The Internet and Current Issues

James Prow and Thomas Johnson

Summary of Research Questions

1) Which issue between literacy rate, gross domestic product, and average life expectancy does information technology affect the most?

    -We will compute the mean-squared-error between a histogram representing the number of internet users per total population of a country and the histograms of each of the following issues: literacy rates, GDP rates, and average life expectancy rates. This will determine which issue most accurately approximates the internet users per total population ratio.

Motivation and Background

In terms of human interaction, the internet has changed everything. We can contact people on other continents and people in the same room at comparable speeds. Overlooking the cultures, lifestyles and challenges of every nation on earth, we can describe every country's value by its GDP figure and store each one in a text file. We can search databases filled with names of people who live halfway across the world, find their information on search engines, and have third parties compile page-long profiles summing up their lives. Unfortunately, wonderful though these luxuries are for those privileged enough to live in a climate of prosperity, there are still people starving, dying of disease and desperate to escape the shackles of tyranny. The information era has certainly made the world smaller, but has it made it a better place to live?

As partners, this difficult question prompted considerable thought. We struggled with defining "the world" and quantifying "better". After a bit of brainstorming, our sights turned to the resources necessary to investigate. It turns out that the same technological advances that birthed our question also birthed the means of answering it. The sheer size of publicly available online datasets is staggering, and by applying some computational knowledge we can attach meaning to what were previously raw numbers. We narrowed "the world" down to its human inhabitants. What we called "better" was a positive change in quality of life over time. Perhaps most importantly, we gathered resources in the form of applicable data.

IBM hosts a web services called Many Eyes that, aside from producing beautiful graphics visualizing world events, compiles and publicly distributes some of their underlying datasets. Of the thousands of sets available, one that caught our eyes reported the number of internet users in every country for the past five years. This, proportional to total population, became

our yardstick for technological investment for a given country. These values can be compared to more typical statistics such as GDP, average life expectancy, and literacy rate to reveal information that could answer our question regarding how the digital age has affected the welfare of human beings in general. We predict that higher internet usage per capita will correlate positively with greater quality of life.

Knowing what impact technology has on society can make the best of said impact. For example, if developing nations have focused on computing and internet access, social groups could respond by providing more aid through these channels. Just as the low-overhead nature of online commerce revolutionized business, so too could it make big changes in welfare. Being able to distribute aid, education and resources electronically is likely less expensive than the alternative, potentially resulting in a more economical and thus further-reaching platform for worldwide improvement of the human condition. Of course, gathering more information will help people help others more effectively and we are happy to do some number crunching for the common man.

Datasets

The program will make use of 5 different data sets, each of which provide different pieces of information that will be used in various computations. The data sets are as follows:

1) Internet World Use
    a) URL: http://www-958.ibm.com/software/analytics/manyeyes/datasets/internet-world-use/versions/1
    b) Source: The World Bank
    c) Description: Gives the number of internet users per country in each of the years between 2007-2011.

2) Total Population of all Countries Between 2008-2011
    a) URL: http://www-958.ibm.com/software/analytics/manyeyes/datasets/total-population-of-all-countries-/versions/1
    b) Source: The World Bank
    c) Description: Gives the total population in each country in each of the years between 2008-2011.

3) Life Expectancy by Years
    a) URL: http://www-958.ibm.com/software/analytics/manyeyes/datasets/life-expectancy-by-years-4/versions/1
    b) Source: CIA World Factbook

c) Description: Gives the average life expectancy of people in each country as of December, 2012.

4) Gross Domestic Product (purchasing power parity) per Country
   a) URL: http://www-958.ibm.com/software/analytics/manyeyes/datasets/gdp-purchasing-power-parity/versions/1
   b) Source: CIA World Factbook
   c) Description: Gives the Gross Domestic Product (purchasing power parity) of each country as of February 2013.

5) Adult Literacy Rate (aged 15 and over)
   a) URL: http://www-958.ibm.com/software/analytics/manyeyes/datasets/adult-literacy-rate-aged-15-and-ov/versions/2
   b) Source: Guardian Data Store
   c) Description: Gives each country and that country's literacy rate in 2009.

