Sec 1. Dataset Setup and Exploration Report

I used a jupyter notebook to conduct the dataset exploration step because it is an efficient way to visualize the data and understand its various configurations. After loading the dataset into a pandas dataframe, I checked the shape and missing values of the dataset first. Then, I took a closer look at the specific of the passages and the question-pair and identified a few interesting findings listed blow:

The dataset contains two different datasets - `passages` and `tests`. In this dataset, there are 3,200 passages and 918 test questions–answer pairs. The passages are stored in a single passage column (no missing values) with size (~1.40 MB) and contain four duplicates. Passage lengths are variable, with a mean of 62.1 words (median 48.0; min 1; max 425). Notably, there is a non-trivial tail of ultra-short passages—185 entries under five words (e.g., “Overviews,” “Travel guides,” “Map of Uruguay”). These very short passages likely provide no useful information for the RAG and would hinder the performance of retrieval precision.

In the question-answer set, questions average 53.1 characters (about 9.1 words), and answers are short (mean 19.2 characters, about 3.4 words). Taking a look at very short answers, we could find that a lot of them are just yes/no answers, one word response, a year to answer the given question.

Data quality is generally strong for the question-answer set, there are no empty questions or answers—but there are minor issues such as 10 questions not ending with a question mark, as there’s a comma at the end of the sentences. There’s also a single very short question (“hard”), which could be a data quality problem.

For a RAG pipeline, the analysis suggests deduplicating passages, normalizing text, and filtering or downweighting ultra-short passages; alternatively, if page structure is available, merging short fragments with adjacent context can preserve coverage. Since we’ll be experimenting with length-aware chunking, it would be better to limit the range of each passage to (100-200 words) to enable efficient retrieval.