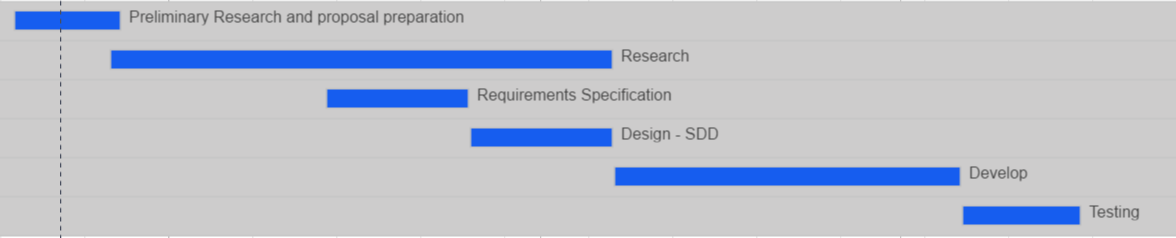
Senior Project Proposal

David Lambertson

1. **Scripture**“Do ye not remember the things which the Lord hath said? – If ye will not harden your hearts, and ask me in faith, believing that ye shall receive, with diligence in keeping my commandments, surely these things shall be made known unto you.” 1 Nephi 15:11  
   If I believe that I shall receive and am diligent in doing the work I have said I would, I know that I will do well on the project and in the class and be successful. Likewise, if I am not diligent, I can know that I will surely fail.
2. **Abstract**The focus of my project is data mining. I want to create a program that goes through social media and tries to predict who the sex of a given user using information accessed from Twitter’s API. The information may include but is not limited to tweets, followers, friends, and those following.
3. **Background**
   1. **Definitions**

* Data Mining – The process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.
* Data – Are any facts, numbers, or text that can be processed by a computer.
* Classes – Stored data is used to locate data in predetermined groups.
* Twitter - Being akin to a free, high-speed, global text-messaging service. In other words, it's a glorified piece of valuable infrastructure that enables rapid and easy communication.
* Folksonomy - Decentralized universe of tags that emerges as a mechanism of *collective intelligence* when you allow people to classify content with labels.
* Tweet - Basic atomic building block of all things Twitter
* Twitter User - Can be anyone or anything. They tweet, follow, create lists, have a home\_timeline, can be mentioned, and can be looked up in bulk.
* Entity - Entities provide metadata and additional contextual information about content posted on Twitter
* Place - Are specific, named locations with corresponding geo coordinates. They can be attached to Tweets by specifying a place\_id when tweeting. Open
* Authorization (OAuth) - An authentication protocol that allows users to approve application to act on their behalf without sharing their password.  
  1. **Why This Topic Interest You**This topic is of interest because I find it interesting to be able to predict certain trends or traits through the things people post on social media. It’s amazing how much people post on social media and what it can actually reveals about social trends in different areas such as fashion, politics, national issues and international issues.
  2. **Prior Work by others**There has been a lot of research put into the Data Mining field. Data mining is employed by many companies who desire to figure out trends in purchases, store visits and the likely causes of this to help increase sales. People even use it to predict outcomes in political races, events and law outcomes. Data mining is well used in different fields.   
     “Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.”

* 1. **Prior Work by Me**There is no prior work done by me on this topic.

