Scikit-Learn Tutorial

Natural Language Processing(NLP)

eliminating punctuation and stopwords

In [ ]:
```
import string
from nltk.corpus import stopwords
```

In [ ]:
```
# View first 10 stopwords
stopwords.words('english')[0:10]
```

In [ ]:
```
# Create a test sentence
test_sentence = 'This is my first string. Wow! we are doing just fine'
```

In [ ]:
```
# Eliminate the punctuation
no_punctuation = [c for c in test_sentence if c not in string.punctuation]
```

In [ ]:
```
no_punctuation = ''.join(no_punctuation)
no_punctuation
```

In [ ]:
```
# eliminate stopwords
clean_sentence = [word for word in no_punctuation.split() if word.lower() not
clean_sentence
```

bag of words

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

- transform text data into numerical data
- It counts how many times each word appears in each document.
- The output is a matrix where rows are documents and columns are words.

In [ ]:
```
vectorizer = CountVectorizer()
```

In [ ]:
```
# create 3 documents
doc1 = "This is first document"
doc2 = "This is second document"
doc3 = "This is third document"
listofdocument=[doc1,doc2,doc3]
```

In [ ]:
```
#fit
bag_of_words = vectorizer.fit(listofdocument)
```

In [ ]:
```
# apply transform method
bag_of_words = vectorizer.transform(listofdocument)
print(bag_of_words)
```

In [ ]:
```

```