



National College of Ireland

BSHCSD4

Software Dev

2022/2023

Aaron Flynn

x19404024

x19404024@student.ncirl.ie

Information Hub

Technical Report

Contents

Executive Summary.....	2
1.0 Introduction	2
1.1. Background	2
1.2. Aims.....	2
1.3. Technology.....	2
1.4. Structure	3
2.0 System.....	3
2.1. Requirements.....	3
2.1.1. Functional Requirements.....	4
2.1.1.1. Use Case Diagram	4
2.1.1.2. Requirement 1 <Name of requirement in a few words>.....	4
2.1.1.3. Description & Priority.....	4
2.1.1.4. Use Case	4
2.1.2. Data Requirements	7
2.1.3. User Requirements	7
2.1.4. Environmental Requirements	8
2.1.5. Usability Requirements.....	8
2.2. Design & Architecture	9
2.3. Implementation	10
2.4. Graphical User Interface (GUI).....	14
2.5. Testing.....	16
2.6. Evaluation	18
3.0 Conclusions	19
4.0 Further Development or Research	19
5.0 References	19
6.0 Appendices.....	19
6.1. Project Proposal.....	19
6.1. Ethics Approval Application (only if required)	26
6.2. Reflective Journals	36
6.3. Invention Disclosure Form (Remove if not completed).....	Error! Bookmark not defined.
6.4. Other materials used	45

Executive Summary

This technical report will go into detail about the reasons behind making a news curation site and what I hope to achieve by making it. The majority of the report will discuss the requirements and go through each of the functional requirements needed for the project to work as a whole. It will end with some snippets of the work that has been completed so far, with a screenshot of some key functionality and another of what the GUI at present looks like, before leaving with some conclusions about the process of development so far.

1.0 Introduction

1.1. Background

I undertook this project after running into the issue of following too many different news sites for different topics. After a while of keeping up with different authors, columnists and sites, you run into the issue of having too many different things open at once and it becomes unmanageable. After trying and failing to find a solution to this issue, I thought of making one myself. There are a number of different solutions out there, particularly when it comes to the partisan politics of America or media personality's news letters. But there's also a more general solution, something akin to Google News, but those are few and far between and lack the more personalized features one would usually hope for. So my main reason was to create a solution to a problem I've encountered that no doubt a numbers of others have also.

Another reason for undertaking it was the challenge of creating it. There are many ways one can go about making a news curation site, relying primarily on RSS akin to the earlier internet, web scraping and others, the choice between a mobile app or a site. I enjoyed the free reign of picking up different technologies to see how one would go about it and arrived on a few I was eager to learn more about and code with.

1.2. Aims

The project hopes to provide a curated selection of news stories and articles to readers in a way that is personalized to their preferences and not too complex to grasp. The site won't focus in on any particular domain, as others like I've mentioned previously, but try to be more inclusive to a general audience. The goal of the site is to help readers stay informed about current events, the different areas that interest them, and present it in an organized and easy to access site. It also serves to be more convenient to the avid reader, presenting news from multiple sources in a single place.

1.3. Technology

The project will include the use of many libraries and technologies. Chief among them will be the use Django web framework for the front end. It handles the scripts being used, database management, user authentication, and the handling of HTTP requests. For the database, I will be using SQLite, as it's a lightweight database engine which is simple to use for a beginner with little experience working with databases. I will also be using pytest for unit testing and along with that, CircleCI for continuous integration. I had previously used Jenkins but when it came to deployment, hosting a Jenkins server on an AWS instance was difficult due to the large dependencies needed to be downloaded, CircleCI didn't have these issues so I decided to change to that pretty late in development. Using the two of them in parallel can make building a more reliable project long term easier. And for article analysis, I will be using TensorFlow & Pytorch for a language processing model, which will be used for the synopsis generation and categorization process, once the articles have been gathered earlier. The web scraping and other article gathering methods will be done with celery, which will schedule tasks so that no one will have to continually set off scripts every day to get new articles.

1.4. Structure

The document from here onwards will discuss the technical aspects of the project, it will begin with an overview of the system requirements, non-functional and functional. There will be elaboration on the functional requirements, going over the use cases of each of them and discussing the scope, description, flow description, alternate flow, exceptional flow, termination, and post condition of each.

2.0 System

2.1. Requirements

There are a broad number of requirements in this application, ranging from functional and non-functional requirements, data requirements, user requirements, environmental requirements, usability requirements and more. We will go into more detail for each of these later in the document, but as a started, we will go over the non-functional requirements, which describe how a system should behave and perform, rather than what it should do.

The chief one that informed a number of development decisions was obtaining copyright holder permission for development of the project. Maintaining legal compliance in the development process was a primary concern. Other non-functional requirements included internationalization and user experience, which will touch upon other requirement headers.

For data requirements, we have the chief among them being the news stories and articles that will be used in the curation process, included in that is the metadata gotten from the articles. But we also have the data we get from the user, their personal data they provide on registration and also the analytics data that the user will create while using the application.

With user requirements, we touch upon a few different areas including user accessibility, personalization, quality of content and user engagement. All these ensure that we meet user standards when using a web application

For environmental requirements, we touch upon browser compatibility and also operating system compatibility. Being a web application, it's important that it can be used across different platforms and devices. We also touch upon security again.

