CSE 535: Information retrieval

Project 3: Search systems using Apache Solr

Version 1.1

Table of Contents

1. Introduction

The purpose of this project is to build a full search system using Apache Solr to solve a real life problem. The aim is to have teams apply the different concepts learnt in the class to some interesting problems while learning how to setup, configure and use Apache Solr. A lot of the details in this project are left open ended to encourage teams to think out of the box and use their problem solving skills to arrive at answers.

The teams would have an option to choose amongst four projects listed here. The rest of the document lists the details for each project, submission and grading guidelines. All deliverables are due on **2nd December 2013 23:59 EST/EDT**. The proposal and design document is due by **10th November 2013 23:59 EST/EDT**.

2. Project descriptions

This section lists the different project specifications. Even though each specification is different from the other, there are some common deliverables that are listed below. Please read them carefully. Also refer to the resources section for pointers on how to download the dataset, setting up Solr etc.

Common requirements

All projects must satisfy the following common requirements:

- The project must be demo-able. It constitutes a large portion of your grade and we would spend 10-15 minutes per team per demo including questions.
- A UI must be provided. However, you will not be penalized for a simplistic UI. We will not allow the Solr UI to be repurposed as the UI for your system.
- Robust error handling: All errors must be handled gracefully and no ugly stack traces or exceptions must be displayed to end-users.

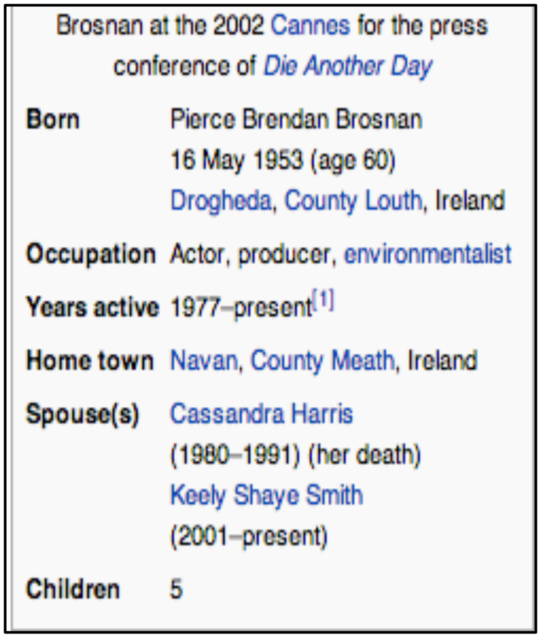
2.1. Question-Answering system

The aim of this project is to develop a QA system using English Wikipedia. We would limit our attention to Wikipedia infoboxes. However, teams are free to use the rest of the content from a page as they deem fit. The figure below shows the screenshot of an example infobox:

This could be used to answer questions about Pierce Brosnan like:

- Where
  - Was he born?
  - Is his hometown?
- What

    - Is his occupation?



Brosnan at the 2002 Cannes for the press conference of *Die Another Day*

Born	Pierce Brendan Brosnan 16 May 1953 (age 60) Drogheda, County Louth, Ireland
Occupation	Actor, producer, environmentalist
Years active	1977–present[1]
Home town	Navan, County Meath, Ireland
Spouse(s)	Cassandra Harris (1980–1991) (her death) Keely Shaye Smith (2001–present)
Children	5

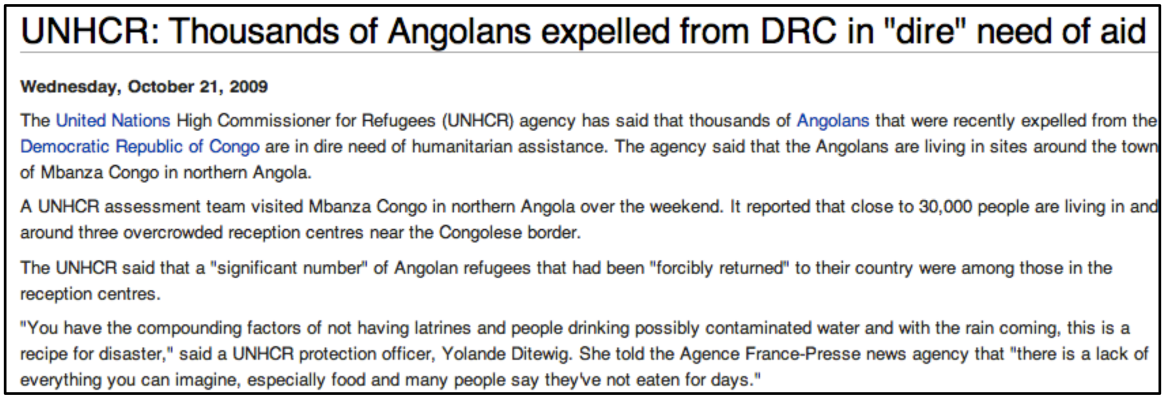
    - Is his full birth name?
- When
    - Was he born?
    - Did he remarry?
- Who
    - Is or was his first wife?
    - Is his second wife?

