentioned when the tweets showed various emotions.

Project 5: F&B Online Business Growth Study

Problem statement:

Many new F&B entrepreneurs are using digital platforms to sell their products such as YouTube, Instagram, Facebook, etc. Especially during the COVID19, there are more F&B supports online business. In this project, you are tasked to utilized various data information online like social media platforms to identify the growth/changes from 2019/2020 to now.

Project 6: Hotel Review Sentiment Analysis

Problem statement:

- Online hotel booking platforms like AirBnb, Booking, Agoda are very popular among many leisure and business travelers.
- The objective of this project is to analyze the hotel reviews left by the travelers.
- These reviews include the review title, review text, review rating, etc.
- We are trying to develop a tool to analyze the reviews from different hotels in Singapore, and identify whether customers satisfy with the hotel service (sentiment analysis) and what are the good and bad aspects of the hotel from the hotel reviews.

Possible datasets:

- You can either use exiting datasets: https://data.world/datafiniti/hotel-reviews
- Or you can crawl by yourself

Detailed tasks to perform:

- Identify several hotels that you aim to study
- Collect the reviews from the right source of platform related to these hotels
- Clean the data you have if needed.
- Perform the sentiment analysis task to understand the emotion/sentiment while they review the hotel
- Identify the good and bad aspects of the hotel from the reviews. This can be done by performing entity detection or topic modeling based on the positive and negative reviews.
- Develop the suitable visualization to show your results.

Project 7: Digital Crime Analyzer

Problem statement:

Digital crime becomes more and more popular. It refers to the criminal activity done against computers and networks or using the computer as a tool to do that activity. It can be in the form of offences against computer data or systems, unauthorized access, modification or impairment of a computer or digital system. In this project, you can develop one tool to help the crime investigators to better investigate the digital crime. In digital crime investigations, the investigators and incident responders often need to examine and analyze log records and disk / memory images or other possible resources. For example, they may utilize a history of actions to reconstruct a chain of past events and decide whether or not a crime has been committed, the circumstances surrounding the crime and the perpetrator. Normally, the online behaviors can be identified from the various log files together with digital trails or artifacts extracted from disk or memory images. So, a good data-driven solution can help crime investigators better understand the digital behaviors with various features.

Possible dataset to use:

CSE-CIC-IDS2018 (https://www.unb.ca/cic/datasets/ids-2018.html)

Project 8: Citizens' response to Governments' public policy on COVID19

Problem statement:

Due to the global pandemic, Singapore has been significantly affected. From the beginning of the COVID19 in early 2020, Singapore government has gone through different phases and each phase has its own precautions measures policy. In this project, you are tasked to understand the citizens' response to government's covid19 policy in each phase.

Detailed tasks:

- crawl the data from social media platforms or government official sites about citizens' response to the precaution measures.
- Use the sentiment analysis API to detect the sentiments of those discussions responding to government's policy
- Analyze the topics within each emotion categorizes using topic modeling or detection.

Project 9: COVID19 analyzer

Problem statement:

COVID19 has become the global pandemic and is affecting all the countries for now. Singapore government and other countries have been working together to fight for this difficult battle together. As IT professionals, your team is tasked to develop one intuitive and intelligent tool for users to better understand the situation of the Covid19 in Singapore or / and other countries. To be innovative, you can design one program with proper information that can be useful from

individual/government/business perspective. You can also integrate the data from various resources into one application to make your tool useful.

Possible datasets to use:

COVID-19 Singapore (https://data.world/hxchua/covid-19-singapore)

Kaggle datasets: (https://www.kaggle.com/datasets)

There are also many other online possible datasets that you can use.

For example, public datasets provide by Singapore government on (https://data.gov.sg/).

Information security datasets:

https://www.loggly.com/blog/ddos-monitoring-how-to-know-youre-under-attack/

https://kukuruku.co/post/some-useful-commands-to-use-during-ddos/

https://ossec-docs.readthedocs.io/en/latest/log_samples/

In this project, we don't restrict you the dataset you can use. You are free to choose the one that can benefit your tool. If needed, you can also develop your own crawler to crawl the data online, where you can learn the skills of the data crawling.

Project 10: Singapore Cyber Security in the age of COVID19

Problem statement:

The COVID-19 pandemic was a remarkable, unprecedented event which altered the lives of billions of citizens globally resulting in what became commonly referred to as the new- normal in terms of societal norms and the way we live and work. Aside from the extraordinary impact on society and business as a whole, the pandemic generated a set of unique cyber-crime related circumstances which also affected society and business. The increased anxiety caused by the pandemic heightened the likelihood of cyber-attacks succeeding corresponding with an increase in the number and range of cyber-attacks [1].

In this project, you are tasked to analyze the COVID-19 pandemic from a cyber-crime perspective and highlight the range of cyber-attacks experienced in Singapore during the pandemic. You can perform a comparison between the cyber-attacks experienced in Singapore before and after the pandemic.

Reference: [1] Cyber security in the age of COVID-19: a timeline and analysis of cyber-crime and cyber-attacks during the pandemic.

https://reader.elsevier.com/reader/sd/pii/S0167404821000729?token=F2430D16034A7A9366027C474F14F31778657C1B2657516ECF9DB994A9BAAECE204093EC704FDA6BD7A995AF583E99C8&originRegion=eu-west-1&originCreation=20210913084931

Task Allocation

Each student will contribute coding of the project.

Extra credit

Extra credits will be given for innovative features and ideas of how the dataset can be used to help the business, such as using machine learning to perform the complex data analytics and prediction.