INDIVIDUAL REPORT

QUESTION ANSWERING SQAUD 2.0 DATASET

Report by
Pon swarnalaya Ravichandran
Team 7
The George Washington University

TABLE OF CONTENTS

INTRODUCTION

A Question Answering (QA) dataset is a collection of questions paired with corresponding answers, designed for training, evaluating, and benchmarking machine learning models in the field of natural language processing (NLP). QA datasets play a crucial role in advancing the development of models that can understand and generate human-like responses to questions.

Key characteristics of QA datasets include:

Diversity of Topics: QA datasets cover a wide range of topics and domains to ensure that models can generalize well across various subject matters.

Question Formats: Questions in these datasets may come in different formats, including multiple-choice questions, open-ended queries, or even prompts that require a model to generate a free-form response.

Annotations: QA datasets are annotated with correct answers to the associated questions. Annotations are typically created by human annotators and are essential for training and evaluating the performance of QA models.

Difficulty Levels: Questions in QA datasets may vary in difficulty, ranging from straightforward factual queries to more complex questions that require reasoning and contextual understanding.

Size and Scale: QA datasets vary in size, with some containing a few thousand examples, while others, like SQuAD (Stanford Question Answering Dataset), contain tens of thousands of question-answer pairs.

Popular QA datasets include:

- SQuAD (Stanford Question Answering Dataset): A widely used dataset consisting of questions posed on a set of Wikipedia articles, with answers provided by human annotators.

- MS MARCO (Microsoft Machine Reading Comprehension): A dataset designed for large-scale machine reading comprehension and question answering, often with longer passages compared to SQuAD.

- TREC (Text REtrieval Conference) QA Track: A series of QA datasets used in the TREC competitions, covering diverse topics and question types.

- CNN/Daily Mail: A dataset constructed from news articles, containing questions about the content of the articles.

DATASET

SOURCE: https://rajpurkar.github.io/SQuAD-explorer/

The dataset we choosed to work on is SQuAD 2.0 QA dataset. SQuAD 2.0 (Stanford Question Answering Dataset 2.0) is a large-scale question answering dataset that has gained significant attention in the field of natural language processing and artificial intelligence. The Stanford Question Answering Dataset (SQuAD) is a popular benchmark dataset used for evaluating the performance of question-answering systems. It was first introduced in 2016 as SQuAD 1.0 and later updated in 2018 as SQuAD 2.0. The main objective of SQuAD is to provide a standardized dataset for evaluating machine comprehension systems, specifically the ability of a machine to understand natural language text and answer questions based on it.

SQuAD 1.0 contains around 100,000 question-answer pairs, while SQuAD 2.0 contains over 150,000 question-answer pairs, including both answerable and unanswerable questions.

The objectives of SQuAD are twofold.

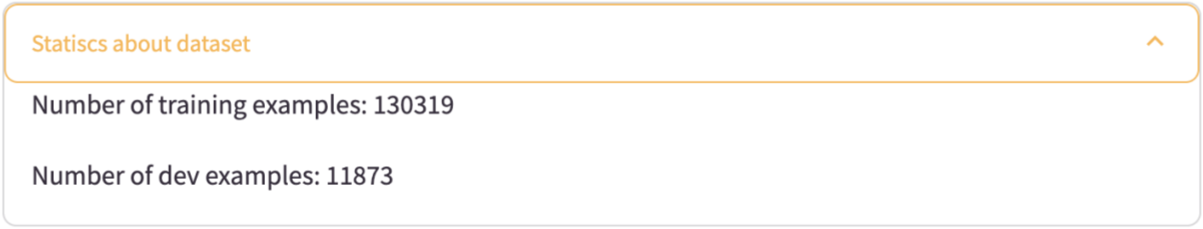
1. It aims to provide a benchmark dataset that can be used to compare the performance of different question-answering systems.
2. It aims to encourage the development of machine comprehension systems that can perform well on real-world problems.

The dataset is designed to be challenging, as it requires systems to understand natural language text, perform accurate text comprehension, and provide precise and accurate answers to a wide range of questions. SQuAD 2.0 takes this challenge a step further by including unanswerable questions, which require systems to be able to recognize when a question cannot be answered based on the given text.

DESCRIPTION OF WORK

My contribution in this project involved the whole design of the streamlit app for the introduction and EDA.

