CS483 Project Proposal Template

Team One

Clara, Alberto alberto.clara@wsu.edu AlSarhi, Saeed saeed.alsarhi@wsu.edu
Wellington, Patrick patrick.wellington@wsu.edu
Young, Justin justin.young@wsu.edu

10/10/2018

1 Introduction

Our project will be creating a search engine using Whoosh over the scraped data-set from Wikipedia website. The data-set is about sci-fi movies from 1920 to present. 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 films based on names of actors/actresses (if I type Harrison Ford, the engine should return all the movies Harrison Ford participated), or producers or writers, etc. 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 'plot' which contains free-form text up to 2000 characters. The TEXT fields are also great to incorporate advanced searching features such as ranking, stemming, n-grams, etc. There are two fields release data and running time that are non-text can be used for searching by date range and number values. As far as ranking algorithm goes, we will use TF-IDF for the plot search, it will return at most the top 10 tuples for each search.

2 Database Summary

Sci-Fi Movie(name, plot, director, producer, screenplay, writer, starring, music, cinematography, editing studio, distributor, released runtime, country, language)

Total of 1500 tuples.
schema = Schema(
name=ID(stored=True),
plot=TEXT(stored=True),
director=TEXT(stored=True),
producer=TEXT(stored=True),
screenplay=TEXT(stored=True),
writer=TEXT(stored=True),
starring=TEXT(stored=True),
music=TEXT(stored=True),
cinematography=TEXT(stored=True),
editing=TEXT(stored=True),
studio=TEXT(stored=True),
distributor=TEXT(stored=True),
released=DATETIME(stored=True, sortable=True),

```
    runtime=NUMERIC(int, 16, decimalplaces=0, signed=False, sortable=True, default=0),
    country=TEXT(stored=True),
    language=TEXT(stored=True)
)
```

3 Indexing

The fields that the database will allow the user to search are name, plot, director, producer, writer, and starring actors. The only field that will be stemmed or stopped is the plot 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

Suggestions based on the users search would be our first feature. So based on the users search we can suggest movies that have the same director, actors, or writer(s) as the top results.

4.2 Feature 2

Have the option to sort the search result per ratings. The result will first be unsorted but the user will have the option to turn this feature on/off. This way the user can get to choose a movie with a good rating based on his search.

5 Member Responsibilities

Member responsibility
Final project proposal: everyone
Progress report: Alberto, Saeed, Justin
Search engine code developer: everyone
Tester: everyone
Slides: Alberto, Saeed
Final Report: everyone
Final Project: everyone