Pradeep Vajrala NextLabs solution

# Part-1

### Question-1 [[*Solution link*](https://colab.research.google.com/drive/1DKE48fuhvUnY9-rQDb3Izd54z6C_pWhe?usp=sharing)]

***Problem statement***

Question was to extract numbers in orange from given JSON format using regex

*{"orders":[{"id":1},{"id":2},{"id":3},{"id":4},{"id":5},{"id":6},{"id":7},{"id":8},{"id":9},{"id":10},{"id":11},{"id":648},{"id":649},{"id":650},{"id":651},{"id":652},{"id":653}],"errors":[{"code":3,"message":"[PHP Warning #2] count(): Parameter must be an array or an object that implements Countable (153)"}]}*

***Description***

Given JSON is a dictionary type of data with nested dictionaries in it. The numbers highlighted in orange are required to be extracted. For that, we need regex library and python programming language.

***Build instructions***

We just need a regex library for this, so import the library using the command **“import re”.** We can now extract our required information.

***Solution***

The given JSON has to be first converted into a string for being able to parse. This can be done using **str()** function.

**data=str({"orders":[{"id":1},{"id":2},{"id":3},{"id":4},{"id":5},{"id":6},{"id":7},{"id":8},{"id":9},{"id":10},{"id":11},{"id":648},{"id":649},{"id":650},{"id":651},{"id":652},{"id":653}],"errors":[{"code":3,"message":"[PHP Warning #2] count(): Parameter must be an array or an object that implements Countable (153)"}]}**

Now we can use regex to get numbers from the input string. For this solution, using **re.findall()** we can find any data as required.

All the highlighted numbers are values of the key “id” and one value has key “code”.

Simply using “**\d+**” gives all numbers in the input but we want numbers only in the value of the key-value pair. The last two numbers ***2*** and ***153*** are in a string form in key-value format. But, these two digits have a character preceding them whereas, the highlighted numbers are preceded by white spaces. So now we can extract required numbers by using **re.findall('\s+(\d+)', data)**

The regular expression **‘\s+(\d+)’** can be read as ***finding all digits preceded by white spaces***. The “capture group ()” has \d+ enclosed in it, so all numbers in string input are matched but \s+ detects all white spaces. So the code can also be read as ***finding all pair of white spaces with digits succeeding immediately***.

The result is an array of extracted numbers

***['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '648', '649', '650', '651', '652', '653', '3']***

### Question-2 [[*Live link*](https://ratingreview.azurewebsites.net/)]

***Problem statement***

There are times when a user writes ***Good***, ***Nice App or any other positive text***, in the review and gives 1-star rating. Goal is to identify the reviews where the semantics of review text does not match rating.

Goal is to identify such ratings where review text is good, but rating is negative- so that the support team can point this to users.

Deploy it using - Flask/Streamlit etc and share the live link.

***Description***

Given data is a collection of reviews given to Chrome app with Review, Star rating, User ID and other information. Our goal is to find reviews given by users where the rating is bad but the review is a positive text. For that, Logistic Regression is used and the app is deployed using StreamLit.

***Build instructions***

This code requires data exploratory libraries like **numpy, pandas;** natural language libraries **nltk,** machine learning library **sklearn.**

The code is deployed on Azure using StreamLit. The live link attached.

***Solution***

1. The entire review dataset is loaded and only required columns, User ID, Text and Star rating are kept while other columns are dropped.
2. An extra column is added to the data to check if the review is Positive or Negative.
3. The remaining required data is checked for any white spaces, null values and other non-word parameters which might interfere with checking the reviews. The captured non-word parameters are deleted and the dataset is converted entirely into small cases for avoiding any conflict, and then checked for any stopwords.
4. The dataset is then randomly split into testing and training datasets for input to the **LogisticRegression()** function.
5. Output of the function is stored in a pickle file for further use in the StreamLit web application.
6. The streamlit web app takes a csv file as input, uses the pre-trained model and displays which text might be potentially good but is given a negative rating.

### Question-3 [[*Solution link*](https://colab.research.google.com/drive/1DKE48fuhvUnY9-rQDb3Izd54z6C_pWhe?usp=sharing)]

***Problem statement***

Ranking Data - Understanding the correlation between keyword rankings with description or any other attribute.

***Description***

The data set is keyword ranking of browsers, their short and long term descriptions. And the goal is to find if any correlation is between different attributes. For this, cleaning of data is required and capturing the column names which were part of data initially.

***Build instructions***

Using **matplotlib.pyplot,** we can visualize data requiredand fine correlations.

***Solution***

1. Import dataset and convert into a dataframe.
2. Remove unwanted columns such as Dates, Country and Language.
3. Add new columns to the dataframe with length of Short and Long Descriptions
4. Plot a graph between required attributes and observe the results.

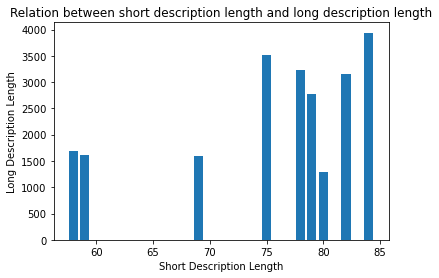
***Questions asked***

1. Is there any correlation between short description, long description and ranking? Does the placement of keywords (for example - using a keyword in the first 10 words - have any correlation with the ranking)?

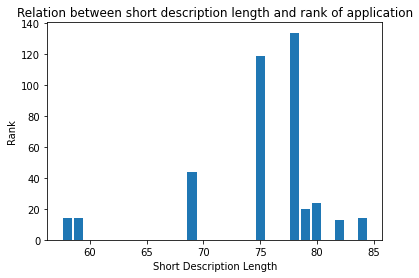
The function **corr()** is used to find correlation between columns in a dataset. Using this function for Short Description Length and Long Description Length, gives us a correlation of **+0.7274** correlation. Which means, there is a strong correlation between the short and long description lengths.

