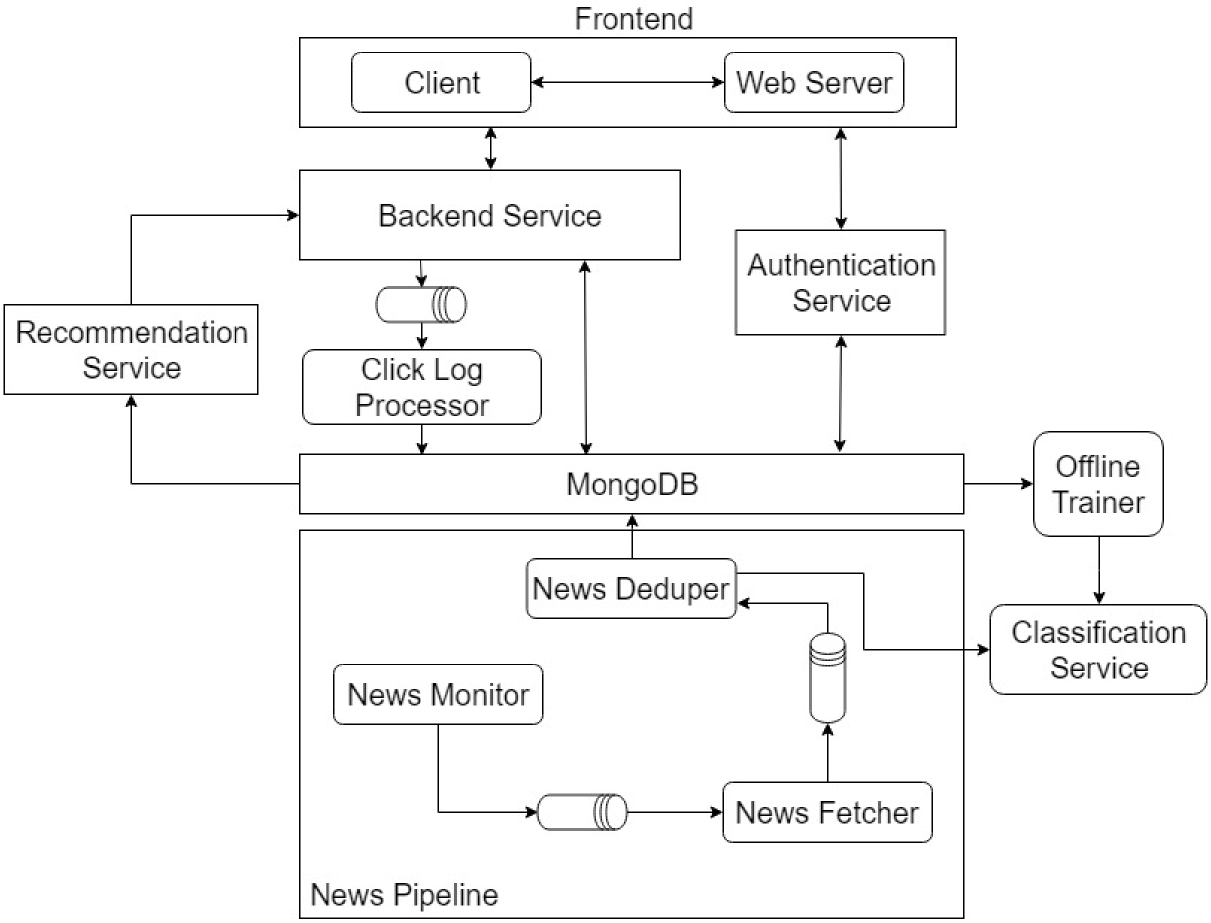
Tap News: Real Time News Scraping and Recommendation System

Shuyan Li

Introduction:

This is a web app that designed and implemented for users to read news from all over the world. Each news is a "card" in the webpage and users can scroll up and down to load more news. Every time the users click the news that they prefer to read, their click records will be saved and a model is build based on the clicks. After finding out the preference for each user, by logging in, the system can then recommend them with their preference news. Since it's a time decaying model, the system cares more on recent user action then their history clicks. The system, however, will not feed only the preference type of news to user, but with a higher probability, based on the frequency the user clicked.

App pipeline:



Implementation Detail

1. fetch news

The resource of the news comes from CNN, ABC, ESPN, and other main news resource There is a API called Newari. Every time I query this API, I get the latest news title and its resource, without content or details. I then implemented a web crawler to fetch from its resource and return the main content, details, images info from the website. After de-dupe the news that already stored in our database by TF-IDF algorithm, I store the news in our database. Since the speed of each service is different, I use RabbitMQ between each service in order to decompose each part.

