**AMRITA SCHOOL OF ENGINEERING, COIMBATORE**

****

**15CSE358 Natural Language Processing**

**Case Study- Problem statement**

**Group Number: 10**

**Automatic Question Tagging System**

**Team Members:**

| **S.No** | **Reg Num** | **Name** |
| --- | --- | --- |
| 1 | CB.EN.U4CSE18012 | Chavalam Nikhitha |
| 2 | CB.EN.U4CSE18037 | Nishita Dash |
| 3 | CB.EN.U4CSE18041 | Parripati Divyasri |
| 4 | CB.EN.U4CSE18502 | Hridhi sethi |

**Abstract:**

Online question and answer forums such as Stack Exchange and Quora are becoming an increasingly popular resource for education. Central to the functionality of many of these forums is the notion of tagging, whereby a user labels his/her post with an appropriate set of topics that describe the post, such that it is more easily retrieved and organized. We propose a multi-label classification system that automatically tags users questions to enhance user experience.

**1) Background:**

Sites that are specifically designed to have questions and answers for their users like Quora and Stackoverflow often request their users to submit five words along with the question so that they can be categorized easily. But, sometimes users provide wrong tags which makes it difficult for other users to navigate through.That is why we propose to design a question tagging system.

**2) Problem statement:**

Designing an automatic question tagging system that can automatically identify correct and relevant tags for a question submitted by the user.

**3) Relevance to NLP:**

NLP is about understanding and inferring from the human language and in our project, our model should understand and infer from the question posed in the forum and identify correct tags for it.

**4) Method overview:**

This model implements a tag suggestion system, automatically inferring and tagging a question that is to be posted in the forum. This process begins by taking in the input text (ie. the question that is to be posted in the forum), this input data has to be preprocessed in order to remove the missing attributes, noise, or outliers, duplicate or wrong data. Then we train the model by merging the questions and tags dataset. Now the ‘score’ is used to filter the posts with better quality and have lots of upvotes. Now text preprocessing like transforming abbreviations, Lemmatizing words, removing punctuations, etc is performed. Here LDA is also used to see if it shows any patterns in words and the main topics. And then the data is put into the classifier for multi-label classification. Finally, the top 10 tags for the related question would be displayed.

**5) Tools and libraries:**

1. Jupyter notebook - An interactive and easy to use IDE (Integrated Development environment).
2. NLTK (Natural Language Toolkit) - The **Natural Language Toolkit** (**NLTK)**, is a suite of libraries and programs for symbolic and statistical natural language processing(NLP). It is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