2.1.1. Functional Requirements

1. Data collection: The system should be able to collect and organize news articles from a variety of sources, primarily the use of online news websites
2. Article analysis: The system should be able to analyse news articles. This includes providing an automated process of extracting the relevant information such as providing summaries, the key points and categories of them
3. User accounts: Users should be able to create and manage their own accounts, allowing them to favourite and save articles, set preferences for categories and generally customize their news feed
4. Search: Users should be able to filter their curation feed based on keywords, categories, columnists, websites and other criteria
5. Security and privacy: The system should have robust security measures to protect user data, such as their emails and addresses
6. Accessibility: The system should be accessible to users with disabilities, including those who use assistive technologies such as screen readers.
7. Analytics and reporting: The system should provide analytics and reporting tools for users to see their reading habits and what kind of material they gravitate towards
8. Multilingual support: The system should be able to support multiple languages, including the ability to translate summaries and headlines into the user's preferred language
9. User feedback: The system should provide a way for users to give feedback and report problems to administrators

2.1.1.1. Use Case Diagram

2.1.1.2. Requirement 1

2.1.1.3. Description & Priority

2.1.1.4. Use Case

Requirement 1. Data collection

Scope: This use case applies to the data collection module of a news curation platform

Description: This use case describes the process of collecting and organizing news articles from the different sites for display on the system

Precondition: The system is not running into any errors and the task queue and associated programs are working as intended

Use Case Diagram:

Flow description:

Precondition: The system is not running into any errors and the task queue and associated programs are working as intended

Activation: This use case starts when the task queue is triggered after a certain period of time of non use

Main Flow:

1. The system's task queue triggers the collection of articles from the specific sources
2. As articles are collected, they are analysed and organized for display within the views, with the category, headline, and date of publication
3. The collected and organized data is stored in the system's database
4. The administrators are notified that the task is complete

Alternate flow:

1. If the articles are unsuccessfully organized and unable to be parsed, and error message is displayed to the administrator's
2. A bug ticket is created and sent to the jira board for administration to fix

Exceptional flow:

1. If the platform is unable to parse data, fails to create a bug ticket, and the method hangs, preventing further articles to be parsed, the allotted timer for the method halts the process and an error message is sent

Termination:

1. The use case terminates when the data collection is complete, or when the allotted timer of the function is cancelled

Post condition:

After the use case has terminated, the platform's database will contain the newly added parsed articles from the different sources, which the users can then use through the platforms search and filtering use cases

Requirement 2. Article Analysis

Description: This use case describes the process of collecting and organizing news articles from the different sites for display on the system

Precondition: This use case describes the process of automatically extracting relevant information from news articles for display on the system

Precondition: The system has collected news articles and they have been stored in the system's database

Activation: This use case starts when the method is called with the `scrape_article` function

Main Flow:

1. The system's analysis program retrieves a news article from the database
2. The program automatically extracts the article's category
3. Pytorch text summarization is used to collect the `article_body` text
4. The extracted data is stored in the relevant variables
5. A new Article model is created with the relevant data incerted
6. The process repeats until the url list provided is exhausted

Alternate flow:

3. If the articles are unsuccessfully organized and unable to be parsed, the for loop on that particularly url is broken and the next url is parsed, or if some fields are unavailable but the bulk are fine, then a 'N/a' is inserted instead of the needed field.

Termination:

2. The use case terminates when the data collection is complete and the url list is exhausted

Post condition:

After the use case has terminated, the platform's database will contain the newly added parsed Article models from the different sources, which the users can then use through the platforms filtering use cases

2.1.2. Data Requirements

There are a number of different data requirements for the creation of a news curation site, those include

1. **News stories and articles:** Obviously the main focus of a news curation site is to present to users a selection of news stories and articles. These are sourced from a number of different sources and the non-functional requirement of getting copyright holder permission, or working within the open-licence terms of others is included
2. **Metadata:** In order to organize and categorize the news stories and articles, and to perform article analysis, the site will require metadata such as the title, author, date of publication and source of the article. This metadata will then be used to aid in the filters and search process
3. **User Data:** During the creation of user accounts, they will be required to enter in an email and password. These are important and fall into the security and privacy requirement, as how we handle this information is very important to maintain a security standard in the interest user security is vital.
4. **Analytics Data:** The site will also collect data on user behaviour, such as page views and category preferences.

2.1.3. User Requirements

There are several different user requirements that need to be considered in order to provide a good experience for the users. These included

1. **Accessibility:** The site should be easy for users to access and navigate, whether this is on desktop, mobile or any other device or web browser they are accessing from. This has to include the implementation of features such as clear navigation between pages and a working search functionality.
2. **Personalization:** The site will also provide personalized experience for users based on their interests and preferences. This will take the form of article recommendation based on topics of interest to the user, and allowing the user to customize their news feed and profile
3. **Quality of Content:** Users will expect a standard to be maintained when it comes to the stories and articles present on the site. We will avoid using non-accredited news

sources in the curation and ensure that the analysis process works to a satisfactory degree. There will also be no bias when it comes to news presentation, a balanced and diverse range of viewpoints will be presented when a controversial issue pops up in the news cycle

4. **User engagement:** Users will be able to report any issues they come across to administration so that recurring issues they run into will be noted and action will be taken

2.1.4. Environmental Requirements

There are a number of environmental requirements to be considered in a news curation web application, for one to function effectively, the following must be considered

1. **Browser compatibility:** The site should be compatible with a range of web browsers, in order to ensure that the site can be accessed by users on different devices and platforms
2. **Operating System:** As an extension of browser compatibility, the application will need to be compatible with the different operating systems, meaning it will have to be able to run on both Windows and iOS and Android. Cross platform usability is an important part of user requirements.
3. **Security:** The site will be designed and built with security in mind in order to protect against potential threats that may arise in the even of a cyber attack or data breach. This will involve implementing measures such as encryption, secure login protocols, and maintenance of site in the form of regular security and library updates when libraries become deprecated and vulnerable to attack.

