Project Documentation

for

Text Analyzer

Student:

Kevin Kolcheck

Teacher:

Andy Harris

Class:

CSCI 240

Date:

3/25/2015

Contents

[1. Project Overview 3](#_Toc91412612)-4

[1.1 Disclosure 3](#_Toc91412614)

[1.2 Description of project 3](#_Toc91412614)

[1.3 Project title 3](#_Toc91412614)

[1.4 Intended user 3](#_Toc91412614)

[1.5 What problem is being solved? 4](#_Toc91412614)

[1.6 Technologies used 4](#_Toc91412614)

[2. Data Design 4](#_Toc91412648)-5

[2.1 The data 4](#_Toc91412614)

[2.2 Data representation 4](#_Toc91412614)

[2.3 Data persistence 4](#_Toc91412614)

[2.4 Data diagram 5](#_Toc91412614)

[3. UI Design 5-6](#_Toc91412648)

[3.1 Elements 5](#_Toc91412614)

[3.2 Layout details 5](#_Toc91412614)

[3.3 Design 6](#_Toc91412614)

[3.4 Classes, Methods, & Functions 6](#_Toc91412614)

[3.5 UI diagram 6](#_Toc91412614)

[4. Algorithm 6](#_Toc91412648)-7

[5. Revision 7](#_Toc91412648)

# Project Overview

## Disclosure

The idea behind the project is not my own. Earlier today I had read a couple of really interesting articles that pertain to text anylsis. The first article was about Microsoft’s recent failed A.I. tweetbot called Tay. <http://www.forbes.com/sites/kalevleetaru/2016/03/24/how-twitter-corrupted-microsofts-tay-a-crash-course-in-the-dangers-of-ai-in-the-real-world/#6c3256cc32cb>. While not something nearly as complicated, the idea that a piece of software interacts and changes with things people write sounded really cool. While that is not what I am trying to accomplish with this program, the process behind analysing their tweets, is.

The second article I read was about an A.I. that wrote a Sci-Fi book (with a input from the creators) that almost won a literary prize <http://www.latimes.com/books/jacketcopy/la-et-jc-novel-computer-writing-japan-20160322-story.html>.

As I was digging for something in this vein that was more on my level, I came across the base idea for this project on this website: <http://www.sciencebuddies.org/science-fair-projects/project_ideas/CompSci_p022.shtml#summary>. When I read this it just sounded like a great fit for something that would be a challenge, but also be a good way to get my feet wet when it comes to mining data and actually using it to make conclusions about something.

## Description of project

The intent of this project will be take a piece of writing and pull it apart to measure things about word choice, sentence length, word length, and so on, with the hope that by showing the program enough writing by an author, it would be able to identify the same author’s footprint even when it is not credited. Information obtained will be saved to a database or .csv file (not yet determined).

## Title

Text Analyzer

## Intended user

The user will likely be a typical adult. The program user interface is very simple and straightforward, in that it can be operated by either clicking and loading in a .txt file, or copy and pasting.

## What problem is being solved?

The project is designed to heavily analyse a piece of writing, so it will be able to extract information about ANY snippet of English writing that would take a person hours to do on their own and even help someone identify an author.

## Technologies used

* Java
* Swing – for UI
* Perhaps MySQL or SQLite for database (if that’s the route I take for data management)

# Data Design

## The data

The data this program handles is: the number of sentences contained in text, number words per sentence, number of letters in each word, how many words per sentence, average number of words per sentence, the average word length, frequency of differing sentence lengths, frequency of prepositions and conjunctions.

## Data representation

Honestly, I’m not sure what would be the best way to represent the data. As I’ve worked on the documentation, I’ve come to realize more and more that this would have been a disaster for me to start without planning. I would like to discuss with you, or someone, about the best way to represent the data as I do not know. I was thinking about using arrays

## Data persistence

Not totally sure what would be the best way to do this. At bare minimum the analysis will be saved to a txt file, or csv, but perhaps it would be better to store all data will be saved to a database or something using either MySQL or SQLite. I’ve not done anything with databases, so I’m really not sure if this is the type of thing one would use a database for.

## Data diagram

I currently do not know the best way to represent the data, thus was struggling to come up with a diagram.

At a higher level it’s not a big deal, every author will have separate works/pieces of writing that will be looked at, but then it starts to get more convoluted to me because every work will be made up sentences, which are made up of words, which are made of characters, and if I’m going to be parsing things on different levels, from sentence length to word length and frequency of word length, how should I represent that?

# UI Design

## Elements

(I have not used Swing or know a great deal about it, some of these elements may need adjusted based on what Swing offers, however this is the basic idea.)

