seven-74

April 20, 2024

Assignment 7 TCOD74 - Bendre Anushka A.

[14]:
```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk.probability import FreqDist
from sklearn.feature_extraction.text import TfidfVectorizer
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Vaibhavi\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\Vaibhavi\AppData\Roaming\nltk_data…
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Vaibhavi\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Vaibhavi\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[14]:
```
True
```

[26]:
```
# Sample document
document = """
Natural language processing (NLP) is a subfield of linguistics,
computer science, and artificial intelligence concerned with the
interactions between computers and human language, in particular how
to program computers to process and analyze large amounts of natural
language data. Challenges in natural language processing frequently
involve speech recognition, natural language understanding, and
natural language generation.
```

```
"""
```

[27]:
```
# Tokenization
tokens = word_tokenize(document)
```

[17]:
```
# POS Tagging
pos_tags = nltk.pos_tag(tokens)
```

[18]:
```
# Stop words removal
stop_words = set(stopwords.words('english'))
filtered_tokens = [word for word in tokens if word.lower() not in stop_words]
```

[19]:
```
# Stemming
stemmer = PorterStemmer()
stemmed_tokens = [stemmer.stem(word) for word in filtered_tokens]
```

[20]:
```
# Lemmatization
lemmatizer = WordNetLemmatizer()
lemmatized_tokens = [lemmatizer.lemmatize(word) for word in filtered_tokens]
```

[21]:
```
# Term Frequency (TF) calculation
tf = FreqDist(lemmatized_tokens)
```

[22]:
```
# Inverse Document Frequency (IDF) calculation
corpus = [document]
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(corpus)
idf = vectorizer.idf_
```

[25]:
```
# Print the results
print("1. Preprocessed Document:")
print("Tokens:", tokens)
print("POS Tags:", pos_tags)
print("Filtered Tokens (Stop words removal):", filtered_tokens)
print("Stemmed Tokens:", stemmed_tokens)
print("Lemmatized Tokens:", lemmatized_tokens)
print("\n2. Term Frequency (TF):")
print(tf)
print("\n3. Inverse Document Frequency (IDF):")
print(dict(zip(vectorizer.get_feature_names_out(), idf)))
```

```
1. Preprocessed Document:
Tokens: ['Natural', 'language', 'processing', '(', 'NLP', ')', 'is', 'a',
'subfield', 'of', 'linguistics', ',', 'computer', 'science', ',', 'and',
'artificial', 'intelligence', 'concerned', 'with', 'the', 'interactions',
'between', 'computers', 'and', 'human', 'language', ',', 'in', 'particular',
'how', 'to', 'program', 'computers', 'to', 'process', 'and', 'analyze', 'large',
```

```
'amounts', 'of', 'natural', 'language', 'data', '.', 'Challenges', 'in',
'natural', 'language', 'processing', 'frequently', 'involve', 'speech',
'recognition', ',', 'natural', 'language', 'understanding', ',', 'and',
'natural', 'language', 'generation', '.']
POS Tags: [('Natural', 'JJ'), ('language', 'NN'), ('processing', 'NN'), ('(',
'('), ('NLP', 'NNP'), (')', ')'), ('is', 'VBZ'), ('a', 'DT'), ('subfield',
'NN'), ('of', 'IN'), ('linguistics', 'NNS'), (',', ','), ('computer', 'NN'),
('science', 'NN'), (',', ','), ('and', 'CC'), ('artificial', 'JJ'),
('intelligence', 'NN'), ('concerned', 'VBN'), ('with', 'IN'), ('the', 'DT'),
('interactions', 'NNS'), ('between', 'IN'), ('computers', 'NNS'), ('and', 'CC'),
('human', 'JJ'), ('language', 'NN'), (',', ','), ('in', 'IN'), ('particular',
'JJ'), ('how', 'WRB'), ('to', 'TO'), ('program', 'NN'), ('computers', 'NNS'),
('to', 'TO'), ('process', 'VB'), ('and', 'CC'), ('analyze', 'VB'), ('large',
'JJ'), ('amounts', 'NNS'), ('of', 'IN'), ('natural', 'JJ'), ('language', 'NN'),
('data', 'NNS'), ('.', '.'), ('Challenges', 'NNS'), ('in', 'IN'), ('natural',
'JJ'), ('language', 'NN'), ('processing', 'NN'), ('frequently', 'RB'),
('involve', 'VBP'), ('speech', 'NN'), ('recognition', 'NN'), (',', ','),
('natural', 'JJ'), ('language', 'NN'), ('understanding', 'NN'), (',', ','),
('and', 'CC'), ('natural', 'JJ'), ('language', 'NN'), ('generation', 'NN'),
('.', '.')]
Filtered Tokens (Stop words removal): ['Natural', 'language', 'processing', '(',
'NLP', ')', 'subfield', 'linguistics', ',', 'computer', 'science', ',',
'artificial', 'intelligence', 'concerned', 'interactions', 'computers', 'human',
'language', ',', 'particular', 'program', 'computers', 'process', 'analyze',
'large', 'amounts', 'natural', 'language', 'data', '.', 'Challenges', 'natural',
'language', 'processing', 'frequently', 'involve', 'speech', 'recognition', ',',
'natural', 'language', 'understanding', ',', 'natural', 'language',
'generation', '.']
Stemmed Tokens: ['natur', 'languag', 'process', '(', 'nlp', ')', 'subfield',
'linguist', ',', 'comput', 'scienc', ',', 'artifici', 'intellig', 'concern',
'interact', 'comput', 'human', 'languag', ',', 'particular', 'program',
'comput', 'process', 'analyz', 'larg', 'amount', 'natur', 'languag', 'data',
'.', 'challeng', 'natur', 'languag', 'process', 'frequent', 'involv', 'speech',
'recognit', ',', 'natur', 'languag', 'understand', ',', 'natur', 'languag',
'gener', '.']
Lemmatized Tokens: ['Natural', 'language', 'processing', '(', 'NLP', ')',
'subfield', 'linguistics', ',', 'computer', 'science', ',', 'artificial',
'intelligence', 'concerned', 'interaction', 'computer', 'human', 'language',
',', 'particular', 'program', 'computer', 'process', 'analyze', 'large',
'amount', 'natural', 'language', 'data', '.', 'Challenges', 'natural',
'language', 'processing', 'frequently', 'involve', 'speech', 'recognition', ',',
'natural', 'language', 'understanding', ',', 'natural', 'language',
'generation', '.']

2. Term Frequency (TF):
<FreqDist with 32 samples and 48 outcomes>

3. Inverse Document Frequency (IDF):
```

```
{'amounts': 1.0, 'analyze': 1.0, 'and': 1.0, 'artificial': 1.0, 'between': 1.0,
'challenges': 1.0, 'computer': 1.0, 'computers': 1.0, 'concerned': 1.0, 'data':
1.0, 'frequently': 1.0, 'generation': 1.0, 'how': 1.0, 'human': 1.0, 'in': 1.0,
'intelligence': 1.0, 'interactions': 1.0, 'involve': 1.0, 'is': 1.0, 'language':
1.0, 'large': 1.0, 'linguistics': 1.0, 'natural': 1.0, 'nlp': 1.0, 'of': 1.0,
'particular': 1.0, 'process': 1.0, 'processing': 1.0, 'program': 1.0,
'recognition': 1.0, 'science': 1.0, 'speech': 1.0, 'subfield': 1.0, 'the': 1.0,
'to': 1.0, 'understanding': 1.0, 'with': 1.0}
```

[ ]: