UNIVERSITY COLLEGE LONDON

Department of Computer Science

British Library Big Data Experiment – Week 5 Progress Report

Date:	7th July 2014
Course Instructor:	Dr. D. Mohamedally
Module:	COMPGS99: MSc SSE/CS Project 2014

Authors:

Nektaria Stavrou

Stelios Georgiou

Wendy Wong

Stefan P. Alborzpour

1. Overview of Activities

LAST WEEK KEY ACTIVITIES

- Developed a high level project plan. The final version was revised after the agreement of requirements with the stakeholders.
- Created a set of UML Designs.
- Nektaria and Stelios prepared and gave an internal presentation for GZ07 (Professional Practice)
- Created a live collaboration document for the stakeholders to review the requirements.
- Setup a Team Foundation Server as the project repository.
- Began work on the user interface of the BL web portal.
- Developed test cases for the front end (user interface testing).
- Meeting with Microsoft for guidance on Azure technologies as well as a review of both project requirements and plan.
- Meeting (Skype) with James Baker to review project progress.

THIS WEEK KEY ACTIVITIES

- Implementation of the feature that retrieves the frequency of specific words in a document.
- Implementation of the feature that allows users to select how many words to be displayed in a list of most frequent words from a document.
- Implementation of the feature that retrieves the frequency of a bigram and a trigram.
- Implementation of the feature that retrieves the adjacent word for the most frequent words from a document.
- Connection to the Azure Storage Account through Visual Studio.
- Established backend that allows connection to database.
- Provisioned a HDInsight cluster.
- Computed a test case within HDInsight framework.
- Meeting with James Baker to review requirements and project progress.
- Meeting with Simon Julier (UCL Lecturer & Wendy's supervisor) to discuss project.
- Attended the British Library Labs project event.

NEXT WEEK KEY ACTIVITIES

- Implement field specific search.
- Implement Boolean search capability.
- Implement wildcard search.

2. Deliverables

Description	RAG Status
External Comms The external comms piece is prepared on behalf of the team to be published to the British Library Digital Scholarship blog website.	**Green**
Word frequency feature The feature that calculates the frequency of words in a text document. This scans the entire text and accumulates the words and their counters into a list.	**Green**
Specific word frequency feature The feature calculates the frequency of a specific word within a text document. The user inputs a specific word and the system returns the number of occurrences.	**Green**
Adjacent to the most frequent word feature This feature finds the adjacent words of the most frequent words. It uses the word frequency feature in order to identify the most frequently occurring words, for each word it then identifies and returns the adjacent words.	**Green**
Word frequency of a bigram feature The feature calculates the frequency of a sequence of two words in a text document. The user inputs two words and the frequency of that bigram within the document is calculated and returned.	**Green**
Word frequency of a trigram feature The feature calculates the frequency of a sequence of three words in a text document. The user inputs three words and the frequency of that trigram within the document is calculated and returned.	**Green**

3. Minutes

Meeting 1 –Requirements review with James Baker

Date	Time	Location
Monday 30th June 2014	11:00 – 12:00	British Library

Attendees	James Baker (JB), Nektaria Stavrou (NS), Wendy Wong (WW), Stefan P. Alborzpour (SA)
Apologies	Stelios Georgiou
Minutes	Stefan P. Alborzpour

Agenda Item – Points Discussed	**Type** Action, Decision or Info
1 – Requirements	
1.01 SA went through each requirement for sign off by JB.	Info
1.02 JB advised for the inclusion of a new should requirement. A user should have the ability to cite the search with the opportunity to store their search, like a string of text, recording the search method.	Info
1.03 JB advised for the inclusion of a new should requirement. The users should be able to package and download desired data.	Info
1.04 JB acknowledged that the boolean search feature is a should requirement and asked for the justification of this to be documented.	Action
2 – Project plan	
2.01 SA provided JB with an explanation of the project plan.	Info
2.02 JB identified some irregularities and asked for them to be corrected.	Action
3 – Project closure	
3.01 JB asked the team about the legacy of the project. SA responded by explaining that the metadata shall reside in Azure marketplace for researchers to access. He also informed JB that the architecture and code created by the team will be available. NS clarified that the team's Azure subscriptions will expire after the project.	Info
3.02 SA referred back to the project plan to explain the team's intention with regard to project hand over and the creation of a project closure document.	Info
4 – Comms piece	
4.01 JB gave feedback on the comms piece draft. Firstly, he suggested the inclusion of some interesting quotes from the focus group. Secondly, he asked for some comments on the data itself, how it is different from what we typically work with and what makes it interesting.	Info

4.02 The team agreed to alter the document.	Action

New Action Items	Owner	Due Date
Justify priority of boolean search feature.	SA	02-07-2014
Correct the project plan.	SA	02-07-2014
Revise comms piece.	WW	10-07-2014

Meeting 2 – Meeting with Simon Julier (UCL Lecturer & WW's supervisor)

Date	Time	Location
Wednesday 2nd July 2014	15:00 – 16:00	Malet Place Engineering Building, room 6.04

Attendees	Simon Julier (SJ), Nektaria Stavrou (NS), Stelios Georgiou (SG), Wendy Wong (WW), Stefan P. Alborzpour (SA)
Apologies	
Minutes	Wendy Wong

Agenda Item – Points Discussed	Type Action, Decision or Info
1 – Project overview	
1.01 Introduction of team to SJ. The team provided a project overview to SJ. SJ asked for a requirements list and example data.	Action
2 – Project discussion	
2.01 SJ raised a lexical analysis point, in older text-based documents, spelling of words were influenced by regional phonetics before the printing press. This means that between texts, the same word may have multiple spellings.	Info
2.02 Contact Will from the Newton Project, a project that dealt with accessibility to available data, based on Newton's works.	Action
2.03 Consideration of the Cohort Effect for the varying use of search engines between older and younger academics. Must keep this in mind when developing the user interface of the portal.	Info
2.04 SA asked about the dissertation format and literature review.	Info
2.05 Emphasis on the importance of conducting user groups.	Info

New Action Items	Owner	Due Date
Send SJ requirements list and example data.	WW	04-07-14
Contact Will (Newton Project)	WW	04-07-14