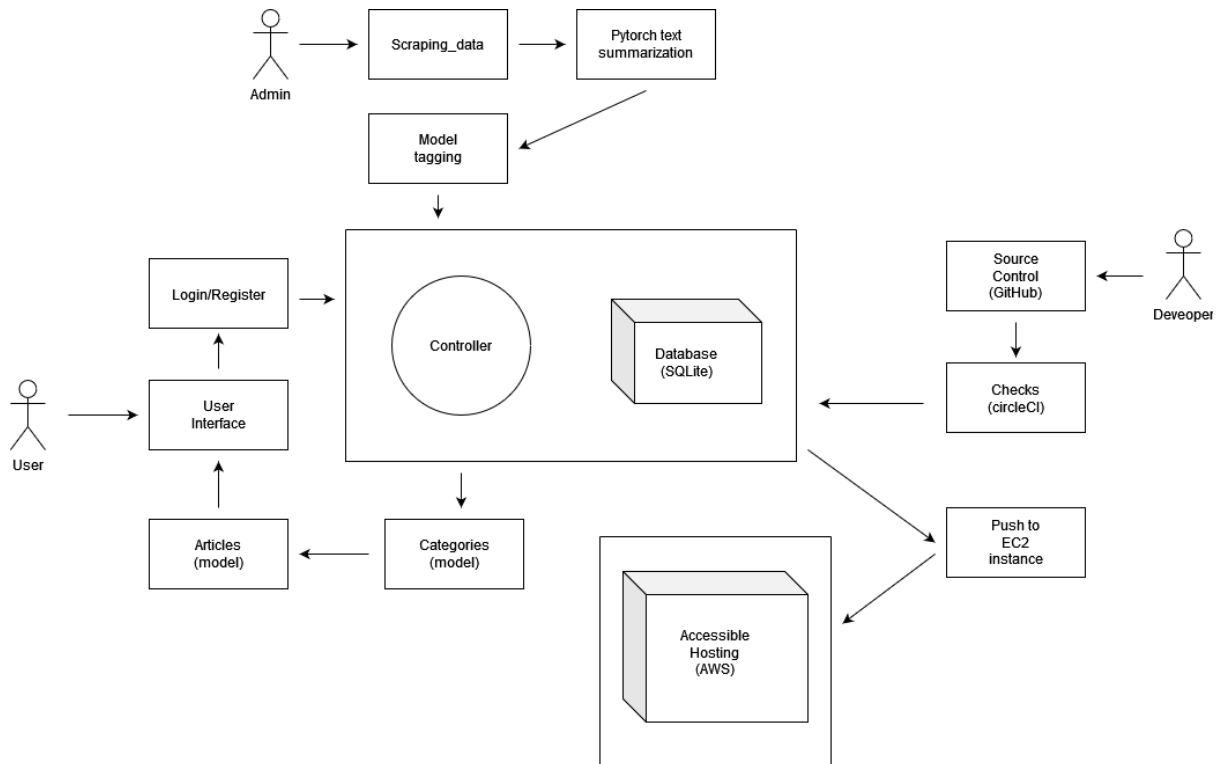
2.1.5. Usability Requirements

Usability requirements for the site are important as providing an easy and intuitive to use application for users is important. There are a number of ways this will be implemented, those including:

1. **Clear and concise navigation:** The application should have a clear and concise navigation structure that makes it easy for users to find the content they are looking for, which plays into the personalization that users will also be given
2. **Consistency:** The application will be consistent in terms of its layout, design and functionality, in order to reduce potential confusion and aid in users being able to understand how to use the application

3. **Error Handling:** The application will handle errors and user mistakes in a clear and user-friendly way, this will include providing helpful error messages in the event something goes wrong while the user is using the application

2.2. Design & Architecture



In the above, you can see the high level of the project. The different actors and components are also there.

To first go over the Developer and explain the components they use, first, for version control, I used github, which connects with the CI/CD pipeline I developed to integrate with my AWS Instance. Once the test script runs, and the necessary checks have been passed, we then run `deploy.sh` and deploy the updated version to be hosted. Screenshot of which can be found below. For further details on testing and the CI/CD pipeline, see 2.5 Testing

```
(venv) ubuntu@ip-172-31-8-204:~/Information-hub$ python3 manage.py runserver 0:3000
Watching for file changes with StatReloader
Performing system checks...

System check identified no issues (0 silenced).
May 14, 2023 - 19:03:53
Django version 4.2.1, using settings 'info_hub.settings'
Starting development server at http://0.0.0.0:3000/
Quit the server with CONTROL-C.
```

Due to the large imports required to run the main branch, I needed to create a separate branch for the production environment that did not have any of the text summarization functionality, something that didn't matter for the tests.

To move onto the Admin and explain the components they use, they first use the scraping function. This is built using bs4 (beautifulsoup) and then incorporates the pytorch text summarization. The text summarization uses the abstractive summarization algorithm built on the t5-base model. This model was trained on the C4 Common Crawl web corpus dataset. Once that work is done, it then moves on to tag and populate the model, in this case, both the Categories and Articles models.

The User on the other hand primarily has access to the user interface. On this page, then can view the categories and subsequent articles, that come from the database to the UI. They can also, however, register a profile with a password which can be loaded and stored into the SQLite database.

2.3. Implementation

```
def category(request, category_id):
    """Show a single category and all its articles"""
    category = Category.objects.get(id=category_id)
    articles = category.article_set.order_by('-date_added')

    for article in articles:
        soup = BeautifulSoup(article.text, 'html.parser')
        headline = soup.find('h3')
        image = soup.find('a href')
        if headline:
            article.headline = headline.text
            article.text = str(soup).replace(str(headline), '')
        if image:
            article.image = image['href']
            article.text = str(soup).replace(str(image), '')

    context = {'category': category, 'articles': articles}
    return render(request, 'info_hubs/category.html', context)
```

The category model below serves as the foundation on which every other function is built on so it is worth talking about. It allows the user to view articles without a new file having to be hard coded in with every article. To do that, I put the majority of that work within the scraping itself. Rather than scraping entire files, getting only what is relevant and being able to parse that for display saves a lot of work. To do this, I used the function above that relies mostly on the BeautifulSoup html.parser to get the relevant tags, such as image and heading, and display to pass that information within the article object.

