

## Capstone 1 Data Wrangling

Bill Bocker

Oct. 10, 2017

The data base for this project will be a list of words scraped from Indeed want ads to determine what Web Development skills are the most required. The client is a web development school.

The following data cleaning steps are required:

1. Read the site html into a BeautifulSoup object.
2. Remove all scripts and style elements.
3. Get the text from the bs4 object.
4. Remove blank lines and ends-of-line.
5. Remove any Unicode characters.
6. Remove all non-alphabetic characters. (This may have to be modified to handle some skills such as Angular2.)
7. Set to lower case and split them into a list.
8. Remove stop-words.