RAMAIAH UNIVERSITY OF APPLIED SCIENCES

ASSIGNMENT

Course Code	19CSC312A
Course Name	Artificial Intelligence
Programme	B. Tech.
Department	Computer Science and Engineering
Faculty	FET

Name of the Student	Bharath Kumar B ; Charith Kumar S ; Deepak R ;
Reg. No	18ETCS002032 ; 18ETCS002039 ; 18ETCS0020041
Semester/Year	6th /2018
Course Leader/s	Dr. Subarna Chatterjee

Declaration Sheet				
Student Name	Bharath Kumar B ;Charith Kumar S ; Deepak R			
Reg. No	18ETCS002032 ; 18ETCS002039 ; 18ETCS002041			
Programme	B. Tech.	Semester/Year	6th /2018	
Course Code	19CSC312A			
Course Title	Artificial Intelligence			
Course Date		to		
Course Leader	Dr. Subarna Chatterjee			

Declaration

The assignment submitted herewith is a result of my own investigations and that I have conformed to the guidelines against plagiarism as laid out in the Student Handbook. All sections of the text and results, which have been obtained from other sources, are fully referenced. I understand that cheating and plagiarism constitute a breach of University regulations and will be dealt with accordingly.

Signature of the Student		Date	
Submission date stamp (by Examination & Assessment Section)			

Signature of the Course Leader and date	Signature of the Reviewer and date

Faculty of Engineering and Technology			
Ramaiah University of Applied Sciences			
Department	Computer Science and Engineering	Programme	B. Tech.
Semester/Batch	6th/2018		
Course Code	19CSC312A	Course Title	Artificial Intelligence
Course Leader(s)	Dr. Subarna Chatterjee, Gp. Capt. Rath and Santoshi Kumari		

Assignment-2

Register No:	**18ETCS002032** **18ETCS002039** **18ETCS002041**	Name of Student	**Bharath Kumar B** **Charith Kumar S** **Deepak R**

Questions		Marking Scheme	Marks: Max Marks	First Examiner Marks	Moderator
Question 1	1.1	Compare recent algorithms of NLP showing the steps to address the issue	03		
	1.2	Identify and explain the appropriate pre-processing techniques.	03		
	1.3	Identify and explain the appropriate NLP based sentiment analysis technique.	04		
		Question 1 Max Marks	10		
Question 2	2.1	Perform pre-processing on the created customer dataset	05		
	2.2	Perform sentiment analysis of the customer.	07		
	2.3	Results and Discussions.	03		
		Question 2 Max Marks	15		
		Course Marks Tabulation **Total Assignment Marks**	25		

Question	First Examiner	Remarks	Moderator	Remarks
1				
Marks (Max 25)				

Signature of First Examiner **Signature of Moderator**

Please note:

1. Documental evidence for all the components/parts of the assessment such as the reports, photographs, laboratory exam / tool tests are required to be attached to the assignment report in a proper order.
2. The First Examiner is required to mark the comments in RED ink and the Second Examiner's comments should be in GREEN ink.
3. The marks for all the questions of the assignment have to be written only in the **Component – CET B: Assignment** table.
4. If the variation between the marks awarded by the first examiner and the second examiner lies within +/- 3 marks, then the marks allotted by the first examiner is considered to be final. If the variation is more than +/- 3 marks then both the examiners should resolve the issue in consultation with the Chairman BoE.

Assignment-2

Instructions to students:

1. The assignment consists of **2** question.
2. Maximum marks is **25**.
3. The assignment has to be neatly word processed as per the prescribed format.
4. The maximum number of pages should be restricted to **15**.
5. The printed assignment must be submitted to the course leader.
6. **Submission Date: 19 Jun 2021 (Saturday)**
7. **IMPORTANT**: It is essential that all the sources used in preparation of the assignment must be suitably referenced in the text.
8. Marks will be awarded only to the sections and subsections clearly indicated as per the problem statement/exercise/question

Preamble

Natural Language Processing (NLP) refers to AI method of communicating with an intelligent systems like Siri, Alexa using a natural language such as English, Hindi. NLP, is an attempt to make a computer understand human language. Computers can easily understand programming languages. How do we make sure that computers are able to understand a natural language?

Processing of Natural Language is required when you want an intelligent system like a robot to perform as per your instructions, or, when you want to hear decision from a dialogue based expert system. The field of NLP involves making computers to perform useful tasks with natural languages that humans use. The input and output of an NLP system can be −

- Speech
- Written Text

Solution for 2.1 Perform pre-processing on the created customer dataset

Tools Used

1. Python
2. Pandas library
3. scikit-learn library
4. Jupyter Notebook as an IDE.

Dataset and task Overview

The dataset contains Amazon baby product reviews.
https://www.kaggle.com/bittlingmayer/amazonreviews

It has three columns: name, review, and rating. Reviews are text data and ratings are numbering from 1 to 5 where 1 is the worst and 5 is the best review.

Our job is to analyze the reviews as positive and negative reviews. Here we used the first five entries to examine the data.

```
import pandas as pd
products = pd.read_csv('amazon_baby.csv')
products.head()
```

name	review	rating
Planetwise Flannel Wipes	These flannel wipes are OK, but in my opinion ...	3
Planetwise Wipe Pouch	it came early and was not disappointed. i love...	5
Annas Dream Full Quilt with 2 Shams	Very soft and comfortable and warmer than it l...	5
Stop Pacifier Sucking without tears with Thumb...	This is a product well worth the purchase. I ...	5
Stop Pacifier Sucking without tears with Thumb...	All of my kids have cried non-stop when I trie...	5

Data Preprocessing

In this dataset, we have to work on these three columns and all three of them are crucial. If the data is not available in any row in a column that row is unnecessary.

```
len(products) - len(products.dropna())
```

We have null values in 1147 rows. Now, check how much total data we have.

```
len(products)
```

We have a total of 183531 data. So, if we delete all the null values, we will still have a sizable data to train an algorithm. So, let's drop the null values.

```
products = products.dropna()
```

We need to have all the string data in the review column. If there is any data that has other types, it will cause trouble in later steps.