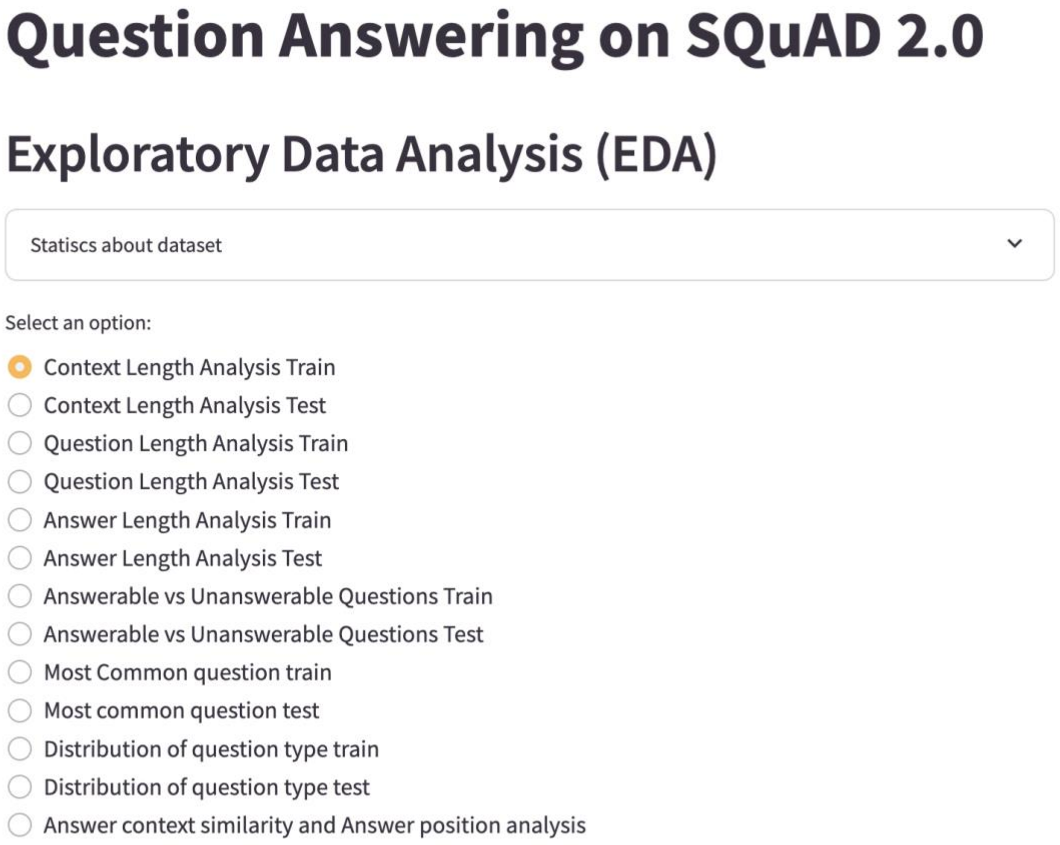
I worked on the Introduction page and encompassed it with the contents and did some of the EDA part as well. I typically used the st.expander and st.radio items on the EDA page.



Statiscs about dataset

Number of training examples: 130319

Number of dev examples: 11873



Question Answering on SQuAD 2.0

Exploratory Data Analysis (EDA)

Statiscs about dataset

Select an option:

- ● Context Length Analysis Train
- ○ Context Length Analysis Test
- ○ Question Length Analysis Train
- ○ Question Length Analysis Test
- ○ Answer Length Analysis Train
- ○ Answer Length Analysis Test
- ○ Answerable vs Unanswerable Questions Train
- ○ Answerable vs Unanswerable Questions Test
- ○ Most Common question train
- ○ Most common question test
- ○ Distribution of question type train
- ○ Distribution of question type test
- ○ Answer context similarity and Answer position analysis

Individual work:

- **EDA**
- **Stream lit designing**

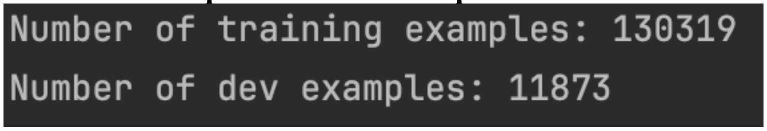
EXPLORATORY DATA ANALYSIS:

- **Fetched the data from these url,**

```
urls = {
   "train": "https://rajpurkar.github.io/SQuAD-explorer/dataset/train-v2.0.json",
   "dev": "https://rajpurkar.github.io/SQuAD-explorer/dataset/dev-v2.0.json",
}
```

- **Some statistics about the dataset:**

#Trainin examples and dev examples



```
Number of training examples: 130319
Number of dev examples: 11873
```

- **The datatype of the dataset is dictionary.**

#PRINTING AN EXAMPLE FROM BOTH THE SETS.

