First, launch the Enthought Canopy application on your computer. Then download the Jupyter Notebook for Exercise 3 from the course website. Open the Notebook in your web browser and work through it. As you work through the Notebook, answer the following questions.

**Task 1: Bag of Words**

In the matrix $X$ above, why are the first two rows identical?

In the matrix $X$ above, what does the 2 in the third row correspond to?

In the 2x17 matrix above, the first row contains three non-zero values (one 2 and two 1’s). What features do these three non-zero values correspond to?

What feature has the lowest weight? Why?
Task 2: Sentiment analysis of Twitter data

How many different features were extracted from the corpus?

Which classifier yielded the highest accuracy on this dataset and what is this classifier’s accuracy?

When creating the Perceptron classifier, it was given an argument of `max_iter=10`. What does this correspond to?

When using 2-grams, how many different features were extracted from the corpus?

When using 2-grams, which classifier yielded the highest accuracy and what is this classifier’s accuracy?
Task 3: To spam or not to spam: that is the question

How many different features were extracted from the corpus?

Which classifier yielded the highest accuracy on this dataset and what is this classifier’s accuracy?

What are the five words with the lowest weight? What are the five words with the highest weight?

Do the three classifiers classify the message as ham or spam? Do the results for any of the classifiers change when executed multiple times on this message?
In the **TIME** column, please estimate the time you spent on this exercise. Please try to be as accurate as possible; this information will help us to design future exercises.

### PART | TIME | SCORE
--- | --- | ---
Exercise |  |  

If you or your partner is using a late coupon, please indicate who is using the coupon and how many coupons.
Late coupons: ____________________________