Practical 7: Text Analytics Cheatsheet

Theory & Concepts

1. Tokenization: Splitting text into words or sentences.
   Example: "He runs fast" -> ["He", "runs", "fast"]

2. POS Tagging: Assigning part-of-speech tags to words.
   Example: pos_tag(["He","runs","fast"]) -> [("He","PRP"),("runs","VBZ"),("fast","RB")]

3. Stop Words Removal: Removing common words that add little meaning.
   Example: ["this","is","a","test"] -> ["test"]

4. Stemming: Reducing words to root form (may be non-dictionary).
   Example: PorterStemmer().stem("running") -> "run"

5. Lemmatization: Reducing words to dictionary form.
   Example: WordNetLemmatizer().lemmatize("better","a") -> "good"

6. TF-IDF: Term Frequency × Inverse Document Frequency.
   TF = term count / total terms; IDF = log(N / df)

Code with Comments

```
import numpy as np
import pandas as pd
import nltk, string
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk import pos_tag
from sklearn.feature_extraction.text import TfidfVectorizer

nltk.download('stopwords')
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('wordnet')

text = \"\"\"Text analytics is the process of deriving insights from text data.
It involves Tokenization, POS Tagging, Stop Words Removal, Stemming, Lemmatization.\""\"

# Sentence Tokenization
sentence = sent_tokenize(text)

# Word Tokenization
words = word_tokenize(text)

# POS Tagging
```

```
pos_tags = pos_tag(words)

# Stop Words Removal
stop_words = set(stopwords.words('english'))
filtered = [w for w in words if w.lower() not in stop_words and w not in string.punctuation]

# Stemming
stemmer = PorterStemmer()
stemmed = [stemmer.stem(w) for w in filtered]

# Lemmatization
lemmatizer = WordNetLemmatizer()
lemmatized = [lemmatizer.lemmatize(w) for w in filtered]

# TF-IDF
vectorizer = TfidfVectorizer()
tfidf = vectorizer.fit_transform([' '.join(sentence)])
df_tfidf = pd.DataFrame(tfidf.toarray(), columns=vectorizer.get_feature_names_out())
print(df_tfidf)
```