Methodology

Combine the data set containing internet users per country with the data set containing each country's total population to obtain an *internet users per person* measure. Notice that the data set containing total populations of each country and the data set containing internet users per country contain data from the years 2008-2011 and 2007-2011 respectively. Be sure to use only the data from 2011 in both data sets. Also notice that the data sets may include different sets of countries. Only include the countries common in both data sets. Then normalize the ratios so that they add up to one.

Compare each of the data sets containing literacy rates, GDP rates, and average life expectancy rates respectively to the internet per person measure. Prior to comparison, the ratings corresponding to these data sets should be normalized so that they sum to one. Make sure that all four measures use the same set of countries.

Compute the mean squared error between the internet per person measure and the other 3 measures to determine which issue (life expectancy, literacy, or GDP) internet usage affects the most. The issue that corresponds to the smallest mean squared error with the internet users per person measure will be the issue most affected by internet usage.

Given the issue which the internet has the greatest impact, determine the top 10 countries with the highest internet per person ratio and plot those countries' internet per

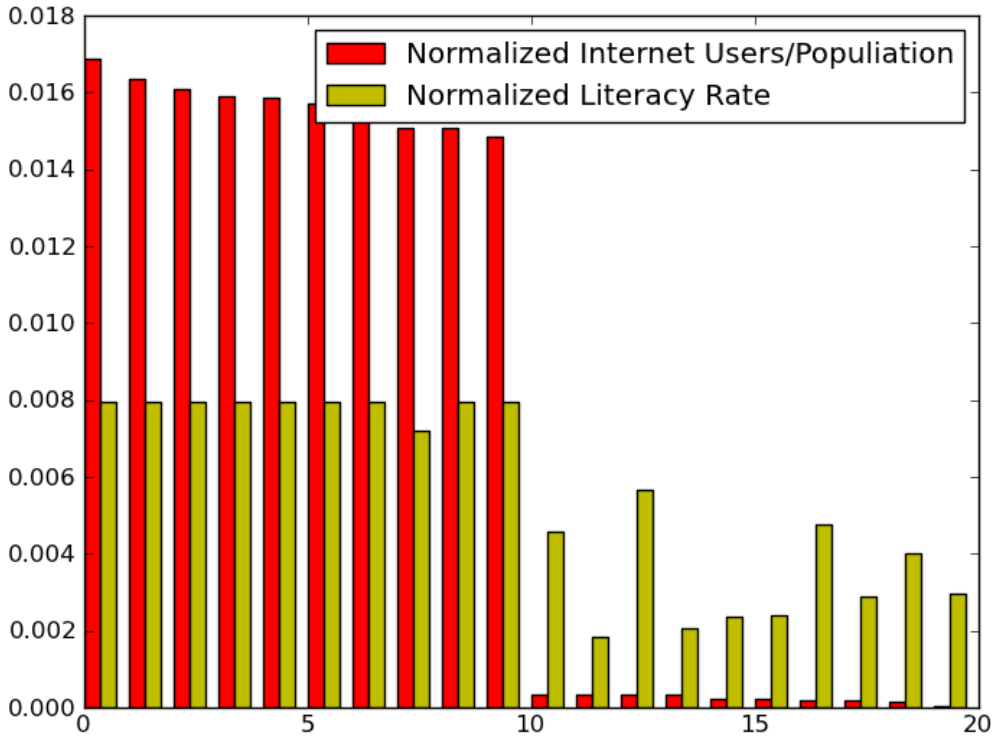
person ratios next to their corresponding ratings in the issue that internet technology affects the most. Then do the same with the 10 countries with the lowest internet per person ratio.

