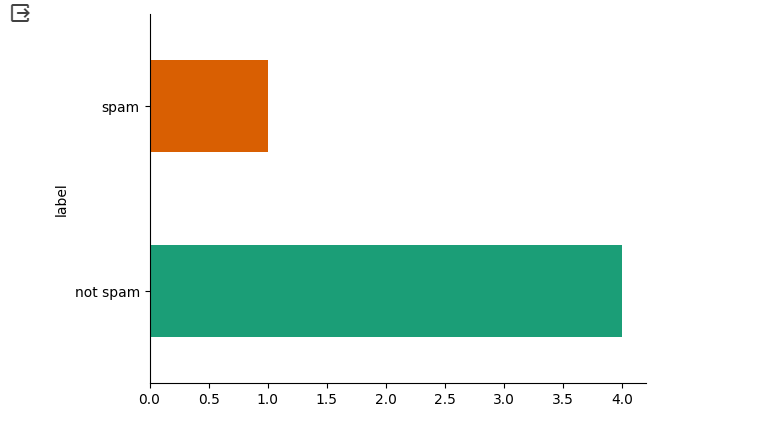
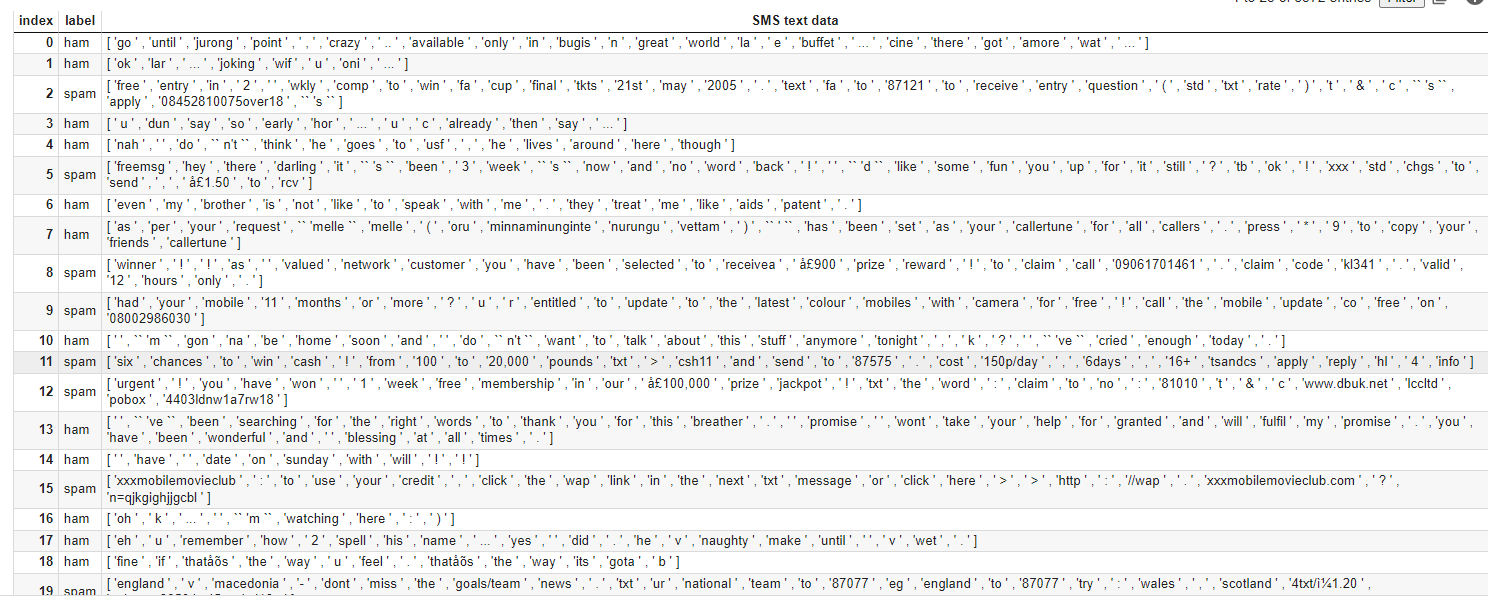
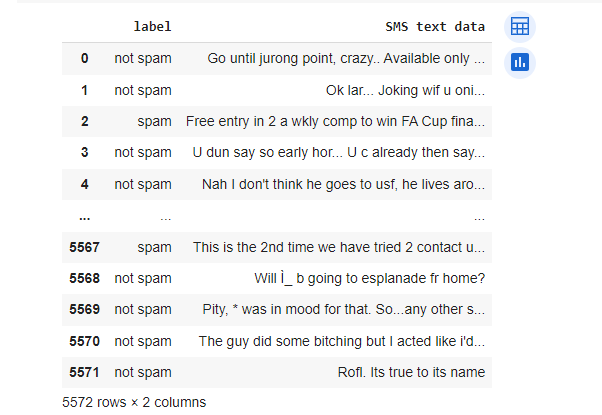
1. **Dataset**

