Technical Task

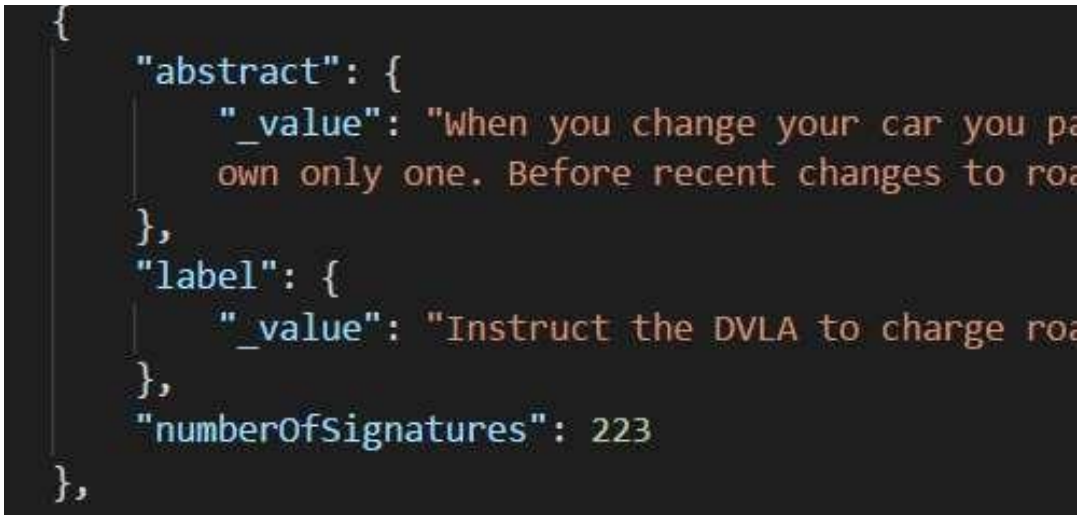
Please complete the task as instructed, and then upload your code and outputs to a public Bitbucket or GitHub repository.

Task

- Use python or pyspark to transform the input data into the desired outputs.
- Include at least one test.
- Keep best practices in mind.
- Place your code and outputs into a Bitbucket repository.

Input Data

The input data is a JSON file containing a set of government petitions. For each petition, we have the label (the title of the petition), the abstract (the main text of the petition), and the number of signatures.



Output

Create a CSV file with one row per petition, containing the following 21 columns:

- petition_id: a unique identifier for each petition (this is not present in the input data and needs to be created)
- One column for each of the 20 most common words across all petitions, only counting words of 5 or more letters, storing the count of each word for each petition.

For example, if "government" is one of the 20 most common (5+ letter) words, one column should be titled government. If the first petition includes the word "government" three times, and the second petition does not mention "government", then the government column should read 3 for the first petition and 0 for the second petition.

petition_words	
PK	petition_id
	government
	economy
	...