Project Proposal for General Assembly's Data Science Class

Noel Carrascal, noelcarrascal@gmial.com, https://github.com/noelcjr/SF_DAT_17_WORK.git

Projects

Google books Spanish N-grams analysis and forensic linguistics N-grams are fixed size tuples of items that are made of words extracted from the Google Books corpus. The n specifies the number of elements in the tuple, so a 5-gram contains five words or characters. The original data set is available from http://books.google.com/ngrams/. Not long ago I heard about a book written anonymously by Harry Potter's author J.K. Rowling under the pen name Robert Galbraith. Forensic linguistics uncovered Rowling as the author. I want to download the Spanish n-grams and use it to do blindfold tests of two writers from analysis of their writings. I will enter labeled text written by each writer and try to identify unlabeled text using machine learning algorithm. I will present probability results on authorship, statistics on Spanish N-grams and statistics on authors language use from N-grams.

Country rankings from 2014's CIA world fact book. This is a database downloaded from http://jmatchparser.sourceforge.net/factbook/. I have deleted and simplified tables in order to focus only on country rankings by 76 fields (i.e. birth and death rates, unemployment, labor force, GDP, debt, internet hosts and users, etc). First, I want to present rankings in an easily digestible way. For example, the ranking of countries' surface areas can be understood in a simpler way by a map of the world. Then, I want to scale the surface areas of countries proportional to the other rankings. For example, Russia is the largest country by surface area, but not the top country by GDP; thus, in a map of the world by GDP, Russia would be scaled down, and richer countries with smaller surface area would be scaled up. This makes it easier to see how countries compare visually for all other rankings in the CIA fact book. Second, I will analyze the spectrum of possible development indexes that can be calculated from the 76 ranking fields. This is important to realize how the GINI coefficient, for measurement income inequality, would compare to all possible indexes from CIA fact book rankings. Third, the CIA rankings are accurate and can be used to analyze peoples' perceptions of facts about countries. I want to do a simple experiment to find out how peoples' common knowledge of rankings compare to actual rankings. Using a web server hosted on my laptop, I want to ask 10 questions to all participants in the class over the LAN. It should take 5 minutes or less. I will show them different, random pairs of countries and ask which country the person thinks is largest in surface area. Using these answers, we can compare results to the actual rankings done by the CIA. How would the map of the world by surface area look according to people's knowledge? How close do they get? After how many questions can we get close to, or reproduce, the actual ranking? Having a measure of people's aggregate knowledge by ranking of countries would be useful as a reference point to asses the accuracy of other rankings for which there is no known answer. It would be like mining minds with a visualization tool based on the scaling of the countries surface areas to condense ranking information.