Classification Activity

Below are a set of activites to be completed next class. You should submit your assignment as a single file named `classification-activity.py`. You will be given the first ~30 minutes of class to complete and submit the assignment.

Task 1

Given the labels below. Calculate precision, recal, accuracy, and f_1 score.

```
actual_labels = ['spam', 'ham', 'spam', 'spam', 'spam',
               'ham', 'ham', 'spam', 'ham', 'spam',
               'spam', 'ham', 'ham', 'ham', 'spam',
               'ham', 'ham', 'spam', 'spam', 'ham']

predicted_labels = ['spam', 'spam', 'spam', 'ham', 'spam',
                    'spam', 'ham', 'ham', 'spam', 'spam',
                    'ham', 'ham', 'spam', 'ham', 'ham',
                    'ham', 'spam', 'ham', 'spam', 'spam']
```

Task 2

Create a sample corpus, by creating an array of sentences like we did in class. Using nltk and sklearn, create a function to do the following.

1. Create a bag of words for each sentence.
2. Create a bag of words using 3-grams.
3. Create a tfidf value for each 3-gram in the sentence.

Task 3

Review the sentiment analysis classifier creation code. Read each tweet using the TfidfVectorizer. Then, Use the sklearn MultinomialNB classifier to classify the sentiment of tweets as positive of negative. Use the `sad.thorn` file from the previous sentiment analysis assignment.

```python
from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB()
clf.fit(X, y) # Add the appropriate test information
clf.predict(Z) # Try to predict new tweets
```