First peer review report (Cohortney)

Polina Pilyugina, Alexey Voskoboinikov

April 23, 2021

1 Are project objectives clearly explained? If not write what was not clear. Provide your opinion on the problem.

Overall, the main idea to create a unified library with a variety of methods for sequences clustering is clear. It would be beneficial to include description of the main Cohortney method in more detail and outline its main characteristics. Additionally, comparison and description of other methods would also help to understand the problem more, in terms of why these particular models are of interest for possible future users. Also it might be useful to include examples of real-life applications of the sequence clustering.

In our opinion, these methodology will be beneficial for many industry applications, such as banking, for example. And these project will allow interested scientist to have a useful tool for their experiments.

2 Is proposed baseline solution relevant to the problem? Could you recommend something to the team?

We believe that in terms of dataset choice it is wise to use the dataset proposed by the collegues, as it is used in Cohortney benchmarking and they are relevant to the problem. We believe that proposed baselines are a good set of models to include in the initial version of the library. As for the baseline solutions, it would be good to add links to the repositories with existing releases of this algorithms as reference, even if they would not use them in their work. If there is no existing releases, then it is better to state it explicitly. Also, it would be great to understand the various Cohortney-based solutions, included in the baselines. In particular, how these models relate to the Cohortney and what is mean by Cohortney in Cohortney-base models.

3 Is it clear from the report how the team is going to test / evaluate the results? (not only metrics but code and bugs). Could you recommend something here?

As for evaluation, the proposed idea to benchmark using existing datasets and results from Cohortney repository is explained. We believe that proposed inclusion of synthetic datasets would also allow for better understanding of how the model works and it would be beneficial for future users. We would like to recommend to additionally add unit tests with minimal examples as a pre-commit hooks for developers version of the library, if the colleagues would find it useful. It may also help for future contributors to the library.

4 Is it possible to guess next project development steps from the report and github repository? What a team is going to do next, provide with 3 next possible steps?

Outline of the project description and readme file in github allow to understand the structure of the project steps. We guess, the next steps of the project development would include:

- Evaluation of existing Cohortney library and its API, as it will serve as a base for other model APIs. Writing the code and API for their own models.
- Writing unit tests for the models to test their functionality and scalability
- Evaluation of the results on benchmarking datasets. Additional code improvements and unit tests if required.

5 Would you recommend to improve the report and how? Or it is all good.

In general, the report provides a nice view on the project structure and main steps. It would be good to include more details about particular models which will be implemented. Additionally, it will be interesting to read more about the challenges behind the differentiation of the models, as was suggested by the authors in the Main Challenges part.

Also, links to the scientific papers and github repositories related to the Cohortney-based models would help reader to better understand problem statement.

6 Is project github repository easy to follow?

As of now it contains mostly the report and the readme description it is hard to judge on the future structure of the code base. However, we see that structuring the code base and API choice is the big first step, so it will appear afterwards.