Practical Natural Language Processing
Instructor: Rasika Bhalerao
Assignment 6
Due November 2 (at the beginning of class at 2pm)

Please submit your solutions using NYU Classes.

For this assignment, we will use the same dataset as from Assignment 2, and the assignment can be turned in the same way (a PDF with code snippets / output and English text explanations). Read the same dataset into a Pandas dataframe as before. It is also recommended to convert the documents to lowercase and filter out stopwords.

None of this will be "by hand." Please use Scikit-learn.
The "(One sentence)" suggestions are suggestions. You will be graded based on evidence that you did the tasks described and understood the intentions behind the questions.

1. Start with k = 10 topics. Fit an LDA object to the set of all news text. Then, examine the top *n* words from each topic (choose a reasonable *n* such as 10 or 20). How well do the topics represent real-world topics? (One sentence)

2. Randomly select 5 real news examples and 5 fake news examples, and examine the topic distributions for each document. Which topics are prevalent in the real news documents? (One sentence) Which topics are prevalent in the fake news documents? (One sentence)

3. Use the LDA vectors for the documents as features in a Linear Logistic Regression classifier to predict whether each document is real news or fake news. According to the resulting coefficients from the regression, which topics are most useful in determining whether or not something is real news or fake news? (One sentence)

4. Pick real news or fake news, whichever is more interesting to you. Then, use the LDA vectors for those news documents to cluster them. You can use KMeans clustering with a reasonable value for K (if you don't have strong feelings for a particular K, I recommend 10).* Then, select 5 news documents from each resulting cluster. Do the clusters correspond to anything? (One sentence)

* If you don't like KMeans, you can use a different clustering method.