#Train

Context: Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".

Question: When did Beyonce start becoming popular?

Answer: ['in the late 1990s']

Is Impossible: False

#dev

Context: The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

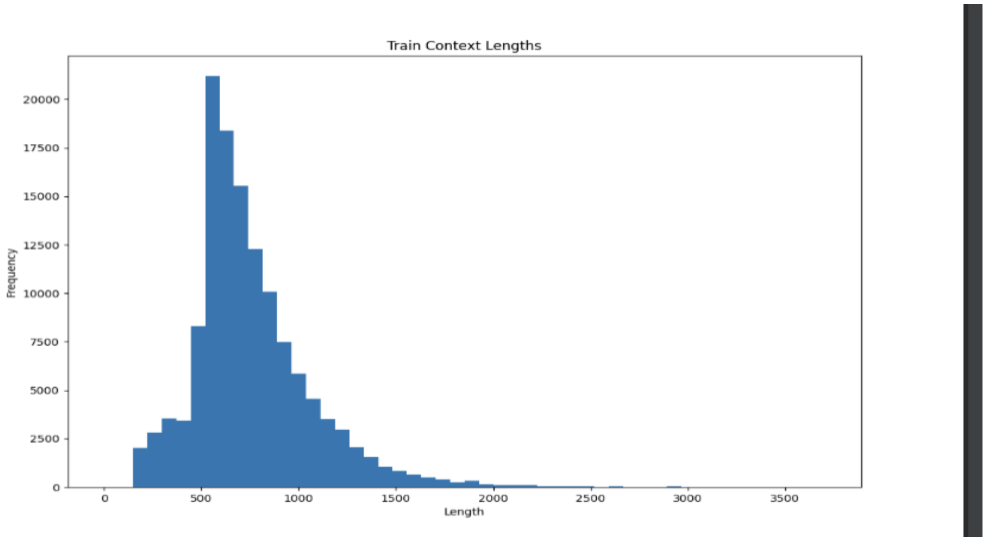
Question: In what country is Normandy located?

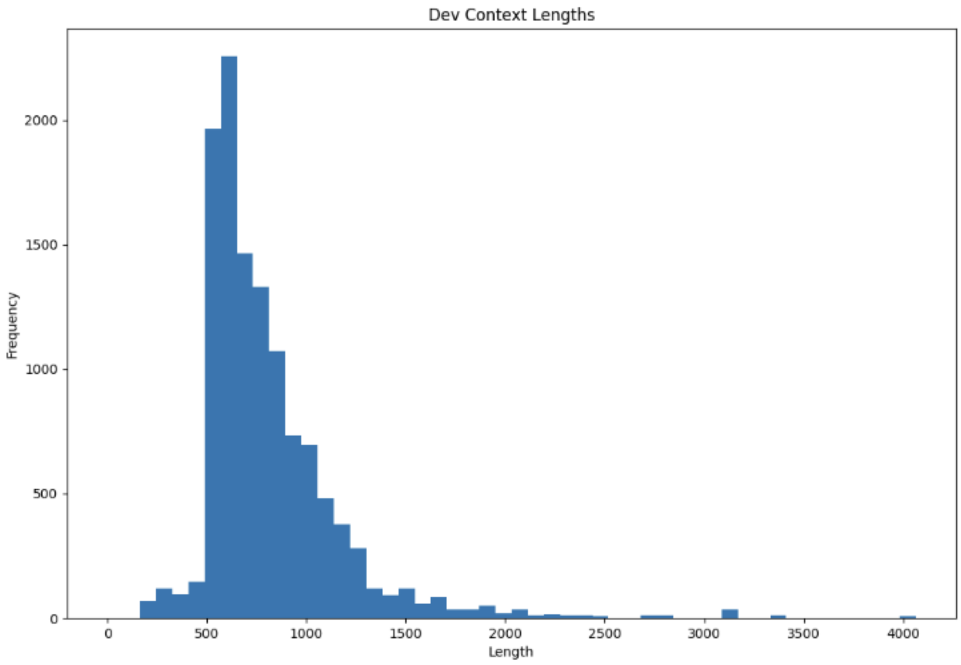
Answer: ['France', 'France', 'France', 'France']

Is Impossible: False

1. Context text analysis:

   To put it simply, a contextual analysis is an examination of a text (in any format, including multi-media) that aids in evaluating the text not just in light of its historical and cultural context but also in terms of its textuality, or the characteristics that identify the work as a text.

