In [1]:

```python
# Once, download the webtext corpus and import libraries
import nltk
from nltk import word_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report

nltk.download('reuters')
nltk.download('punkt_tab')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package reuters to
[nltk_data]     C:\Users\stefa\AppData\Roaming\nltk_data...
[nltk_data]   Package reuters is already up-to-date!
[nltk_data] Downloading package punkt_tab to
[nltk_data]     C:\Users\stefa\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\stefa\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[1]:

```
True
```

In [2]:

```python
# For a simple multi-class problem, we filter to only those documents that have a single category
def get_single_label_docs(fileids):
    docs = []
    labels = []
    for fid in fileids:
        cats = nltk.corpus.reuters.categories(fid)
        if len(cats) == 1:  # Only use documents with one category
            docs.append(nltk.corpus.reuters.raw(fid))
            labels.append(cats[0])
    return docs, labels

# Get Reuters file IDs and split into training and test sets
train_ids = [doc_id for doc_id in nltk.corpus.reuters.fileids() if doc_id.startswith("training")]
test_ids = [doc_id for doc_id in nltk.corpus.reuters.fileids() if doc_id.startswith("test")]

# Get the documents and labels for the training and test sets
train_docs, train_labels = get_single_label_docs(train_ids)
test_docs, test_labels = get_single_label_docs(test_ids)
```

In [3]:

```python
# Preprocess the documents with various techniques
def preprocess(docs):
    stemmer = PorterStemmer()
    lemmatizer = WordNetLemmatizer()

    stemmed_texts = []
    lemmatized_texts = []
    stemmed_lemmatized_texts = []

    for text in docs:
        tokens = word_tokenize(text)
        stemmed_text = " ".join([stemmer.stem(token) for token in tokens])
        lemmatized_text = " ".join([lemmatizer.lemmatize(token) for token in tokens])
        stemmed_lemmatized_text = " ".join([lemmatizer.lemmatize(stemmer.stem(token)) for token in tokens])
        stemmed_texts.append(stemmed_text)
        lemmatized_texts.append(lemmatized_text)
        stemmed_lemmatized_texts.append(stemmed_lemmatized_text)

    return stemmed_texts, lemmatized_texts, stemmed_lemmatized_texts

# Preprocess the training and test documents
train_stemmed, train_lemmatized, train_stemmed_lemmatized = preprocess(train_docs)
test_stemmed, test_lemmatized, test_stemmed_lemmatized = preprocess(test_docs)

# Save different versions, so we can compare them later
with open("train_docs.txt", "w") as f:
    f.write(train_docs[0])
with open("train_stemmed.txt", "w") as f:
    f.write(train_stemmed[0])
with open("train_lemmatized.txt", "w") as f:
    f.write(train_lemmatized[0])
with open("train_stemmed_lemmatized.txt", "w") as f:
    f.write(train_stemmed_lemmatized[0])
```

In [4]:

```python
def train_and_compare(train_docs, train_labels, test_docs, test_labels):
    if len(train_docs) == 0 or len(test_docs) == 0:
        print("No documents to classify. Exiting.")
        return

    if len(train_docs) != len(train_labels):
        print("Number of training documents does not match number of training labels. Exiting.")
        print(len(train_docs), len(train_labels))
        return

    if len(test_docs) != len(test_labels):
        print("Number of test documents does not match number of test labels. Exiting.")
        print(len(test_docs), len(test_labels))
        return
    print(f"Number of training documents: {len(train_docs)}")
    print(f"Number of test documents: {len(test_docs)}")

    # Convert the text data to TF-IDF features
    vectorizer = TfidfVectorizer(stop_words='english', max_df=0.5)
    X_train = vectorizer.fit_transform(train_docs)
    X_test = vectorizer.transform(test_docs)

    # Train a Multinomial Naive Bayes classifier
    clf = MultinomialNB()
    clf.fit(X_train, train_labels)

    # Predict on the test set
    predictions = clf.predict(X_test)

    # Evaluate the classifier
    accuracy = accuracy_score(test_labels, predictions)
    print(f"Accuracy: {accuracy:.4f}")
    print("Classification Report:")
    print(classification_report(test_labels, predictions))
```

In [5]:

