Homework 3

We will be implementing different text processing techniques and feature vector models and sending them through the Naïve Bayes classifier. This will allow us to identify which combination of text processing and feature vectors provides the “best” results for our classification task.

The dataset we are using for this homework is a collection of articles from the Huffington Post in CSV format. The original dataset had 42 categories, which I have narrowed down to the 5 most popular categories to reduce the size of the dataset. The dataset has the following attributes, where the starred attributes are the ones relevant to our task:

* \* category: category in which the article was published
* \* headline: the headline of the news article
* authors: list of authors who contributed to the article.
* link: link to the original news article.
* \* short\_description: Abstract of the news article
* date: publication date of the article.

The *category* attribute defines the labels (y) we are trying to predict.Headline and short\_description are the text data we will use to train our model. I have provided the code to get the y and X data since there were some complications with this specific dataset.

# Part 1

Set up 10-fold cross-validation experiments using Naïve Bayes for each of the following combinations of stemming/lemmatization, and vector encoding:

1. Bag-of-Words vectorization after applying
   1. No stemming/lemmatization
   2. PorterStemmer
   3. SnowballStemmer
   4. WordnetLemmatizer
2. TF-IDF vectorization after applying
3. No stemming/lemmatization
4. PorterStemmer
5. SnowballStemmer
6. WordnetLemmatizer
7. LSA reduction with TF-IDF vectorization after applying
   1. No stemming/lemmatization
   2. PorterStemmer
   3. SnowballStemmer
   4. WordnetLemmatizer

This means you are running 12 different 10-fold cross-validation experiments. **This will take a while, that is normal!!**

For each of the experiments, gather the accuracy and micro- and macro-averaged F1-scores.

# Part 2

Answer the following questions based on your results:

1. Look at the category data
   1. What are the five categories we are classifying?
   2. How many instances of each class are there?
   3. What type of dataset are we dealing with based on these numbers?
2. Look at the results for vectorization
   1. Which of the 3 vectorization methods performs “best”?
   2. Why do you think this method performs so well in conjunction with Naïve Bayes? Think about the assumptions being made by the Naïve Bayes model as well as the distribution of classes.
   3. Explain why it makes sense that the other two vectorization methods have worse performance. Hint: for each of the vectorization models we discussed potential pitfalls in class, and we still want to bear the Naïve Bayes assumptions in mind.
3. Look at the results for stemming and lemmatization
   1. Is there an improvement when applying stemming and lemmatization as compared to not pre-processing the data? Which method had the highest improvement?
   2. Explain why you think this is the case.