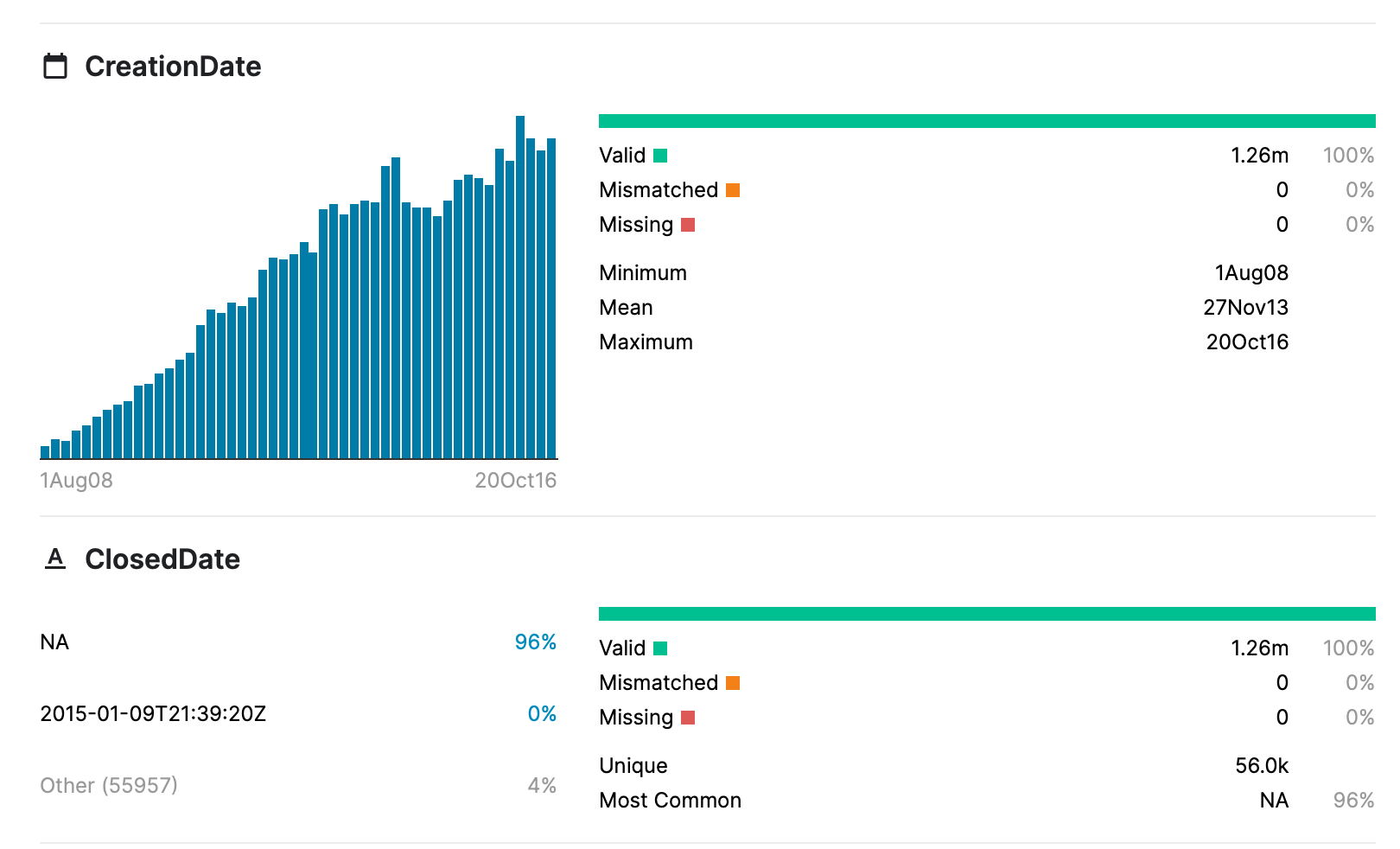
