Assignment for position: Data scientist

Frontiers runs a number of open access journals in several scientific fields. Authors can submit their articles for publication to one of these journals. However, in some cases the authors may not be aware of the journal that best matches the scope of their paper. If the wrong journal is chosen, it may result in delays or even rejection. To this end, we are developing a feature that suggests to the authors the three most relevant journals to their manuscript, to choose from.

You are tasked to build a text classifier for this feature that, given some input text, can recommend the most suitable Frontiers journals to it.

You have at your disposal a .jsonl file containing

- Article identifier
- Body text
- Frontiers journal name

for all articles published by Frontiers in January 2020. You can find it here: https://drive.google.com/file/d/1es3EX0MdDAeolwFl_K_fS3RP0JFRxE2U/view?usp=sharing

Remarks:

- The solution should be coded in Python.
- You can use any Python library you may find useful.
- Together with the code you should also provide a report where you describe your approach and present the results.
- You are particularly encouraged to discuss the choice of the evaluation metric(s) and how this translates to business value.
- (last but not least) As you write code for this assignment, keep in mind that it will be reviewed (and in real life, put in production) by other colleagues. Clean code, a modular structure, python packaging, testability, explicit dependencies, documentation, are all things that can facilitate the team!

Please email your solution in .zip format to davide.fiocco@frontiersin.org and be prepared to discuss it in the next interview stage.