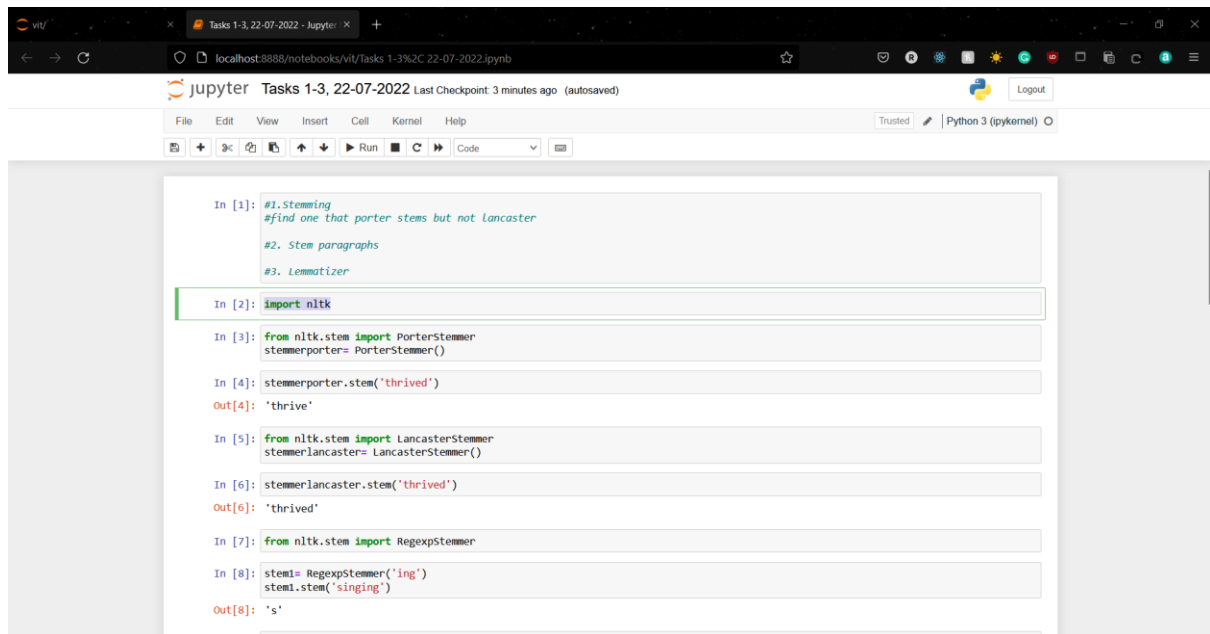


19BCE2446

Rizwan Hussain

Tasks 1-5, 22-07-2022

1. Task -1: Stemming



```
In [1]: #1. Stemming
        #find one that porter stems but not Lancaster
        #2. Stem paragraphs
        #3. Lemmatizer

In [2]: import nltk

In [3]: from nltk.stem import PorterStemmer
        stemmerporter = PorterStemmer()

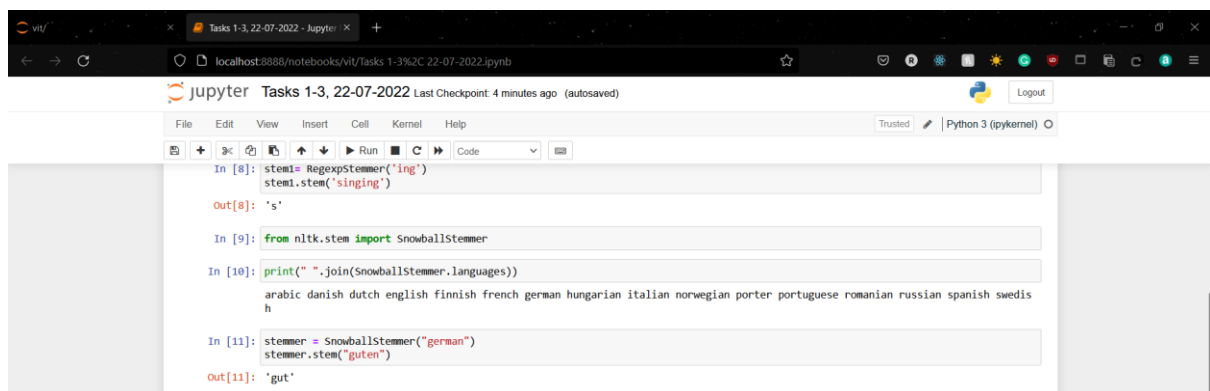
In [4]: stemmerporter.stem('thrive')
Out[4]: 'thrive'

In [5]: from nltk.stem import LancasterStemmer
        stemmerlancaster = LancasterStemmer()

In [6]: stemmerlancaster.stem('thrive')
Out[6]: 'thrived'

In [7]: from nltk.stem import RegexpStemmer

In [8]: stem1 = RegexpStemmer('ing')
        stem1.stem('singing')
Out[8]: 's'
```