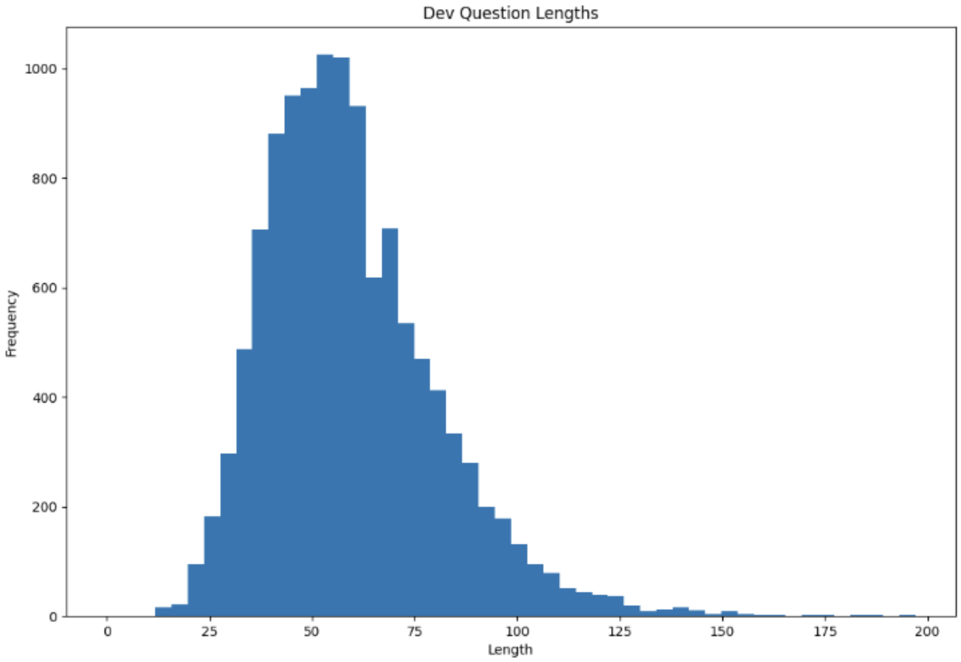


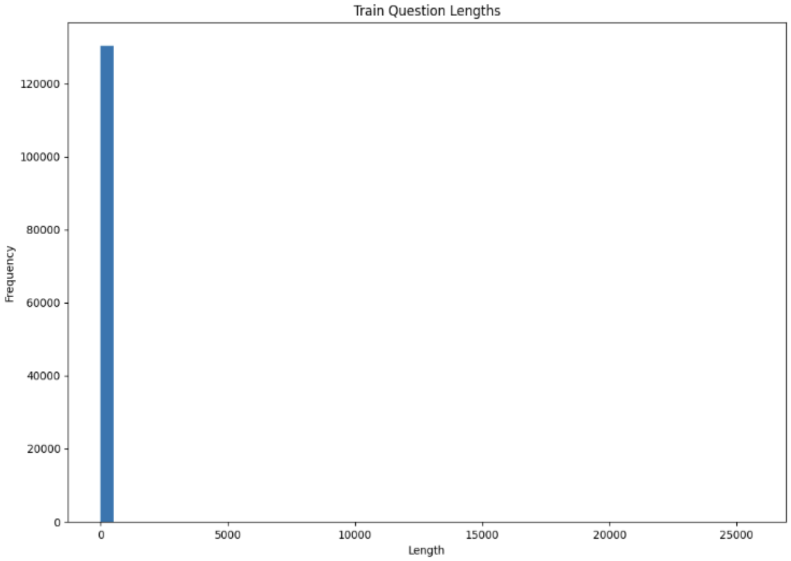


Through the above plot we are able to see the frequency of the context in both the train and dev set of SQuAD2.0.

