

Name: Venkata Krishnarjun Vuppala

SRN: PES2UG19CS451

Semester: 6

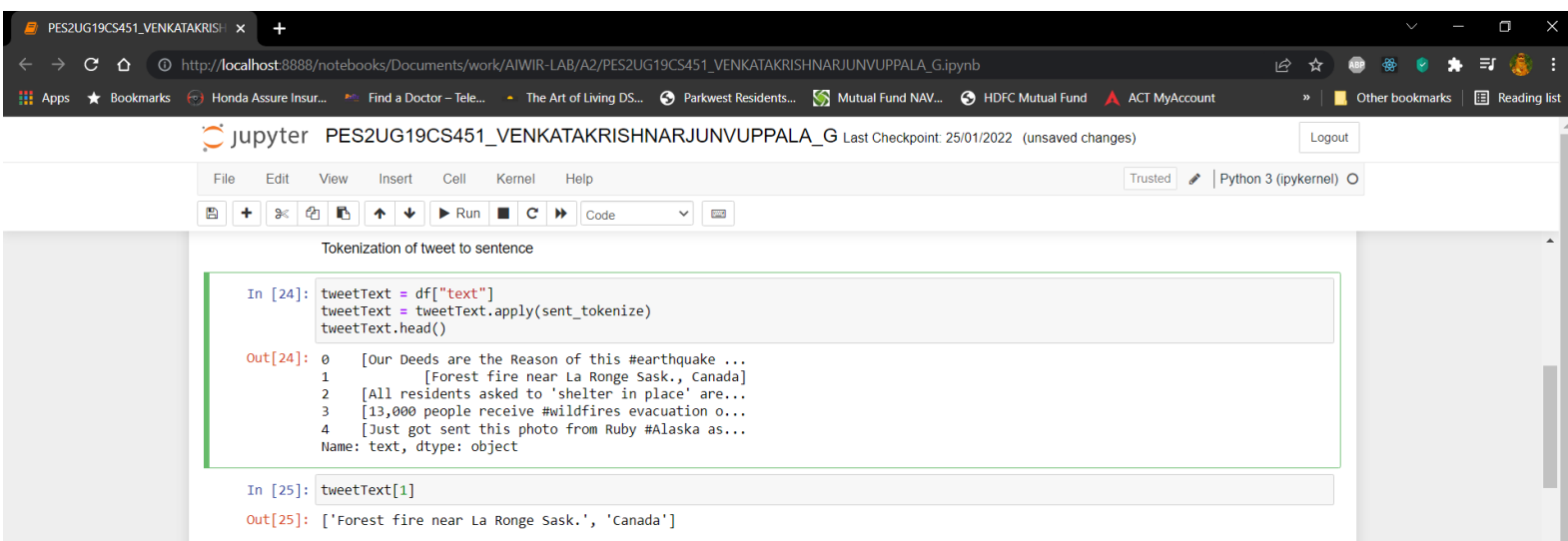
Section: G

Subject: Algorithms for Intelligence Web and Information Retrieval

Assignment 2

1. Tokenize each Tweet into sentences

Code and Output:



The screenshot shows a Jupyter Notebook interface in a web browser. The browser's address bar displays the URL: `http://localhost:8888/notebooks/Documents/work/AIWR-LAB/A2/PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G.ipynb`. The notebook's title bar indicates the file name: `PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G`, with a last checkpoint of 25/01/2022 and unsaved changes. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Help) and a toolbar with icons for file operations, running cells, and code execution. The notebook content shows two code cells. The first cell, labeled 'In [24]:', contains the following Python code:

```
tweetText = df["text"]
tweetText = tweetText.apply(sent_tokenize)
tweetText.head()
```