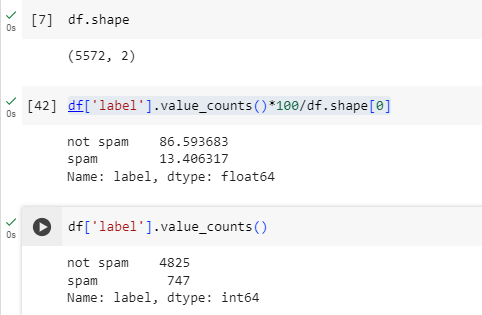
The dataset that I used for this project is obtained from Kaggle, called the “SMS Spam Collection” dataset. This dataset sparked my curiosity since I wanted to further explore how to detect spam text messages and develop a model to help with this. With the rise of spam text messages, I thought this would provide the most valuable impact and this is something that many of us receive occasionally and has practical uses. The two categories within my dataset are spam and not spam.. For this dataset, the categories have already been pre-defined. In my dataset, I’ve defined and divided a “document” to refer to one SMS text message. It contains 5752 SMS messages, of which 4825 are non-spam messages and 747 as spam



Below is a depiction of the dataset; it contains two columns: label categorizes the data as either spam or not spam. The SMS dataset has been tokenized and







1. **Methodology**

My approach to this project was…

The preprocessing steps that I conducted were lowercasing, punctuation removal and

removing stopwords. Initially, during pre-processing, I removed the special characters and punctuation throughout the dataset, however I realized later that this was impacting the data since website links were no longer distinguishable.

*Describe the steps that you performed and what informed your decisions along the way. For example, if you decided not to lowercase the text because it gave your better results at some later stage, include that decision and your reasoning for doing that. This should include your preprocessing steps (in detail, e.g., lowercasing, stemming, etc., do not just say “each document was preprocessed”) and the kinds of analysis that you performed. For any steps that you didn’t implement yourself (e.g., topic modeling), mention which package/library you used.*

1. **Results & Analysis**

The main takeaways from this graph is ….

*Present your results as formatted tables/figures (they should not just be listed in the body of a paragraph). This must include at least the results of the required steps, but may also include any other interesting findings you came across (for example, you could show the results of topic modeling both with and without a certain preprocessing step that you noticed made a large difference in the quality of the results). For each table/figure, include a description of your main takeaways or findings*

1. **Discussion**

Through the pre-processing, some of the insights that I’ve gained about the dataset include:

Some of the lessons that I learned from this project include

*Include 2 subsections in your discussion. The first should cover what you learned about your dataset – you might imagine that you are describing what your results showed, at a high level, to a friend who doesn’t have NLP experience but is interested in the corpus that you chose. The second subsection should cover what lessons you personally learned during the completion of the assignment. You might write about finding and processing data, preprocessing and its effects on topic modeling results, limitations you noticed with the approaches used, or anything else*

*Formatting: The specific formatting of the report is up to you, but part of your grade will be based on having a well-organized and professionally presented document. Please avoid things like blurry, low-resolution/poorly-cropped screenshots, submitting one long paragraph with no subsections/formatting, or 5 copying and pasting long strings from your program output that have no formatting applied. There is no minimum/maximum page limit for the report, but in a single-column format similar to the one that this document is written in, around 3-5 pages is the expected length (including figures and tables).*