Now, we will check the datatype of the review data of every row. If there is any row having data in any other type than string we will change that to a string.

```
for i in range(0,len(products)-1):
    if type(products.iloc[i]['review']) != str:
        products.iloc[i]['review'] =
str(products.iloc[i]['review'])
```

As we are doing sentiment analysis, it is important to tell our model what is positive sentiment and what is a negative sentiment.

In our rating column, we have ratings from 1 to 5. We can define 1 and 2 as bad reviews and 4 and 5 as good reviews.

Solution for 2.2 Perform sentiment analysis of the customer

We will denote positive sentiments as 1 and negative sentiments as 0. Let's write a function 'sentiment' that returns 1 if the rating is 4 or more else return 0. Then, apply the function sentiment and create a new column that will represent the positive and negative sentiment as 1 or 0.

```
def sentiment(n):
    return 1 if n >= 4 else 0
products['sentiment'] = products['rating'].apply(sentiment)
products.head()
```

	name	review	rating	sentiment
1	Planetwise Wipe Pouch	it came early and was not disappointed. i love...	5	1
2	Annas Dream Full Quilt with 2 Shams	Very soft and comfortable and warmer than it l...	5	1
3	Stop Pacifier Sucking without tears with Thumb...	This is a product well worth the purchase. I ...	5	1
4	Stop Pacifier Sucking without tears with Thumb...	All of my kids have cried non-stop when I trie...	5	1
5	Stop Pacifier Sucking without tears with Thumb...	When the Binky Fairy came to our house, we did...	5	1

Look, we have the 'sentiment' column added at the end now!

First, we need to prepare the training features. Combine both 'name' and 'review' columns and make one single column. First, write a function 'combined_features' that will combine both the columns. Then, apply the function and create a new column 'all_features' that will contain the strings from both name and review columns.

```
def combined_features(row):
    return row['name'] + ' '+
row['review']products['all_features'] =
```

```
products.apply(combined_features, axis=1)
products.head()
```

	name	review	rating	sentiment	all_features
1	Planetwise Wipe Pouch	it came early and was not disappointed. i love...	5	1	Planetwise Wipe Pouch it came early and was no...
2	Annas Dream Full Quilt with 2 Shams	Very soft and comfortable and warmer than it l...	5	1	Annas Dream Full Quilt with 2 Shams Very soft ...
3	Stop Pacifier Sucking without tears with Thumb...	This is a product well worth the purchase. I ...	5	1	Stop Pacifier Sucking without tears with Thumb...
4	Stop Pacifier Sucking without tears with Thumb...	All of my kids have cried non-stop when I trie...	5	1	Stop Pacifier Sucking without tears with Thumb...
5	Stop Pacifier Sucking without tears with Thumb...	When the Binky Fairy came to our house, we did...	5	1	Stop Pacifier Sucking without tears with Thumb...

You can see the ‘all_features’ column at the end. Now, we are ready to develop the sentiment classifier!

Develop the sentiment classifier

Here is the process step by step.

We need to define the input variable X and the output variable y.

X should be the ‘all_features’ column and y should be our ‘sentiment’ column

```
X = products['all_features']
y = products['sentiment']
```

We need to split the dataset so that there is a training set and a test set.

The ‘train_test_split’ function from the scikit-learn library is helpful. The model will be trained using the training dataset and the performance of the model can be tested using the test dataset.

'train_test_split' automatically splits the data in 75/25 proportion. 75% for the training and 25% for the testing. If you want the proportion to be different, you need to define that.

```
from sklearn.model_selection import train_test_splitX_train,
X_test, y_train, y_test = train_test_split(X, y, random_state=0)
```

I am going to use 'CountVectorizer' from the scikit-learn library.

CountVectorizer develops a vector of all the words in the string. Import CountVectorizer and fit both our training, testing data into it.

```
from sklearn.feature_extraction.text import CountVectorizercv =
CountVectorizer()
ctmTr = cv.fit_transform(X_train)
X_test_dtm = cv.transform(X_test)
```

Let's dive into the original model part. This is the most fun part. We will use the Logistic Regression as this is a binary classification. Let's do the necessary imports and fit our training data in the model.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
model = LogisticRegression()
model.fit(ctmTr, y_train)
```

The logistic regression model is trained with the training data.

Solution for 2.3 Results and Discussions

Use the trained model above to predict the sentiments for the test data. If we pass the test features, it will predict the output y that is the sentiment data.

```
y_pred_class = model.predict(X_test_dtm)
```

output:

```
array([1, 1, 1, ..., 1, 1, 0], dtype=int64)
```

Here is the output for the test data. As we remember, we used 1 for good reviews and 0 for a bad review.

Use the accuracy_score function to get the accuracy_score of the test data. So, it will compare the predicted 'sentiment' with the original 'sentiment' data to calculate the percentage of accuracy.

```
accuracy_score(y_test, y_pred_class)
```

The accuracy score I got for this data on the test set is 84%, which is very good.

Conclusion

This simple sentiment analysis classifier can be useful in many other types of datasets. It can be used in real-world projects and businesses as well. The dataset we used here resembles a real business dataset.

References

Data Set from
https://www.kaggle.com/bittlingmayer/amazonreviews