Working with Text Lab Sheet

1. Work through the `10-Working with Text`, `10-Text Classification and 10-Feature Selection on Text` notebooks.
2. Choose three different categories from the 20 NewsGroup dataset (used in the `10-Text Classification` notebook) and compare Naïve Bayes and SVM classifiers. Show results, which algorithm works best?
3. Change the feature set to include word bigrams – does this improve performance?
4. Include stop words in the pre-processing of the data. Does this improve performance?
5. Reduce the dimensionality of the data by using Document Frequency reduction. What reduction in dimensionality can happen without changing the performance of the model?
6. Use filter feature selection to see what performance you get for a reduced feature set of 500 features. Does this improve or disimprove performance?