CS483 Example

Steven Call

February 3, 2017

1 The reason for this program and subject.

The purpose of this program was to use python and familiarize ourselves with web scraping. We were allowed to use Wikipedia's API so that we would not get in trouble trying to take information we shouldn't be. Once this information and data was collected it was to be stored into a database so that we can easily use it later for our final project.

1.1 The reason for this topic.

The reason I decided to choose this topic was because I love rock and roll music especially older rock and was interested in older artists/groups and when they originated. I was not able to parse and collect all the information I wanted but am hoping that I can continue this program even after this assignment to more thoroughly gather information on these interesting artists for this genre of music.

1.2 What was used.

I created a program that took a specific title that I had given it and searched Wikipedia's API for that same title then accessed the content on that page. Then this program would find and save specific information from links on that page such as the year the artist/band started music, the year the artist was born, and the year the artist died. Then this data would be saved into an xml file for later use.

2 Issues occurred with my scrapping.

While working through this program I had multiple issues trying to capture the correct information and not collect the incorrect information. This was overcome by categorizing data that was not needed and the capture would find everything and collect it except what was specified as not needed.

2.1 Wikipedia formatting issues.

Another issue that I came across was trying to capture both dates of years active. Unlike the birth date and death date, the years active was only held by one tag. I tried to capture the first date a dash (-) then the second date but this caused issues as well because lots of band had a date-

present. Since I could not find a way to properly capture this different cases I decided that the information I would collect would just be the start date of when the band or artist started making rock and roll music.

2.2 Etree used for xml.

Since our professor went over etree in class and gave us sample code that started our understanding of etree I used it to contain my xml files that I created. This seemed like the most logical choice due to the fact that I have never used MySQL or SQLight and have not been taught how to use these things in our databases class yet.