The scraping is another large amount of code, snippets of which can be found below.

```
def scrape_data(request):
    if request.method == 'POST':
        # URL of the news article to scrape
        headers = {
            'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)'
                          ' AppleWebKit/537.36 (KHTML, like Gecko)'
                          ' Chrome/58.0.3029.110 Safari/537.36'
        }
        url = request.POST.get('url')
        domain_name = tldextract.extract(url).domain

        response = requests.get(url, headers=headers)
        soup = BeautifulSoup(response.text, 'html.parser')

        if domain_name == 'rte':
            maincontent_div = soup.find('div', {'id': 'primary-nav-global'})
            cat_params = ['news', 'sport', 'entertainment', 'business', 'lifestyle', 'culture']
        elif domain_name == 'theguardian':
            maincontent_div = soup.find('ul', {'class': 'pillars'})
            cat_params = ['international', 'commentisfree', 'sport', 'culture',
                          'lifeandstyle'] # The guardian cat_params aren't needed but the rte o
        elif domain_name == 'theamericanconservative':
            maincontent_div = soup.find('div', {'class': 'c-featured-posts__posts'})
        elif domain_name == 'dailykos':
            maincontent_div = [soup.find('div', {'class': 'top-news__primary_news'}),
                              soup.find('div', {'class': 'top-news__secondary_news'}),
                              soup.find('div', {'class': 'top-news__more_news'})]

        links = []
        if isinstance(maincontent_div, list):
            for div in maincontent_div:
                links += div.find_all('a')
        else:
            links = maincontent_div.find_all('a')

        urls = []
        for link in links:
            href = link.get('href')
            full_url = urljoin(url, href)
```

```

def parse_category(url, domain_name):
    # A dictionary to map terms to categories
    category_mapping = {
        'us-news': 'News',
        'world': 'News',
        'football': 'Sport',
        'lifeandstyle': 'Culture',
        'commentisfree': 'Opinion',
        'film': 'Entertainment',
        'tv-and-radio': 'Entertainment',
        'music': 'Culture',
        'science': 'Science',
        'christian-aid-today': 'Opinion',
    }

    path = urlparse(url).path
    categories = list(Category.objects.order_by('date_added'))

    # Convert all categories to lowercase
    for cat in categories:
        cat.text = cat.text.lower()

    # Convert category text to lowercase before comparison
    if domain_name == 'theamericanconservative':
        category_text = 'opinion'
    elif domain_name == 'dailykos':
        category_text = 'news'
    else:
        category_text = next((cat.text for cat in categories if cat.text in path.lower() and cat.text != 'news'), None)

    # Check if a category with the same name already exists
    existing_category = None
    if category_text:
        # Match categories regardless of case sensitivity
        existing_category = next((cat for cat in categories if cat.text == category_text.lower()), None)

```

In the above methods, we see the starting point that kicks off the rest of the methods in the `info_hubs/views.py` file that will gather the data to be processed. Due to their being 4 main data sources that each have their own html to parse, it required setting up a domain name variable that would direct the course of each url through the methods.

The `parse_category` method too was very important to the storing of files with the database, that required a lot of work to get working correctly, the `category_text` often being taken as a string instead of the designated category object. Using a lot of error checking and standardizing within the different urls, I was able to get it working so that this issue doesn't occur when parsing the data.

```

def summarize(article):
    # Load model and tokenizer
    model_name = "t5-base"
    model = AutoModelForSeq2SeqLM.from_pretrained(model_name)
    tokenizer = AutoTokenizer.from_pretrained(model_name)

    # Define summarization pipeline
    summarizer = pipeline(
        "summarization",
        model=model, tokenizer=tokenizer,
        framework="tf", device=device,
        max_length=50, min_length=24,
        num_beams=4, length_penalty=2.0,
        early_stopping=True, no_repeat_ngram_size=2,
        num_return_sequences=1,
        top_p=0.92, top_k=40,
        temperature=0.8,
    )

    # Load text and preprocess
    sentences = preprocess(article)

    # Generate summaries for each sentence
    summaries = []
    summary_text = summarizer(article)[0]['summary_text']
    summary = postprocess(summary_text)
    summaries.append(summary)

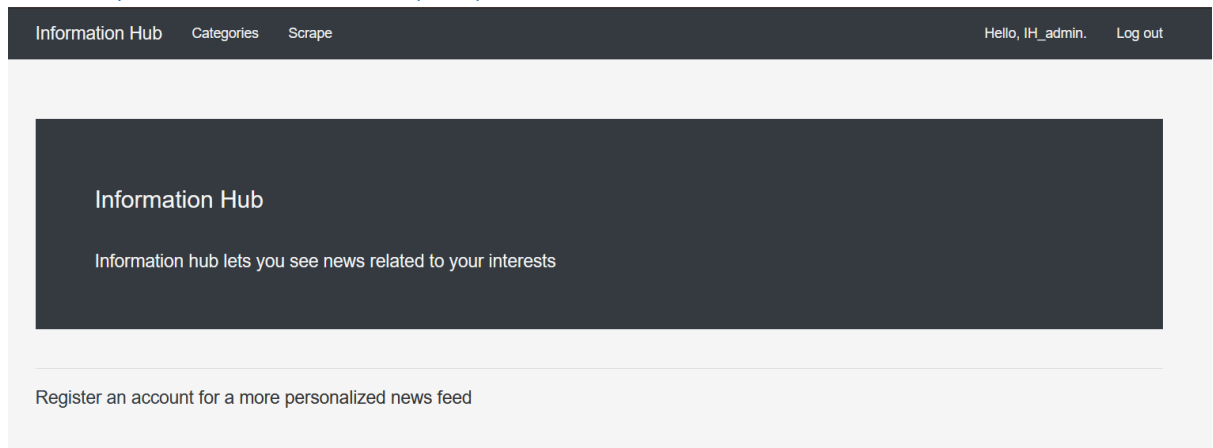
    # Join summaries into a single text
    summary_text = ' '.join(summaries)

    # Print summary
    print(summary_text)
    return summary_text

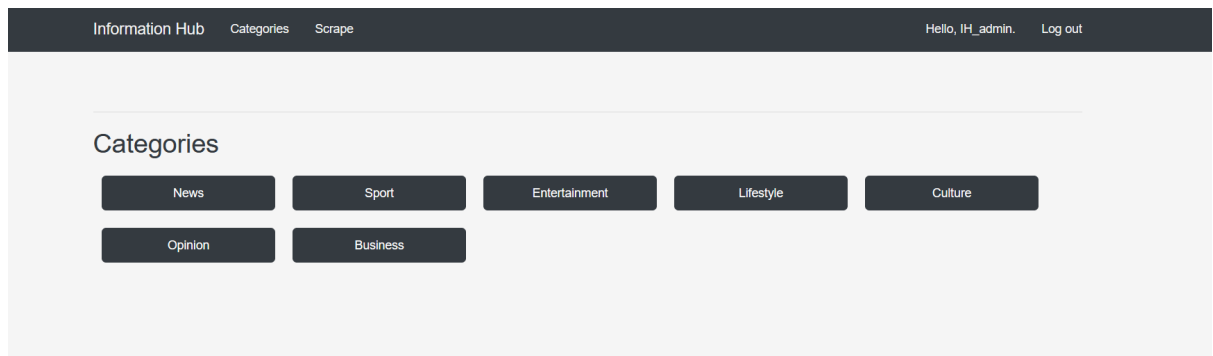
```

