THE SIMONS FOUNDATION

Candidate Project

PART 1)

Download the nucleotide from 30271926 NIH's nucleotide database using their API:

https://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch

Using at least the following parameters and values when calling the EFetch endpoint:

- **db**: use the "nucleotide" database
- **id**: use "30271926" (~30KB)
- **rettype**: use a retrieval type of "fasta"
- **retmode**: use a retrieval mode of "xml"

Build a simple web application (it doesn't have to look pretty) that takes as input a regular expression pattern representing the nucleotide sequence the user wants to search for. For example,

"(AATCGA|GGCAT)"

The web application should allow the user to be able to navigate through and/or visualize the results. *(For example, you may want to consider showing a list of positions within the data that the pattern was found along with the found pattern)*

The application should be built using:

- Django: https://www.djangoproject.com/

Optional Packages that may be helpful:

- multiprocessing: https://docs.python.org/3/library/multiprocessing.html
- Postgress: https://www.postgresql.org/
- Django-restframework: https://www.django-rest-framework.org/
- Celery: https://github.com/celery/celery (for distributed processing)
- Django-filters: https://django-filter.readthedocs.io/
- Memcache: https://memcached.org/
- Virtualenv: https://virtualenv.pypa.io/en/latest/
- React: https://opensource.fb.com/projects/react/
- Or any other packages you wish to use [but please do NOT use any cloud-services].

PART 2)

For the second part of your project please do the same thing but this time write a command line utility and use nucleotide “224589800” (~238MB) instead. You can write a django management command or just write a standalone python script. The script should print out the results to stdout or write them to an output file.

Other Guidelines:

Your project should include any management commands that you may need as well as a requirements.txt that includes all the packages you used.

Please commit your work to a github account of your choice and share the url. The github project should include a README.md that includes your architecture and design decessions and explain why you took certain approaches and why you may have chosen to use certain packages. What problems there may be with your solution and how enhancements could be made in the future if you had more time. For example, would your solution from Part 1 work with the data set from Part 2?

Note: Even if your project is not complete at the end of the time-frame your work will still be considered.