2. QUESTION LENGTH ANALYSIS:

Question length analysis refers to the examination and evaluation of the length of questions in a given dataset or context. This analysis involves studying the number of words, characters, or other relevant metrics that make up a question. The purpose of question length analysis is to gain insights into the characteristics of questions and their potential impact on various natural language processing (NLP) tasks, including question answering and text summarization.

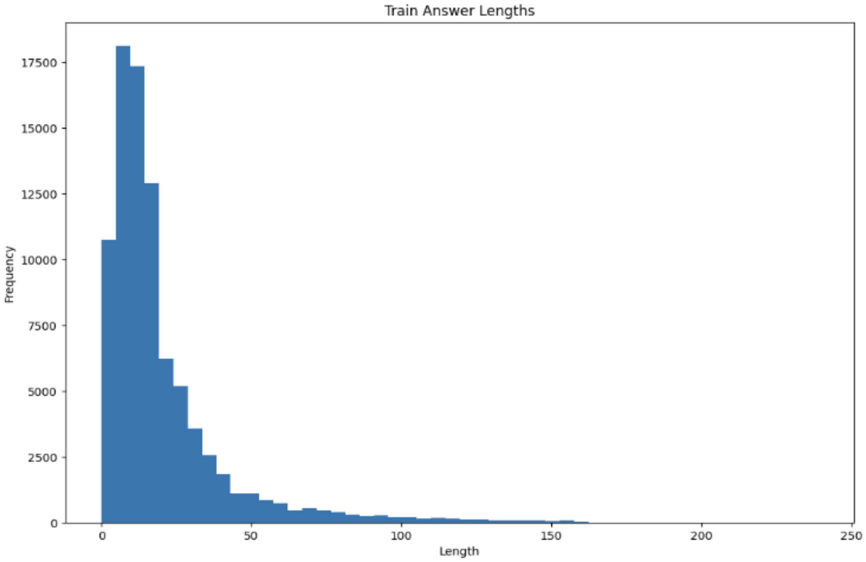


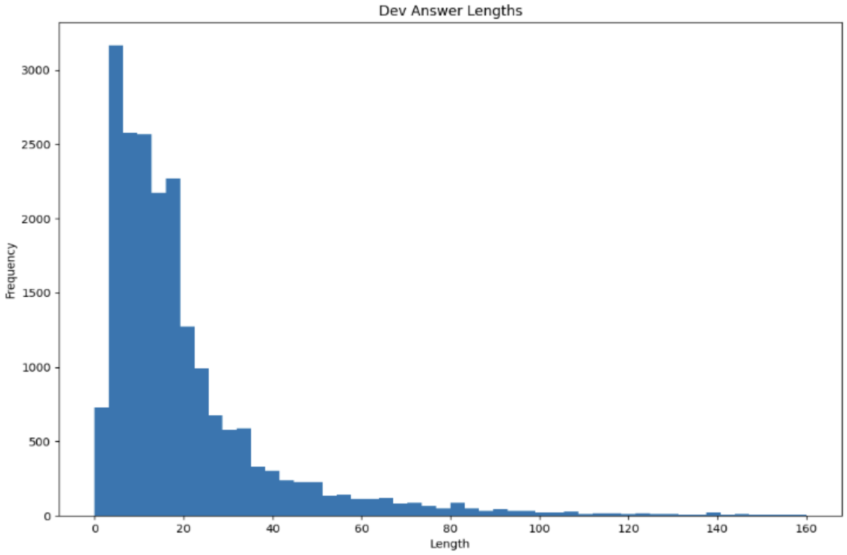


A plot of length and frequency in question length analysis typically represents the distribution of question lengths in a dataset. In this type of plot, the x-axis usually represents the length of questions (measured in terms of words, characters, or some other unit), and the y-axis represents the frequency or count of questions with a particular length. Such a plot is commonly known as a "length-frequency distribution" or "histogram."

3. ANSWER LENGTH ANALYSIS:

This analysis involves studying the number of words, characters, or other relevant metrics that make up the answers provided for a given set of questions. The goal is to gain insights into the characteristics of answers and understand how they vary in length across different contexts.

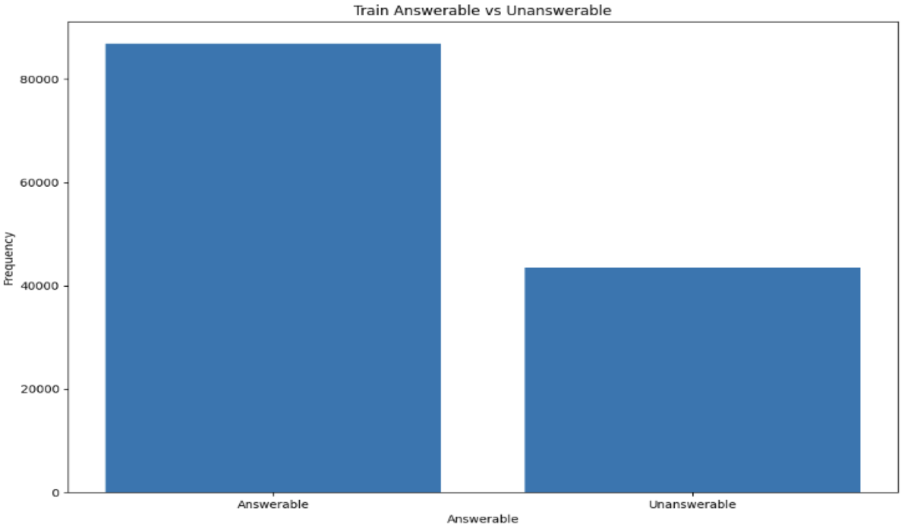


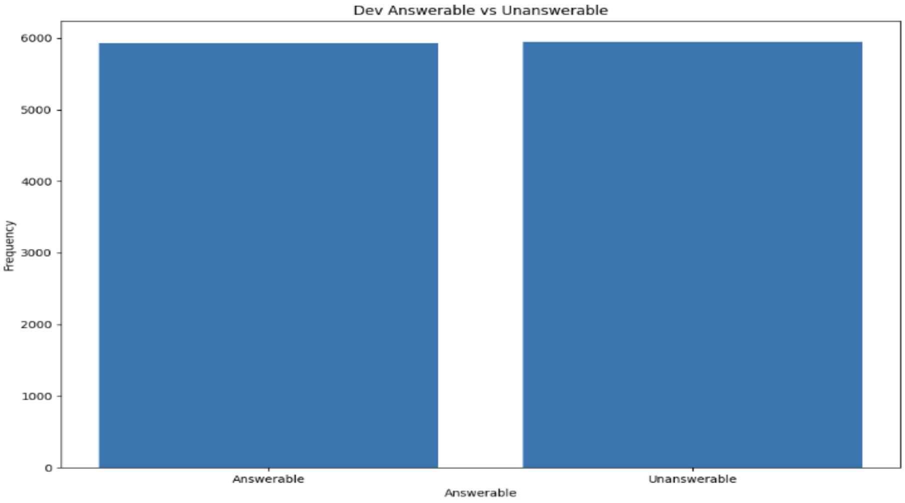


An answer length analysis plot typically involves visualizing the distribution of answer lengths in a dataset.

4. ANSWERABLE VS UNANSWERABLE:

Typically, the SQuAD 2.0 Is different from the earlier SQuAD 1.1 because of this complexity and the inclusion of the unanswerable questions in it. So let us draw some insight about the ans vs unans.