```
# Train and compare the stemmed and lemmatized versions of the documents
print("Original:")
train_and_compare(train_docs, train_labels, test_docs, test_labels)
print("Stemmed:")
train_and_compare(train_stemmed, train_labels, test_stemmed, test_labels)
print("\nLemmatized:")
train_and_compare(train_lemmatized, train_labels, test_lemmatized, test_labels)
print("\nStemmed and Lemmatized:")
train_and_compare(train_stemmed_lemmatized, train_labels, test_stemmed_lemmatized, test_labels)
```

```
Original:
Number of training documents: 6577
Number of test documents: 2583
Accuracy: 0.7151
Classification Report:
                 precision    recall  f1-score   support

            acq       0.68      0.94      0.79       696
           alum       0.00      0.00      0.00        19
            bop       0.00      0.00      0.00         9
        carcass       0.00      0.00      0.00         5
          cocoa       0.00      0.00      0.00        15
        coconut       0.00      0.00      0.00         1
         coffee       1.00      0.27      0.43        22
         copper       0.00      0.00      0.00        13
         cotton       0.00      0.00      0.00         9
            cpi       0.00      0.00      0.00        17
            cpu       0.00      0.00      0.00         1
          crude       0.96      0.42      0.59       121
            dlr       0.00      0.00      0.00         3
           earn       0.72      0.99      0.83      1083
           fuel       0.00      0.00      0.00         7
            gas       0.00      0.00      0.00         8
            gnp       0.00      0.00      0.00        15
           gold       0.00      0.00      0.00        20
          grain       0.00      0.00      0.00        10
      groundnut       0.00      0.00      0.00         2
           heat       0.00      0.00      0.00         4
            hog       0.00      0.00      0.00         1
        housing       0.00      0.00      0.00         2
         income       0.00      0.00      0.00         4
    instal-debt       0.00      0.00      0.00         1
       interest       1.00      0.04      0.07        81
            ipi       0.00      0.00      0.00        11
     iron-steel       0.00      0.00      0.00        12
            jet       0.00      0.00      0.00         1
           jobs       0.00      0.00      0.00        12
           lead       0.00      0.00      0.00         4
            lei       0.00      0.00      0.00         3
      livestock       0.00      0.00      0.00         6
         lumber       0.00      0.00      0.00         4
      meal-feed       0.00      0.00      0.00         1
       money-fx       0.67      0.18      0.29        87
   money-supply       0.00      0.00      0.00        28
        naphtha       0.00      0.00      0.00         1
        nat-gas       0.00      0.00      0.00        12
         nickel       0.00      0.00      0.00         1
         orange       0.00      0.00      0.00         9
       pet-chem       0.00      0.00      0.00         6
       platinum       0.00      0.00      0.00         2
         potato       0.00      0.00      0.00         3
        propane       0.00      0.00      0.00         1
       reserves       0.00      0.00      0.00        12
         retail       0.00      0.00      0.00         1
           rice       0.00      0.00      0.00         1
         rubber       0.00      0.00      0.00         9
           ship       0.00      0.00      0.00        36
strategic-metal       0.00      0.00      0.00         6
          sugar       1.00      0.12      0.21        25
            tea       0.00      0.00      0.00         3
            tin       0.00      0.00      0.00        10
          trade       0.86      0.57      0.68        76
        veg-oil       0.00      0.00      0.00        11
            wpi       0.00      0.00      0.00         9
            yen       0.00      0.00      0.00         6
           zinc       0.00      0.00      0.00         5

       accuracy                           0.72      2583
      macro avg       0.12      0.06      0.07      2583
   weighted avg       0.63      0.72      0.63      2583

