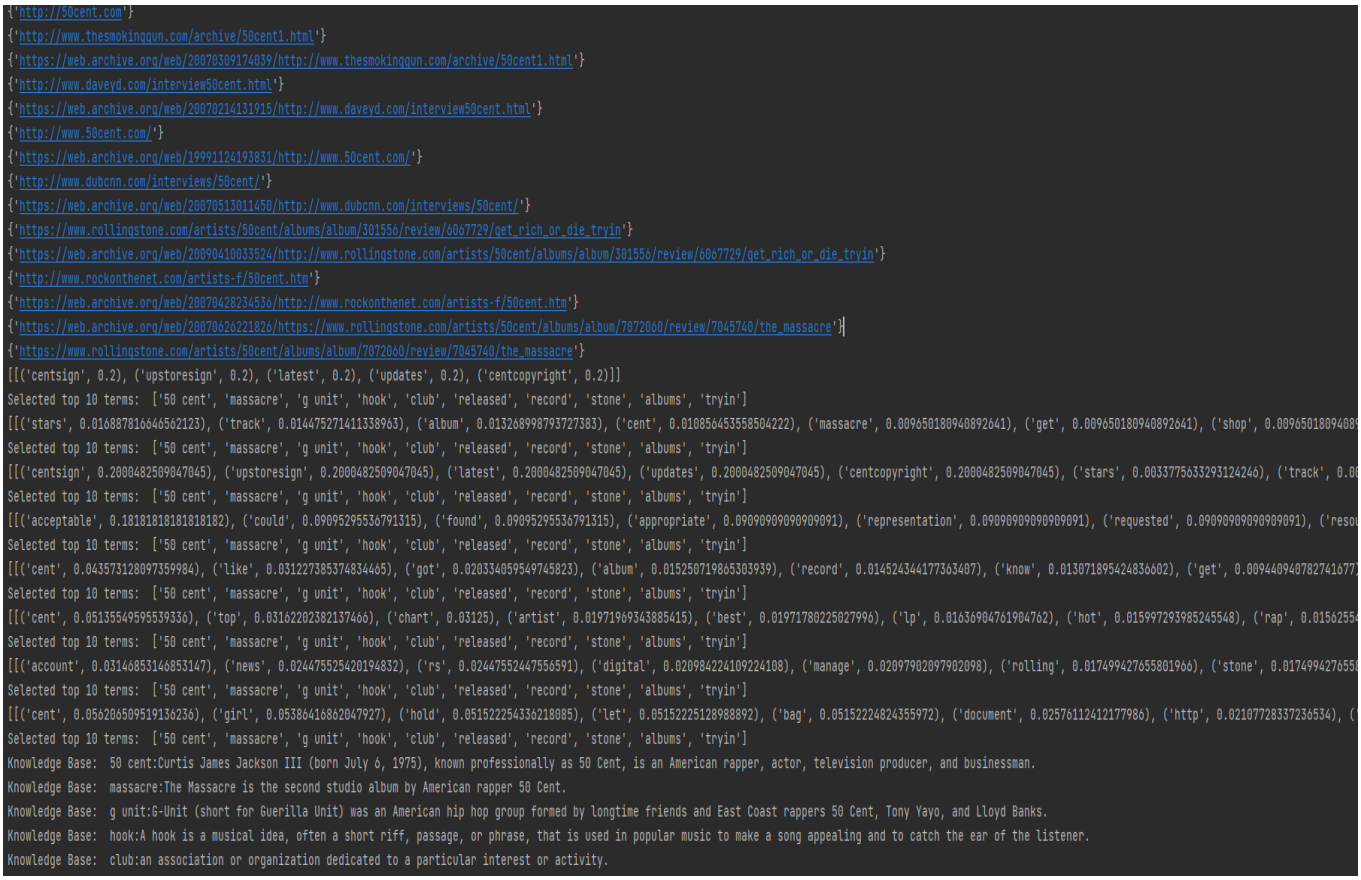
Alan Perez
AXP200075
CS4395.001

Portfolio: Finding or Building a Corpus

For the knowledge base, I extracted the top terms from each individual site and selected the most appropriate ones for the artist's album/music. I manually looked up the definition of each term and any relevant information regarding the artist/celebrity. I specifically looked at the top frequent terms and created a hardcoded dictionary for the knowledge base. My initial approach was to do it by automatically extracting the definition from the existing information within the URLs but ran into issues trying to implement it this way. Considering the small amount of data, I believe manual entry was the best option since it allows more control over the content.

I made the mistake of choosing a topic that didn't have enough data or rich text for the corpus, had I gone with a topic that was a lot broader then I'd have had sufficient data



Top 10 terms: '50 cent', 'massacre', 'g unit', 'hook', 'club', 'released', 'record', 'stone', 'albums', 'tryin'

Example of Chatbot Dialog:

User: Can you finish the rest of this lyric? "Hate it or love it underdogs on top:
Chatbot: "And I'm gon' shine, homie, until my heart stop"

Alan Perez
AXP200075
CS4395.001

User: What is 50 cent's best selling album?
Chatbot: "Get Rich or Die Tryin"
User: What group was 50 cent a part of in the 2000s?
Chatbot: g unit