Requirements Analysis Document

Text Analysis and Visualization Tools

Team G

Preface:

This document outlines the requirements of the Text Analysis tools. This document is written for the designers and client of the project.

Project Members:

William Blake, Kable Monck, Yuki Osada, Callum Webb, Garry Woo, Christopher Wright (manager)

Meeting Schedule:

Date	Time	Venue
08/08/2011	10:00	CSSE Tutorial room
15/08/2011	10:20	CSSE Tutorial room
22/08/2011	10:00	CSSE Seminar room

Future meetings are to be held every Thursday at 1pm, and when required any following Monday at 10am.

Sign Off:

Signature ____

Date ____

1 General Goals

1.1 Purpose of the System

To create a set of textual analysis tools that will work in an existing system that gathers information about plays and poems from the Shakespearean era. These tools will be an extension to an existing interface that will provide a more detailed analysis on texts. The tools will be used to show statistical data, such as word and letter frequencies, visualisation data such as charts and word clouds, as well as more advanced analytical tools such as doubling charts. The system to be developed must work with the existing architecture, which includes a webpage users visit to request and view texts.

1.2 Objectives and Criteria

- Create a set of tools that will provide a detailed analysis of texts derived from already existing XML files

- Create a web interface that users can visit to use the tools

2 Current System

Users can visit a webpage where they can view a formatted version of a play, showing elements such as acts, scenes and the line number as it was in the original text. When the user selects a play to view, a request is made to a database that contains a XML version of the play. The XML files have been manually written based on the original texts, and are parsed to a styler which transforms the XML to a readable version that is displayed to the user on a webpage.

There are modern versions of plays, as well as the original diplomatic transcriptions. The user can jump to a specific act and scene, or line number. Each play is formatted for aesthetics on a webpage. This includes italics for characters and stage directions, and large headings for each act.

3 Proposed System

3.1 Overview

The system that our group is creating is an arrangement of tools to assist researchers and scholars in finding certain features of English period play texts. This will include features based around the words in the play as well a doubling and props chart which will be able to show when characters are on stage. This will enable scholars to use statistics found in the play as well as allow for troupes of actors to see visually what character are on stage at a time.

3.2 Functional Requirements

Functional Requirements	**Description**	**Priority**
Word Frequency Table	To produce a table of words with their corresponding frequencies for a given range. This range could be the whole text, an act, a scene or character.	**1**
Letter frequency table	To produce a table of letter frequencies for a given range. This range could be the whole text, an act, a scene, a character or a signature range.	**1**
Word Cloud	To produce a word cloud visualisation of the word frequency.	**1**
Word Tracker	To show the result of a word frequency in context for a given word by producing the sentences it appears in. This could also include the act and scene in which it appears and the character speaking the word.	**2**
Letter Tracker	To show the result of a letter in context by listing all the words with a given letter in.	**2**
Letter Cloud	To produce a letter cloud visualisation of the letter frequency.	**2**
Doubling chart	To produce a doubling chart. A doubling chart shows which characters are in scene and how many words they say in each scene.	**3**
Prop Chart	To produce a chart similar to a doubling chart for when props are on scene.	**3**
Cloud frequency	To hover over a word or letter in the word or letter cloud, and show its frequency.	**4**
Character Algorithm	To build an option to define boys and mature actors to generate a list of possible doublings. This will then tell us the minimum number of actors needed to perform a play.	**5**

Value Estimation Ratio

Function	Value	Estimated Time	VAR
Word Frequency Table	$13.89	**10**	$1.39
Letter frequency table	$13.89	**10**	$1.39
Word Cloud	$13.89	**15**	$0.93
Word Tracker	$11.11	**18**	$0.62
Letter Tracker	$11.11	**18**	$0.62
Letter Cloud	$11.11	**19**	$0.58
Doubling chart	$8.33	**30**	$0.28
Prop Chart	$8.33	**30**	$0.28
Cloud frequency	$5.56	**50**	$0.11
Character Algorithm	$2.78	**80**	$0.03

3.3 Non-functional Requirements

3.3.1 User Interface and Human Factors

This system will be typically used by people who are studying Renaissance drama texts. The system will provide useful statistical and contextual data to aid the study of these dramas. These tools include concordances, doubling charts, prop charts and visualization tools. A web based frontend will be used with these tools running on a server backend.

The user interface for these tools should remain as simple as possible. It can be assumed that there will a range of users and some will have a computer illiterate background. Also users are assumed to have no prior knowledge of the operation of these tools. The interface must therefore provide an explanation and guide before the actual tool. The web front should also provide a more detailed section with an advanced guide and full documentation of the tools.

The web interface will be built to be compatible on all major browsers such as Internet Explorer, Firefox, Chrome, Safari and Opera. The web interface will run on all of these browsers and thus will be cross platform. Java applets will be avoided and all processing will be done on the web server rather than on the client computer. This will ensure maximum compatibility across all operating systems. Since the site is about Renaissance drama, it will be assumed that the default language for the website will be in English and only English.

The interfaces for the different tools will allow for text inputs or selections. The text inputs will be the main source of human error. The text inputs will give sufficient error prompts and deny certain character/symbol inputs. Browser timeouts will interruptions in the interface thus all the tools must complete their processing within this time limit.

The concordance tool will be used to analyze the text for letter and word frequencies. The tool will also present the adjacent text to the word or letter. The user interface of the system will consist of a text input (a word or a letter) and a display of its frequency and use in the text. There will be several parameters that the user can adjust to give different concordances. One parameter will be the range of lines that the search will be performed. Also a parameter for the amount of adjacent text displayed can be set.

