Artificial Intelligence (CS 571)

Assignment-6: News Headlines classification using Decision Tree

(Read all the instructions carefully & adhere to them.)

Date: 11-10-2018 **Deadline: 29-10-2018**

1. **Problem Statement:** Given the headline of a news, the objective is to identify the category of news. (Note: Use only headline as input and predict category)
   For example:
   Short_description: ....
   Headline: There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV
   Date : **....**
   Link : **....**
   Authors: **....**
   Category: CRIME

   Consider the following categories only: Business, Comedy, Sports, Crime, Religion

2. **Dataset:** **news-category-dataset.json**

3. **Classification Algorithm:**
   Decision tree: Install Weka data mining package http://www.cs. waikato.ac.nz/ml/weka/(See WekaManual-3-7-7.pdf for more details) and use Decision Tree algorithm (C4.5 / J48).

4. **Features:**
   a. **Lexical Features:** Word n-gram**.**
   b. **Syntactic Features:** Parts of speech tag unigrams.

   Implement n-gram (n=1,2 and 3) features for each question instance. You may choose only the most frequent n-gram to provide as a features for your model. For n=1, use 500 most frequent 1-gram, similarly use 300 and 200 most frequent n-gram,for n=2 and 3 respectively.

   For parts of speech tag unigrams, first you need to get POS tag for each question instance. Use can use any library like Stanford POS tagger see https://nlp.stanford.edu/software/tagger.shtml, NLTK POS tagger see http://www.nltk.org/book/ch05.html etc. Similar to lexical feature use use 500 most frequent 1-gram to build the model.

5. **Result and Evaluation:** Perform 10-fold cross-validation and report the performance of the classification model (Decision Tree) using the following formula.

   Accuracy = # of correct predictions / Total no. of predictions

Instructions:

1. You can use any programming language for the implementation. But Python should be the preferred choice.
2. Please submit all the relevant documents including code, input files and a brief description about the code in a compressed file.