IntelliNews Sprint 2

Prepared By: Eleanor Haas, Yunis Hussein, and Jack Ogden

Virginia Tech, CS 5934, Dr. Mengistu

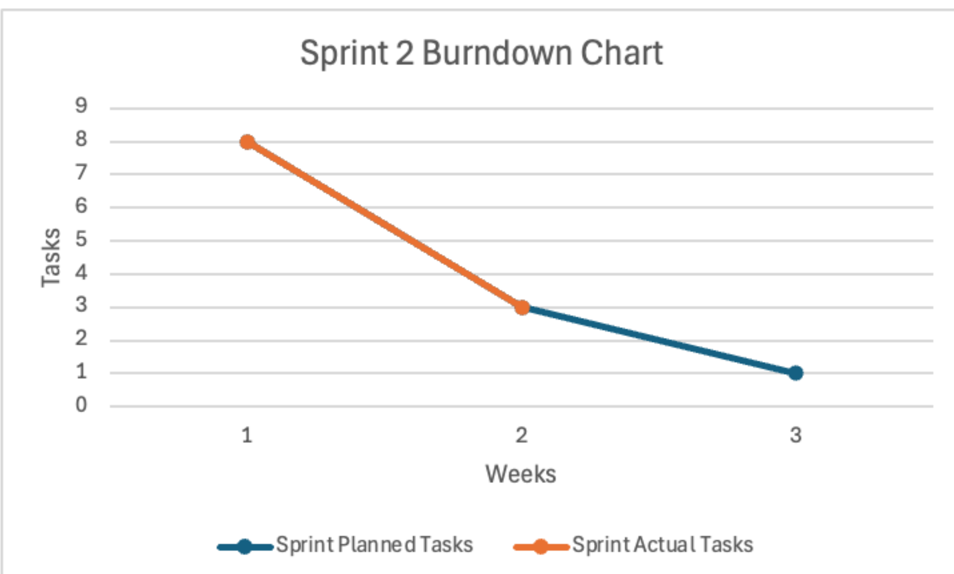
06/17/24

1. Scrum

a. Sprint 2 Backlog/Burndown Chart:

Sprint 2 Backlog:

Item	Priority	Completed
Fix headline scraper URL's	High	✅
Fully automate article scraping	High	✅
Integrate headline and article scraper	High	✅
Develop database for article data	High	✅
Clean up code	Med	✅
Enable selection between real-time and 24-hour trends	Med	✅
Enable date range selection for article search	Low	
Enable country selection for trends	Low	