\*All the numbers indicate total characters in the description, not words in description

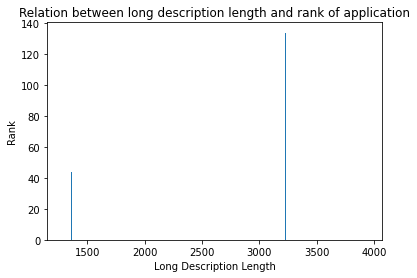
Plotting a graph between these two also indicates the correlation. In most cases, longer the short description is, the longer the long description is given.

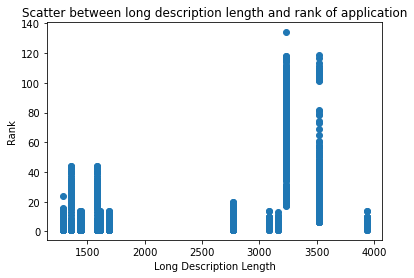


Relation between short description length and ranking of application is also plotted, which suggested that not necessarily the shorter the length, better app ranking will be, but having keywords in the short description can help the app rank better by App Store Optimization(ASO)



Finally, relation between long description length and ranking of respective applications has also been plotted using a bar graph and also a scatter plot for better understanding. Though it does not mean longer the description, better the ranking, having keywords in the description helps it in better ranking at Search Engine Optimization(SEO)





1. Does APP ID (Also known as package name) play any role in ranking?

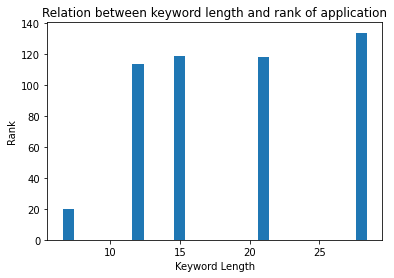
Yes. It certainly does play a role in App Store Optimization(ASO) and Search Engine Optimization(SEO). APP ID is a keyword used for ranking apps in specific search query using words in the title/app name. APP ID has the biggest weight in ranking of application and having keywords in the first half of title is more important as they are visible to users and are not truncated or cut-off, so they play a major role in ranking. And also, using the exact title match for the search increases the ranking.

Also, in Google SEO, Exact Domain Match (EDM) plays an important role, meaning, having keywords in URL plays a direct role in better ranking of App.

1. Any other patterns or good questions that you can think of and answer?

One question I was interested in was how will the number of keywords affect the keyword ranking and overall ranking of the application?

For this, a new column of “Keyword Length” is added to the dataframe, and a bar plot is plotted to observe the results. We see that the lower the keyword length is, the higher the rank is. But it also refers that not just having any keyword would improve the search result. Keywords with higher keyword ranking help in better visibility of the app thus improving SEO and ASO.



# Part-2

Question-1

***Problem statement***

Check if the sentence is Grammatically correct: Please use any pre-trained model or use text from open datasets. Once done, please evaluate the English Grammar in the text column of the below dataset.

***Description***

Using a model to determine if the given sentence is grammatically correct or not.

***Build instructions***

**Language-tool** is the library used to evaluate the dataset. So, install and import it. Clean the data by removing repeated and unwanted spaces.

Question-2

Write about any difficult problem that you solved. (According to us difficult - is something which 90% of people would have only 10% probability in getting a similarly good solution).

In the field of Computer Vision, human pose estimation is one of the problems that is relatively difficult to solve. It is so because we have to locate pre-determined human joints, limbs and other body parts for estimating poses. By locating the joints, the visibility of joints plays an important role as they are used to calculate angles between other joints and thus are used for calculating angles and combination of these angles gives us pose angles.

In a project I’ve done, the results were very accurate with a 93.1% accuracy rate of detecting human poses regardless of their skin tone, background noises, picture quality to an extent, clothes they were wearing in the given input image.

Question-3

Formally, a vector space V' is a subspace of a vector space V if

* + V' is a vector space
  + every element of V′ is also an element of V.

Note that ordered pairs of real numbers (a,b) a,b∈R form a vector space V. **Which of the following is a subspace of V?**

* + The set of pairs *(a, a + 1)* for all real *a*
  + *The set of pairs (a, b) for all real a ≥ b*
  + *The set of pairs (a, 2a) for all real a*
  + *The set of pairs (a, b) for all non-negative real a,b*

Given, (a,b) is an ordered pair of real numbers and forms a vector space V. Then,

1. Set of pairs (a,a+1) for all Real a (a∈R)

Let x=(a1,a1+1) and y=(a2,a2+1)

x+y = (a1+a2, a1+a2+2) is a real number.

x-y = (a1-a2, a1-a2) is also a real number. So (a,a+1) forms a subspace of V.

1. Set of pairs (a,b) for all Real a>=b

Let x=(a1,b1) and y=(a2,b2)

x+y = (a1+a2, b1+b2) is a real number.

x-y = (a1-a2, b1-b2) is also a real number. So (a,b) for all a>=b forms a subspace of V.

1. Set of pairs (a,2a) for all Real a (a∈R)

Let x=(a1,2a1) and y=(a2,2a2)

x+y = (a1+a2, 2a1+2a2) is a real number.

x-y = (a1-a2, 2a12-a2) is also a real number. So (a,2a) forms a subspace of V.

1. Set of pairs (a,b) for all Real non negative a&b (a,b∈R)

Let x=(a1,b1) and y=(a2,b2)

x+y = (a1+a2, b1+b2) is a real number.

x-y = (a1-a2, b1-b2) is also a real number but not necessarily a non-negative number. So (a,b) for non-negative a&b does not form a subspace of V.