Stemmed:
Number of training documents: 6577
Number of test documents: 2583
```

```
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

```
Accuracy: 0.7116
Classification Report:
                 precision    recall  f1-score   support

            acq       0.65      0.95      0.77       696
           alum       0.00      0.00      0.00        19
            bop       0.00      0.00      0.00         9
        carcass       0.00      0.00      0.00         5
          cocoa       0.00      0.00      0.00        15
        coconut       0.00      0.00      0.00         1
         coffee       1.00      0.14      0.24        22
         copper       0.00      0.00      0.00        13
         cotton       0.00      0.00      0.00         9
            cpi       0.00      0.00      0.00        17
            cpu       0.00      0.00      0.00         1
          crude       0.98      0.45      0.62       121
            dlr       0.00      0.00      0.00         3
           earn       0.74      0.99      0.84      1083
           fuel       0.00      0.00      0.00         7
            gas       0.00      0.00      0.00         8
            gnp       0.00      0.00      0.00        15
           gold       0.00      0.00      0.00        20
          grain       0.00      0.00      0.00        10
      groundnut       0.00      0.00      0.00         2
           heat       0.00      0.00      0.00         4
            hog       0.00      0.00      0.00         1
        housing       0.00      0.00      0.00         2
         income       0.00      0.00      0.00         4
    instal-debt       0.00      0.00      0.00         1
       interest       1.00      0.02      0.05        81
            ipi       0.00      0.00      0.00        11
     iron-steel       0.00      0.00      0.00        12
            jet       0.00      0.00      0.00         1
           jobs       0.00      0.00      0.00        12
           lead       0.00      0.00      0.00         4
            lei       0.00      0.00      0.00         3
      livestock       0.00      0.00      0.00         6
         lumber       0.00      0.00      0.00         4
      meal-feed       0.00      0.00      0.00         1
       money-fx       0.72      0.15      0.25        87
   money-supply       0.00      0.00      0.00        28
        naphtha       0.00      0.00      0.00         1
        nat-gas       0.00      0.00      0.00        12
         nickel       0.00      0.00      0.00         1
         orange       0.00      0.00      0.00         9
       pet-chem       0.00      0.00      0.00         6
       platinum       0.00      0.00      0.00         2
         potato       0.00      0.00      0.00         3
        propane       0.00      0.00      0.00         1
       reserves       0.00      0.00      0.00        12
         retail       0.00      0.00      0.00         1
           rice       0.00      0.00      0.00         1
         rubber       0.00      0.00      0.00         9
           ship       0.00      0.00      0.00        36
strategic-metal       0.00      0.00      0.00         6
          sugar       1.00      0.12      0.21        25
            tea       0.00      0.00      0.00         3
            tin       0.00      0.00      0.00        10
          trade       0.87      0.43      0.58        76
        veg-oil       0.00      0.00      0.00        11
            wpi       0.00      0.00      0.00         9
            yen       0.00      0.00      0.00         6
           zinc       0.00      0.00      0.00         5

       accuracy                           0.71      2583
      macro avg       0.12      0.06      0.06      2583
   weighted avg       0.63      0.71      0.62      2583


Lemmatized:
Number of training documents: 6577
Number of test documents: 2583
```

```
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

```
Accuracy: 0.7127
Classification Report:
                 precision    recall  f1-score   support

            acq       0.68      0.94      0.79       696
           alum       0.00      0.00      0.00        19
            bop       0.00      0.00      0.00         9
        carcass       0.00      0.00      0.00         5
          cocoa       0.00      0.00      0.00        15
        coconut       0.00      0.00      0.00         1
         coffee       1.00      0.23      0.37        22
         copper       0.00      0.00      0.00        13
         cotton       0.00      0.00      0.00         9
            cpi       0.00      0.00      0.00        17
            cpu       0.00      0.00      0.00         1
          crude       0.98      0.41      0.58       121
            dlr       0.00      0.00      0.00         3
           earn       0.72      0.99      0.83      1083
           fuel       0.00      0.00      0.00         7
            gas       0.00      0.00      0.00         8
            gnp       0.00      0.00      0.00        15
           gold       0.00      0.00      0.00        20
          grain       0.00      0.00      0.00        10
      groundnut       0.00      0.00      0.00         2
           heat       0.00      0.00      0.00         4
            hog       0.00      0.00      0.00         1
        housing       0.00      0.00      0.00         2
         income       0.00      0.00      0.00         4
    instal-debt       0.00      0.00      0.00         1
       interest       1.00      0.02      0.05        81
            ipi       0.00      0.00      0.00        11
     iron-steel       0.00      0.00      0.00        12
            jet       0.00      0.00      0.00         1
           jobs       0.00      0.00      0.00        12
           lead       0.00      0.00      0.00         4
            lei       0.00      0.00      0.00         3
      livestock       0.00      0.00      0.00         6
         lumber       0.00      0.00      0.00         4
      meal-feed       0.00      0.00      0.00         1
       money-fx       0.67      0.16      0.26        87
   money-supply       0.00      0.00      0.00        28
        naphtha       0.00      0.00      0.00         1
        nat-gas       0.00      0.00      0.00        12
         nickel       0.00      0.00      0.00         1
         orange       0.00      0.00      0.00         9
       pet-chem       0.00      0.00      0.00         6
       platinum       0.00      0.00      0.00         2
         potato       0.00      0.00      0.00         3
        propane       0.00      0.00      0.00         1
       reserves       0.00      0.00      0.00        12
         retail       0.00      0.00      0.00         1
           rice       0.00      0.00      0.00         1
         rubber       0.00      0.00      0.00         9
           ship       0.00      0.00      0.00        36
strategic-metal       0.00      0.00      0.00         6
          sugar       1.00      0.16      0.28        25
            tea       0.00      0.00      0.00         3
            tin       0.00      0.00      0.00        10
          trade       0.89      0.53      0.66        76
        veg-oil       0.00      0.00      0.00        11
            wpi       0.00      0.00      0.00         9
            yen       0.00      0.00      0.00         6
           zinc       0.00      0.00      0.00         5