The above method uses pytorch to get the text summarization from the parsed articles gotten from the web scraping.

2.4. Graphical User Interface (GUI)



The above is the index page for the site, showing the banner and relevant options for an admin user



The above is the categories page for the site and allows the user to specify which type of articles they want to read.

Information Hub Categories Scrape Hello, IH_admin. Log out

Opinion

What happens when leaders disregard the truth? Putin and Trump are about to find outPeter Pomerantsev

Evidence was meant to destroy wrongdoers as sunlight does a vampire . find the evidence, the logic went, and the powerful could be shamed and brought to justice...

Link: <https://www.theguardian.com/commentisfree/2023/may/14/what-happens-when-leaders-disregard-truth-putin-trump-about-to-find-out>
Author: Peter Pomerantsev
Site: The Guardian

Kate Moss is now into gardening? I love it when ravers become boringEmma Beddington

How to spend it rebranded as "HTSI" last year, presumably in spirit of "quiet luxury" John sutter: I am a mere rubbernecker in the HTSI universe....

Link: <https://www.theguardian.com/commentisfree/2023/may/14/kate-moss-is-now-into-gardening-i-love-it-when-ravers-become-boring>
Author: Emma Beddington
Site: The Guardian

Westminster forgot its promises to 'coastal communities', and left them to rotJohn Harris

A decade ago, politicians and journalists were suddenly confronted with an issue that had always festered at the edge of the national conversation . the dire state of england's seaside towns, and their deep social problems....

Link: <https://www.theguardian.com/commentisfree/2023/may/14/britain-seaside-towns-deprived-areas-investment-levelling-up>
Author: John Harris
Site: The Guardian

The above shows the category page that specifies which articles to show to the user. In the above case, it is the opinion page.

Information Hub Categories Scrape Hello, IH_admin. Log out

Scrape

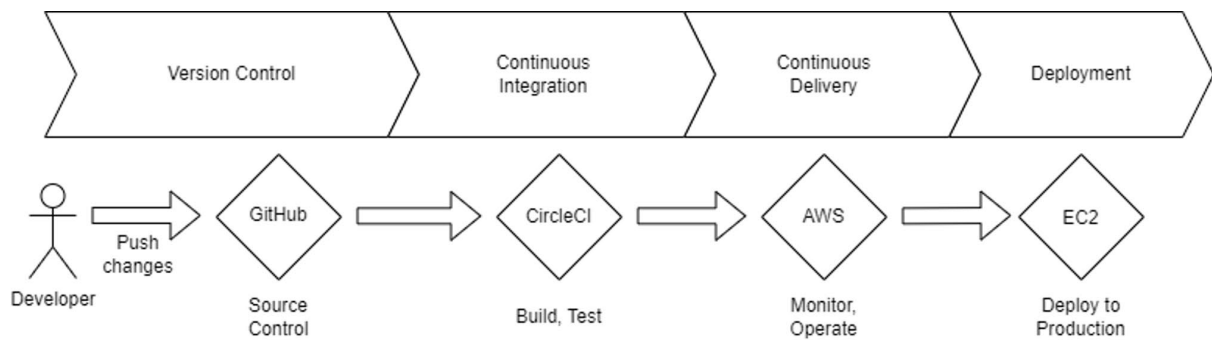
Enter the URL to scrape:

The above is the scrape page where an admin can specify a url to scrape, thus getting the data for the methods to run.

