News Aggregator with Python and Machine Learning
1st Audit

Client: Thomas Griffith
Team Members: Yuliang Zhang, Jiahua Liang, Chen Zhang, Luokun Gong, Vishnu Vardhan Jasti, Jun Yang

Background

News Channel Problem:

- Traditional Media Recession
- New Media Fragmentisation & Vulgar

Difficult to get high-value news efficiently

Solution:

- A News Aggregator Focusing on High-Value News
- Using Web Crawler and Machine Learning Tools

http://gpu.jkl.io:5000/

Continuous Project Problem:

- Vague User Scope
- Rely on Algorithm, difficult to estimate user satisfaction
- Insufficient news channels
- Existing Code Bugs
  - Bad BERT Results
  - Crawler Stops Occasionally
  - Only 1 news source could be stored in data base

Solution??

Client’s Expectation & Requirements

More Canberra Focused:

- More Extraction Work on Canberra Based News Channels
- Improve Algorithm for Canberra residents

User Surveys:

- Conduct 2-3 User Surveys
- Improve our work according to user feedback

Improve Web UI

Stakeholders

Client: Thomas Griffith

Provide: development resources
- server, web frame, data, etc

Expect:
- A Canberra Focused, operable news website
- 2-3 User Survey Reports

Team Members

Provide:
- Time and effort

Expect:
- Improve code skills
- Improve teamwork skills
- Get good marks

Users

Provide:
- Time to use the website
- Time to do user survey

Expect:
- A efficient website for getting high-value news

Resources & Costs

Resources:
• Kosmos Dataset (news dataset including more than 300,000 articles)
• GPU servers (used for training machine learning model)
• The code source and products from last semester’s work
• The Bootstrap or other frames for web development
• Some Canberra focused news websites.

Costs:
Development related costs will be solved by our client.

Risks

Risks:
• Fail to achieve the expected accuracy and efficiency.
• Fail to achieve final integrate of different work groups.
• Timeouts for debugging and final deliverable.
• The risk of potential dissatisfaction from the users
• The risk of failing to respond to the feedback from the user survey

Risk Management:

- Review work every 2 weeks. If any progress is falling behind, try: switch methods/ add workload/ adjust team structure…
- Conduct first user survey early (week 5), so we can have sufficient time to response to survey results.
- Communicate with client and tutors timely.

Tools & Constrains

Development Tools

Github

Python3

Google Doc

MongoDB

Flask

Bootstrap

Google Form

Gensim

Keras

Pytorch

Tensorflow

Communication Tools

Slack
With Client

WeChat
With Team members

- In-Person Client Meeting every 2 weeks.
- Online Client Meeting all other weeks.
- Team Meeting whenever necessary.