       accuracy                           0.71      2583
      macro avg       0.12      0.06      0.06      2583
   weighted avg       0.63      0.71      0.62      2583


Stemmed and Lemmatized:
Number of training documents: 6577
Number of test documents: 2583
```

```
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

```
Accuracy: 0.7081
Classification Report:
                 precision    recall  f1-score   support

            acq       0.65      0.94      0.77       696
           alum       0.00      0.00      0.00        19
            bop       0.00      0.00      0.00         9
        carcass       0.00      0.00      0.00         5
          cocoa       0.00      0.00      0.00        15
        coconut       0.00      0.00      0.00         1
         coffee       1.00      0.14      0.24        22
         copper       0.00      0.00      0.00        13
         cotton       0.00      0.00      0.00         9
            cpi       0.00      0.00      0.00        17
            cpu       0.00      0.00      0.00         1
          crude       0.98      0.44      0.61       121
            dlr       0.00      0.00      0.00         3
           earn       0.73      0.99      0.84      1083
           fuel       0.00      0.00      0.00         7
            gas       0.00      0.00      0.00         8
            gnp       0.00      0.00      0.00        15
           gold       0.00      0.00      0.00        20
          grain       0.00      0.00      0.00        10
      groundnut       0.00      0.00      0.00         2
           heat       0.00      0.00      0.00         4
            hog       0.00      0.00      0.00         1
        housing       0.00      0.00      0.00         2
         income       0.00      0.00      0.00         4
    instal-debt       0.00      0.00      0.00         1
       interest       1.00      0.02      0.05        81
            ipi       0.00      0.00      0.00        11
     iron-steel       0.00      0.00      0.00        12
            jet       0.00      0.00      0.00         1
           jobs       0.00      0.00      0.00        12
           lead       0.00      0.00      0.00         4
            lei       0.00      0.00      0.00         3
      livestock       0.00      0.00      0.00         6
         lumber       0.00      0.00      0.00         4
      meal-feed       0.00      0.00      0.00         1
       money-fx       0.71      0.14      0.23        87
   money-supply       0.00      0.00      0.00        28
        naphtha       0.00      0.00      0.00         1
        nat-gas       0.00      0.00      0.00        12
         nickel       0.00      0.00      0.00         1
         orange       0.00      0.00      0.00         9
       pet-chem       0.00      0.00      0.00         6
       platinum       0.00      0.00      0.00         2
         potato       0.00      0.00      0.00         3
        propane       0.00      0.00      0.00         1
       reserves       0.00      0.00      0.00        12
         retail       0.00      0.00      0.00         1
           rice       0.00      0.00      0.00         1
         rubber       0.00      0.00      0.00         9
           ship       0.00      0.00      0.00        36
strategic-metal       0.00      0.00      0.00         6
          sugar       1.00      0.08      0.15        25
            tea       0.00      0.00      0.00         3
            tin       0.00      0.00      0.00        10
          trade       0.87      0.43      0.58        76
        veg-oil       0.00      0.00      0.00        11
            wpi       0.00      0.00      0.00         9
            yen       0.00      0.00      0.00         6
           zinc       0.00      0.00      0.00         5

       accuracy                           0.71      2583
      macro avg       0.12      0.05      0.06      2583
   weighted avg       0.63      0.71      0.62      2583
```

```
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
c:\VScode Git\DS2\.venv\Lib\site-packages\sklearn\metrics\_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```