Project 2: insurance reviews

Dataset

Train dataset
https://drive.google.com/file/d/1r3ZwNtY1f7lWrYKgC-hk0GJCcoxSuTej/view?usp=sharing

Test dataset
https://drive.google.com/file/d/1qsapCTYbmWFQdBWOCsXdwVlvU1-gOYHn/view?usp=sharing

This dataset contains reviews published by insurance customers. There are different columns

- date: it contains the date of the publication of the review and the period of the review experience, and it has to be cleaned.
- note: it is the number of stars given by the customer (it should be predicted in the test dataset)
- auteur: the id of the customer
- avis: the review
- assureur: the name of the insurance
- produit: the type of insurance

Exploratory data analysis

You can explore the data to better understand it:

- you can study and visualize the evolution of the number of stars (by insurance, etc.)
- wordcloud to visualize the frequency of the terms used in reviews (by insurance, etc.)

You can explore the words used in the review and do some cleaning.

Unsupervised learning

You will create an unsupervised model to better understand the reviews, and create segmentations that you can interpret.

Bonus: train a word embedding model and analyze similar words and word analogies.

Supervised learning

You will create different models to predict the number of stars of the review with the training dataset.

And then you will predict the number of stars of the reviews in the test dataset.
You will submit your results that will be evaluated with the RMSE metric.

The best projects will get a bonus mark !

Bonus, again: train a text generation model and generate review samples.

Unsupervised learning part must explain what topics are present in the dataset and the supervised learning part must predict a rating given a review.
Your report must explain what technics/approaches you use, how you use them and the results obtained. If an approach doesn't work as planned you can show and explain (It will be very appreciated).
You can work in pairs of students. Your report must contain the names of students involved.
Your report must explain the logic of your approaches and results.
You can write in English or French.
Your report must contain your link to your Colab Notebook.
Your report must be deposited on DVO before **7 january 2022**.