UoNSE Project - Proposal

Note: Please see links at the bottom of this page for Contact Details. This document outlines the motivation and goals for this project, as well as some forwarding points about choosing a role for yourself.

Contents:

- Introduction
- Motivation
- Aim
- Criteria for design
- /Current Design
- Technology to be Used
- Licensing
- What can you do?
- Links
  - Google Group:
  - Google Docs:
  - Contact Details:
  - Github:

1. Introduction

The UoNSE Project is a group project of around 30 individuals (at the time of this writing) who are interested in working on something exciting and different from what the current academic curriculum provides. Alexander Mendes 'commissioned' this project in response to a selection of students who wanted to gain more from their studies. The chosen area: 'text mining', fulfills both a need from the Bioinformatics Research Group, as well as being a fitting learning environment for interested students to attain new skill sets. Mendes also speculated that he can only provide support to either Java or C++ languages, which is important to note because it may influence the project development.

2. Motivation

Approximately eighty-percent of global data is held in unstructured form[1], which means that the importance of data mining is assured. The power found in being able to retrieve information from a given source and output it in a meaningful form is that it allows us to assert relationships between things of importance to us, sometimes this is obvious, but more often it is subtle. In order to access and understand data in unstructured formats we need a complex software framework that supports this function; therein lies the challenge presented to you by this proposal.

Several text mining frameworks exist at present, for example: Apache Hadoop, mongoDB and several other implementations of the MapReduce paradigm, these are scalable and targeted for large-data-centric-companies. The benefit in designing our own framework is that you will gain experience in software design, teamwork, embracing new technologies and ultimately getting a better understanding than simply calling one of the aforementioned frameworks.

The notion of "standing on the shoulders of giants" stems from the nature of the academic world, where everything that you learn builds on top of something much larger than just yourself. With this in mind, this project provides an opportunity to give back to the academic community that gave you knowledge, and notwithstanding that, the applications of this project

are extensive.

There are also many advantages conferred to your respective careers as Software Engineers | Computer Scientists | IT professionals, garnered from working in a large team environment and especially for the experience it adds (which will aid you in future group-projects, by learning the pitfalls and the strengths of group decision making). Additionally this is something that will look great on your resume.

[1] - Secondary quote from IBM, at 'Information on Demand' (2010).

3. Aim

The aim of this project is to design and build a software system capable of text mining a resource and returning possible relationships inferred by the resource.

4. Criteria for design

The system should have an interface for query input, which could for example consist of a target resource and a search term. Modularity is emphasised as the system needs to approach queries in a generic manner. Since the data being processed is 90% of the time unstructured the system must perform text mining analytics in order to generate inferred relationships. The output results must be displayed intuitively such as a weighted graph. We need to design modules correctly with predetermined communication protocols such that programming of each component can happen concurrently. To do this, the following need to be determined before coding starts: (a) function names and parameters already determined, (b) function return types already determined and (c) a consistent API.

5. Current Design

The primary focus at present is to sort everyone into one of these primary ‘activity’ groups (see Figure 1). Each individual has the discretion to choose a group that they are interested in.

Input Processing - Group 1

This group will have the responsibility of defining how input is used to collect ‘useful’ data from the web or other resource and silo it into a generic form (ie. known flexible form). They will need to regularly discuss with Group 2 to ensure consistency in design and to make goals meet.

Data Processing - Group 2

This group defines how to process the data within the generic data format (from Group 1) to find meaningful relationships and the output format, resulting from the text mining. In addition to having discussions with Group 1, they must also work closely with Group 3 so that the design remains consistent.

Presentation - Group 3

This group is responsible for displaying the output in a meaningful and expressive form. They will need to negotiate with group 2 on the design of the output format, which will affect how easy the output is to graph.

GUI - these will evolve later, since they are inconsequent to the design of the core-framework or can be changed independently.

