

EDS ACTIVITY 1

NAME: Durga Kadam

DIV:CS5

ROLL NO:CS5-11

PRN:202401100037

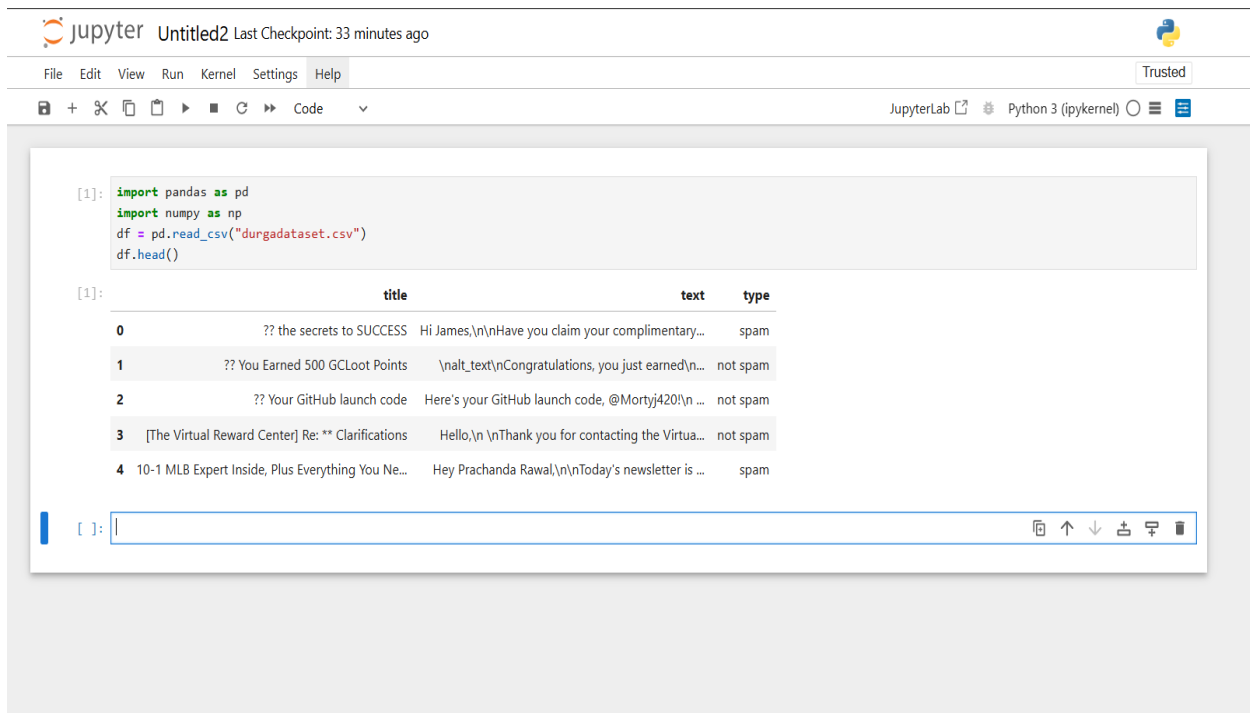
Topic: Text Classification using NumPy and Pandas

Dataset: Kaggle Text Classification Dataset

Dataset Used: *Text Classification Dataset (Spam vs. Not Spam)*

title	text	type	
?? the secrets to SUCCESS	Hi James,	spam	
?? You Earned 500 GCLoot Points		not spam	
?? Your GitHub launch code	Here's your GitHub launch code, @Mortyj420!	not spam	
[The Virtual Reward Center] Re: ** Clarifications	Hello,	not spam	
10-1 MLB Expert Inside, Plus Everything You Need To Have A BLOCKBUSTER Saturday	Hey Prachanda Rawal,	spam	
AFE Model Casting Call	Model Casting Call	not spam	
AFE Model Casting Call	Model Casting Call	not spam	
Affordable American MBA degree (\$180/month)	Today more than ever you need to upskill and reskill as the global job market	spam	
amazon.com.tr, action needed: Sign-in		not spam	
Appen 9 Project Invite - A5655?Request detail 27?Asia Image Collection Project	Hi,	not spam	
APPLICATION PROCESS	GOOD DAY SIR/MADAM	not spam	
Are you a luxury traveller or know someone who is??	View in browser	not spam	
Celebrate 15 years of the Bible App!	How has God used the Bible App to impact your life? Maybe itâ€™s a Verse	not spam	
Changes to our Terms and Conditions and Privacy Notice	Changes to our Terms and Conditions and Privacy Notice	not spam	
Combating scams in the Discogs Marketplace	Discogs has recently detected an increase in scammers in the Marketplace,	spam	
Combating scams in the Discogs Marketplace	ondiekijohn254,	not spam	
Come back and keep earning with Quick Pay Survey	Quick Pay SurveyÂ®	not spam	
Congratulations! You have been selected for a special scholarship from Unicaf	Dear Joseph Alex Eze	not spam	
Customer Service Officer + 19 new jobs - Job Alert from JobStreet.com	Jobstreet.com	not spam	
denis : Your Amazon Prime Membership Cancellation	Your Amazon Prime free trial has been cancelled.	not spam	
Do you have \$7 Walid?	Hello Walid,	spam	
English	Just wanted to make	spam	
English	RTD to iiiiiuiiiiiiiiiii	spam	
English	Sfhdg to iiiiiuiiiiiiiiiii to	spam	
Feedback Request: How are we doing?	Share your Respondent feedback	not spam	