We leave it to the teams to decide what queries they support including if they want to limit the format supported etc. It is neither mandatory to support all queries that are listed above nor in the form as listed above. However, the following requirements apply in addition to the common requirements listed in the following section.

- You must index at least 5000 different infoboxes / pages.
- You must support queries on at least three of the following types:
    - People
    - Films / Songs / Albums
    - Places
    - Organizations / Clubs / Teams
    - Events / Matches
- The queries must be full and valid English sentences

2.2. Location based news search

The aim of this project is to allow users to search for news items subject to certain geographic restrictions. We would limit our attention to English wikinews. A sample news page is shown below:



UNHCR: Thousands of Angolans expelled from DRC in "dire" need of aid

Wednesday, October 21, 2009

The United Nations High Commissioner for Refugees (UNHCR) agency has said that thousands of Angolans that were recently expelled from the Democratic Republic of Congo are in dire need of humanitarian assistance. The agency said that the Angolans are living in sites around the town of Mbanza Congo in northern Angola.

A UNHCR assessment team visited Mbanza Congo in northern Angola over the weekend. It reported that close to 30,000 people are living in and around three overcrowded reception centres near the Congolese border.

The UNHCR said that a "significant number" of Angolan refugees that had been "forcibly returned" to their country were among those in the reception centres.

"You have the compounding factors of not having latrines and people drinking possibly contaminated water and with the rain coming, this is a recipe for disaster," said a UNHCR protection officer, Yolande Ditewig. She told the Agence France-Presse news agency that "there is a lack of everything you can imagine, especially food and many people say they've not eaten for days."

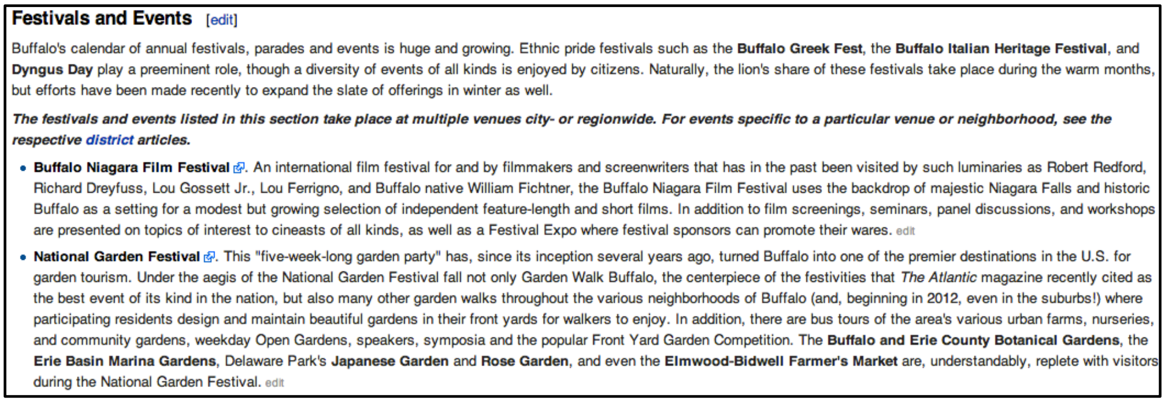
Note that the article simply lists a date and no publish location is specified. It is expected that teams figure out the referenced locations in a given text and handle them as such. You could use online or downloadable gazetteers to do this task.