1. **Description**Provide the details of your project.
   1. Provide a description of your project.  
      I will write a GUI that is very simple. It allows the user to input a username for Twitter. Using this information, I will hit the Twitter API to allow myself to pull information from Twitter. Using this data, the program will determine if the user is male or female.
   2. Provide a clear description of what constitutes success for your project (e.g., 90% accuracy).  
      The project is a success if the program can accurately predict the sex of the user correctly 80% of the time.
   3. Tasks
      1. Preliminary Research and proposal preparation
         1. Familiarizing myself with Twitter
         2. Determining what programming language to use
         3. Determining if and what IDE I will use
         4. Writing the proposal
      2. Research
         1. Figuring out what how to connect to the Twitter API and get data
         2. Figure out how their data is set up
         3. How to store the data so I do not have to pull it every time
         4. How I want to store the results
         5. Different methods to figure out the difference between male and female
         6. Quirks about males and females
         7. What I should tailor my questions around to determine the sex
      3. Requirements Specification
         1. Writing of the SRS
      4. Design - SDD
         1. How I want to display results
         2. What the user interface will look like
         3. How I want the code organized
         4. Determining the questions and polishing them up.
      5. Develop
         1. The interface
         2. Polishing of the Code
      6. Testing
2. **Scope** 
   1. **What is Included**The scope of this project is it will only use Twitter. My project will only try to determine gender, nothing else.
   2. **What is not Included**This will not include any other social media platforms. I also will not be trying to determine anything else such as age, location, or any other information.
3. **Tasks and Schedule**
   1. Preliminary Research and proposal preparation  
      All these tasks need to be finished before the proposal is submitted and approved. This will take me a week from the semester starting.
      1. Familiarizing myself with Twitter – Getting familiar with Twitter and its API. This is where a lot of the time has come from.
      2. Determining what programming language to use – Deciding on what language to use once I understand the Twitter API some more.
      3. Determining if and what IDE I will use – Once I choose the language, what IDE would best fit for me to use.
      4. Writing of the Proposal – This is the actual writing of the project proposal.
   2. Research  
      These tasks should be finished within five to seven weeks from the proposal being approved. They will be started a few days after the requirements document is started. While I am doing research is when I will start working on the requirements and design so I could be doing multiple things at once.
      1. Figuring out what how to connect to the Twitter API and get data – Once the proposal has been approved, I want to start to get acquainted with Twitter’s API and learn how to connect to it. This could include a prototype to help myself get use to using the API in the language I choose. This needs to be finished within two weeks of my proposal being approved.
      2. Figure out how their data is set up – Through the prototype, learning how the returned data from the API is organized so I can parse it and grab what I desire. This will tie closely with 6.2.1. They have the same schedule.
      3. How to store the data so I do not have to pull it every time
      4. How I want to store the results
      5. Different methods to figure out the difference between male and female
      6. Quirks about males and females
      7. What I should tailor my questions around to determine the sex
      8. Requirements Specification
   3. Requirements Specification  
      This document should be finished within a week and a half.
      1. Writing of the SRS
   4. Design – SDD  
      Design should be finished a week and a half from the requirements being written and submitted.
      1. How I want to display results
      2. What the user interface will look like
      3. How I want the code organized
   5. Develop  
      I have allocated the most time to this section. I have given myself roughly four to five weeks.
      1. The interface
      2. Polishing of the Code
   6. ****Testing – This should take a week or two. There will be small points of testing happening during development, but this is the main check for any major flaws.  
        
      According to the chart in Figure 1, this project should take me the majority of the semester. Working roughly 10-12 hours a week, I should end up with roughly 140 hours.

Figure

1. **Deliverables**
   1. Proposal
   2. Software Requirements Specification
   3. Software Design Document
   4. Source code
   5. Executable
2. **Applicability**The knowledge which I will be utilizing that I have learned previously is that of Java and JavaFX to write the code and create the GUI. The knowledge I have not obtained and hope to have by the end of the project is how to use Twitter to data mine and successfully predict trends/determine fact from data mining.
3. **Required Resources with Cost**As of right now, there are no costs associated with this project. I have access to the software needed for free and accessing the data at Twitter is through their API and it does not cost money to use. There might be some costs if I decide to buy a book to help with learning how to mine social media effectively.
4. **References**

* http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
* **http://chimera.labs.oreilly.com/books/1234000001583/ch01.html#social-websites-rage**
* [**https://dev.twitter.com/overview/api/users**](https://dev.twitter.com/overview/api/users)
* [**https://dev.twitter.com/overview/api/entities**](https://dev.twitter.com/overview/api/entities)
* [**https://dev.twitter.com/overview/api/tweets**](https://dev.twitter.com/overview/api/tweets)
* [**https://dev.twitter.com/overview/api/places**](https://dev.twitter.com/overview/api/places)
* [**https://dev.twitter.com/overview/api/twitter-libraries**](https://dev.twitter.com/overview/api/twitter-libraries)
* [**https://dev.twitter.com/**](https://dev.twitter.com/)