Project 2

- **import nltk**

- **from nltk.corpus import stopwords**
- **from bs4 import BeautifulSoup**
- **import urllib.request**

- **import plotly.io as pio**

- **page = urllib.request.urlopen('https://en.wikipedia.org/wiki/Narendra_Modi')**

- **html_plain = page.read()**

Project 2

- **print(html_plain)**
- **soup = BeautifulSoup(html_plain,'html.parser')**
- **soup_text = soup.get_text(strip = True)**

print(soup_text)

- **ready_text = soup_text.lower()**
- **print(ready_text)**

Project 2

```
tokens = []
for t in ready_text.split():
    tokens.append(t)
print(tokens)
len(tokens)
#nltk.download()
```

```
stop_words = stopwords.words('english')
clean_tokens = tokens[:]
for token in tokens:
    if token in stop_words:
        clean_tokens.remove(token)
print(clean_tokens)
```

Project 2

- len(clean_tokens)

- freq = nltk.FreqDist(clean_tokens)

- for key, val in freq.items():
- print('Word: ' + str(key) + ', Quantity:' + str(val))

- high_freq = dict()
- for key, val in freq.items():
- if (val > 10):
- high_freq[key] = val

Project 2

- print(high_freq)

- fig = dict({
- "data": [{"type": "bar",
- "x": list(high_freq.keys()),
- "y": list(high_freq.values())}],
- "layout": {"title": {"text": "Most frequently used words in the page"}, "xaxis": {"categoryorder":"total descending"}}
- })
- pio.show(fig)