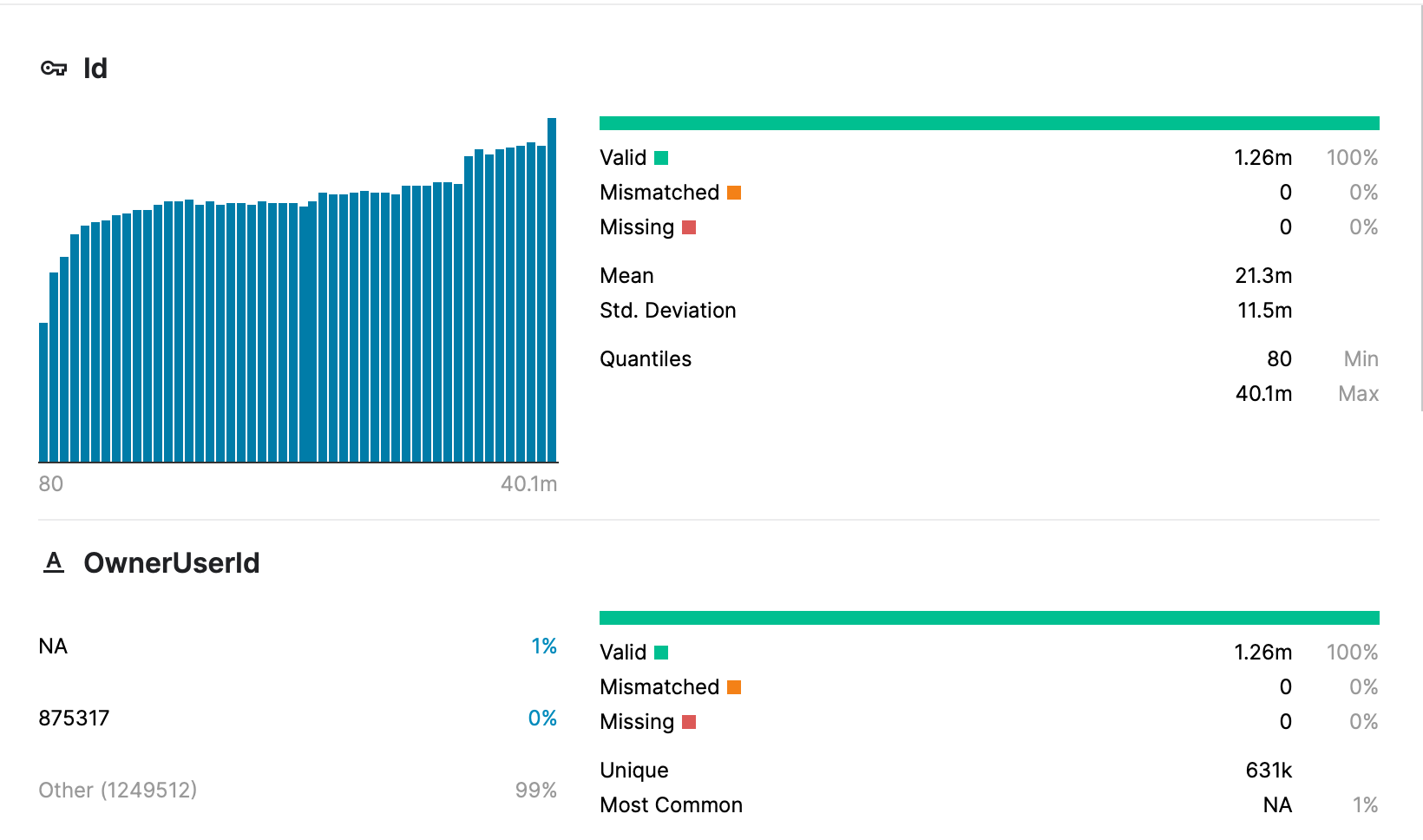
**6) Dataset**

