News Aggregator with Python and Machine Learning Semester 2

Statement of Work

Created by TechLauncher's News Aggregator with Python and Machine Learning Team

1. Background

1.1 Problem

Traditional media are facing comprehensive challenges, and new media are also criticized for many problems, such as the vulgar content catering public interest. Even though the website from last semester solves the problem to some extent, we still have some problems. The first one is a vague user scope. Previously, the project did not reach a conclusion on user scope, which would make our work unrealistic. The ultimate goal of this project is to provide users with a satisfactory experience, so this year we will delimit a reasonable range of users. The second problem is our classification is based on complex algorithms, so the result is kind of not predictable, then it makes us hard to estimate user satisfaction. The third problem is our news channels is not sufficient, so some important news may be missed because of this. Finally, we have some existing code bugs in our program, as the crawler stops occasionally, and the database cannot conclude all news channels inside.

1.2 Solution

First of all, we initially defined the user range as Canberra residents. Canberra has a large number of government employees and school-related personnel, and its people have a relatively high level of education. Therefore, there is a relatively strong demand for high-quality news. Secondly, in order to cater to Canberra residents' preferences, we will adjust the extracting and ranking algorithm to give Canberra-related or Canberra residents interested news a higher weight. Thirdly, we may make some adjustments to the UI of the website. Finally, in order to truly confirm residents'satisfaction, we will conduct 2-3 user surveys, and adjust and improve the project according to the results of the survey.

2. Tools and Skills related

2.1 Team Management Tools
Slack, Trello, Github, Google Doc

2.2 Machine Learning Tools
Tensorflow/ Keras, PyTorch, scikit-learn, scikit-image, OpenCV, NLP tools (Gensim, spacy.io, bert-as-a-service).

2.3 Web Development Tools
Python web frameworks(Bootstrap), minimal JavaScript/HTML/CSS required.

2.4 User Survey Tools
Google forms.

3. Requirement

3.1 Product Functions

The primary goal of this project is to modify and update the previously implemented machine learning models to improve upon the accuracy of classification and ranking from various news-sources. Furthermore, extend the previous project by implementing a user feedback system to modify/correct our algorithms and possibly show customized content for users based on their choice of interests. This project converges mainly on Canberra based news sources to have a better understanding of user feedback and improve upon location-based importance.

3.2 General Constraints

- The back-end of the program should be based on Python 3.
- The web-based front-end should utilize HTML, CSS and JavaScript
- The program should connect to the company's database.
- The rank system should be developed based on machine learning or neural networks.
- The input data are pure texts of news.

3.3 Specific Requirements

3.3.1 Identification Methods

- The model developed should be able to assign a better ranking to the news articles that are not part of the training set.
- The program should be able to recommend new articles to an unregistered user based on the ranking system and location.
- The program possibly could provide customized content based on a registered user's interest

3.3.2 System Output

- The program shall provide a ranked list of the local news as well as international news.
- The program shall display the recommended news on the front-end.
- The program shall refresh the existent user interface.

3.3.3 System Input

- The program shall be able to process the textual data from new articles.
- The testing data should use a web crawler to get the testing news from other news webpages
- The model developed should use the existing data set of news as the training set provided by the client.

3.3.4 Algorithms

Our team should figure out and build another algorithm (like BERT) which is different from the algorithm the client has used (like k-means).

3.3.5 Optional Function

If time permits, we can integrate a user's upvotes and downvotes to the algorithm to show suggested news articles.

3.4 Minimum Viable Product (MVP)

3.4.1 An Operational Website

The program should at least implement

- The requirement in 3.3.2 System Output to assign better ranks of the inputting news and list them on the front-end.
- The requirements in 3.3.3 System Input, be able to accept the latest news and use the crawler to form such a test set.

3.4.2 Appropriate Algorithms in Deep Learning

Our team should improve the current deep learning model which are based on some current open-source algorithm.