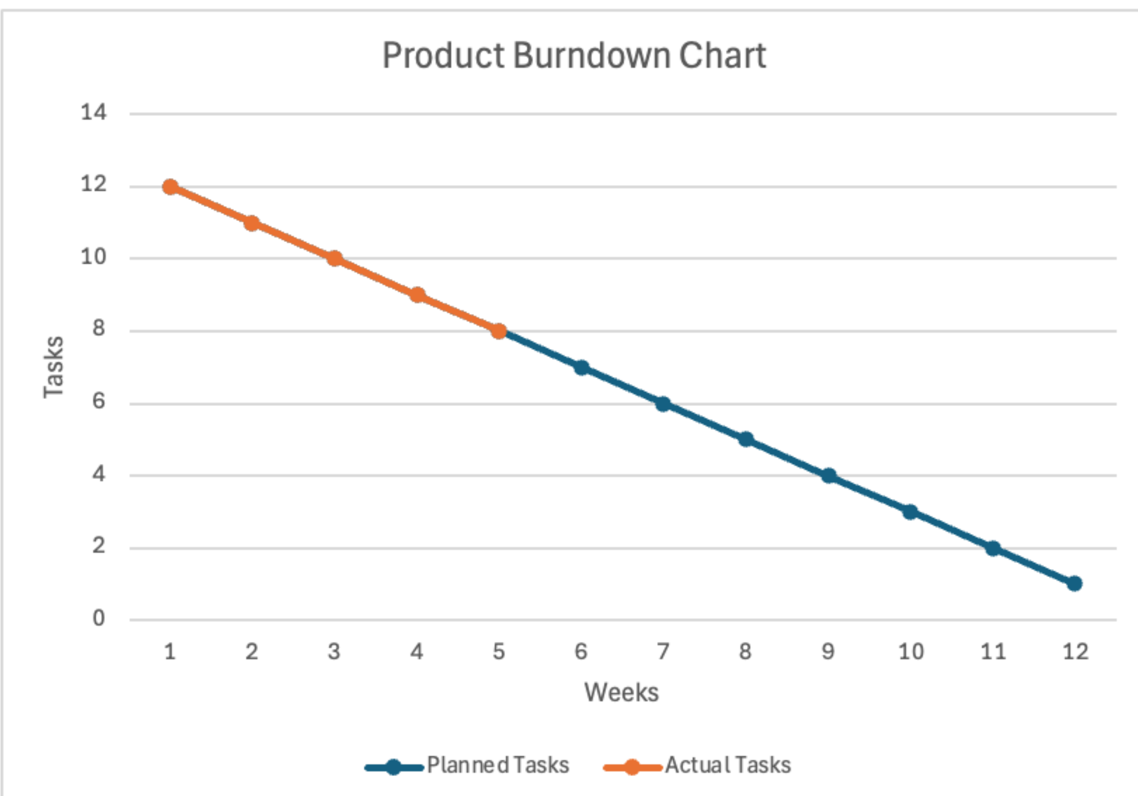


Commentary on the sprint burndown chart: Based on the sprint 2 burndown chart (which was created based on our product roadmap and sprint backlog) the project is on time and efficiency is at an optimal state thus far.

b. Product backlog (updated user stories)

- As a user outside of the United States, I would like the headlines to be more relevant to me.
- As a user, I would like to switch between live trending stories and stories trending over the past 24 hours.
- As a user who does not check the news frequently, I would like to see stories from a larger timeframe.
- As a user, I would like to see the sources corresponding to every statement in the summary.
- As a developer, I want to store the data scraped from the web in a database that is easily understood by a human and is easily accessible for natural language processing.

Overall Product Backlog/Burndown Chart:



Commentary on the product burndown chart: Based on the product burndown chart (which was created based on our product roadmap) the project is on time and efficiency is at an optimal state thus far.

c. Issue Tracking

The issue log for our project can be found here.

2. Documentation and code for completed tasks

The code for the completed web scraper can be found here. It is contained in the 'final_web_scraper.py' file and the output is contained within a file named 'articleBodyText.csv'. Below is a sample image of what a row of the output file looks like:

articleBodyText

	Headlines	Links	ArticleBodyText
0	Canada, Surrey, India	https://www.theindianpanorama.news/indians-abroad/indians-abroad-indians-abroad/4-arrested-for-indian-origin-mans-killing-in-canada/amp/	Yuvraj Goyal was found dead by the police when they were responding to the call of a shooting in Surrey on Friday, June 7 morning SURREY, BC (TIP): A 28-year-old Indian-origin man was shot dead in a suspected "targeted killing" in Canada's British Columbia province, according to the police, who have arrested and charged four people with first-degre Yuvraj Goyal was found dead by the police when they were responding to the call of a shooting in Surrey on Friday, June 7, morning, the homicide unit of the Royal Canadian Mounted Police said in a statement. Goyal worked at a car dealership in Surrey, Global News reported, quoting his sister, Charu Singla. She said the family had no idea why he was killed and that they had no indication that he had any ties to organized crime. Goyal's brother-in-law, Bawandeep, said the victim was on the phone with his mother, who lives in Inc "He came back from his gym, (his) daily routine, stepped out of his car, and was shot," Bawandeep said. Police arrested four suspects — 23-year-old Manvir Basram of Surrey, 20-year-old Sahib Basra of Surrey, 23-year-old Harkirat Jhutty of Surrey, and 20-year-old Keilon Francois of Ontario — shortly after the shooting. They ha In a statement released on Saturday, the homicide unit said authorities were alerted to a vehicle fire shortly after the shooting. On Saturday, the four were charged with first-degree murder, the Integrated Homicide Investigatic Police said that while initial evidence suggests it was a "targeted shooting, investigators are working to determine the reason that Goyal, a 28-year-old community member with no history of police contact, was killed." The homicide unit also urged anyone with information or people driving in the area with dash-camera footage to contact the police. (Source: PTI)

It contains the relevant headline being scraped, the link to the article, and its contents.

3. Demo Working Product System

Below is a sample run of our completed web scraper. It is only done on a small subset of data as it would take extremely long to run on the whole dataset.

4. For Progress report discussion

This and the past sprint were focused on finishing the first component of the application, the web scraper. The user stories we originally wrote had to be rewritten for specificity, as shown below:

- As a developer I want to have a program that takes news articles from the web so they can be fed to an LLM and condensed.
- As a developer I want my web scraper to copy the article's text into a database so it is easily organized by topic.

We did implement most of the elements in these user stories. Currently the 'database' system for the text is a .csv file. This works for the testing of the web scraper, but it may have to be changed into a text file for feeding the text to the LLM. We also added several potential enhancements to the issues tracker, which may be added to the backlog to improve the final product.

One of the issues we encountered with finding news sources and articles was finding ones without a paywall. While we know of some ways for users to get around paywalls, it is different to expect a machine to be able to. Using the HTML alone helped, but this is an issue that will continue to need to be addressed.

The goal for this coming sprint will be to begin on the LLM condenser for the articles. We have a short list of possible open-source and free LLMs to use, and our first task will be choosing one for certain. After that we will begin training the model and testing it to ensure that the output is expected and correct.