Testing Assignment

Senior Data Scientist (N.Rich)

The problem

You work as an outsource Data Scientist for the job board company (e.g. seek.com). The product manager wants to group vacancies into different categories depending on the job title. The only limitation is that the actual production data of the company is not available yet. So you're asked to provide your solution based on the provided open-source data.

Note: you don't need to spend more than 6 hours on the assignment. Please provide your best solution in the given timeframe.

The data

The training data consists of augmented texts with the associated category code. The occupation data include names of the categories associated with each code.

The data structure is:

- train_df.csv (augmented data)
    - Code - the occupation category
    - Title - the augmented job title
- test_df.csv (augmented data)
    - Code - the occupation category
    - Title - the augmented job title
- occup_df.csv (occupation data)
    - Code - the occupation category
    - Occupation - the name of the category

The task

Based on the data provided:

1) Construct an interpretable ML model + for each occupation category extract the most important words (from to 2 to 5 words is sufficient) affecting the decision of the model that the title belongs to the certain category.
2) Omitting interpretability, construct an ML model that gives the best performance. The metrics shall include the category-wise metrics as well as the overall metrics. Dump the best model.

Note: the first task assumes that the model can have relatively bad performance, the main point is the extraction of the most important words for each category.

Output

The code should be written in Python (.py, .ipynb), and all intermediate files (e.g. model dump) should be also included. The code shall be reproducible so the assessor can run it easily.