. The output, labeled 'Out[24]:', displays a list of five tweets, each as a list of tokens. The second cell, labeled 'In [25]:', contains the code `tweetText[1]`, and its output, labeled 'Out[25]:', shows the tokenized text for the second tweet: `['Forest fire near La Ronge Sask.', 'Canada']`.

```
Tokenization of tweet to sentence

In [24]: tweetText = df["text"]
          tweetText = tweetText.apply(sent_tokenize)
          tweetText.head()

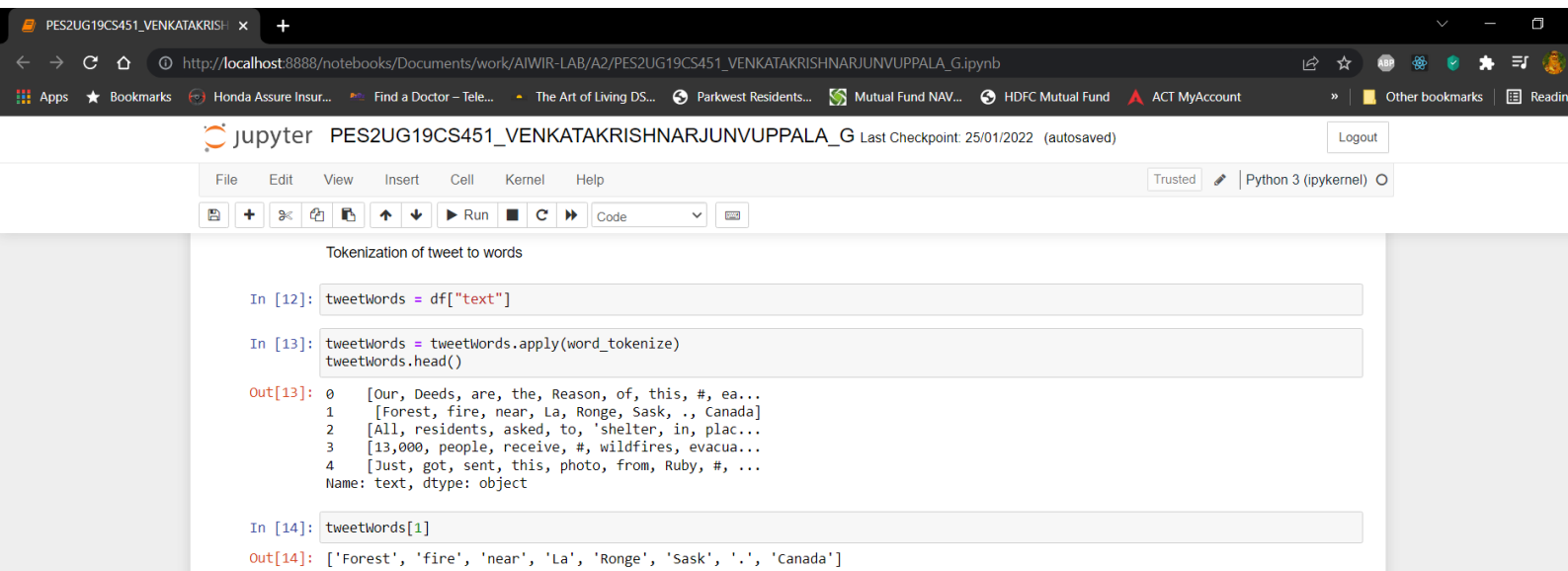
Out[24]: 0    [Our Deeds are the Reason of this #earthquake ...
          1    [Forest fire near La Ronge Sask., Canada]
          2    [All residents asked to 'shelter in place' are...
          3    [13,000 people receive #wildfires evacuation o...
          4    [Just got sent this photo from Ruby #Alaska as...
          Name: text, dtype: object

In [25]: tweetText[1]

Out[25]: ['Forest fire near La Ronge Sask.', 'Canada']
```

2. Tokenize each tweet into words

Code and Output:



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
Tokenization of tweet to words

In [12]: tweetWords = df["text"]

In [13]: tweetWords = tweetWords.apply(word_tokenize)
tweetWords.head()