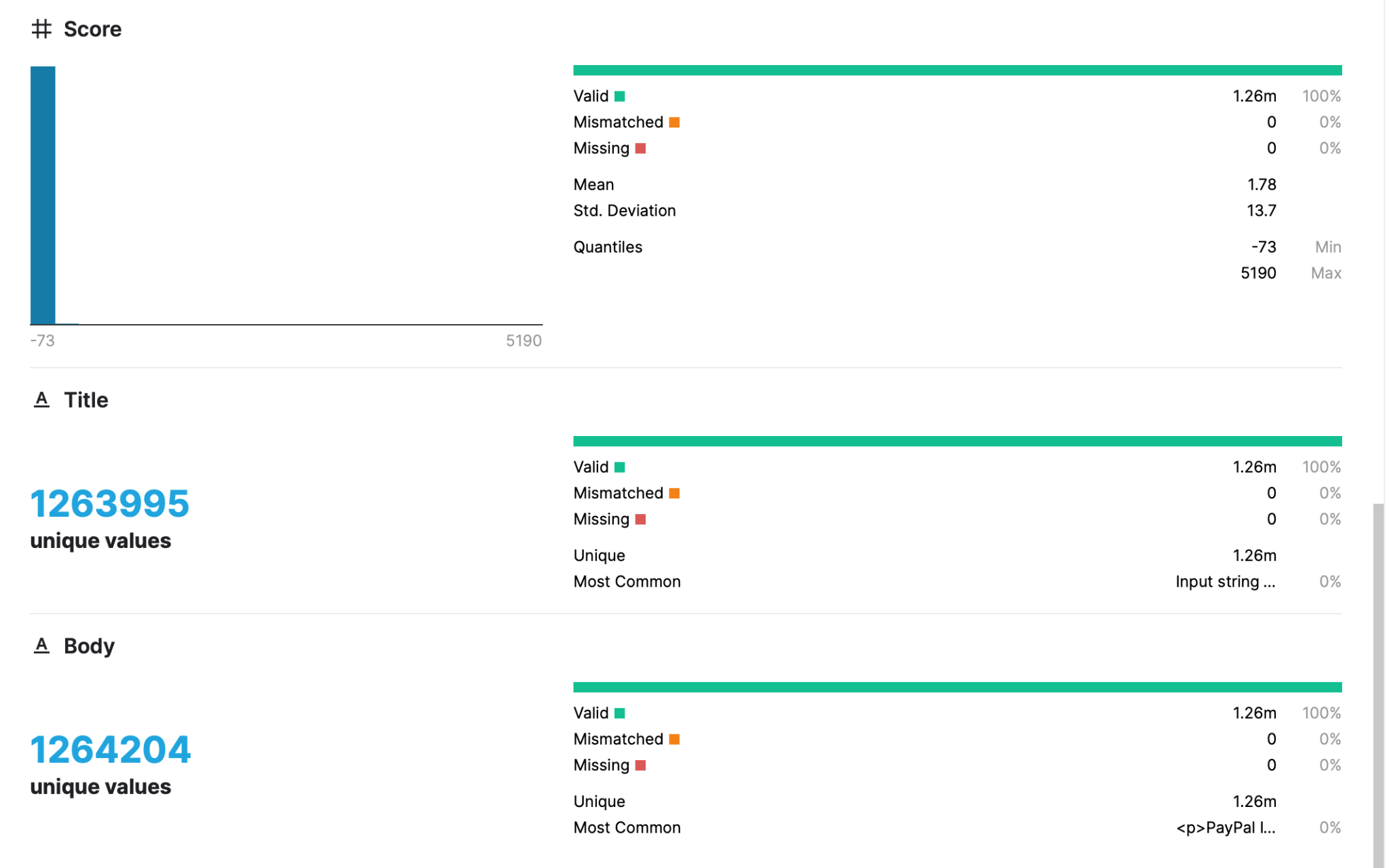
**Dataset Description**:

