In [7]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import re

# Step 1: Load the Data
spam_data = pd.read_csv("spam-data.csv")
features = spam_data[['Number of Words', 'Number of Links', 'Number of Capitalized
labels = spam_data['Class']  # Ensure 'Class' is the column name for labels

# Step 2: Build and Train Logistic Regression Model
X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.2
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)
predictions = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, predictions))

# Step 3: Parse the `emails.txt` File and Extract Features
with open("emails.txt", 'r') as file:
    emails = file.read().split('----------------')

email_features = []
for email in emails:
    if email.strip():  # Check if the string is not just whitespace
        num_links = len(re.findall(r'http[s]?://\S+', email))
        num_words = len(re.findall(r'\w+', email))
        num_capitalized_words = len(re.findall(r'\b[A-Z]{2,}\b', email))
        num_spam_words = len(re.findall(r'\b(free|credit|offer|loan|winner|win|urge
        email_features.append([num_words, num_links, num_capitalized_words, num_spa

# Step 4: Check Emails for Spam
email_features_df = pd.DataFrame(email_features, columns=['Number of Words', 'Numbe
email_spam_predictions = model.predict(email_features_df)
for i, prediction in enumerate(email_spam_predictions):
    print(f"Email {i+1} is {'spam' if prediction else 'not spam'}")

# Step 5: Analyze the `spam-data.csv` File for Feature Importance
feature_importance = np.abs(model.coef_[0])
print("Feature Importance:\n", list(zip(features.columns, feature_importance)))
threshold = 0.1  # Arbitrary threshold for low importance
less_important_features = [features.columns[i] for i in range(len(feature_importanc
print("Less Important Features:", less_important_features)
```

```
Accuracy: 0.9310344827586207
Email 1 is spam
Email 2 is not spam
Email 3 is spam
Feature Importance:
 [('Number of Words', 0.06945489225574682), ('Number of Links', 1.0684307821133459),
('Number of Capitalized Words', 0.4759124370163708), ('Number of Spam Words', 1.2946
08569856821)]
Less Important Features: ['Number of Words']
```

Model Accuracy: The logistic regression model has an accuracy of approximately 93.1%. This is a fairly high accuracy rate, indicating that the model is performing well in distinguishing between spam and non-spam emails based on the features you've used.

Email Classification Results:

Email 1 is classified as spam. This could be due to high counts of typical spam indicators such as the number of links, spam words, or capitalized words. Email 2 is classified as not spam. This suggests that the characteristics of this email did not match those typically found in spam, according to your model's learning. Email 3 is classified as spam. Similar to Email 1, this might exhibit more of the typical spam features. Feature Importance:

Number of Links (1.0684) and Number of Spam Words (1.2946) are the most influential features in predicting spam. This suggests that emails with more links and typical "spammy" words (like "free," "winner," etc.) are more likely to be classified as spam. Number of Capitalized Words (0.4759) also plays a significant role but is less impactful compared to the number of links and spam words. It might indicate that emails with many capitalized words are more attention-grabbing and potentially suspicious, a common characteristic in spam emails. Number of Words (0.0695) is identified as a less important feature. This indicates that the sheer quantity of words in an email isn't a strong predictor of whether it's spam. Rather, the content and context (like links and specific words) are more critical. Less Important Features:

'Number of Words' being less important suggests that the length of the email does not significantly influence whether an email is considered spam by model. It's more about what the words are and how they're presented (links, capitalization, and specific spam-indicative words).