|  |  |
| --- | --- |
| **Junior Data Engineer Programme - Interim Project** | |
| **Project Description** | Twitter is a social media platform where people communicate with one another using 280-character tweets, images, videos, and hashtags. Its popularity as a fast information dissemination platform has led to applications in various domains (e.g., business, disaster recovery, intelligent transportation, smart cities, military scenarios, etc.). Users on Twitter are generating about half billion tweets every day. Some of these tweets are available to researchers and developers through Twitter's public APIs. In this project, you are assigned a task from your line manager to **collect different types of data from Twitter, build your own Twitter data crawler, and store the crawled data in the database**.  Specification of the crawler:   * The crawler can collect the user's profile information from Twitter user- Joe Biden (@JoeBiden) * The crawler can collect the user's social network information from the Twitter user- Joe Biden (@JoeBiden). * The crawler can collect the tweets using the following two keywords: [Coronavirus, Vaccination].   **The database schema, data structure design, and data storing in PostgreSQL or SQL server should also be included in this project.**  You can select open-source libraries (e.g., Tweepy) for building the Twitter data crawler. You may consider using some GUI tools (e.g., pgAdmin) to view the data in PostgreSQL or SQL server.  **Reference:**   * Twitter Sentiment Analysis, using Python: <https://youtu.be/o_OZdbCzHUA> * Tweet Visualization and Sentiment Analysis in Python - Full Tutorial <https://www.youtube.com/watch?v=1gQ6uG5Ujiw> * Scrape and Store Tweets using Python 3: <https://medium.com/analytics-vidhya/scape-and-save-tweet-using-python-3-5a186777c388> * Step by Step: Twitter Sentiment Analysis in Python: <https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d> |
| **Project Objectives** | 1. Use Python to develop a Twitter data crawler through Twitter's public APIs to collect user’s profile information, user’s social network information and tweets data. 2. Explore the web and select appropriate open-source libraries to help solve the above project. 3. Install the PostgreSQL or SQL server in your environment and design the database schema, tables and entity-relationship model(s). 4. Apply SQL in Python to insert, update and view the crawled data in the database. |
| **BSM** | All (see [BSM Evaluation Rubric](https://docs.google.com/document/d/1TFShS188dINbDR3SZU2KZBdqHnfJIJ_9Ylxe0CleGH8/edit?usp=sharing)) |
| **Duration** | * Interim Project: 10 hours + Presentation: 20 - 30 minutes per team |
| **Project Team** | * Small group of 3 - 4 members |
| **Project Structure** | The project is organized into **3 different phases**:   |  |  |  | | --- | --- | --- | | **Project Phase** | **Focus** | **Session Duration (hours)** | | **Phase 1**  (2 hours) | Project Briefing and Team Formation | 0.5 | | Brainstorming | 0.5 | | Bridging Python and SQL programming knowledge with Project | 0.5 | | Exploring the appropriate open-source libraries | 0.5 | | **Phase 2**  (5 hours) | System Design   * Data Structure Design * Crawler Design | 1  4 | | **Phase 3**  (3 hours) | System Integration Testing | 1 | | User Acceptance Test | 1 | | Presentation Prep - Creation of presentation deck and rehearsal | 1 | |
| **Project Deliverables/ Outputs** | 1. Your **project portfolio** should include the following:  |  |  |  | | --- | --- | --- | | **Phase 1** | **Phase 2** | **Phase 3** | | * A summary of open-source libraries comparison linked to project * A list of ideas generated, consisting of conceptual sketches, function and key components | * A document about the database schema, table structure, and entity-relationship model * The source code of your crawler to finish this task | * A user guide about how to use crawler * Twitter data crawler prototype |  1. **Presentation slides**    * Guide: [Guidelines to Create a Presentation Deck](https://docs.google.com/document/d/1UwFH_8KzTti7XCvyw06qfLWhdznMgnLuRRvxnTFUsBM/edit?usp=sharing) |
| **Presentation/**  **Showcase** | * Time is allocated for presentation prep and rehearsal before the actual presentation.   + Guide: [Tips to Deliver a Presentation](https://docs.google.com/document/d/1pWhRsFhhHjx66AJix9wTh8ZNgXFJHcjWNg7nPrUtfBU/edit?usp=sharing) * Each team will take turns to present the completed project and showcase their prototype Twitter data crawler to an assessment panel. Every team member is expected to present in the presentation |
| **Assessment** | * **Formative Assessment of Project Deliverables** -   + Specific qualitative feedback will be provided by instructor(s) for each team’s performance (technical)/ deliverables throughout all phases.     - Download to fill up and submit this [Deliverables Checklist](https://docs.google.com/document/d/1DunFzAL-Gdzcgj7al-dOC9VNPecu8s-kkVmRHpOtYWk/edit?usp=sharing) at the end of each project phrase:   + Team score will be given for presentation based on a rubric.     - [Presentation evaluation rubric](https://docs.google.com/spreadsheets/d/1yALgVTOIBghvTHMviuvUzP24aeUtkpOQ7cZnLmXlt20/edit?usp=sharing) * **Peer Assessment** - Participants will assess their team members’ BSM and collaboration in the project.   + Download to fill up and submit this [Peer Assessment BSM Form](https://docs.google.com/document/d/14vxwpstyNORWcnvWEBcbBMNEvfOchZgo4PdcDq5AAY4/edit?usp=sharing) by the end of the project. * **Self-Reflection** - Participants will complete a self-reflection form to self-assess how they have demonstrated BSM in their project teams.   + Download to fill up and submit this [Self-Assessment BSM Form](https://docs.google.com/document/d/1NbQsyCLVMDf_fAqEfS5QG7AtsaoRDEQZzgjPUM3pldU/edit?usp=sharing) by the end of the project. |