Dataset with the text of 10% of questions and answers from the Stack Overflow programming Q&A website.

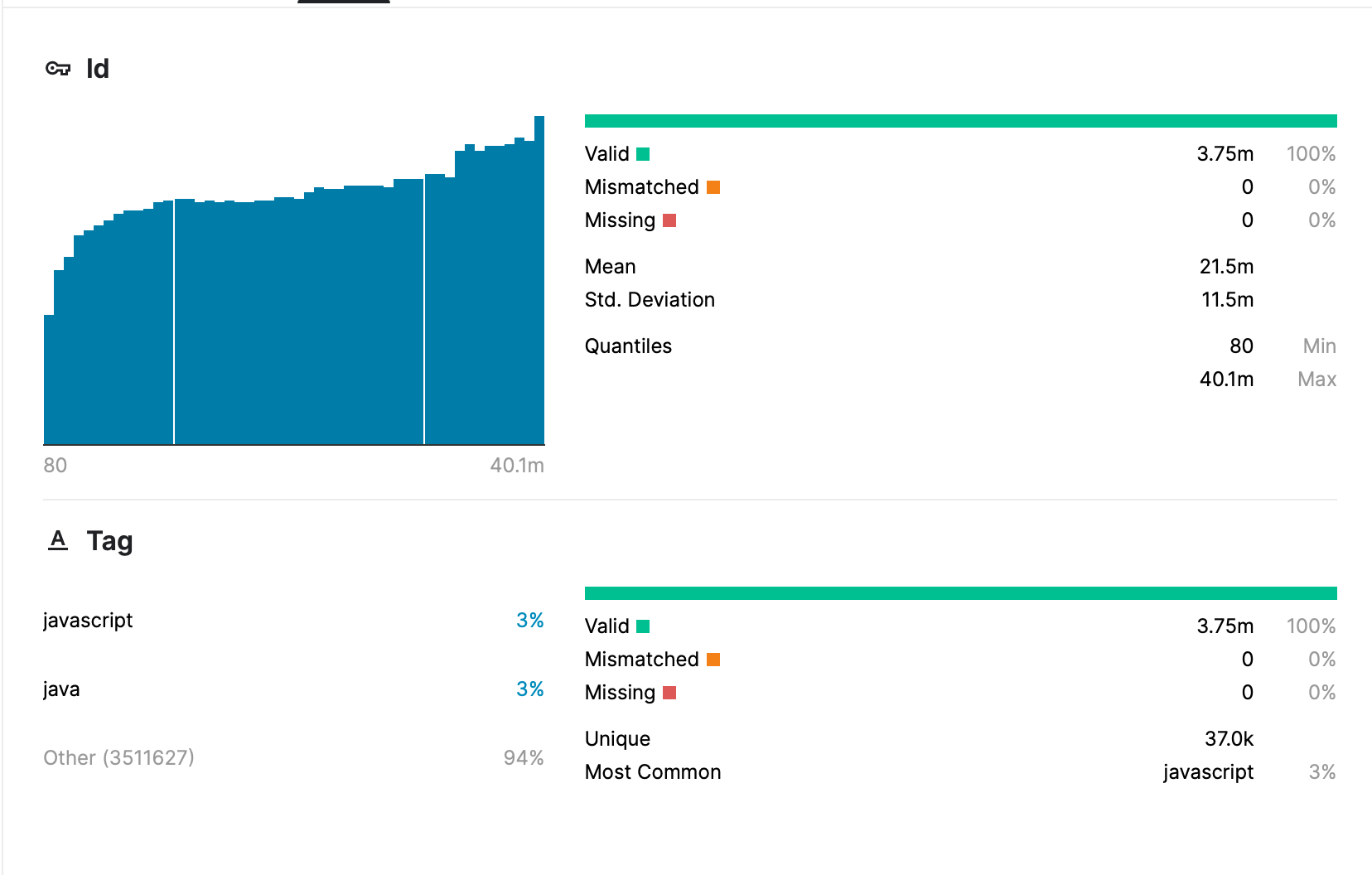
This is organized into three tables:

* Questions contain the title, body, creation date, closed date (if applicable), score, and owner ID for all non-deleted Stack Overflow questions whose Id is a multiple of 10.