The expectation is to support queries that allow searching for news items at and within a specified radius from a given location. For example, "scandals in UK" and "fires within 100 miles of Buffalo" for the first and second types. There is no restriction on how the system expects these to be worded as long as they are supported. This includes specifying a specialized query language if needed. Other requirements are as follows:

- You must index at least 5000 news articles.
- You must give more importance to recent news articles than older ones.

2.3. Travel destination search

The aim of this project is to suggest travel destinations to users based on their preferences. We would limit our attention to English wikivoyage for this purpose. A snapshot of a wikivoyage page is given below.



Festivals and Events [edit]

Buffalo's calendar of annual festivals, parades and events is huge and growing. Ethnic pride festivals such as the **Buffalo Greek Fest**, the **Buffalo Italian Heritage Festival**, and **Dyngus Day** play a preeminent role, though a diversity of events of all kinds is enjoyed by citizens. Naturally, the lion's share of these festivals take place during the warm months, but efforts have been made recently to expand the slate of offerings in winter as well.

The festivals and events listed in this section take place at multiple venues city- or regionwide. For events specific to a particular venue or neighborhood, see the respective district articles.

- **Buffalo Niagara Film Festival**. An international film festival for and by filmmakers and screenwriters that has in the past been visited by such luminaries as Robert Redford, Richard Dreyfuss, Lou Gossett Jr., Lou Ferrigno, and Buffalo native William Fichtner, the Buffalo Niagara Film Festival uses the backdrop of majestic Niagara Falls and historic Buffalo as a setting for a modest but growing selection of independent feature-length and short films. In addition to film screenings, seminars, panel discussions, and workshops are presented on topics of interest to cineasts of all kinds, as well as a Festival Expo where festival sponsors can promote their wares. edit
- **National Garden Festival**. This "five-week-long garden party" has, since its inception several years ago, turned Buffalo into one of the premier destinations in the U.S. for garden tourism. Under the aegis of the National Garden Festival fall not only Garden Walk Buffalo, the centerpiece of the festivities that *The Atlantic* magazine recently cited as the best event of its kind in the nation, but also many other garden walks throughout the various neighborhoods of Buffalo (and, beginning in 2012, even in the suburbs!) where participating residents design and maintain beautiful gardens in their front yards for walkers to enjoy. In addition, there are bus tours of the area's various urban farms, nurseries, and community gardens, weekday Open Gardens, speakers, symposia and the popular Front Yard Garden Competition. The **Buffalo and Erie County Botanical Gardens**, the **Erie Basin Marina Gardens**, Delaware Park's **Japanese Garden** and **Rose Garden**, and even the **Elmwood-Bidwell Farmer's Market** are, understandably, replete with visitors during the National Garden Festival. edit

Each page has similar subsections describing activities to do, stats about climate, getting there etc. The search system must be capable of using these as search parameters and giving results. There are no restrictions on how the results are to be ranked and are open to interpretation. The following requirements apply however:

- You must index at least 5000 different pages.
- You must allow the users to specify multiple parameters, at least as ANDed conditions.
- There must be a filter and/or drill down facility allowing users to apply / remove preferences or automatically discover new facets within their results to drill down on.

2.4. Multi-source search

The aim of this project is to present results from multiple sources for a given query. For this purpose, you could pick at least three from the following sources:

- English Wikipedia
- English Wikiquotes
- English Wikinews
- English Wikivoyage
- English Wikibooks

Each source must be maintained in a separate index and results combined from multiple sources when being displayed to the user. You are expected to create something like a portal page for a given query, showcasing different results from different sources. This also means ensuring that the different results are in fact somewhat related to each other. It is left to the users to determine how to present the results, group them, rank them etc. Some other requirements that apply:

- You must index at least 2000 pages per source, i.e., at least 6000 in total.
- You must provide an autosuggest option (like in Google search) as the user is typing in his query.
- Temporal ordering must be maintained where applicable, i.e, recent events or posts must appear higher.
- Results from different sources must not be merged while presenting.

3. Deliverables

We expect the following deliverables for this project per team. The following subsections provide more details about each deliverable as applicable.

- A short proposal and design document detailing high-level implementation details, mockups and work distribution.
- Code deliverables including solr configuration files and document processing programs etc.

- A final report expanded over the proposal document detailing final implementation details.
- A short demo of the working application.