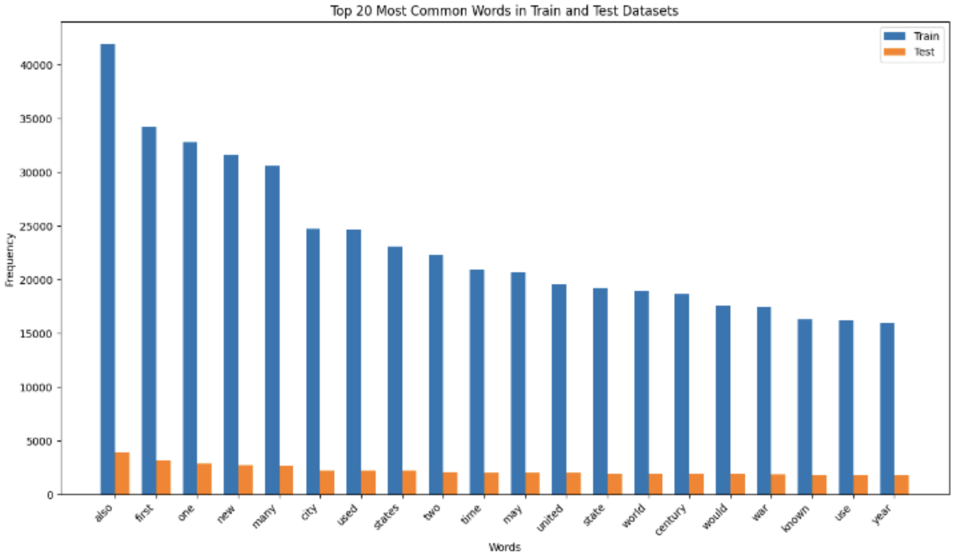




The above plot representing the statistic between the answerable and unanswerable shows that the train set has more than 80000 answerable questions and more than 40000 unanswerable questions. Whereas in test set, the answerable and unanswerable questions are equal which makes it more precise for the prediction.

5. Word length analysis:

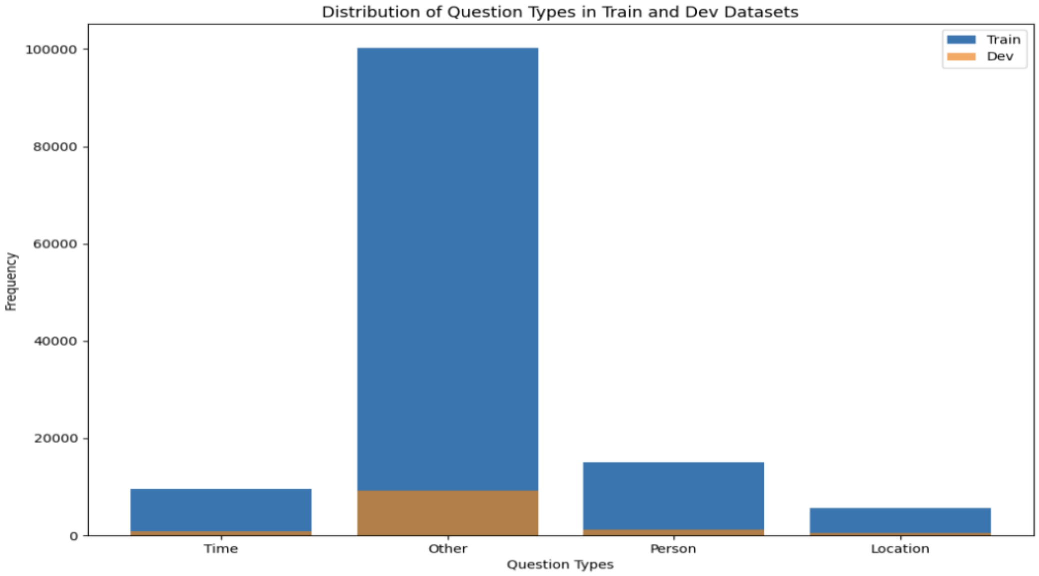
Analyzing word length entails looking at how different word lengths are distributed within a text or dataset. This research sheds light on the language's properties, such as word length variability, average word length, and possible patterns that could be important in a range of linguistic contexts.One can distinguish between the two plots below: the right-skewed distribution predicts longer words, while the left-skewed distribution suggests shorter words.



This is the statistics of the words and frequency of them in the dataset.

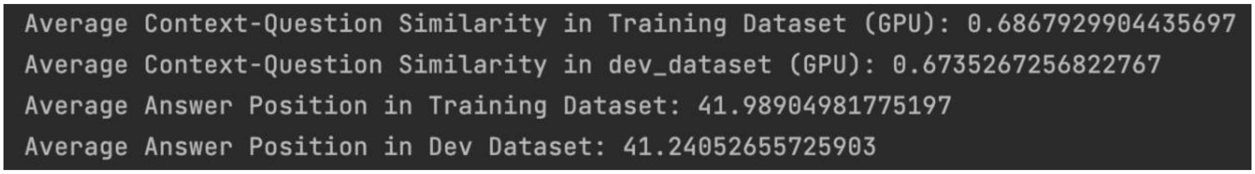
6. Distribution of question type:

The distribution of question types refers to the analysis of the different categories or types of questions present in a dataset or a body of text. Understanding the distribution of question types is crucial in natural language processing (NLP) tasks, especially in question answering systems and information retrieval applications.