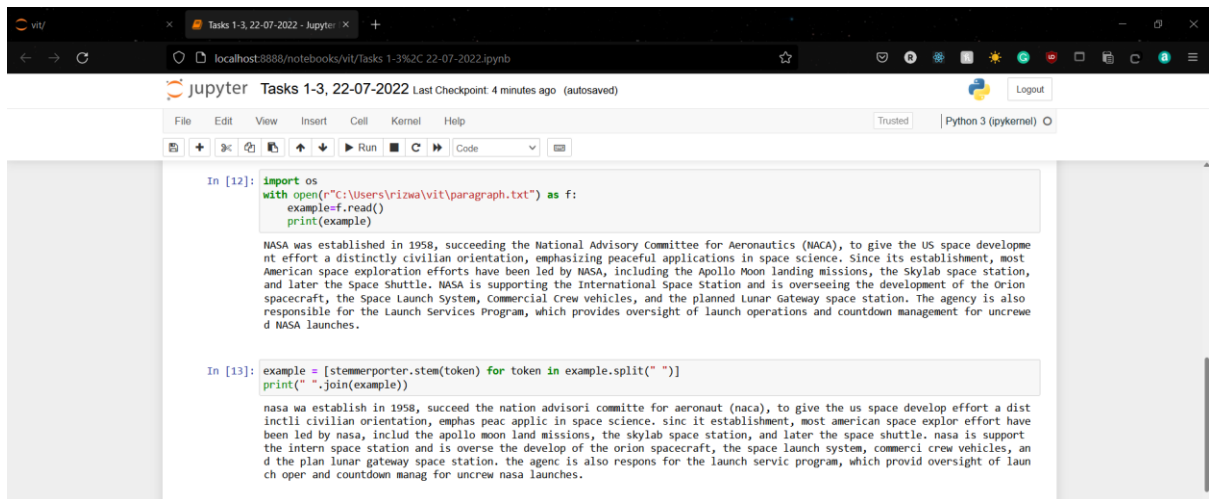
```
In [8]: stem1 = RegexpStemmer('ing')
        stem1.stem('singing')
Out[8]: 's'

In [9]: from nltk.stem import SnowballStemmer

In [10]: print(" ".join(SnowballStemmer.languages))
arabic danish dutch english finnish french german hungarian italian norwegian porter portuguese romanian russian spanish swedis
h

In [11]: stemmer = SnowballStemmer("german")
        stemmer.stem("guten")
Out[11]: 'gut'
```

2. Task- 2: Stemming a Paragraph



The screenshot shows a Jupyter Notebook interface with two code cells. The first cell imports the Porter stemmer and reads a file named 'paragraph.txt'. The second cell uses the stemmer to process the text from the first cell. The output of the first cell is a long paragraph about NASA. The output of the second cell is the same paragraph with all words stemmed.

