CSE 411: Advanced Programming Techniques
Fall 2017

Programming Assignment 4 : Simple Text Analysis
Due: Friday, Dec 15th, 11:59pm.

Introduction

The idea behind this assignment is to write a simple program in Scala that does some text processing.

Procedure

1. Create a folder for an sbt project. In the build.sbt file, set the Scala version to 2.12.3. Put the source code for the program in the src/main/scala folder.
2. In this project we'll be analyzing the top40 database of pop songs posted on Coursite.
3. Read in the top40.sql file posted on Coursesite into a list of Strings. Feel free to edit the file first to remove the DDL code at the top, leaving just the insert lines.
4. Run through the list of insert statements and map it into a list of triples (Tuple3) that hold the actual data values, 1) Artist name, 2) Song Title (split into a list of words with punctuation removed), 3) Number one flag.
5. Perform a series of transformations (map, etc) and actions (sum, count, etc) that allow you to calculate some statistics (at least the following, but feel free to calculate other interesting values):
   a) The most frequent non trivial words (not *the*, *and*, etc.) in the song titles. Make up your own list of trivial words to test against.
   b) A list of the artists with the most 1) top 40 songs 2) number one songs
   c) Assuming P(A|B) = the probability of A given B. Calculate:
      - P(a song being number one| the title contains a particular frequent word)
      - P(the title contains a particular frequent word | a song being number one)
6. Display the results to standard out.