CS483 Project Proposal Template

Team One

Clara, Alberto - alberto.clara@wsu.edu
AlSarhi, Saeed - saeed.alsarhi@wsu.edu
Wellington, Patrick - patrick.wellington@wsu.edu

10/16/2018

1 Introduction

Our project will be creating a search engine using Whoosh over the scraped data set from the web page IMDB. The data set would contain a large set of movies to search from. We will index the queries on a large CSV file containing the scraped data. The engine will allow search on all attributes of the data set. For example: search movies based on there names, year of release, genre, and description. Most of the fields are going to be of type TEXT that's where Whoosh is great at searching. One of the field in the data set is the 'Description' which contains free-form text that briefly explains the theme of the movie. The TEXT fields are also great to incorporate advanced searching features such as voting, stemming, stopping, n-grams, spelling correction etc. The year of release is a non-text attribute that can be used for searching by the year.

2 Database Summary

Movie(Name, Year_of_release, Description, Rating, Genre, IMDB_Url, Votes_Count)
The Data Set size is around 1500 tuples.
schema = Schema(
    Name = ID(stored=True),
    Year_of_release = TEXT(stored=True),
    Description = TEXT(stored=True),
    Rating = TEXT(stored=True),
    Genre = TEXT(stored=True),
    IMDB_URL = TEXT(stored=True),
    Votes_Count = TEXT(stored=True),
)

3 Indexing

The fields that the database will allow the user to search are name, year of release, rating, genre, and descriptions. The only field that will be stemmed or stopped is the description field, because everything else being indexed is a name and there is nothing to stem or stop in a name and it is important to preserve the exact information represented in a name. The stemmed and stopped plot data will be stored in an inverted index for use in TF-IDF, and the original data will be stored to be displayed in search results.

4 Features

4.1 Feature 1

Watching trailers is one of the best ways to find a movie you might enjoy. So we decided to implement the ability to redirect the user to the IMDB movie web page. The IMDB web page would contain the movie trailer and detailed info as well. This feature will help find all the info the user is looking for.

4.2 Feature 2

All movies in the database have a rating that shows how people enjoyed a movie. So we Decided to implement a feature that gives the user the choice to rate a movie on a scale of (0 to 10) based on how much they enjoyed watching it.

5 Member Responsibilities

Member responsibility
Final project proposal: everyone
Progress report: everyone
Search engine code developer:

- Web scraper - Saeed - Patrick
- Web interface - Alberto - Saeed
- Movie voting mechanism - Patrick
- Core Whoosh implementation - Alberto - Saeed
- Stemming and Stopping - Patrick

Testing: everyone
Slides: everyone
Final Report: everyone