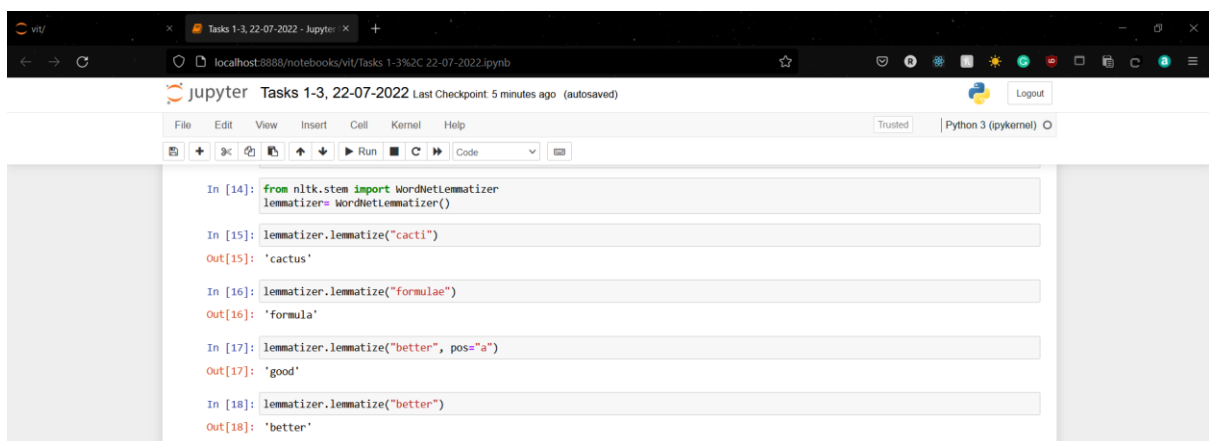
```
In [12]: import os
with open(r"C:\Users\rizwa\vit\paragraph.txt") as f:
    example=f.read()
    print(example)

NASA was established in 1958, succeeding the National Advisory Committee for Aeronautics (NACA), to give the US space developme
nt effort a distinctly civilian orientation, emphasizing peaceful applications in space science. Since its establishment, most
American space exploration efforts have been led by NASA, including the Apollo Moon landing missions, the Skylab space station,
and later the Space Shuttle. NASA is supporting the International Space Station and is overseeing the development of the Orion
spacecraft, the Space Launch System, Commercial Crew vehicles, and the planned Lunar Gateway space station. The agency is also
responsible for the Launch Services Program, which provides oversight of launch operations and countdown management for uncrew
ed NASA launches.

In [13]: example = [stemmerporter.stem(token) for token in example.split(" ")]
print(" ".join(example))

nasa wa establish in 1958, succeed the nation advisori committe for aeronaut (naca), to give the us space develop effort a dist
inctli civilian orientation, emphas peac applic in space science. sinc it establishment, most american space explor effort have
been led by nasa, includ the apollo moon land missions, the skylab space station, and later the space shuttle. nasa is support
the intern space station and is overse the develop of the orion spacecraft, the space launch system, commerci crew vehicles, an
d the plan lunar gateway space station. the agenc is also respons for the launch servic program, which provid oversight of laun
ch oper and countdown manag for uncrew nasa launches.
```

3. Task- 3: Lemmatize



The screenshot shows a Jupyter Notebook interface with four code cells. The first cell imports WordNetLemmatizer from nltk.stem. The next three cells demonstrate the lemmatization of the words 'cactus', 'formulae', and 'better' (with pos='a'). The outputs are 'cactus', 'formula', and 'better' respectively.

```
In [14]: from nltk.stem import WordNetLemmatizer
lemmatizer= WordNetLemmatizer()

In [15]: lemmatizer.lemmatize("cacti")
Out[15]: 'cactus'

In [16]: lemmatizer.lemmatize("formulae")
Out[16]: 'formula'

In [17]: lemmatizer.lemmatize("better", pos="a")
Out[17]: 'good'

In [18]: lemmatizer.lemmatize("better")
Out[18]: 'better'
```

4. Task- 4: Jieba (for segmentation)



The screenshot shows a Jupyter Notebook interface with two code cells. The first cell imports jieba. The second cell uses jieba.cut to segment the sentence '你好吗' and prints the result. The output is a list of characters: ['你', '好', '吗'].

```
In [19]: import jieba

In [20]: seg= jieba.cut("你好吗")
print(" ".join(seg))

Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\rizwa\AppData\Local\Temp\jieba.cache
Loading model cost 0.450 seconds.
Prefix dict has been built successfully.

你好 吗

In [ ]:
```

5. Task- 5: Tokenize

[illegible]