2.5. Testing

```
▼ ✓ run-tests
30 OK
31 Destroying test database for alias 'default'...
32 Name                               Stmts  Miss  Cover
33 -----
34 info_hub/__init__.py                 0      0  100%
35 info_hub/settings.py                 19      0  100%
36 info_hub/urls.py                     4      0  100%
37 info_hubs/__init__.py                 0      0  100%
38 info_hubs/admin.py                   4      0  100%
39 info_hubs/apps.py                     4      0  100%
40 info_hubs/migrations/0001_initial.py  6      0  100%
41 info_hubs/migrations/0002_rename_synopsis_article_and_more.py  4      0  100%
42 info_hubs/migrations/0003_alter_article_options_and_more.py  4      0  100%
43 info_hubs/migrations/0004_alter_category_options.py            4      0  100%
44 info_hubs/migrations/0005_alter_article_category.py            5      0  100%
45 info_hubs/migrations/__init__.py     0      0  100%
46 info_hubs/models.py                  21      0  100%
47 info_hubs/tests.py                   96      1   99%
48 info_hubs/urls.py                     4      0  100%
49 info_hubs/views.py                   244    155   36%
50 manage.py                              12      2   83%
51 users/__init__.py                     0      0  100%
52 users/admin.py                         1      0  100%
53 users/apps.py                           4      0  100%
54 users/migrations/__init__.py           0      0  100%
55 users/models.py                         1      0  100%
56 users/tests.py                         33      0  100%
57 users/urls.py                           5      0  100%
58 users/views.py                         19      0  100%
59 -----
60 TOTAL                                 494    158   68%
61 CircleCI received exit code 0
```

For testing, I primarily relied on CircleCI to run the testing scripts already developed. The testing files within the project contain both unit tests and integration tests, covering all of the functionality within users and most of the risky functionality in info_hubs/views.py. info_hubs is primarily low due to the extensive error checking and potential negative testing, the extent of which to cover in unit tests would be unneeded. Some of the primary testing done in



In the above, you can see the pipeline diagram I used in the project. The circle CI script can also be found below, which triggers the build and test tasks.

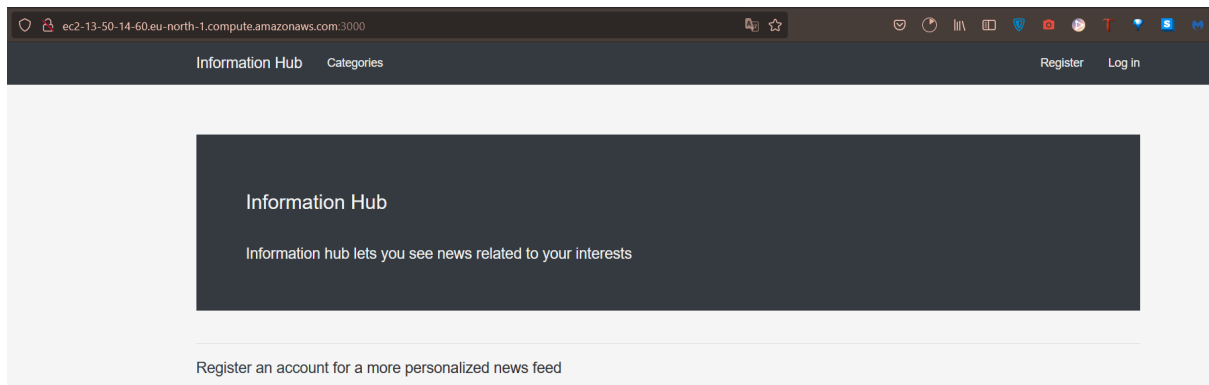
```

jobs:
  build:
    working_directory: ~/Information-hub
    docker:
      - image: circleci/python:3.9.6-buster
    steps:
      - checkout
      - run:
          name: update-pip
          command: |
            sudo apt-get update && \
            sudo apt-get install -y python3-pip && \
            sudo pip3 install --upgrade pip
      - run:
          name: install-virtualenv
          command: sudo pip3 install virtualenv
      - restore_cache:
          key: dependency-cache-{{ checksum "requirements.txt" }}
      - run:
          name: install-dependencies
          command: |
            virtualenv venv
            . venv/bin/activate
            pip install -r requirements.txt
      - save_cache:
          key: dependency-cache-{{ checksum "requirements.txt" }}
          paths:
            - ./venv

```

```
test:
  working_directory: ~/Information-hub
  docker:
    - image: circleci/python:3.9.6-buster
  steps:
    - checkout
    - restore_cache:
      key: dependency-cache-{{ checksum "requirements.txt" }}
    - run:
      name: run-tests
      command: |
        . venv/bin/activate
        coverage run manage.py test
        coverage report
    - save_cache:
      key: dependency-cache-{{ checksum "requirements.txt" }}
      paths:
        - ./venv
```

Once these are run, we can then see the working and deployed version of the project available at the following link: <http://ec2-13-50-14-60.eu-north-1.compute.amazonaws.com:3000/>



2.6. Evaluation

3.0 Conclusions

There are some advantages to the project over other news curation platforms but also some limitations that could be patched but at the time of writing, have not been, to start off with the good news.

The web scraping in particular took a lot of work to get to a state of functioning the way it is at present, the mix between the work on the category model and the integration of it with the scraping of the 4 different sites required a lot of trial and error but for it to work as seamlessly as it does now was quite a feat.

Another part that came out good was the testing and the CI/CD work. After a prominent issue with the instance running into errors, both from the initial Jenkins implementation and then the pytorch text summarization, being able to make a production environment that the tests can use was an advantage to the development going forward.

Some of the limitations for the project include the limited functionality within the User model. I had wanted to develop more user specific methods and actions, but those relying on the initial work of the scraping, parsing, ML work etc. made that work pretty consistently distant.

4.0 Further Development or Research

Given more time to work on the project, my main focus would be on setting up a task scheduler through MRQ to turn the web scraping and article loading into a daily function. Along with that, I would also write a method to delete all articles present in the database that are a week old or more, and run that as a daily scheduled task too.

