Snapppt Programmer Assessment ReadMe

1. Running the Code

- Run the following commands:
  - **bundle install**
  - **ruby data_massager.rb** [To run the data massaging script]
  - **ruby data_massager_test.rb** [To run the unit tests of the data massaging script]

Note: Upon running the massaging script, **10 new files** will be created with **100 objects**, each with massaged data from the **original user's json file.**

2. Algorithm and Solution Approach

Solution Approach

I have used the **gem 'json-streamer'.** The decision to use this gem targets the **bonus point of reading and handling large json files**.

- **json-streams:** https://github.com/thisismydesign/json-streamer
  It uses file streaming and reads the objects one by one.

To create the test cases for the script, I have used the **gem 'test-unit'**
- **test-unit:** https://test-unit.github.io/

And, to write the massaged json object into a file, the **gem 'json'** is used.
- **json:** https://rubygems.org/gems/json

Algorithm

- A file naming counter is initialized along with an empty massaged json_array.
- The file is passed into the json-streamer parser.
- Objects are taken into memory one by one from the json file.
- Each object is sanitized, i.e., _id field removed, empty arrays removed, and alpha-numeric characters are removed from the bio.
- Once the object is ready, it is pushed into the json_array object.
- After the array has 100 objects, a method is called which dumps the array objects into a newly created .json file.
- Then, the json_array is empty, and the file naming counter is incremented.

- Hence, the process repeats until all the data is parsed and dumped into separate files.

3. Test Cases

I have used the **test-unit** gem to test out everything. You can run the test cases using the **ruby data_massager_test.rb** command.

- Test 01: checks if the empty array is removed from the given data.
- Test 02: checks if the _id field is removed from the given data.
- Test 03: checks that the sanitized object is the same as the passed data when correct data is passed.
- Test 04: checks if alpha-numeric characters are removed from the biofield.
- Test 05: checks if a new file is created successfully to dump the chunked data.
- Test 06: checks if the data inside the newly created files is accurate and does not have empty arrays, _id field, and alpha-numeric characters in the biofield.

4. Developer Notes

- I have tried to fulfil the requirements and the **bonus points** given in the assessment documents.
- **Also, I have tested it with 10k+ data and it works smoothly without in-memory consumption. You can also stream data live on runtime and it will work. (Test it by making the program sleep for 5 seconds and add more data on runtime in the file)**
- Please let me know if anything is missing from the requirements or if additional functionality is required.
- **Email:** aslam.bilal0011@gmail.com

The End