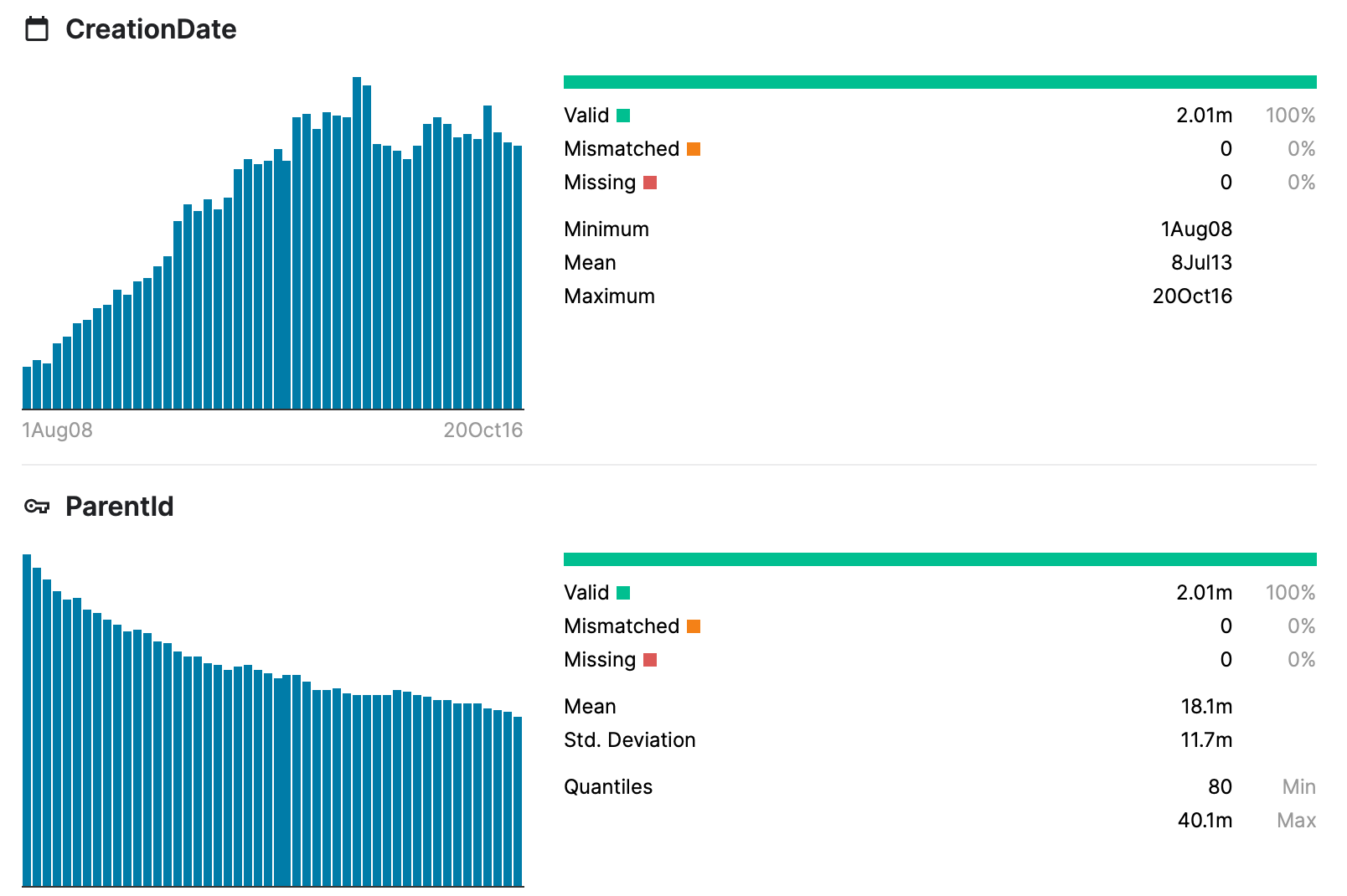
