Natural Language Processing:

Assignment 4: What are you asking me?

Jordan Boyd-Graber

Out: **2. September 2014**
Due: **3. October 2014**

Introduction

As always, check out the Github repository with the course homework templates:

`git://github.com/ezubaric/cl1-hw.git`

The code for this homework is in the `hw4` directory.

The aim of this assignment is to do text classification on trivia questions, sorting them into their appropriate category. We'll be using the Naïve Bayes classifier provided by NLTK.

Unlike previous assignments, the code provided with this assignment has all of the functionality required. Your job is to make the functionality better by improving the features the code uses for text classification.

NOTE: Because the goal of this assignment is feature engineering, not classification algorithms, you *may not* change the underlying algorithm.

About the Data

First, visit the Kaggle site and download the two csv files with the training and test data and place them in your `hw4` directory.

Quiz bowl is an academic competition between schools in English-speaking countries; hundreds of teams compete in dozens of tournaments each year. Quiz bowl is different from Jeopardy, a recent application area. While Jeopardy also uses signaling devices, these are only usable after a question is completed (interrupting Jeopardy's questions would make for bad television). Thus, Jeopardy is rapacious classification followed by a race—among those who know the answer—to punch a button first.

Here's an example of a quiz bowl question:

> Expanding on a 1908 paper by Smoluchowski, he derived a formula for the intensity of scattered light in media fluctuating

densities that reduces to Rayleigh's law for ideal gases in The Theory of the Opalescence of Homogenous Fluids and Liquid Mixtures near the Critical State. That research supported his theories of matter first developed when he calculated the diffusion constant in terms of fundamental parameters of the particles of a gas undergoing Brownian Motion. In that same year, 1905, he also published On a Heuristic Point of View Concerning the Production and Transformation of Light. That explication of the photoelectric effect won him 1921 Nobel in Physics. For ten points, name this German physicist best known for his theory of Relativity.

ANSWER: **Albert Einstein**

Two teams listen to the same question. Teams interrupt the question at any point by "buzzing in"; if the answer is correct, the team gets points and the next question is read. Otherwise, the team loses points and the other team can answer.

Classifying Category

There are many kinds of questions asked in these tournaments: science (as above), literature, history, etc. The goal of this project is to create an automated system that predicts the category of a question as accurately as possible.

These data will be the subject of the final project (you'll help to answer the questions), so this will be a useful warmup to help you get to know these data a little bit better.

Submission

In addition to turning in your code on Moodle, you'll also need to submit your predictions on Kaggle, an online tournament site for machine learning competitions.

In addition, please turn in a file called `explanation.txt` explaining your process of creating additional features. Make sure you state your username there.

Your username should be of the form `CU_Firstname.Lastname` so that we can easily map it to your grade.

How this Assignment is Graded (35+ points)

You'll get full credit on this assignment (35 points) if you can significantly improve on the baseline system (as reported by the Kaggle system). If you can do much better than your peers, you can earn extra credit (up to 15 points).

Questions / Hints

- Don't use all the data until you're ready. Use the `--subsample` option to use a subset of the data to see how you're doing on smaller datasets.