7. AVERAGE CONTEXT-QUESTION AND ANSWER POSITION ANALYSIS:

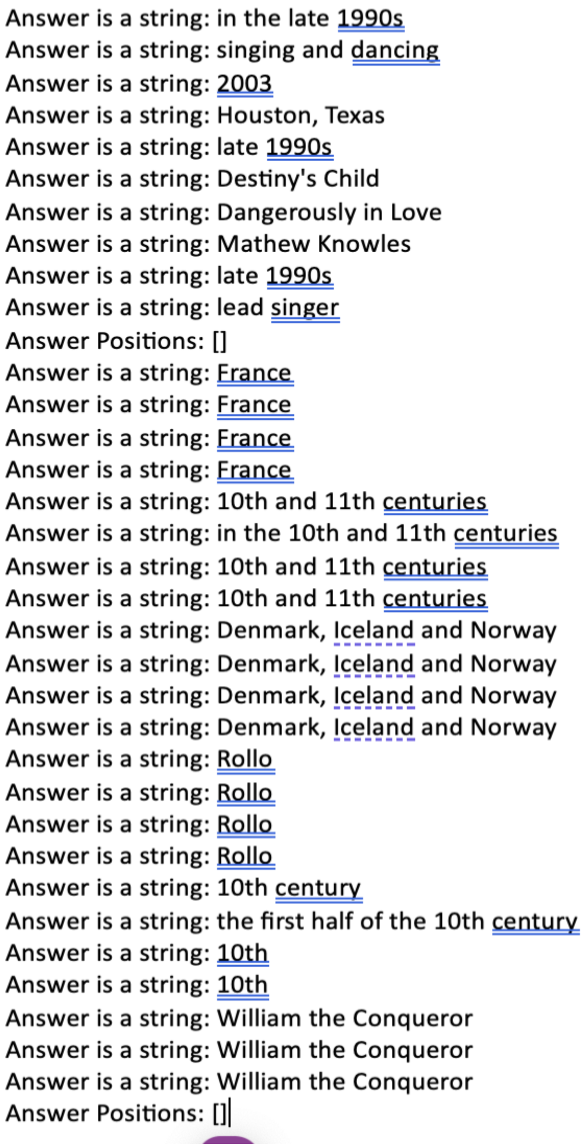
The analysis of the average context-question and answer position involves examining the relative positions of questions and answers within a dataset or a set of documents. This analysis can provide insights into the organization and structure of the data, particularly in the context of question answering tasks and document retrieval scenarios.



```
Average Context-Question Similarity in Training Dataset (GPU): 0.6867929904435697
Average Context-Question Similarity in dev_dataset (GPU): 0.6735267256822767
Average Answer Position in Training Dataset: 41.98904981775197
Average Answer Position in Dev Dataset: 41.24052655725903
```

8. ANSWER POSTION ANALYSIS:

the process of determining the optimal or most relevant location of the correct answer within a given set of documents or text. It involves developing algorithms and techniques to identify the position of the answer in order to enhance the precision and accuracy of responses generated by NLP model



```
Answer is a string: in the late 1990s
Answer is a string: singing and dancing
Answer is a string: 2003
Answer is a string: Houston, Texas
Answer is a string: late 1990s
Answer is a string: Destiny's Child
Answer is a string: Dangerously in Love
Answer is a string: Mathew Knowles
Answer is a string: late 1990s
Answer is a string: lead singer
Answer Positions: []
Answer is a string: France
Answer is a string: France
Answer is a string: France
Answer is a string: France
Answer is a string: 10th and 11th centuries
Answer is a string: in the 10th and 11th centuries
Answer is a string: 10th and 11th centuries
Answer is a string: 10th and 11th centuries
Answer is a string: Denmark, Iceland and Norway
Answer is a string: Denmark, Iceland and Norway
Answer is a string: Denmark, Iceland and Norway
Answer is a string: Denmark, Iceland and Norway
Answer is a string: Rollo
Answer is a string: Rollo
Answer is a string: Rollo
Answer is a string: Rollo
Answer is a string: 10th century
Answer is a string: the first half of the 10th century
Answer is a string: 10th
Answer is a string: 10th
Answer is a string: William the Conqueror
Answer is a string: William the Conqueror
Answer is a string: William the Conqueror
Answer Positions: []
```

Conclusion:

The data was stable and balanced to workon. After EDA, the data was used for modelling.