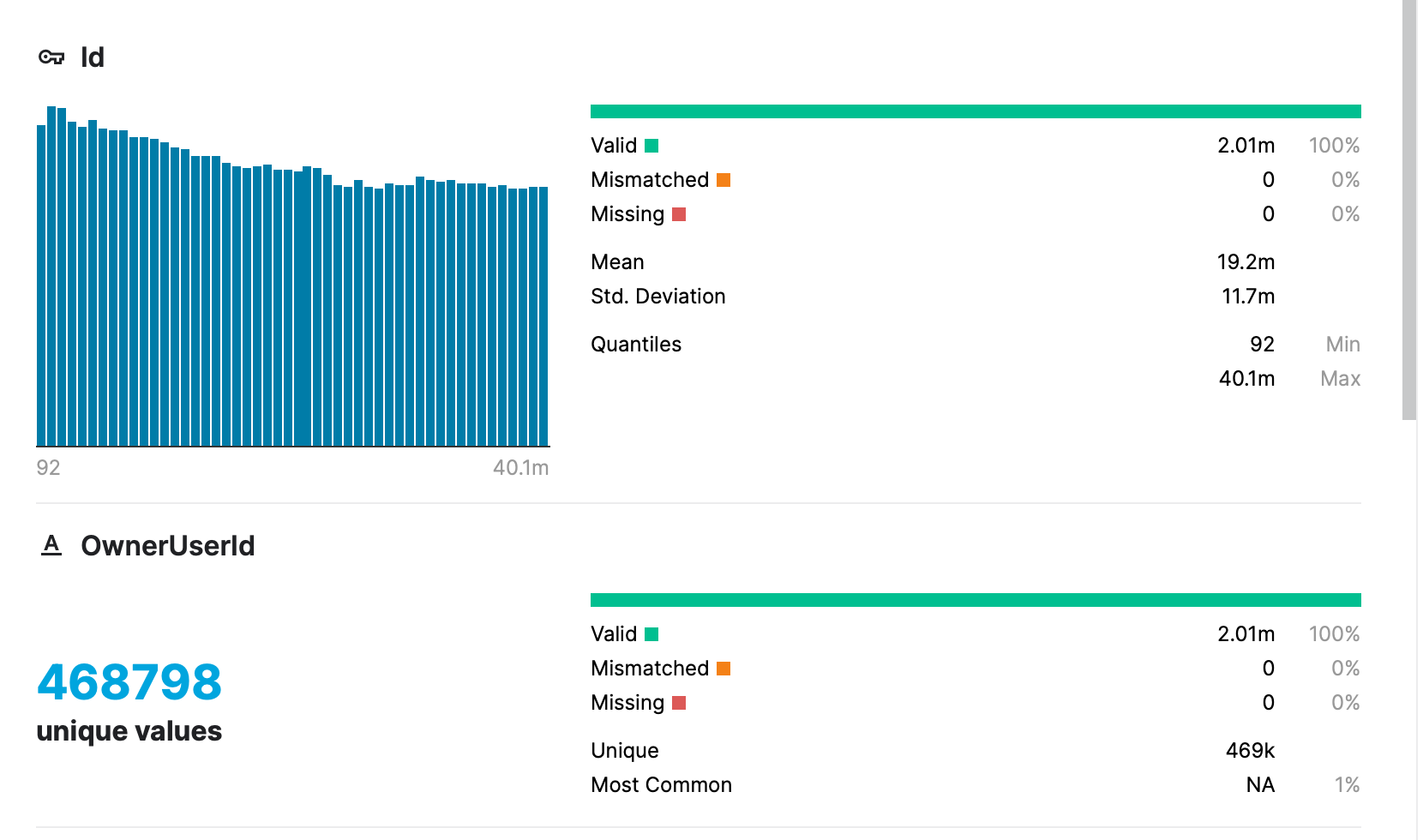


