Homework #3. Download and understand a Dataset

Context

Now, you're starting your road as a Computational Social Scientist. In the scope of this project, you'll analyze a person's behaviour patterns based on internet activity. Hereby internet, we mean messaging, more precisely, your activity in the Telegram app (if you have less than 20 dialogues please - communicate with me directly).
It's sensitive data, so it's good for you to feel responsible for the matter. We will analyze data and understand the person (or persons) behind the data.

Please remember - you should do the task in your head first. I mean, go throw the pipeline described in the task section imaginary and become comfortable with it.

Task #1

In this task, you'll download and prepare data for the project. Calculate the time you spent on this task.

1. Install a tool (download, go through the steps in the **How to run** section)
   https://github.com/SanGreel/telegram-data-collection

2. Get Telegram API credentials
   https://my.telegram.org/apps

3. Set credentials (api_id, api_hash) in *config/config.json* (should be based on the *config_example.json*)

4. Download all dialogues, save the time spent on this activity.
   `python 0_download_dialogs_list.py --dialogs_limit -1`

5. Download dialogues data, save the time spent on this activity (at least **100k messages per chat**).
   `python 1_download_dialogs_data.py --dialogs_ids -1 **--dialog_msg_limit 100000`**
   ***penalty for less than 100k msgs parameter value (it's ok if you don't have such amount of data in the chats)**

* PRs with fixes and improvements are appreciated (they give you extra points).

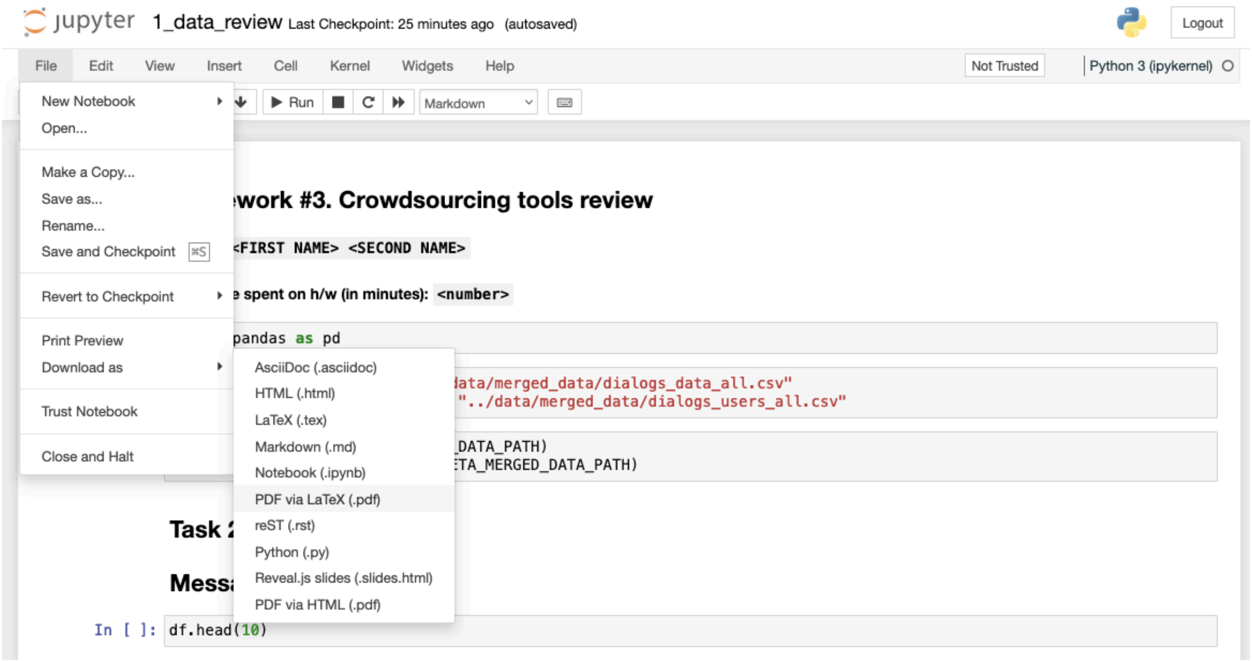
Caveats

1. Telegram can block activity from your IP and account.
   If the problem is faced you should record 1-minute video, you can use Zoom to record a call.
2. If so “TG propaganda dataset can be used” - contact Andrew Kurochkin to receive the data.
3. If there are any concerns for you with such type of data - contact me Andrew Kurochkin, please.

Task #2

During this task, you’ll take an initial look at the data. Calculate the time you spent on this task.

1. Fork and clone repo on your local machine - https://github.com/SanGreel/telegram-dialogs-analysis-v2.
2. Run all cells in the file `0_merge_data.ipynb`.
3. Open `1_data_review.ipynb`, do all tasks in the file, download your notebook as pdf (with all screenshots, output, and any additional text you want to add in the report).



work #3. Crowdsourcing tools review

<FIRST NAME> <SECOND NAME>

e spent on h/w (in minutes): <number>

```
pandas as pd
```

```
lata/merged_data/dialogs_data_all.csv"
 "../data/merged_data/dialogs_users_all.csv"
```

```
_DATA_PATH)
TA_MERGED_DATA_PATH)
```

Task 2

Mess

```
In [ ]: df.head(10)
```

Private msgs to fix in the work

Expected Outcome

1. PDF report (file named “3_<your_id>.pdf”) exported from the file `3_data_review.ipynb`.
2. 1 csv file named “3_<your_id>.csv” with all fields filled.
   ATTENTION: naming is critical as I’m using an automated approach to merge your

answers for the review.
File with fields to fill can be downloaded here - https://docs.google.com/spreadsheets/d/12yFMNEo57zBYM2-5TWUSpubEfgi-kmC67YVBSOCiIsA/edit#gid=0.

Your ID can be found here - https://docs.google.com/spreadsheets/d/1U9hFib_R1OxEo8G6ZNNNJEFeBzjIdL27WdAVLY4qct8/edit?usp=sharing

Deadline

26/10/2024, 23:59 (max: 10pt)