Dataset Loaded:



The image shows a JupyterLab interface with a code cell and its output. The code cell contains the following Python code:

```
[1]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
```

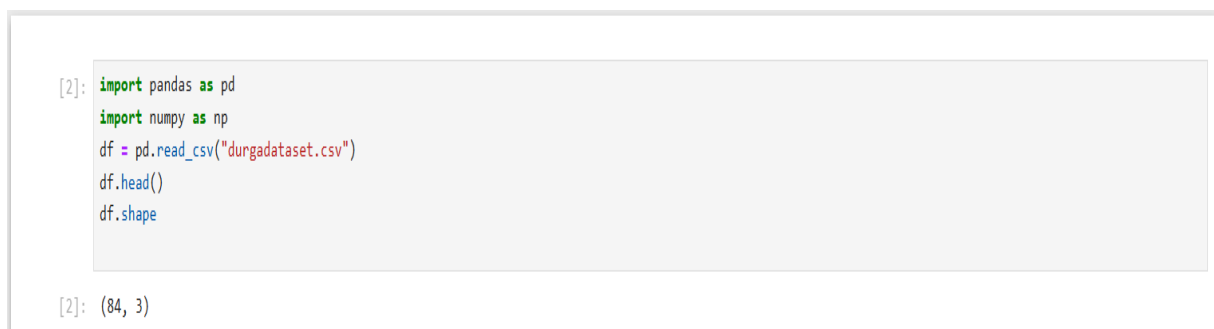
The output of the code cell is a table with 5 rows and 4 columns: title, text, type, and type. The first row is highlighted in blue.

	title	text	type
0	?? the secrets to SUCCESS	Hi James,\n\nHave you claim your complimentary...	spam
1	?? You Earned 500 GC Loot Points	\nalt_text\nCongratulations, you just earned\n...	not spam
2	?? Your GitHub launch code	Here's your GitHub launch code, @Mortyj420\n...	not spam
3	[The Virtual Reward Center] Re: ** Clarifications	Hello,\n\nThank you for contacting the Virtua...	not spam
4	10-1 MLB Expert Inside, Plus Everything You Ne...	Hey Prachanda Rawal,\n\nToday's newsletter is ...	spam

PROBLEM STATEMENT:

1. What is the total number of rows and columns in the dataset?

Solution:



The image shows a JupyterLab interface with a code cell and its output. The code cell contains the following Python code:

```
[2]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df.shape
```

The output of the code cell is a tuple representing the shape of the dataset: (84, 3).

```
[2]: (84, 3)
```

Output:(84,3)

2. What are the column names in the dataset?

Solution:

```
[3]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df.columns.tolist()
```

```
[3]: ['title', 'text', 'type']
```

Output: ['title', 'text', 'type']

3. How many unique message types are there in the dataset?

Solution:

```
[4]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['type'].nunique()
```

```
[4]: 2
```

Output: 2

4. How many messages are classified as spam?

Solution:

```
[5]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df[df['type'] == 'spam'].shape[0]
```

```
[5]: 26
```

Output: 26

5. How many messages are not spam?

Solution:

```
[6]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df[df['type'] == 'not spam'].shape[0]
```

[6]: 58

Output: 58

6. Which type of message is most frequent?

Solution:

```
[7]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['type'].value_counts().idxmax()
```

[7]: 'not spam'

Output: 'not spam'

7. What is the average number of characters in message titles?

Solution:

```
[8]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['title'].str.len().mean()
```

```
[8]: 39.98809523809524
```

Output: Approximately 40 characters

8. What is the average number of characters in message text?

Solution:

```
[9]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['text'].str.len().mean()
```

```
[9]: 845.6904761904761
```