Visualization tools will be used to display particular results from the concordance tool. These will include word clouds of the most frequent words, letter clouds for the most frequent letters and graphs of the frequencies. This interface should be very simple with only a selection of the range of words/letters to be displayed (e.g. 20 or 100 most frequent words/letters). The interface will simply then display a visualization for the selected range.

The doubling charts will be used to produce a chart that shows where characters/actors are on stage and the line numbers that must be spoken. This tool will be used to determine the possible doubling of characters and the minimum set of actors required to present the drama. Actors may play multiple roles but no character can be played by different actors. The user interface will display this chart visually, where doubling are shown by act and scene. The user will be able to select which text they wish to view the doubling charts for. Also there will be the option to filter the charts to view only a set of characters or characters for a particular actor.

The prop charts will be used similarly to the doubling charts. They will show which props are on stage in a particular scene and produced the minimum set of props required for a particular drama.

3.3.2 Documentation

Documentation will be very important for this system since it is assumed that users will have little knowledge about computing. A user manual will provide a thorough description of what the tool does and how to use it. The manual will provide a step by step guide on how and where to input data to gather statistical data on the Renaissance texts. Also a description of the allowable input will help users to avoid errors using the system.

Further documentation of the Java source code will prove important since the tools will be further developed by other programmers after this project is over. Javadocs will be used for all methods and classes. This will generate a standard documentation for the source code. A document and UML class diagram will be written to describe the current system architecture.

3.3.3 Hardware Consideration

The system is running on a powerful web server so there'll be high computational resources as well as a lot of storage space. This platform is not anything unusual so there shouldn't be any specific hardware concerns.

3.3.4 Performance Characteristics

The tools to be developed will only be used by a minority of the visitors, and the possible input texts are limited to the ones stored on the server, so as long as the system can handle an approximate upper bound there shouldn’t be performance issues. However, the system should be efficient, but large scalability issues can likely be ignored.

As the system will be accepting queries in real-time as part of a web page, it must have a quick response time. If necessary, it may be possible to pre-process the input data offline to allow faster and easier extraction of information from the input.

3.3.5 Error Handling and Extreme Conditions

Errors will mainly occur when users input incorrect keywords or values into the particular tools. These errors may occur when a user inputs non-alphanumeric characters in a text field in the concordance tools or generating visualizations of non-existent concordances. The system must be built as simple as possible so prompts suggesting corrections to erroneous inputs will be used to guide users.

Tooltips will be used to guide users of the different aspects of input, results and processes of the tools. They will provide information on what the tools do, what the results mean and the types of allowable inputs.

Since some of the tools require significant searching and processing, restrictions must be set to limit the resources used by each user. These include multiple clicks of a button wont spawn many searches. Also a limit on the size of the word clouds or other visualization tools may be required. Since these tools will eventually be run on a server backend, one user must not be allowed to contend the resources.

3.3.6 System Interfacing

Requests may be coming from external sources, but the system should only need to interface with the web server components which will be requesting data from the system. The actual I/O appears to be low priority, as it is higher priority to get the actual analysis tools done, so it will the design will be simple but flexible for now. If external libraries are to be used then the formats required by them also need to be considered.

3.3.7 Quality Issues

The system must report any analytical faults to the user to prevent false data being unknowingly used by the client. It would be reasonable to assume that the system must be easily reset to recover from any failure in a matter of seconds or minutes (not counting failure of supporting backend or web page systems which are outside the immediate scope of the proposed project.)

3.3.8 System Modifications

It is likely that the client will be interested in extra statistical and visual tools for textual analysis beyond the immediate specified requirements after these are met. To facilitate further development by another party, clear and extensive documentation will be required.

3.3.9 Physical Environment

NA

3.3.10 Security Issues

The system may analyse documents protected by copyright and as such must be secure from unauthorised access to these texts. It may be necessary to allow different levels of access to different users.

3.3.11 Resource Issues

The client will be responsible for all backups and data management. Our team will be developing standalone statistical and visual tools only, as there are systems already in place for the storage and management of backend data. Installation and maintenance of the proposed system will be carried out by another party which is responsible for administration of the pre-existing systems.

3.4 Constraints

The tools being developed in this project must be compatible with the pre-existing platform already in use for much of the back-end processing of texts, images and tagging scripts. Much of the meshing together of our tools and the current backend will be performed by others working in the scope of the project-at-large and will be undertaken after our proof-of-concept statistical analysis tools have been developed.

Within the scope we have to work with, the main constraint on programming language and development environment is that the tools must be accessible via a basic web front-end interface. Java has been agreed upon as the language to be used and is able to fit this requirement.

The use of open-source tools already available has been encouraged; as they will mainly be used for auxiliary tasks once the output from our statistical analysis tools has been generated (eg. feeding letter occurrence frequencies into tag-cloud generating software).

Glossary

Concordance – a list of frequencies and context of the words and letters in a particular text

Doubling Chart – a chart that displays the act/scene of a play on one axis and the actors/characters on the other axis. The spoken line numbers are written for each actor/character for a particular act/scene. The chart thus shows when actors/characters are on stage.

Javadocs – an automatically generated documentation of the methods and classes of a Java class/program

Prop Chart – a chart that displays the act/scene of a play on one axis and the props on the other axis. The chart is marked when props are used for a particular act/scene to show when props are on stage.

Renaissance – the movement or revolution human thinking/ideas that occurred in Europe from the 14th to the 17th century. Great reforms in literature, science, art, religion, politics and education were made in this period.

Tooltip – a prompt giving helpful hints or explanations of a particular part of the interface. The prompt usually only appears when the cursor is hovering over the particular part of the interface.

Wordcloud – a visualization of a set of words in a the shape of a cloud. Words with higher frequency/value are displayed larger than words with lower frequency/value in descending order.