COMP9321 Semester 2, 2017

Assignment 1: Data Curation

Aims and Background

This project aims to give students hands-on experience in designing and implementing a Web application on their own.

Big data analytics is firmly recognized as a strategic priority for modern enterprises. At the heart of big data analytics lies the data curation process, consists of tasks that transform raw data (unstructured, semi-structured and structured data sources) into curated data, i.e. contextualized data and knowledge that is maintained and made available for use by end-users and applications. To achieve this, the data curation process may involve techniques and algorithms for extracting, classifying, linking, merging, enriching, sampling, and the summarization of data and knowledge.

To facilitate the data curation process and enhance the productivity of researchers and developers, UNSW.SOC research group identify and implement a set of basic data curation APIs and make them available as services to researchers and developers to assist them in transforming their raw data into curated data. The curation APIs enable developers to easily add features - such as extracting keyword, part of speech, and named entities such as Persons, Locations, Organizations, Companies, Products, Diseases, Drugs, etc.; providing synonyms and stems for extracted information items leveraging lexical knowledge bases for the English language such as WordNet; linking extracted entities to external knowledge bases such as Google Knowledge Graph and Wikidata; discovering similarity among the extracted information items, such as calculating similarity between string and numbers; classifying, sorting and categorizing data into various types, forms or any other distinct class; and indexing structured and unstructured data - into their data applications.

The curation APIs are available as an open source project on GitHub. https://github.com/unsw-cse-soc/Data-curation-API

Download Technical Report (HERE) and Research Paper (HERE).

Requirements

1. Dataset

Governments Open Data helps guide business investment, foster innovation, improve employment opportunities, and spur economic growth. Governments Open data is

free, publicly available data that anyone can access and use without restrictions. Examples include USA (https://www.data.gov/) and Australia (http://data.gov.au/) Government Open Data.

In this assignment we will use the Oregon Newsroom XML file (Download) which follows a Simple schema. The full XML file has 2176 records where each record contains elements such as agency, headline, date, city and content.

2. The Homepage

The welcome page of your application, should display the list of records in the XML file in a table. By default, you will display a list (approx. 10 records) chosen at random. You will need to enable the user to choose the number of records (e.g. 10, 100, 1000 and All) to be displayed. You will also need to provide a **MENU** on the welcome page to redirect the user to the Search, Sitemap (a list of pages of a web site accessible to crawlers or users) and ContactUs pages.

3. Search

When the use click on the Search item in the MENU, this forwards to "search.jsp" which has a form with the following elements:

- text fields to search for the agency, headline, date, city and content etc.;
- a submit button labelled "Search";

Advanced Search: The user should be able to search for the Keywords and Named Entities (Poeple, Organizations and Locations) in the content of the elements.

4. Search Results

instructions:

1. The search functions forwards the users to the results page ("results.jsp"). The page has a list of entries that matched the search criteria, a submit button;
2. If the search has turned up empty, the results page must display "Sorry, no matching datasets found!";
3. Otherwise, the results page displays the *headline* for each entry.
   The *headline* is a hyperlink to a page that gives full information about the entry (e.g. agency, headline, date, city and content etc). This page contain 4 button: "Extract Keywords", "Extract People", "Extract Organizations" and "Extract Locations". When the use clicks on a button, list of extracted items (from the Content element) will be illustrated;
4. The results page should show only 10 results at a time. If there are more than 10 results then, at the bottom of the page are two navigation links Previous

and Next that allow the user to navigate the results, 1 page at a time. Ensure that the Previous and Next links are not shown on the first page and the last page of results respectively.

Assignment Execution

Submission Requirements

The due date for this assignment is (end of Week 6): **Sunday, Sep 3 2017, 23:59:00**. You **MUST** demonstrate the assignments in Week 7 during the lab times.

1. After testing, generate a war file from your project. In Eclipse, this is Right-Click on project name --> Export --> WAR file. Make sure that the "Export Sources" checkbox is checked.
2. email the WAR file to the following email address:

```
To: unsw.cse.comp9321@gmail.com
From: [Your unsw email address]
Subject: COMP9321-Ass1-S2_17
Body: [Your Full Name]
      [Your Student number]
Attachment: War-File
```

Important Notes:

1. This email address (unsw.cse.comp9321@gmail.com) is ONLY for submitting your assignemts. For other inquiries please contact your lecturer.
2. You do not need to email the XML file. The WAR file will be sufficient.
3. You can demonstrate your assignments using CSE lab computers as well as your own laptops.
4. The following additional libs are accepted: JSTL. And for front-end purposes, JQuery and Bootstrap. Additionally, any other Javascript libs that are simply referenced (i.e. without having to add to the build path) in your HTML/JSP page is accepted.

Evaluation and Marking

Your assignment will be evaluated by tutors independently (DOWNLOAD Assignment 1 Marking Scheme). We will test the functionality of the search application as well as the handling of errors such as invalid parameters and operations. This is an **individual** assignment worth **15 marks**. You will lose marks for any failed tests. Standard late penalties also applies.

Please use the ***message board*** for resolving other doubts.