



DIGITAL ASSIGNMENT-5

BAG OF WORDS

Practice Questions

December 20, 2024

SAI VARUN AITHA

saivarun.aitha18103@gmail.com

```
In [1]: from sklearn.feature_extraction.text import CountVectorizer
```

```
In [4]: text = """I have three visions for India. In 3000 years of our history, people f
the world have come and invaded us, captured our lands, conquered our minds.
From Alexander onwards, the Greeks, the Turks, the Moguls, the Portuguese, the B
the French, the Dutch, all of them came and looted us, took over what was ours.
Yet we have not done this to any other nation. We have not conquered anyone."""
```

```
In [5]: import re
text_cleaned = re.sub(r'^\w\s', '', text.lower())
```

```
In [6]: vectorizer = CountVectorizer()
bow_matrix = vectorizer.fit_transform([text_cleaned])
```

```
In [7]: features = vectorizer.get_feature_names_out()

bow_array = bow_matrix.toarray()

print("Feature Names (Vocabulary):")
print(features)

print("\nBag of Words Matrix:")
print(bow_array)
```

Feature Names (Vocabulary):

```
['3000' 'alexander' 'all' 'and' 'any' 'anyone' 'british' 'came' 'captured'
'come' 'conquered' 'done' 'dutch' 'for' 'french' 'from' 'greeks' 'have'
'history' 'in' 'india' 'invaded' 'lands' 'looted' 'minds' 'moguls'
'nation' 'not' 'of' 'onwards' 'other' 'our' 'ours' 'over' 'people'
'portuguese' 'the' 'them' 'this' 'three' 'to' 'took' 'turks' 'us'
'visions' 'was' 'we' 'what' 'world' 'years' 'yet']
```

Bag of Words Matrix:

```
[[1 1 2 2 1 1 1 1 1 2 1 1 1 2 1 4 1 1 1 1 1 1 1 1 2 2 1 1 3 1 2 1 1
 8 1 1 1 1 1 1 2 1 1 2 1 1 1 1]]
```

```
In [ ]:
```