3.4.3 User Survey Report

The user survey's play an important role in the development of both frontend and backend and how the algorithms functions. For this project, we aim to get feedback from various age groups on how they consume news and their opinion on current age new media, and other aspects which will help us in designing the project and improving the experience of the user on the website.

4. Identification of resources, risks, potential costs

4.1 Resources

• Kosmos Dataset (multiple broadsheets and public broadcasters of world news from 2014.01 to 2016.06, including more than 300,000 articles)
• GPU servers (used for training machine learning model)
• Team members' laptops used for news data processing.
• The code source and products from last semester's work
• The Bootstrap or other frames for web development
• Some Canberra focused news websites.

4.2 Risks

• The risk of failing to achieve the expected accuracy and efficiency.
• The risk of failure to final integrate due to the disunity of work division.
• The risk of timeouts which leads to the failure of debugging and final deliverable.

• The risk of potential dissatisfaction from the users when they find the news on the website is not ranking in their ideal order.
• The risk of being unable to make an improvement due to the lack of training data set.
• The risk of failing to respond to the feedback from the user survey

4.3 Risk Management

• Review work every 2 weeks. If any process is falling behind, try switch methods/ add additional workload/ adjust team structure…
• Conduct First user survey early (in week5), so the team will have sufficient time to deal with user feedback.
• Communicate with stakeholders timely.

4.4 Potential Costs

• A budget used for data training devices. The client will bear it.
• Budget for entity legal documentation and licenses. The client will bear it.
• Budget for purchasing web development frames. The client will bear it.
• Time used for the whole project. All team members will bear it.
• The commuting and depreciation fee for meeting and development. All team members will bear it.

5. Role and Responsibility

Name	uid	Technical Role	Progress Management Role
Jiahua Liang	u6162679	UI & User	Spokesperson 1
Luokun Gong	u5917339		Minutes Taker
Xiangyun Kong	u6556183		Progress Tracker
Yulinag Zhang	u6782445	Extraction & Algorithm	Spokesperson 2
Vishnu Vardhan Jasti	u6611697		Clerical Assistant
Jun Yang	u6767560		Clerical Assistant
Chen Zhang	u6745297		Progress Tracker

