Assignment 2: Extracting and Linking Attributes

All of your work must be done in and submitted through GitHub. (note: you can't store the data files on GitHub as they will be too large, but put any scripts and documents in git).

Done in groups of 3 to 5 (Please send an announcement, if you need a group member or can't find a group). I will expect more of larger groups.

Preliminary Goals and Questions due Feb 15th at 11:59 (1 mark per day late)
Due on March 5th at 11:59pm (1 mark per day late)

Goals and Questions (maximum 1 page)

Due Feb 15th at 11:59 (1 mark per day late)

Imagine that you are the manager of the Google Chrome project. Write three questions that you would like to study to ensure that Chrome is running smoothly. For each of these questions do the following: describe how you will measure the **outcome** (eg quality, usability), describe 1 to 3 **direct measures** that will help you answer your question, describe 2 to 4 measures that will help you control for **confounding factors**, and describe the expected relationship between the outcome and direct measures as a **hypothesis**.

For example: **How does increased testing affect the number of bugs found in a file?**
As the manager of Chrome, it is important for us to understand if increased testing will lead to fewer defects seen by the customer. Our **outcome** measure of quality will be the number of of customer reported defects per file. Our **direct** measure will be the number of tests that fail per file. We will also include controlling measures such as the level of expertise of the developer and number of reviews per file to control for **confounds**. We suspect that higher expertise and more reviewing will confound testing making increased testing less important. Ultimately we will test the **hypothesis**: An increase in the number of tests will lead to a decrease in the number of customer reported defects per source file.

To get a sense of the possible problems facing Chrome and the available data, look at

- Chrome bug reports
    - https://code.google.com/p/chromium/issues/list

Extracting Attributes (8 marks)

From the three questions you suggest above, choose the most promising **one** and start extracting the attributes that will help you answer this question. Note: Don't wait to finalize your questions in part 1 before you start seeing if you can extract relevant attributes.

For each attribute you extract, in max two sentences, describe why you need it. Extract the attributes from

- Dump of Chrome bugs (.5 GB)
    - https://www.dropbox.com/sh/8a3l6gxxxdny4rk/AABUNAceyxu5_e8tH6UWgUB2a?dl=0
- Git repository (3.1GB)
    - uploading ...

Linking Attributes (8 marks)

Describe any relationships that exist among the attributes. For example, “In the commit log when a bug is referenced, such as ‘bug: 9874’, we are able to link the commit to the bug, which in turn allows us to link …” You can use any database that you’d like. I usually use postgresql, but sqlite might be easier.

1. Create a database schema indicating the attributes and relationships between them.
2. Import the attributes into the database
3. Create a summary table that groups the attributes. Tell use the scale for these attributes. For example, instead of having the name of each file a each developer has modified (categorical scale), we could count the number of files that each developer has modified (absolute scale).