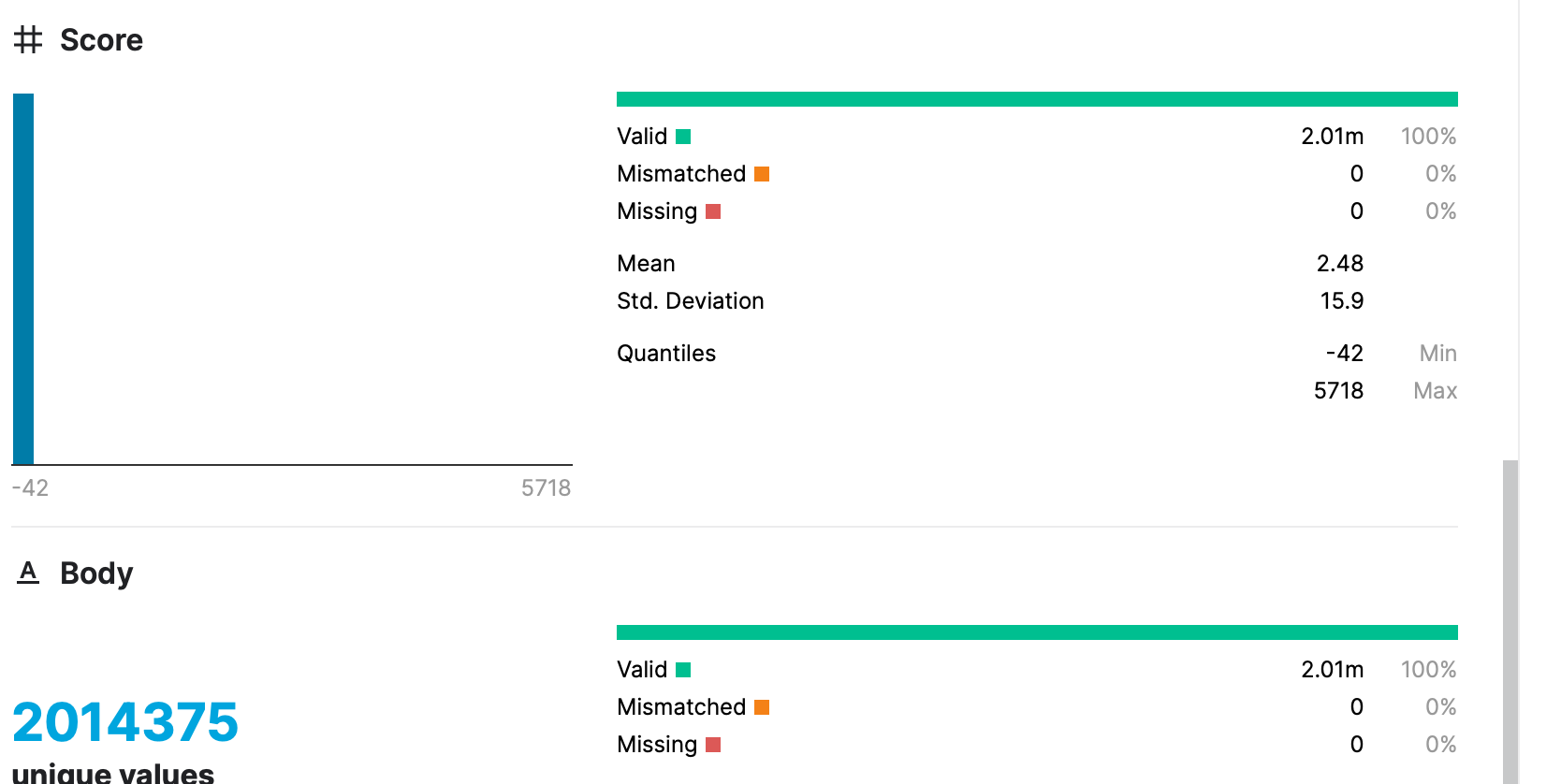


* Tags contains the tags on each of these questions



* Answers contain the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.





**URL**: [StackSample: 10% of Stack Overflow Q&A](https://www.kaggle.com/stackoverflow/stacksample)

**7) Status of implementation:**

<https://colab.research.google.com/drive/1cYDnUI2s214cLUgItbuRSJXO5P3AmmuZ?usp=sharing>

* Data preprocessing
* EDA
* Data Preparation
* Checking various models
* Parameter tuning using Gridsearchcv
* Results and Analysis

**8) Conclusion:**

We will develop an automatic question tagging system that will give correct and relevant tags for the question submitted by the user.

**References**

1. [#ML #NLP: Autonomous Tagging of Stack Overflow Questions](http://cs229.stanford.edu/proj2014/Mihail%20Eric,%20Ana%20Klimovic,%20Victor%20Zhong,MLNLP-Autonomous%20Tagging%20Of%20Stack%20Overflow%20Posts.pdf)
2. <https://www.analyticsvidhya.com/blog/2021/07/nltk-a-beginners-hands-on-guide-to-natural-language-processing/>
3. <https://www.kite.com/python/docs/nltk.toktok>
4. <https://pythonprogramming.net/stop-words-nltk-tutorial/>
5. <https://www.geeksforgeeks.org/python-lemmatization-with-nltk/>