6. Schedule and Due date

6.1 Schedule

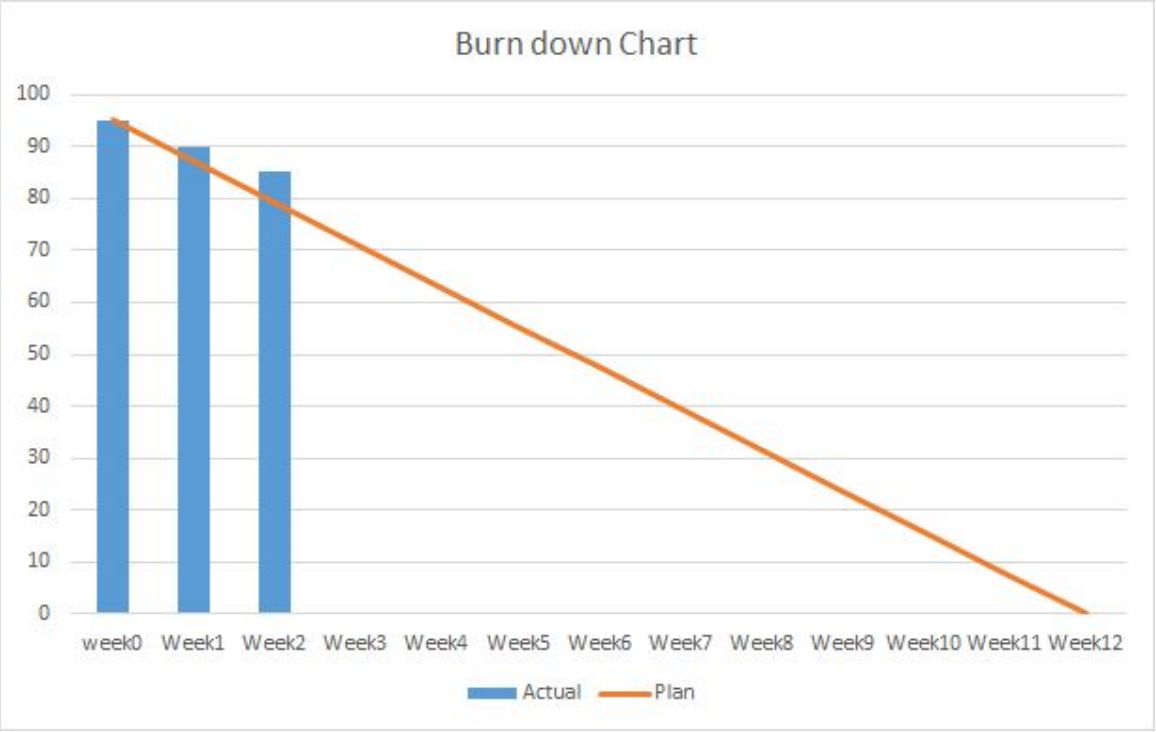
Time	Plan
wk1	1. Form team 2. Review previous codes
wk2	1. Divide roles 2. Confirm requirements 3. Decide schedule and milestones 4. Risk management
wk3	1. Audit 2. Finish legal issues 3. Set up the development environment (data, server, IDE, related frame or libraries) 4. Search for related information
wk4	1. Improve web UI design 2. Improve web crawler 3. Improve machine learning algorithm
wk5	1. Deploy improved website functions 2. Conduct the first user survey
wk6	1. Audit 2. Analyse survey report 3. Reflect to survey report, keep improving functions
wk7	1. Compare among different machine learning algorithms 2. Test & Debug 3. Adjusting UI and functions according to reflection
wk8	1. Conduct the second user survey 4. Analyse and reflect to survey report 5. Keep improving the machine learning algorithms according to reflection
wk9-10	1. Audit 2. Accomplish final deliverable 3. Design poster 4. Showcase (3rd user survey) 5. Adjusting UI and functions according to reflection
wk11-12	1. Finish handover documents 2. Report to client