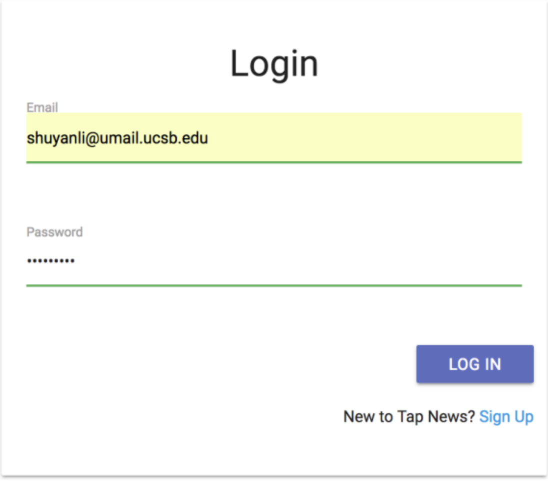
2. Get the news type

Before saving the news to DB, the raw news I get has no “types”, for example, sports, world, health, etc. By using Convolutional Neural Network and machine learning, I first label 1500 news in our database by our human eyes as training and testing data, then train a model using two-layers CNN for the contents of the news. The accuracy of our model is about 60% right now. If I label more training data in the future, I can get much higher accurate. Now every time the fetcher returns a news, I label the news based on the model I trained. After that, I save the news into the database.

3. Recommendation service.

I classified all news into 8 main types. Each news should have only one type for now. After a user sign up and log in, a brand-new model will be stored in the database that the probability of each type is equal. Every time the user clicked on one news card, the number of the clicks for this type of news will be changed, and the model that stored in database will be changed as well. Now the probability of each news is different, and the system will know the preference of this user. As the time goes by, this “time decay” model will gradually set the probability to equal if no more click is obtained.

Demo:



Login

Email

shuyanli@umail.ucsb.edu

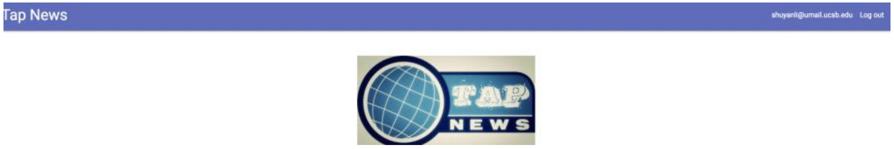
Password

•••••••••

New to Tap News? Sign Up

After logged in, the news card will be rendered on the screen. Feel free to click!

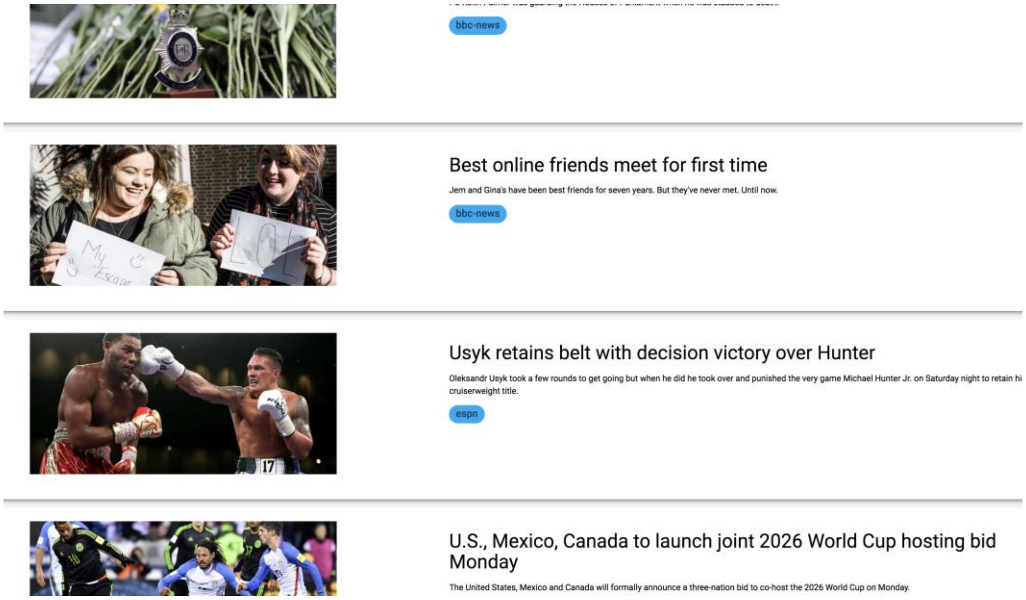
User info on top right:



Tap News

shuyanli@umail.ucsb.edu

Scroll down to see the news card:



Best online friends meet for first time

Jem and Gina's have been best friends for seven years. But they've never met. Until now.

Usyk retains belt with decision victory over Hunter

Oleksandr Usyk took a few rounds to get going but when he did he took over and punished the very game Michael Hunter Jr. on Saturday night to retain hi cruiserweight title.

U.S., Mexico, Canada to launch joint 2026 World Cup hosting bid Monday

The United States, Mexico and Canada will formally announce a three-nation bid to co-host the 2026 World Cup on Monday.

I used the pagination to load the news by “page”. Once the user scrolls down to the bottom, more news will be loaded. I added a “debouncer” here to prevent huge amount of request when scrolled to the bottom. Just like other Web apps like Facebook and LinkedIn, if you scroll down to the bottom, it takes a second to load more.