For all documentation that you submit, each team member must sign it. Most recent Word and Adobe Reader versions support digital signatures by images or drawing them using styluses etc. This is only to avoid arguments between team members. If an unsigned document is submitted, we would assume all members consented.

3.1. Design document

This document is intended to give us an insight into how you are proposing to implement your system. We would be using this as a checkpoint to make sure you are on track and have most of your bases covered. It should include the following mandatory details:

- A high level description of your entire system. Include all components and illustrate how a given dump would be processed, your Solr system, query processing modules if any and any additional subsystems you may be developing that you feel are pertinent.
- Solr features being used. List the high level features with short descriptions of one or two paragraphs on how they would be used within your system and why they are needed.
- Work distribution: Details on which team member is responsible for which component(s).
- UI mockups: Rough design sketches or wire frames of how your UI would look. The final implementation may change and this is meant for illustration purposes only.

Command: **submit_cse535 <teamname>_proposal.pdf**

3.2. Code deliverables

These are meant to validate that the given system has indeed been implemented as demonstrated. We may try and reproduce the system at our end if we feel the need and hence, the expectations below must be followed accordingly. Please read them VERY CAREFULLY.

- Solr deliverables: solrconfig.xml and schema.xml. If you have multiple cores, submit files for each core.
- Processing executables: The code that you are using to convert the given xml file dumps into solr readable xml files. Depending on the programming language used, these could be executable jars, source code with make files etc.
- A readme file: Please provide a readme detailing instructions on how to build, compile, install or run your code as applicable.

3.3. Final report

This is the final representation of your work apart from the demo. This should be used to highlight any features or salient points within your implementation. You should also document any special tweaking that you did for Solr to setup your system. On the whole, the following are expected in the report:

- System diagram and description. You can reuse this as-is from the proposal.
- Features / USPs for your implementation
- Configuration details and tweaks. Document details on schema here, detail each field type and analyzers / filters used.
- UI screenshots showing sample usage.
- Solr stats. Include details and screenshots for the following from your production system. Make sure this is as latest as possible. Each of these corresponds to a specific page within the Solr admin app.
    - Index size
    - Stats on performance: Average response time etc.
    - Stats on caching: Total hits, queries, etc.
- Future work: Any improvements you may have done or would do if given more time.
- Member contributions: Final contributions by each team member.

Please package final deliverables as follows. Follow the given directory structure and naming conventions:

Root folder – compress this entire folder as **<teamname>_final.zip**

- Readme.txt: A readme with any instructions as needed
- code: Directory to hold all code deliverables
    - solr: Sub directory to hold all solr deliverables
        - Create subfolders for each core in case of multiple cores
    - processor: Bundled processing code. Please provide instructions in readme as applicable.
- report: Directory to hold the final report
    - To be named as **<teamname>_report.pdf**

Command: **submit_cse535 <teamname>_final.zip**

3.4. Demo

Each time would be allotted a slot to present a working demo of their system. This would be in the last week of classes and after all deliverables have been submitted. This contributes to a large portion of your grades.

4. Grading guidelines, tips and resources

The final grade would depend on ALL deliverables, thus not submitting one or more of these would affect your grades. However, the focus here is to reward teams for innovation and original thought. Some pointers on doing well on this project:

- Don't be afraid to be bold, wacky or out-of-the-box. As long as you try something new whilst satisfying the specified minimum requirements, you would be rewarded.
- Think of additional resources beyond those provided to you. Although we do not encourage crawling or scraping, you are free to use additional data, APIs etc. to make a better system.
- A lot in this project depends on planning and coordination. Plan in detail, set up integration points and raise issues as early as possible. We would not allow reorganization of teams and hence, the stress on clearly listing individual contributions. Spend enough time on choosing your teammates and most importantly communicate!
- Spend enough time in planning your demo as well as documenting. We would not give you a second chance beyond what you present and submit.

Finally, all Wikipedia related content can be downloaded as full dumps from here: http://dumps.wikimedia.org/backup-index.html

Refer to the Wikimedia websites on how to find the download you are looking for. Each site uses the same markup as you have used for Project 1. You are of course free to use your parsers from that project or use other open source software. A separate Solr guide would be published soon as an appendix to this document.