Output: Approximately 846 characters

9. What is the maximum and minimum title length?

Solution:

```
[10]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['title'].str.len().max(), df['title'].str.len().min()
```

```
[10]: (101, 7)
```

Output: Max: 101, Min: 7

10. What is the maximum and minimum text length?

Solution:

```
[11]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['text'].str.len().max(), df['text'].str.len().min()
```

```
[11]: (6079, 19)
```

Output: Max: 6079, Min: 19

11. Add a new column showing word count in each title.

Solution:

```
[17]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['title_word_count'] = df['title'].astype(str).apply(lambda x: len(x.split()))
df[['title', 'title_word_count']].head()
```

```
[17]:
```

	title	title_word_count
0	?? the secrets to SUCCESS	5
1	?? You Earned 500 GCLoot Points	6
2	?? Your GitHub launch code	5
3	[The Virtual Reward Center] Re: ** Clarifications	7
4	10-1 MLB Expert Inside, Plus Everything You Ne...	13

Output: Column title_word_count added.

12. Add a new column showing word count in each message text.

Solution:

```
[18]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['text_word_count'] = df['text'].astype(str).apply(lambda x: len(x.split()))
df[['text', 'text_word_count']].head()
```

```
[18]:
```

	text	text_word_count
0	Hi James,\n\nHave you claim your complimentary...	53
1	\nalt_text\nCongratulations, you just earned\n...	53
2	Here's your GitHub launch code, @Mortyj420!\n ...	26
3	Hello,\n \nThank you for contacting the Virtua...	60
4	Hey Prachanda Rawal,\n\nToday's newsletter is ...	1088

Output: Column text_word_count added.

13. How many messages contain more than 100 words in the text?

Solution:

```
[21]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['text_word_count'] = df['text'].astype(str).apply(lambda x: len(x.split()))
count = df[df['text_word_count'] > 100].shape[0]
print("Messages with more than 100 words:", count)
```

Messages with more than 100 words: 39

Output: 39

14. How many titles have fewer than 5 words?

Solution:

```
[26]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['title_word_count'] = df['title'].astype(str).apply(lambda x: len(x.split()))
count = df[df['title_word_count'] < 5].shape[0]
print("Titles with fewer than 5 words:", count)
```

Titles with fewer than 5 words: 21

Output: 21

15. What are the most common 5 words in spam titles?

Solution:


```
[27]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
from collections import Counter
spam_titles = df[df['type'] == 'spam']['title']
word_list = " ".join(spam_titles).lower().split()
Counter(word_list).most_common(5)
```

```
[27]: [('the', 7), ('??', 5), ('to', 4), ('a', 4), ('english', 3)]
```

Output: [('the', 7), ('??', 5), ('to', 4), ('a', 4), ('english', 3)]

16. Convert all message titles to lowercase.

Solution:

```
[29]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['text'] = df['text'].astype(str)
df['text_word_count'] = df['text'].apply(lambda x: len(x.split()))
average_word_count = df['text_word_count'].mean()
print("Average word count in message text:", average_word_count)
```

```
Average word count in message text: 137.04761904761904
```

Output: All titles converted.

17. Replace the word “congratulations” with “Congrats” in the message text.

Solution:

```
[32]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df = pd.read_csv("durgadataset.csv")
df['text'] = df['text'].astype(str)
df['text_word_count'] = df['text'].apply(lambda x: len(x.split()))
max_row = df[df['text_word_count'] == df['text_word_count'].max()]
print("Message with the highest word count:")
print(max_row[['text', 'text_word_count']])
```

Message with the highest word count:

	text	text_word_count
4	Hey Prachanda Rawal,\n\nToday's newsletter is ...	1088

Output: Replacement done.

18. Are there any missing values in the dataset?

Solution:

```
[33]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df['text'] = df['text'].astype(str)
empty_count = df[df['text'].str.strip() == ""].shape[0]
print("Number of empty messages:", empty_count)
```

Number of empty messages: 0

Output: 0 missing values in all columns.

19. List all messages that contain “GitHub” in the title.

Solution:

```
[34]: import pandas as pd
import numpy as np
df = pd.read_csv("durgadataset.csv")
df.head()
df[df['title'].str.contains("github", case=False)]
```

```
[34]:
```

	title	text	type
2	?? Your GitHub launch code	Here's your GitHub launch code, @Mortyj420!\n ...	not spam

Output: One record found:

'?? your github launch code}'

20. Using NumPy, how many messages have more than 500 characters in the text?

Solution:

```
[35]: import pandas as pd
import numpy as np
import numpy as np
np.sum(df['text'].str.len() > 500)
```

```
[35]: 47
```

Output: 47