Process Management Tool

Trello

https://trello.com/b/vxJ6hSCv/news-aggregator-sem2

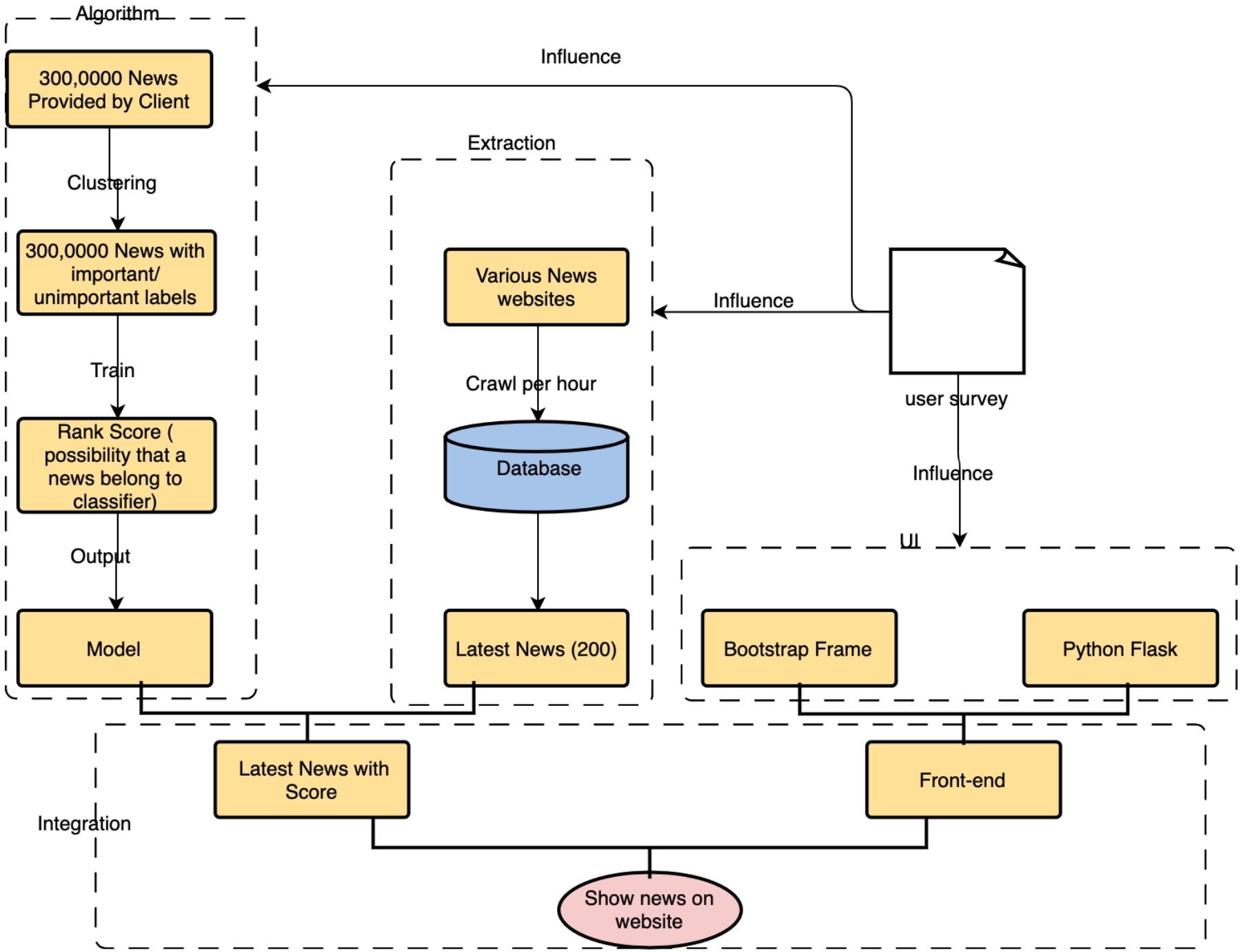
Teamwork

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

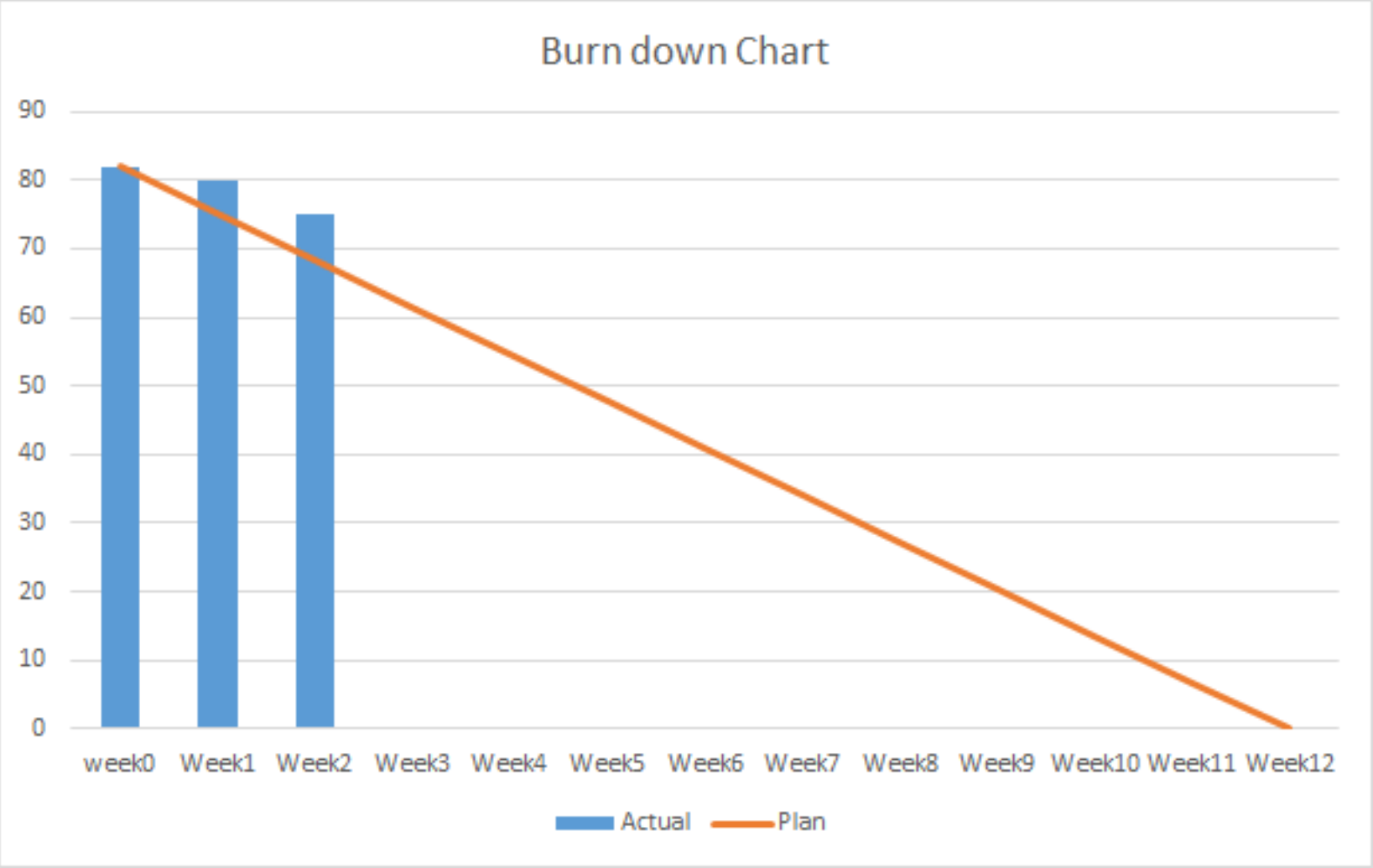
Schedule

Week 1-2	• Form Team • Divide Roles • Set up development environment	Week 7-8	• Second user survey • Reflect to user survey results
Week 3-4	• Improve extraction work • Improve algorithm • Improve UI	Week 9-10	• Poster • Final deliverable • Showcase (3rd user survey)
Week 5-6	• First user survey • Reflect to user survey results	Week 11-12	• Handover documents • Report to client

Workflow


Algorithm
300,0000 News Provided by Client
Clustering
300,0000 News with important/ unimportant labels
Train
Rank Score (possibility that a news belong to classifier)
Output
Model
Influence
Extraction
Various News websites
Crawl per hour
Database
Latest News (200)
Influence
user survey
Influence
UI
Bootstrap Frame
Python Flask
Integration
Latest News with Score
Front-end
Show news on website