A different part of functionality I would add is a daily brief with the users. The user model at present is pretty bare bones and having the option to do more for those models and the site in general would be beneficial.

5.0 References

6.0 Appendices

6.1. Project Proposal

National College of Ireland

Project Proposal

Information hub: web application for news
curation

30/10/2022

BSHCSD4

Software Development

Academic Year 2022/2023

Aaron Flynn

x19404024

x19404024@student.ncirl.ie

Contents

1.0	Objectives.....	21
2.0	Background	21
3.0	State of the Art.....	22
4.0	Technical Approach.....	22
5.0	Technical Details	23
6.0	Project Plan	24
7.0	Testing.....	25

7.0 Objectives

The aim of the project is to build a web application that curates news articles based on topics that a user will enter when setting up their profile. This project will serve several functions that would aid a user when searching for information on the web. With a large portion of the news being curated for us, through large social media sites like twitter and facebook, many of us go through great lengths to find unbiased reporting, which typically involves a lengthy process of opening many sites, preferring certain publications for different areas of reporting.

This application will curate the many different sites a user may have preferences for, or a paid subscription for, and display the headlines and text based on the categories a user will pick when setting up their profile. This will solve a number of issues that I, and a number of others experience when trying to consume news. It is a workflow that I think will solve several issues currently facing people today.

8.0 Background

I chose to undertake this project as when looking for problems in the real world that could be solved, I focused in on smaller details that many people would face. Reading the news being one of them, particularly as the process of how people consume news has so radically

shifted since the internet and commentary has taken over a large part of the demand once taken by legacy news sites where reporting and journalism took precedence over bias and marketing. In effect, it is an area in the current way people consume news media that could do with an overhaul.

I plan to meet the objectives set out in section 1.0 by developing a web stack using a few different technologies, a number of which I have identified, such as Django, celery, flask, TensorFlow, and others which require more research to land on, like a distributed task queue for scraping the web where APIs don't exist. Later on in the proposal you can see a more thorough exploration of the project plan with timelines for development and a further discussion of the technologies and languages going to be used for building the project.

9.0 State of the Art

Other news curation platforms typically follow the newsletter approach, meaning on a daily or weekly basis there is a manual approach to getting the highlights of the week and giving it to the end user that way. This is a common practice for platforms that aim to capture a certain demographic, such as reclaim the net or other politically minded platforms, or those seeking a broader audience like the morning brew.

My project doesn't plan to follow the newsletter approach, as that would lack any real technological implementation, but rather a web-based application that will curate news based on user input, such as categories or areas of focus that would interest them, say technology over market commentary, with a plan to use an ML solution such as TensorFlow to categorise where an article would fit.

10.0 Technical Approach

I will be using the agile development framework as an approach to the project as it is the framework I have the most familiarity with, and also the most success using. With that, I will have dedicated research stories to determine requirements of the project, there will be a certain amount of flexibility with deadlines as a result of the nature of the project with certain areas of implementation being on the short side and others requiring research, comparison use and finally implementation and review. For example, I have chosen to use

TensorFlow as my machine learning library, but I am leaving myself open to the possibility that another library, such as PyTorch, may be better suited to my development goals.

In identifying requirements like chief libraries, datasets, web scraping technologies and so on, this will chiefly be done by research into existing technologies in the sphere, recommendations, how tutorial heavy certain pieces are and so on. I have a chief idea of what kind of web stack I need to have for a project such as this but am open to change if certain requirements are better met with different solutions.

In breaking down the project into tasks, milestones and activities, I have chosen to use Jira as it is the issue tracking solution I had learnt to use during my internship and it allows you to see where an issue currently is in the process of development, what needs to be done and also get a feel for the state of the project at a glance with the roadmap feature. The attached roadmap can be found in the project plan.

11.0 Technical Details

This project will be written chiefly in Python, but will use other languages like HTML and CSS for the web design elements, and it's principal libraries will be Django, celery, flask, tensorflow.

I decided to use Python as I had a lot of experience with it over the internship and used, although briefly, the libraries under consideration,

I decided to use Python as it would allow for a quicker development time and there are a number of tutorials out there for coding this type of project, a concern was that at a certain level or area of development in another language, I would be a bit stranded with little resources available to work through, but Python and broader web application design using Django in congruence with HTML and CSS is a common approach so a number of the pitfalls and potential problems I could run into have been documented and seen prior.

Celery is an asynchronous task queue that is used for lining up scheduled tasks, which will be a major part of getting news articles from the web to be processed and displayed for the end user.

Flask is a micro web framework that I plan to use in congruence with Django if certain elements I aim to write in Django are not supported, it should fill in the gaps where Django can't perform adequately to the ends of the project.

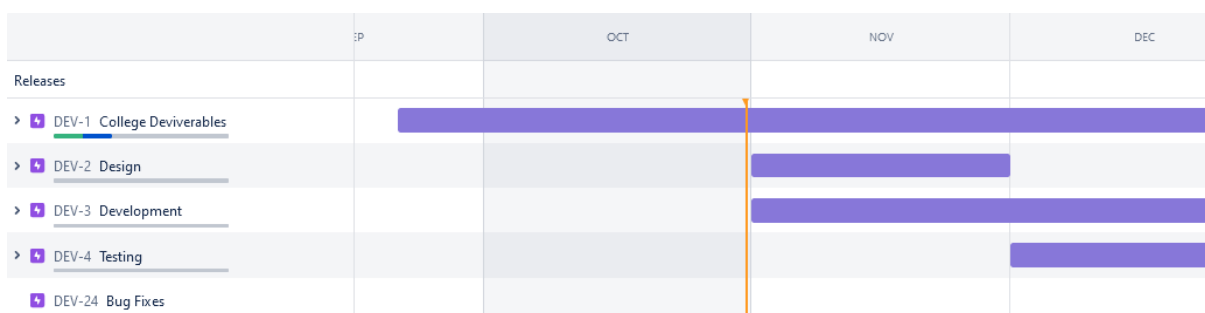
TensorFlow is a Machine Learning library I plan to implement in the project and will serve as the primary way on how different pages will be categorized. This will likely be done with a language processing model that I can train with public data sets and is one of the more common and more reliable uses of machine learning, but having little experience in this area, there may be another model or library that would better suit my needs, so this is open to change but is the library I'm focusing on in the earlier stages of development.

