Design Document for Celebrities Project

by Jiaao He, 2016011279

Introduction

Purpose

Write a network crawl and a web app to serve celebrity searching service

Scope

Programming training

System overview

A crawl is written to fetch data from *wikipedia*. A web app, after that, is programmed to provide searching service.

System Architectural Design

Architectural Design

This project is divided into three parts.

1. The crawl;
2. The web server which provides files and index;
3. The browser-side app which presents the information.

Decomposition Design

Crawl

NodeJS is used to write this part, as its outstanding async performance which is hundreds and thousands times better than that of python.

Raw data of *infobox* tables are stored into html files.

Web server

Django is used in the server to provide data using specific API URL interfaces.

BeautifulSoup soup is used to decomposite the tables and generate index dictionaries stored in *json* files which is loaded into the memory during the server is started.

Web app

AngularJS togather with *jQuery* is used as the frameset. *uiRouter* is used to provide smooth browsing performance. *Bootstrap* stylesheets are used to make the view acceptable.

Data Design

The raw data are stored in html files which contains a `table` element. All pages are given an ID due to its MD5 hash value.

Keywords in `th` labels are extracted, and the sentences in `td` labels are separated into words as keys.

Everytimes a searching requirement is submitted, the server will lookup all the *name* fileds in order to provide high accuracy searching in names. Then keywords are searched by indexes to ensure the speed of searching. Every found element are valued by its frequency of appearing. The value of a person is the sum of all his keywords.

Persons are sorted by the value on client side.

Human Interface Design

Overview

Angular is used.

Screen Images

Skipped.