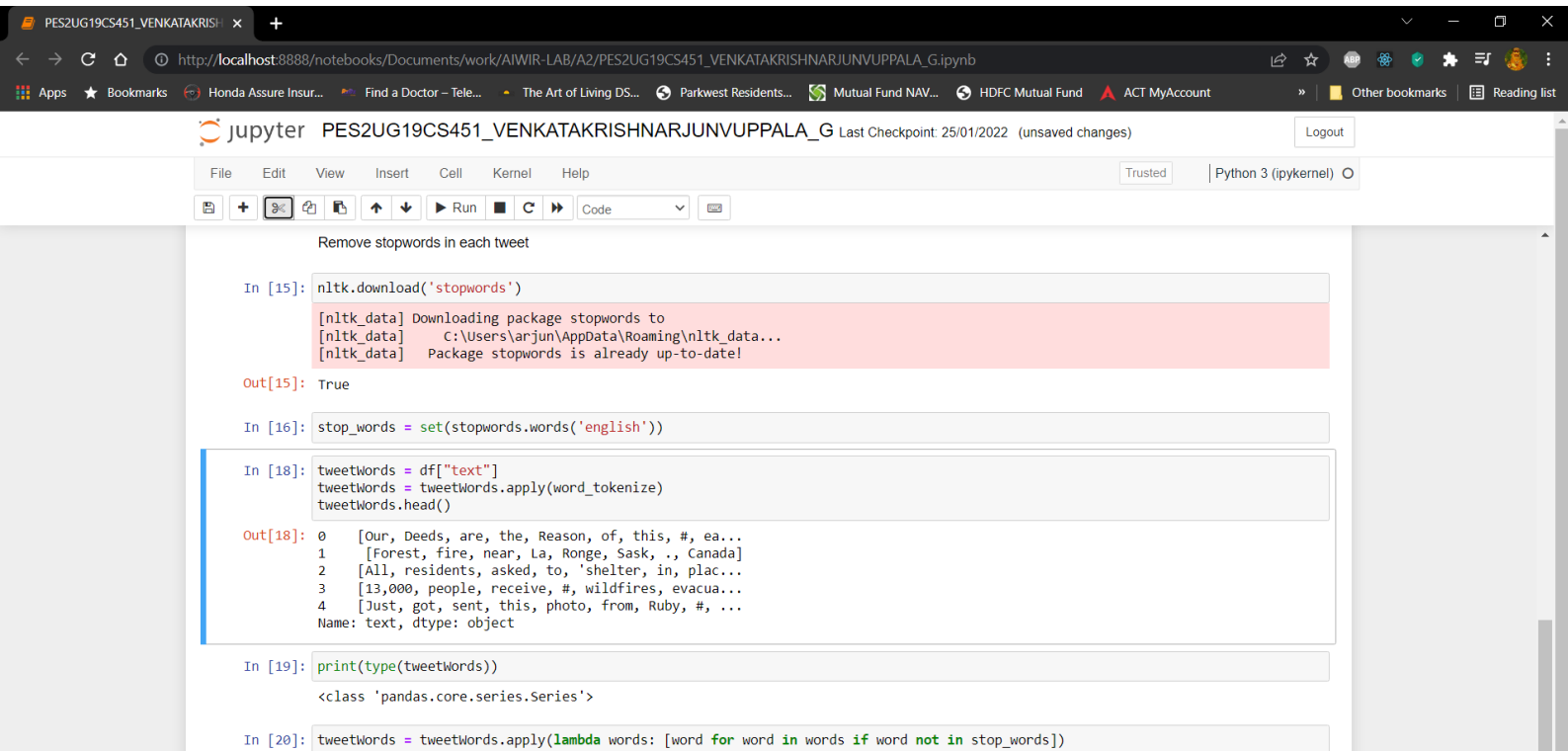
Out[13]: 0    [Our, Deeds, are, the, Reason, of, this, #, ea...
1    [Forest, fire, near, La, Ronge, Sask, ., Canada]
2    [All, residents, asked, to, 'shelter, in, plac...
3    [13,000, people, receive, #, wildfires, evacua...
4    [Just, got, sent, this, photo, from, Ruby, #, ...
Name: text, dtype: object

In [14]: tweetWords[1]

Out[14]: ['Forest', 'fire', 'near', 'La', 'Ronge', 'Sask', '.', 'Canada']
```

3. Remove stopwords in each tweet - NLTK library

Code and output:



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
Remove stopwords in each tweet

In [15]: nltk.download('stopwords')

[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\arjun\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!

Out[15]: True

In [16]: stop_words = set(stopwords.words('english'))

In [18]: tweetWords = df["text"]
tweetWords = tweetWords.apply(word_tokenize)
tweetWords.head()

Out[18]: 0    [Our, Deeds, are, the, Reason, of, this, #, ea...
1    [Forest, fire, near, La, Ronge, Sask, ., Canada]
2    [All, residents, asked, to, 'shelter, in, plac...
3    [13,000, people, receive, #, wildfires, evacua...
4    [Just, got, sent, this, photo, from, Ruby, #, ...
Name: text, dtype: object

In [19]: print(type(tweetWords))

<class 'pandas.core.series.Series'>

In [20]: tweetWords = tweetWords.apply(lambda words: [word for word in words if word not in stop_words])
```

PES2UG19CS451_VENKATAKRISHNARJUNVUPPALA_G Last Checkpoint: 25/01/2022 (unsaved changes) Logout

File Edit View Insert Cell Kernel Help Trusted Python 3 (ipykernel)

In [21]: `tweetWords.head(5)`

```
Out[21]: 0    [Our, Deeds, Reason, #, earthquake, May, ALLAH...
1    [Forest, fire, near, La, Ronge, Sask, ., Canada]
2    [All, residents, asked, 'shelter, place, ', no...
3    [13,000, people, receive, #, wildfires, evacua...
4    [Just, got, sent, photo, Ruby, #, Alaska, smok...
Name: text, dtype: object
```

In [22]: `tweetWords[1]`

```
Out[22]: ['Forest', 'fire', 'near', 'La', 'Ronge', 'Sask', '.', 'Canada']
```