Big Data Science project 2020

Goal: Formulate a research or business question regarding the ongoing COVID-19 pandemic, and solve your question using Twitter Data (gathered by using the Twitter API, see Lab 6). You can use additional data sources as well. The business question and its result should be created with the right audience in mind: a C-level executive in a firm. This person is smart but not a data scientist so communication is key!

The second part of the project focuses on interacting with a different role in a company: data engineers / data architects. Here you should turn your solution into a small product. This can mean a number of different things: you can create a nice ETL-pipeline, a well-designed API, a stand-alone data visualization (ex. d3 or dash.py, or commercial ones such as PowerBI), do something with a graph database,... The main goal here is that your fellow data engineer likes what you did and you deliver a reusable app.

You can use all the methods that you have seen in the BDS labs/classes.

You will be working in teams of two, pairings will be decided by the teachers. The expected workload is about 4 to 5 full days per person. Use GitHub to collaborate; this also provides a timeline for the evaluation.

Some examples of general research questions:

- Predictive analysis: Try to predict certain aspects of the pandemic: the evolution, infection density based on the tweeting behaviour.
- Trending topics/hashtags: Do some sort of Sentiment Analysis (positive, neutral, negative) on hashtags, try to find positive messages among all the news on the COVID-19 virus, try to identify trending topics, analyse the evolution of certain topics through time since the start of the corona crisis, ...
- Identify Spammers: Try to identify fake accounts, for example, accounts that are used for propaganda/fake news/....
- Identify the top influencers: identify people whose tweets *influenced* (define this in a meaningful way by yourself) a lot other people.
- Recommender system: Come up with an algorithm that, based on your Twitter account, recommends you hashtags/pages/people tweeting about the virus.

- Visualization: Do any sort of interesting visualization of the results of the above algorithms, or visualize interesting aspects of the data in an informative/interesting way (e.g. by creating a Dashboard).

Take your time to pick a business questions and to choose a data product approach. To guide you in this process you can send an email to dieter.dewitte@ugent.be with your idea.

Project deliverables

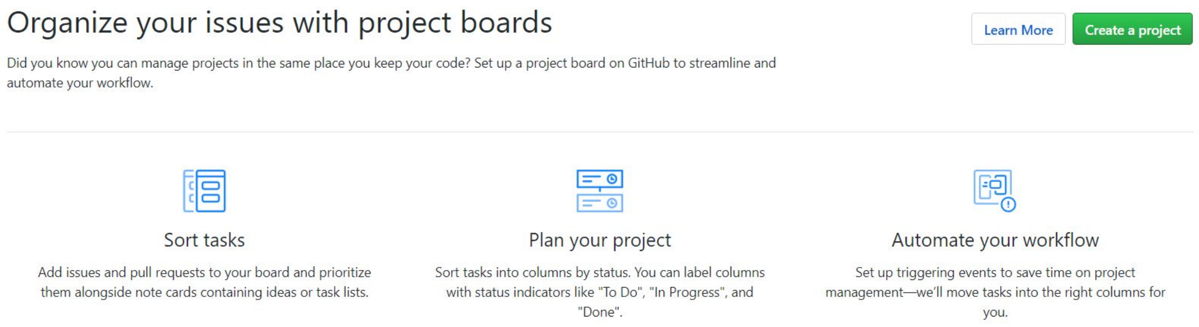
The deliverables of this project are:

A. **a 2-minute video,** in which you present your **business question**, your conceptual **solution** and the key results of your work; (audience: CEO)
B. **a 3-minute video**, in which you present a walkthrough of your demo (audience: data engineer)
C. **your codebase**, as a link to a Git repository. We expect one or several documented Python notebooks (or similar) with all your code that is not part of your demo, i.e. your experiments, problem analysis, data loading or processing, ...
D. **a demo application** (e.g. implemented in a Python notebook). This should be a self-explanatory standalone demo of your solution. Ideally you add a `/demo` folder to your github project with a decent README.md such that we can run your demo without any additional questions. Other code, e.g. data analysis to understand the problem, experiments, data collection/processing (that is not part of your demo), ... should be submitted as well. We recommend choosing an intuitive folder structure.
E. **project management.** To understand how you handled the project we ask you to use the Github project board (see Fig. 1). Use it to divide the tasks, track progress,...

Each team member should present in one video (A or B).

Links to your demo application, codebase and project board should be uploaded on Ufora before **20 May 23:59:59**.

Links to the videos should be submitted before **27 May 23:59:59**.



Organize your issues with project boards

Did you know you can manage projects in the same place you keep your code? Set up a project board on GitHub to streamline and automate your workflow.

Sort tasks

Add issues and pull requests to your board and prioritize them alongside note cards containing ideas or task lists.

Plan your project

Sort tasks into columns by status. You can label columns with status indicators like "To Do", "In Progress", and "Done".

Automate your workflow

Set up triggering events to save time on project management—we'll move tasks into the right columns for you.

Fig. 1: To manage your project we highly recommend the use of the GitHub project board. You can create subtasks (= issues), discuss them with your project member, track progress,...