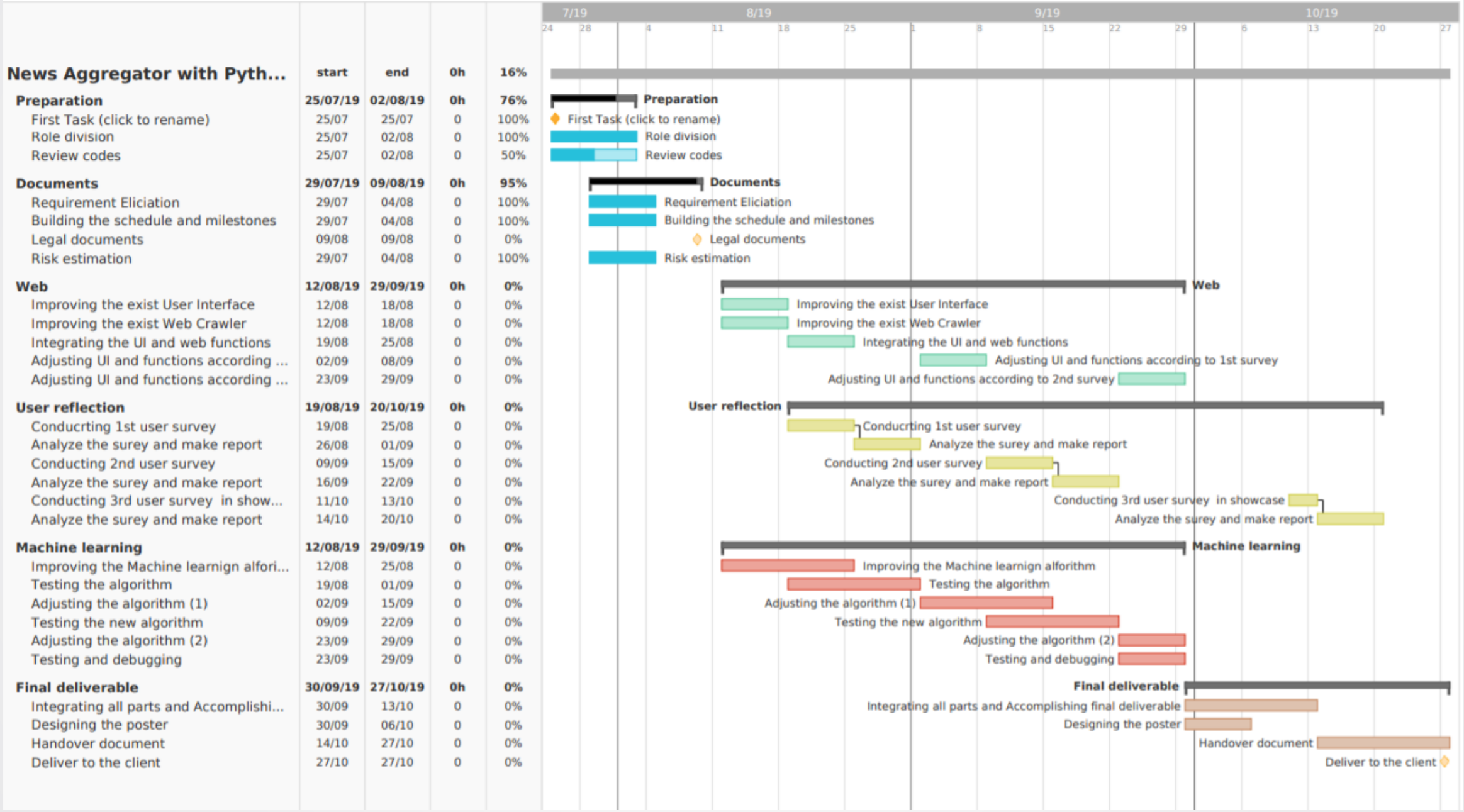
Burndown Chart



The Burndown Chart is calculated according to our Trello record:

