Relyance Data Science Challenge[1]

Purpose

We would like you to work with our data and produce a model to classify sections of text. You will get a feel on the type of data we deal with and we can assess your data science approach. We would like you to use Jupyter notebook and python to do your work.

Background

Our platform reads documents such as privacy policies and extracts specific sections of text to provide insights on the contents of the privacy policy. We have provided a sample of our datasets for you to use. We have simplified the problem by restricting the types of sections to predict and extract.

Objective

1. Write python code that would:
    a. Read the files data.csv described below
    b. Analyze the dataset, do minimal data imputation as needed
    c. Develop a model to classify the text sections and assess accuracy
    d. Allow you to present to a Data Science team
2. Guidelines on coding:
    a. Comment your code for us to follow your logic
    b. Do not spend a lot of time of formatting/fonts
3. Per our NDA, the notebook should not be posted anywhere, but emailed to ds_challenge@relyance.ai as a compressed file. We will install any required open source packages.

Data Description

The csv file, data.csv contains a list of sections and annotations from multiple documents. You can choose to use the columns as you see fit in developing a model to predict the section. Below is additional information on the dataset.

- doc_id: A unique id of the document from which the sections were originally extracted
- text: This is a portion of the document which need to be classified into a section
- section: This is the correct section the 'text' belongs to
- h1, h2, h3, h4, h5: These are hierarchical headers for the text