6.2 Due Date

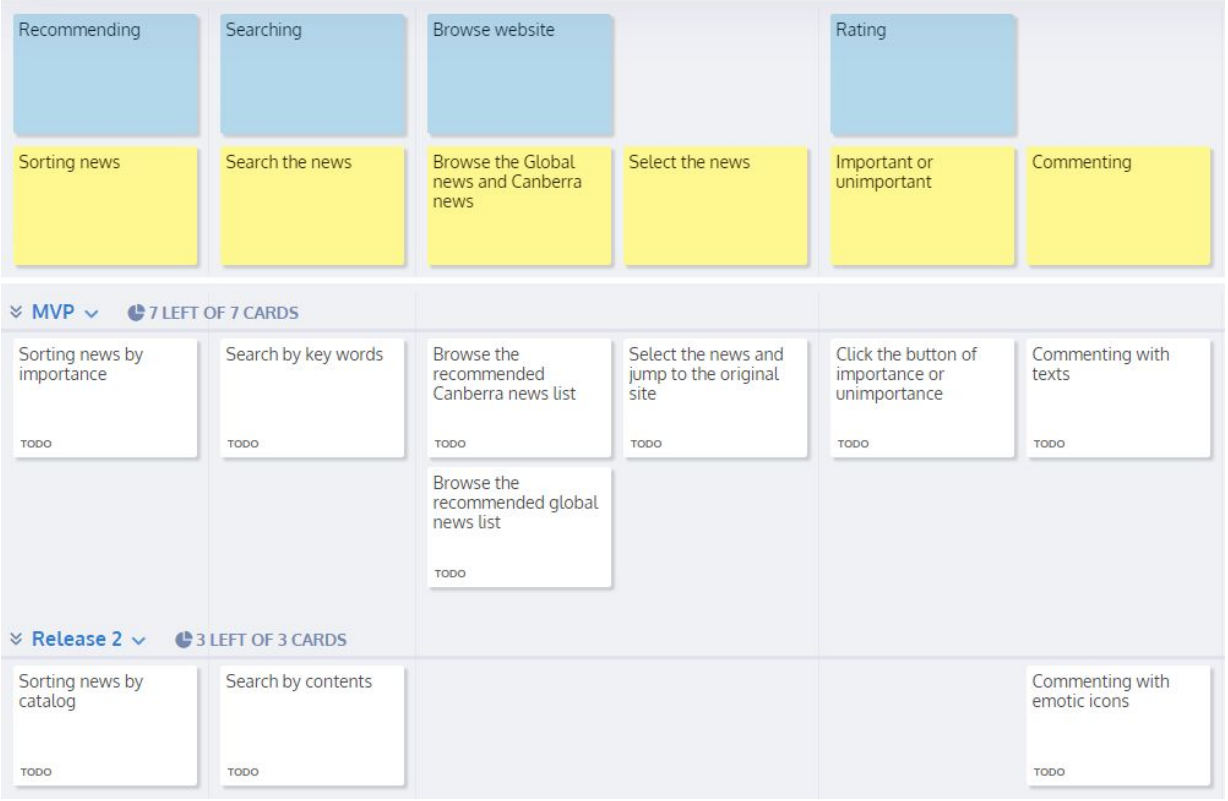
The client does not propose any due day so we will follow the due day of this course to arrange our project (https://cs.anu.edu.au/TechLauncher/dates/)

Client Signature:
Date:

Appendix A. Burndown Chart



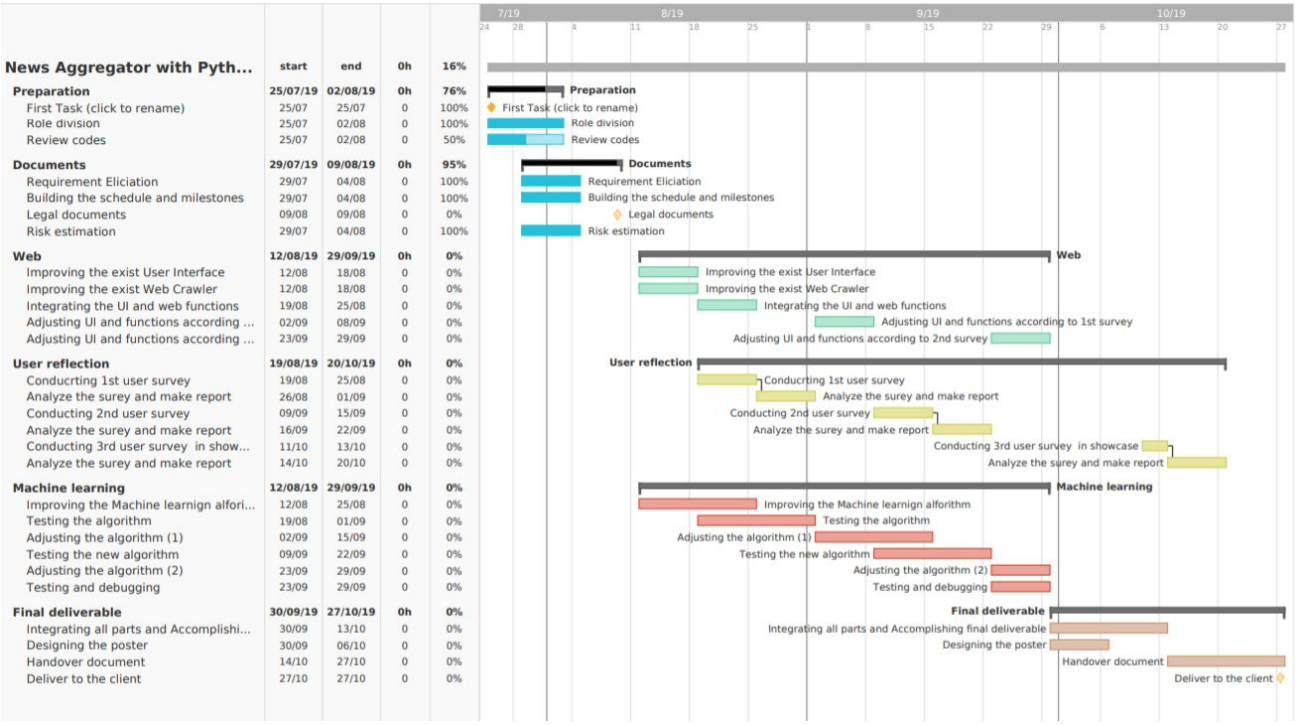
Appendix B. User Story Map



Appendix C. User Story Point Matrix

Story points	User story
5	As a visitor, I would like to select the news so that I can read its details on its original site. As a visitor, I would like to click the buttons of important and unimportant so that I can give feedback on the news
10	As a visitor, I would like to browse the Canberra news list, so that, I can know the events happening in Canberra. that more likely to be related to me. As a visitor, I would like to browse the world news, so that I can know the important events happening in the world. As a visitor, I would like to write texts to comment on the news, so that I can describe my feeling about the news more specifically.
15	As a visitor, I would like to search the specified news by keywords, so that I do not need to waste time to scroll down the news list to find the news.
30	As a visitor, I would like to only read the most important news at the top of their lists, so that I can avoid wasting time on looking for the most important news.

Appendix D. Gantt Chart



News Aggregator with Pyth...	start	end	0h	16%
Preparation	25/07/19	02/08/19	0h	76%
First Task (click to rename)	25/07	25/07	0	100%
Role division	25/07	02/08	0	100%
Review codes	25/07	02/08	0	50%
Documents	29/07/19	09/08/19	0h	95%
Requirement Eliciation	29/07	04/08	0	100%
Building the schedule and milestones	29/07	04/08	0	100%
Legal documents	09/08	09/08	0	0%
Risk estimation	29/07	04/08	0	100%
Web	12/08/19	29/09/19	0h	0%
Improving the exist User Interface	12/08	18/08	0	0%
Improving the exist Web Crawler	12/08	18/08	0	0%
Integrating the UI and web functions	19/08	25/08	0	0%
Adjusting UI and functions according ...	02/09	08/09	0	0%
Adjusting UI and functions according ...	23/09	29/09	0	0%
User reflection	19/08/19	20/10/19	0h	0%
Conducrting 1st user survey	19/08	25/08	0	0%
Analyze the surey and make report	26/08	01/09	0	0%
Conducting 2nd user survey	09/09	15/09	0	0%
Analyze the surey and make report	16/09	22/09	0	0%
Conducting 3rd user survey in show...	11/10	13/10	0	0%
Analyze the surey and make report	14/10	20/10	0	0%
Machine learning	12/08/19	29/09/19	0h	0%
Improving the Machine learnign alfori...	12/08	25/08	0	0%
Testing the algorithm	19/08	01/09	0	0%
Adjusting the algorithm (1)	02/09	15/09	0	0%
Testing the new algorithm	09/09	22/09	0	0%
Adjusting the algorithm (2)	23/09	29/09	0	0%
Testing and debugging	23/09	29/09	0	0%
Final deliverable	30/09/19	27/10/19	0h	0%
Integrating all parts and Accomplishi...	30/09	13/10	0	0%
Designing the poster	30/09	06/10	0	0%
Handover document	14/10	27/10	0	0%
Deliver to the client	27/10	27/10	0	0%