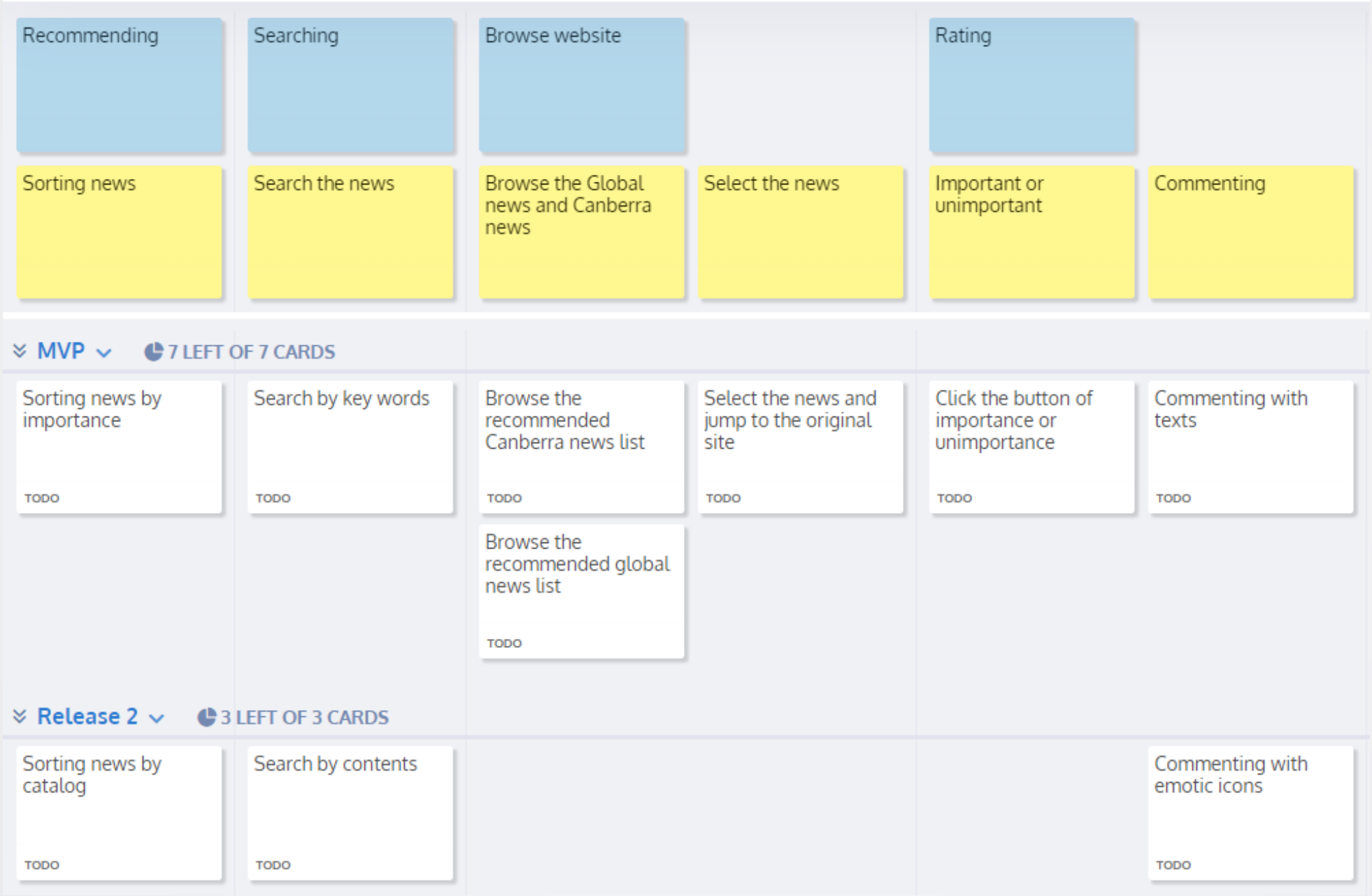
https://trello.com/b/vxJ6hSCv/news-aggregator-sem2

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%

User Story Map



Thank you