The idea of the project is to be able to successfully use machine learning cocepts and build a document classification model. A document labeled dataset is provided, that has 62204 entries of documents. Each entry has a document label and an OCR output of the document. All the words of the document was obscured using fixed length hashing.

The dataset looks something like this:

CANCELLATION NOTICE, 586242498a88 21e314d3afcc 818a7ff3bf29 4e43b72d46c0 578830762b27........ 43565b1afa44 5f6653c869fc

Step 1: Understanding the data

The very first approach was to look at the data and understand what they mean at a very high level. The document can be labelled into 14 different categories and the following number of records:

|  |  |  |
| --- | --- | --- |
| BILL | 18968 |  |
| POLICY CHANGE | 10627 |  |
| CANCELLATION NOTICE | 9731 |  |
| BINDER | 8973 |  |
| DELETION OF INTEREST | 4826 |  |
| REINSTATEMENT NOTICE | 4368 |  |
| DECLARATION | 968 |  |
| CHANGE ENDORSEMENT | 889 |  |
| RETURNED CHECK | 749 |  |
| EXPIRATION NOTICE | 734 |  |
| NON-RENEWAL NOTICE | 624 |  |
| BILL BINDER | 289 |  |
| INTENT TO CANCEL NOTICE | 229 |  |
| APPLICATION | 229 |  |