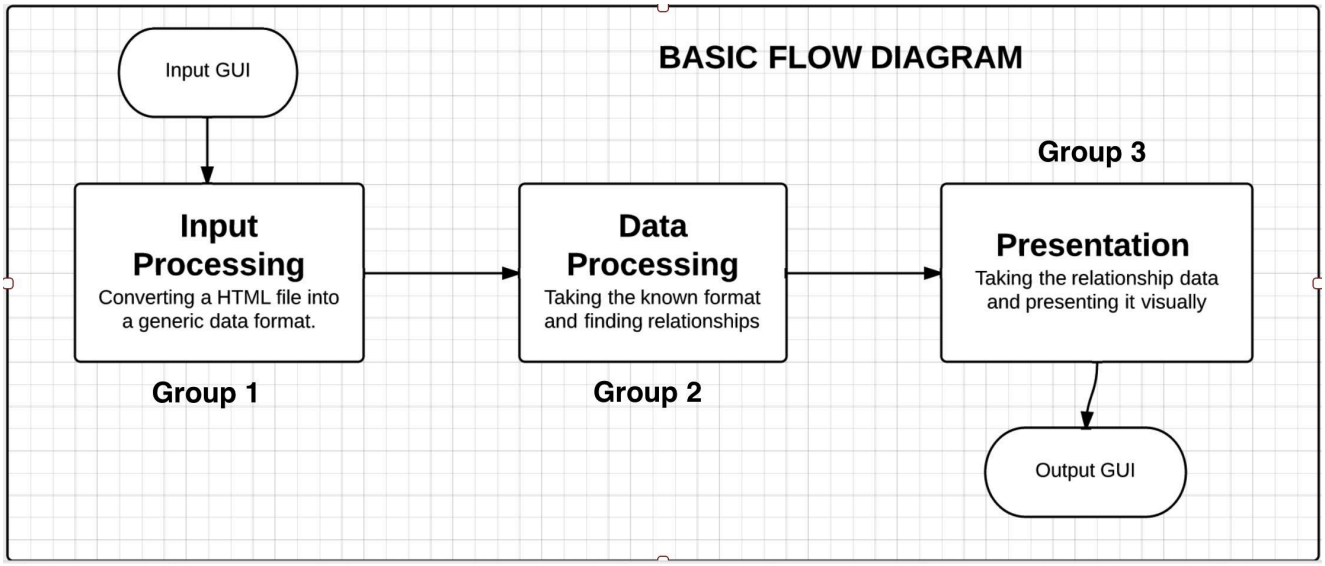

Figure 1 - Basic Flow Diagram. This shows the three initial groups that the project requires.

6. Technology to be Used

The language and other technologies have not currently been decided, though it is envisioned that both the Input-GUI and Output-GUI will be some kind of web-interface. The reason for this is simple: we do not want to limit our options and consequently any libraries or frameworks that are relevant until we have finalised UML diagrams for each respective section.

As mentioned earlier, Alex Mendes can only provide support to either Java or C++ languages. If taking Java as the base of our codebase, several libraries have been suggested, including Java Persistence API, Java NIO API, and Stanford NLP Library. On the other hand using C++ provides <regex> (C++11) which offers the same PERL-style regular expressions that are ubiquitously found in Java. Additionally, BOOST libraries contain BOOST::regex which provides more functionality.

Each activity-group will have different problems to solve and thus they may require different language-features to solve these. However it is important that consistency is maintained and that wrappers are used where different languages would meet. Ultimately the language requires much further discussion and indeed as the design specification changes during the evolution of the project, so to may this influence our choice of language features.

7. Licensing

Overwhelmingly, people within the group have thus far voted for an LGPL licensing scheme.

8. What can you do?

- You can start by reading the links embedded in this document as well as the post history of the UoNSE project facebook-page.
- Enter your details in the Contact Details document (link at bottom of this page) - note that Skype is preferred, but other contact details are presumably ok.
- Research into the components of our current design to learn more about the subproblems that we will need to develop solutions for - this should also inform your decision to choose an activity-group to (at least initially) join and be apart of.
- Participate in discussions, in all domains, eg. facebook, skype, in-person : with other people in the group - your questions/suggestions help keep this project fresh.
- Activity-groups will need to organise themselves regarding keeping in contact, however skype, facebook and email are likely to be useful. Additionally, members will be required to actively engage in independent research so that they may bring new information to the table.
- Learn something; have fun!

9. Links

Google Group:

https://groups.google.com/forum/#!forum/uonse

Google Docs:

https://drive.google.com/a/uonse.org/#folders/0BwMLgTqkvB_7b2hhSzgwWWpqRjQ

Contact Details (Follow the link and add your details):

https://docs.google.com/document/d/1Nz-gZ6KXkTyr0bGlZsdg66FW3JvxmNQJ06F83srWG-4/edit

Github:

https://github.com/UoNSE