Report for Hakerrank task:
Craigslist Post Classifier: Identify the Category
https://www.hackerrank.com/challenges/craigslist-post-classifier-the-category

by Wojciech Fabian

Final score: 76.95 / 100

My task was to predict the category of Craigslist post given 3 types variables for each observation:
- city (string)
- section (string)
- heading (string)

Heading is just some text, so I used Naive Bayes classifier, just as recommended in the problem description.

As I wanted it to be a short task, I treated 'city' and 'section' as just additional words for each observation. Probably they might have been of a better use if given the higher priority during classification, but I wanted to keep the task short.

Tools used:
- Python 2.7 with scikit-learn library (for Machine Learning)

As seen in the code:
1. I read the training data from the file to local variables (category and the text separately)
2. Then I extracted features: first creating the bag-of-words, preprocessing, filtering stop-words, and tokenising text. Afterwords I used a function to calculate frequencies: tfidf, so "Term Frequency times Inverse Document Frequency",
3. Training classifier, using Naive Bayes classifier for multinomial models.
4. Reading input data (I commented out the code for reading the external file in the directory, which I used on my machine for testing)
5. Transforming input data (same way as during feature extraction) and predicting categories
6. Printing the output

What I could have improved if having more time:
- stemming (I actually tried SnowballStemmer, but it was very time-consuming and did not improve the final score)
- regular expressions to remove form the headings gibberish such as "♚"
- maybe closer analysis of the data would give me some additional ideas