Results

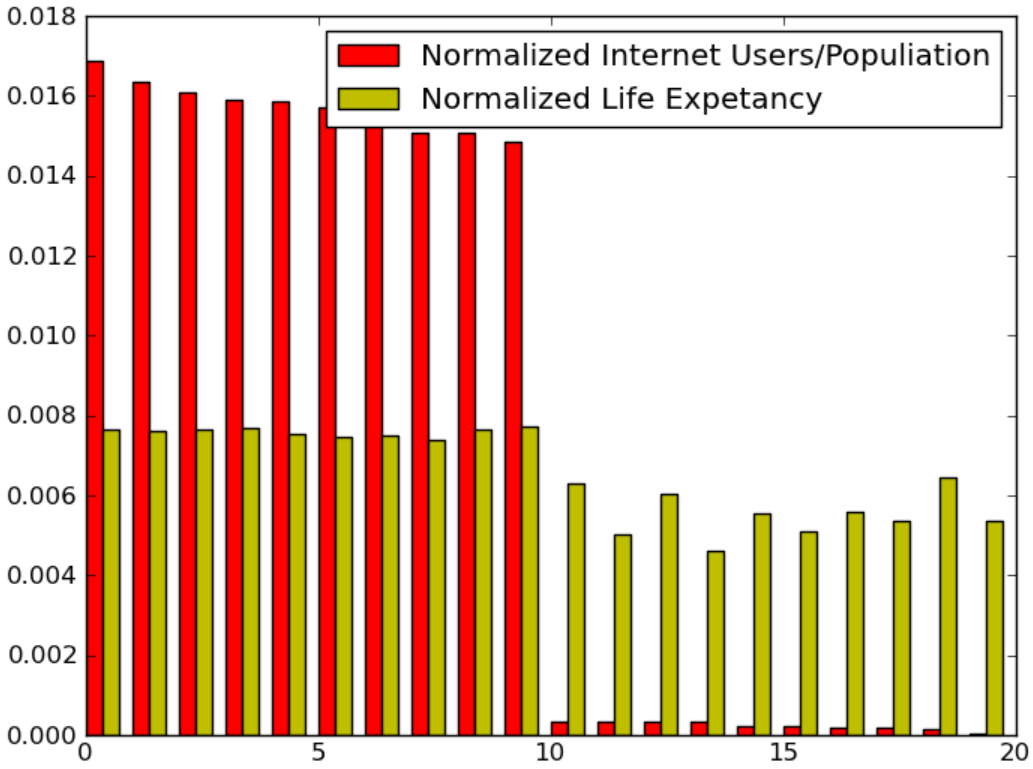
The mean squared errors are as follows: 0.0777 between internet users per population against GDP, 0.0026 between internet users per population against literacy rate, and 0.0029 between internet users per population against life expectancy rate. The latter two figures caught our attention as they were relatively low, potentially indicating some sort of relationship. However, knowing that MSEs aren't the most robust correlational figures, we calculated Pearson's r values for the same data. The Pearson's r values are as follows: 0.1875 between internet users per population against GDP, 0.6812 between internet users per population against literacy rate, and 0.7747 between internet users per population against life expectancy rate. At this point, we were suspicious of a strong correlation between internet users per population and both literacy rate and life expectancy.

We decided to further examine these relationships with more visually stimulating representations of the figures. Using pyplot, we made bar graphs for both of the relationships of interest. These graphs include comparisons between the two issues of interest for 20 countries, the 10 with the highest internet users per population (left) and the 10 with lowest (right). These appear below:

Internet Users/Population and Literacy Rate (Normalized):



Internet Users/Population and Life Expectancy (Normalized)



After analyzing these two graphs, we were not satisfied with the second one. Normalized life expectancy values seemed to display a ceiling effect. In other words, all of the life expectancy values are too high to be meaningful in this statistical situation. However, the literacy rate graph showed a marked difference between the normalized literacy values on the left and right. Therefore, we concluded that there is indeed a correlation between internet use rate and literacy rate.

It is important to note that we cannot draw causal conclusions from a correlational figure. Without more sophisticated analysis, we cannot be certain that an increase in a country's internet usage rate causes an increase in literacy rate, vice versa, or that there are other variables causing both of these effects. However, this phenomenon is certainly worth further investigation. Being able to manipulate and understand it could be very valuable for developing nations and the aid groups who work to help their populations thrive.

Reproduction:

Navigate to the containing folder “johnson_prow_finalproject” via OSX terminal/Windows command line and run the file “internet_effects.py” with Python. The console will output the MSEs and Pearson’s r values. Graphs will be displayed in separate windows, and the x-axis ticks will be displayed as an ordered list in the console (country names were too long and did not fit on the graph). The graphs will also be created in the containing folder when the program is run for future access.

Collaboration

We did not collaborated with anyone other than each other.

Reflection

We learned that coming up with a good data set can be a tough and tedious process. We would advise other students working on a research project to start looking for data early.