ICT1002 Programming Fundamentals Python Projects

(Confidential)

Asst. Prof. Daniel, Wang Zhengkui

Below are the project ideas for your teams to select.

Project 1 – 6: Talent Recruitment Value Difference Study via LinkedIn Data

Problem statement:

Linkedin is one of the most popular platforms for companies to post their job openings. The objective of this project is to analyze the job descriptions related to the "sales" position from different countries. 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 how different job position levels of sales value individualistic or collectivistic competence differently.

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

Detailed tasks to perform:
- Develop a data crawler to crawl the job descriptions related to sales positions from different countries that are freely and publicly available on LinkedIn. 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.
- Crawl the job descriptions from below job locations. Try to make sure the number of job descriptions in each location and each level is as many as possible. This project can support multiple teams.
  Project 1 to study on job locations: USA, China, Singapore, Russia
  Project 2 to study on job locations: Hong Kong,New Zealand, Philippines, Indonesia, Taiwan
  Project 3 to study on job locations: Argentina, Austria, Brazil, Denmark, Ecuador, Finland, France
  Project 4 to study on job locations: Greece, Guatemala, Hungary, India, Ireland, Israel, Italy
  Project 5 to study on job locations: Japan, Korea (South), Malaysia, Mexico, Morocco, Netherlands, Poland
  Project 6 to study on job locations: Spain, Sweden, Switzerland, Thailand, Turkey, Venezuela

- To manage your data, you can put the job descriptions in each location 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 or key words extraction. 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
- Among all the extracted topics or keywords, identify those keywords or topics belonging to individualistic competence or collectivistic competence. Individualistic competence includes solution, knowledge, operation, project, implement, deliver, technology, degree, education etc. Collectivistic competence includes relationship, team, customer, client, partner, group, communication, support, member, people, player etc. The detailed dictionary will be provided later.

Project 7: Talent recruitment value difference study via Indeed or other platforms

Problem statement:
Similar to Linkedin, Indeed is another popular platform allowing companies to post their job openings. In this project, you are tasked to perform similar tasks as the LinkedIn job description analyzer. However, one challenge is that the job descriptions (JDs) may not be categorized well as LinkedIn. Therefore, it is your task to figure out the JD belongs to which level, such as assistant, associate, executive, etc. Therefore, the below information is necessary for the project: job levels, job description, job locations. They should be written in English. If you can crawl these three pieces of information, you can use indeed data.

Project 8: 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 9: 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 10: 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 11: Recession Prediction

Problem statement:

Stocks market is charting new highs, despite the covid-19 pandemic ravinging the world over. Will the bull market continue or is there a looming recession? Is there a holy grail indicator to predict recessions?

Detailed tasks to perform:
Students are encouraged to explore and experiment with data from various data sources. While students may use some of the well known indicators, more credits will be given to project teams who come up with novel indicators.

Project 12: Carpark Finder
Problem statement:
What is your biggest challenge when parking? To find a parking space? To find a free parking space? With the growing amount of vehicles on our roads, parking is becoming a more and more demanding task. Don't waste your time cruising around looking for a car parking spot, if you can get access to real-time information on nearest available carpark.

Suggested datasets to be used:
These are sample data that can be used, students can also use other type of data to add value to the application.
https://data.gov.sg/dataset/hdb-carpark-information
https://data.gov.sg/dataset/carpark-availability

Detailed tasks to perform:
- Develop an application to display list of available carparks based on sets of filters, such as areas, carpark type, free parking, etc
- Students may use API to retrieve more information, such as carpark availability in real time
- Build your own use case from the datasets, perform some meaningful analytics or provide visualisation interface.

Project 13: Price Monitor

Problem Statement:

Almost every e-commerce site has a feature that allows customers to receive alerts on the price of a particular product. The alert can be customized so that it is triggered periodically, or when the advertised price drops below the threshold and can be restricted to vendors within a geographical region. The task here is to build a similar price monitor in Python.

Suggested starting point:
You can start with this tutorial for a slightly different task. It has useful code that you could borrow (and properly acknowledge if you do!) https://www.blog.datahut.co/post/how-to-develop-a-price-comparison-tool-in-python
You can also restrict yourself to a single e-commerce site.

Detailed tasks to perform (note that these tasks can be done in parallel so the tasks can be split among members in the group):
1. You will need to be able to get data from an e-commerce site
2. A user-interface is needed to gather the requirements about the query
3. A suitable presentation strategy should be developed to provide the alert in the most helpful way (to facilitate the user's purchase, for example)
4. If the e-commerce site already has a similar alert feature, you can do a comparison with your own tool.

Project 14: Video Clip Annotator in Python

Problem Statement:

Illegal smoking is a serious health hazard in Singapore society due to the high degree of urbanization resulting in risk of exposure to 2nd hand smoke. Video cameras can passively monitor public areas and their recordings can be analyzed to identify instances of illegal smoking. The analytical methods used for this are machine-learning-based and rely on manually labelled examples. The task here is to build a Python tool that will help label (i.e. annotate) videos by clearly identifying the clips where illegal smoking is happening.

Suggested starting point:
You can start with this codebase for the task for labelling single static images:
https://github.com/tzutalin/labelImg
Images for testing and evaluating your tool will also be provided to you.

Detailed tasks to perform (note that these tasks can be done in parallel so the tasks can be split among members in the group):

You will need to understand how to extend the single-image annotator to sequences of images. A user-interface is needed to display each video in a manner that facilitates efficient and easy annotation. The annotations need to be saved in PASCAL VOC XML format (sample attached). You will need to extend this output format to include information such as start frame and end frame of the labelled activity. An interface is needed to review annotated videos for visual validation. To demonstrate the evaluation, you must label the sample data using two or more different annotators and compute the inter-annotator agreement.

Sample PASCAL VOX XML format:

<annotation>
<folder>n00007846</folder>
<filename>n00007846_6247</filename>
<source>
<database>ImageNet database</database>
</source>
<size>
<width>500</width>
<height>333</height>
<depth>3</depth>
</size>
<segmented>0</segmented>
<object>
<name>n00007846</name>
<pose>Unspecified</pose>
<truncated>0</truncated>
<difficult>0</difficult>
<bndbox>
<xmin>161</xmin>
<ymin>52</ymin>
<xmax>285</xmax>
<ymax>247</ymax>
</bndbox>
</object>
</annotation>

Project 15: 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 16: 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 17: 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 18: 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=F2430D16034A7A93660 27C474F14F31778657C1B2657516ECF9DB994A9BAAECE204093EC704FDA6BD7A995AF5 83E99C8&originRegion=eu-west-1&originCreation=20210913084931

Project 19: International flights Changes during COVID19

Problem statement:

Singapore is a connection hub of the airflights. However, the flights have been reduced since early/mid 2020 after the COVID19 starts. In this project, you are tasked to learn the changes on the number of flights in the key airports (Singapore and internally). You can select several airports as the examples. You may obtain the flights information before and after the covid19, and perform an analysis on the changes on the number of flights.

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.