Label: header, instructions

Scrollbar: yScrollLoad, yScrollResults (this was a Python thing, not sure how Swing does scrollbars for textboxes)

Textbox: txtboxLoad, txtboxResults

Buttons: btnLoadTxt, btnAnalysis

## Layout details

* Total columns: 2
* Total rows: 6
* Header label will stay in column 0
* Instructions label will be in row 1, column 0
* txtAuthor will be on row 3, column 0
* BtnLoadTxt will import a .txt file into txtboxLoad. It will be on row 2, column 1.
* txtTitle will be on row 3, column 0
* txtboxLoad needs to be sticky east and west to fill the width of the window. Attach the yScrollLoad scrollbar to the txtboxLoad. This is where the writing to be analysed will go. Grid to Row 4, column 0.
* txtboxLoad needs to be sticky east and west to fill the width of the window. Attach the yScrollLoad scrollbar to the txtboxLoad. This is where the writing to be analysed will go. Grid to Row 4, column 0.
* Set btnAnalysis as a button that will initiate the method startAnalysis, which will examine txtboxLoad and print results to txtboxResults. Grid to row 5, column 0

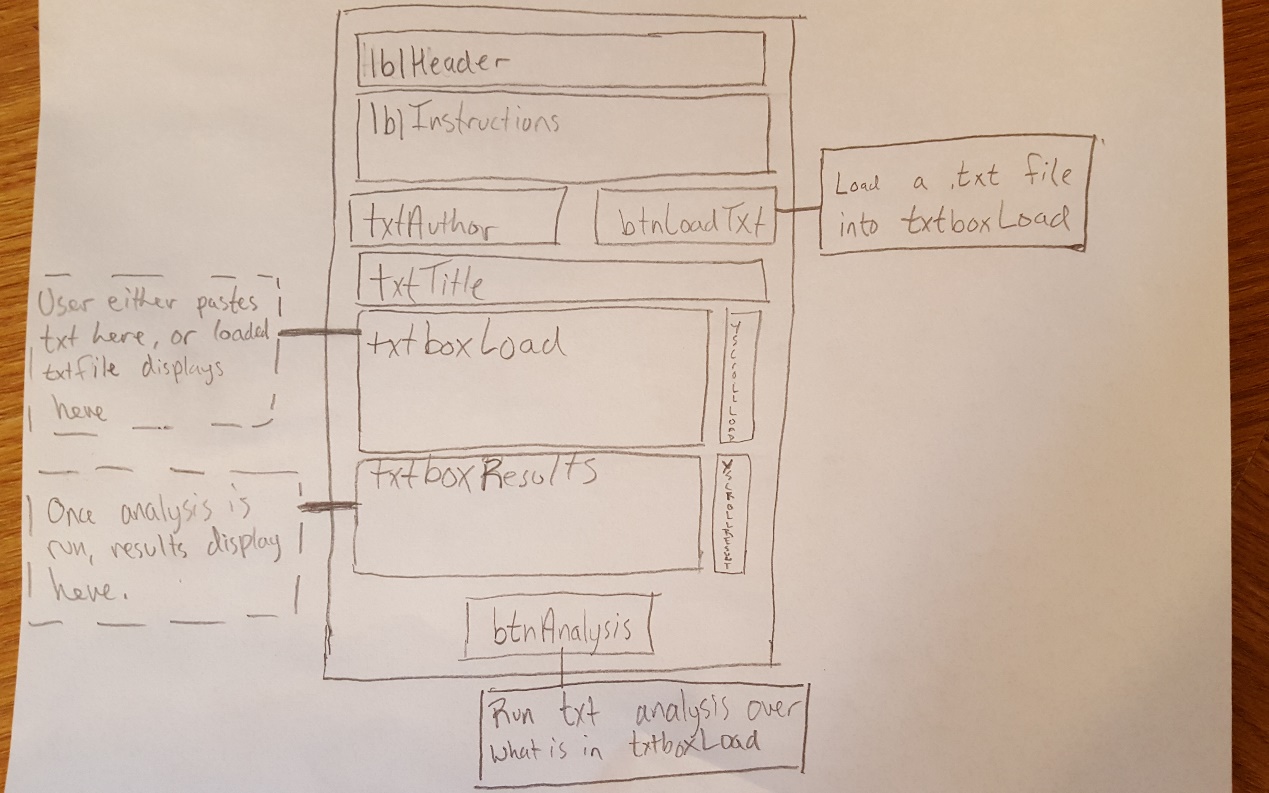
## Design elements

Header Font – 14pt, Helvetica, bold

## Classes & Methods

This is where I’m starting to feel things are half-baked with this idea. Since I’m not really sure what all I need to do with the data, I’m not sure about what classes and methods I’ll need in order to get the job done. I don’t feel that I can answer this without fully understanding the data first, At the very least I need a TextAnalyzer class (main class) to house the program. I’ll need an initializer, and ActionListener methods.

## UI diagram



# Algorithm

This section is going to be very basic, in part, due to not knowing the data and what my classes/methods will be.

//Create TextAnalyzer class (which will be the main class) as an extension of jframe implementing the ActionListener

//create Jbuttons, labels, panels, and whatever else I might need for layout

//Create main (public static void main…), create an instance of TextAnalyzer

//Create initializer for TextAnalyzer, which sets up all of the UI and layout information for rows and columns

//Create method for ActionListeners

Beyond setting up the layout, I’m not sure which classes, methods I will need to accomplish my goal.

# Revision

|  |  |  |
| --- | --- | --- |
| Ver. | Description | Date |
| 1.0 | original proposal | 2016-03-25 |
|  |  |  |