The project will be done in several stages so a few of these libraries, like TensorFlow, likely won't be used extensively or will exist only in a minor form when it comes to the mid point implementation, but further details on how these will be implemented will be found in the project plan when they are addressed.

12.0 Project Plan

The chief of the project plan is done using Jira, the roadmap can be found below and the link to the page is

<https://computing-project-nc.atlassian.net/jira/software/c/projects/DEV/boards/1/roadmap>



The roadmap is not complete as we are still in the early stages of development, and a number of stories will pop up as development continues, whether bug fixes or certain key areas of development not being, as of now, foreseen.

The project plan is broken down into 5 main epics comprised of stories that would need to be done in roughly the timeline given, though there are lee ways given for development and

testing due to the nature of tickets depending on work done previously and some work not yet broken down.

The first epic is the project deliverables, that serve as more placeholders for information and will be developed based on, after the completion of the proposal, work done in Design and Development.

The second epic involves the Design of the project and the stories there at present involve UX and UI design and research. Having only a surface level of how a site should work or the main things to look out for when crafting a good user experience, there would necessarily be research to be done before getting pen to paper. Once design is done, we go onto development which is a large epic so it is tracked from now to the end of April to give enough time for final testing and bug fixes that will pop up.

The third epic will be the largest, as mentioned, and will comprise of research into the different technologies, libraries and so on that will be required for the project to be completed. There are a few I have established as necessary, such as Django, while others are still in the research phase, such as which distributed task queue I should use, whether mongoddb would be the best choice for a document-oriented database program, and others. This being the largest epic, it is necessary that it will be directly linked to the next two epics, testing and bug fixes.

The fourth epic is testing, which will cover testing the functionality done during development, and the creation of a testing suite to cover the foundations of the project. With the project, it will not take too long to get the barebone skeleton done, to have the primitive functionality and web page up. So, I have given the date of December first to start that, as there will be enough done by then to start work on a testing suite, and the development of unit tests to cover the functionality. I have given the date of May 14th for the end of testing, as that is the point of the final Implementation, and I am open to the possibility that the entire project will not have 100% unit test and integration test coverage.

The final epic is bug fixes which is empty at the moment, but will no doubt be quite large by the completion of the project. This will be associated with both tickets in development and tickets in testing, as in both cases, bugs will be found and will have to be cordoned off and made note of for fixing. I have given bug fixes the end date of May 14th also, as just like testing, I doubt I will be able to cover every bug found in the process of testing.

And to reiterate the epics are filled with as much detail as I have at present but will continue to be populated and given detail further on into the mid point implementation and beyond into the final implementation.

13.0 Testing

The testing of this project will be handled primarily by extensive unit testing and integration testing. A number of features present in any web-based application will be sensitive to

breaking when new code is added, even small changes to a dataset or optimization can cause an entire page to break so extensive unit testing can be carried out along side development to check whether an addition or deletion causes issues or not. I plan to use Jenkins as my testing automation suite which will allow me to test each stage of development, independent of requirements on my system. This will be an important part of developing the project as the local development environment may return false positives or false negatives depending on factors of oversight.

For the development of integration testing, this too will be done with a local development suite, monitored using Jenkins and tag-based building with github.

I will also follow general quality assurance processes with manual testing being done on the front end before a feature is released, and that certain core parts of the system are being automated with scripts using selenium and other automated web browser testing.

The development of a throughout QA suite is beyond the scope of this project but key parts that ensure that the system continues to run as intended and that an unsecure piece of code doesn't break the project will be implemented to ensure nothing will cause the project to be unusable at key points in development, such as approaching the midpoint implementation or finishing up with the final implementation.

13.1. [Ethics Approval Application \(only if required\)](#)

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: Aaron Flynn

Student ID: X19404024

Programme BSHCSD4

Year: 4

Module: Computing Project

Project Title: Information Hub: web application for news curation

Please circle (or highlight) as appropriate

This project involves human participants	Yes / No
--	----------

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N
Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes / No
The project makes use of public secondary dataset(s)	Yes / No
The project makes use of non-public secondary dataset(s)	Yes / No
Approval letter from non-public secondary dataset(s) owner received	Yes / No

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) **does not need ethics permission**, but **needs a letter/email from the copyright holder** regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.


An approval letter/email from the owner (e.g. institution, company, etc.) **of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration**. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

For use of RTE's articles to be used in the news curation application, I got the following letter of agreement. These articles will be used to populate the front page of the application. A synopsis will be provided, along with categorization, and credit to the source material will be provided.

Re: Copyright Request

 **Info** <Info@rte.ie>
11/25/2022 1:50 PM

To: Aaron Flynn

CAUTION: DO NOT CLICK links or attachments unless you recognize the sender and know the content is safe..

Hi Aaron,

Thank you for your email.

Please see response below from the News Online team -

Dear Aaron, RTE News is happy for you to use its online material for the purposes of the described academic work, providing RTE News is credited at all times and the assets are only used for the remit of the project.

Kind regards, Celine

RTE Information Office
Email: info@rte.ie Website: www.rte.ie

From: Aaron Flynn <x19404024@student.ncirl.ie>
Sent: 24 November 2022 13:52
To: Info <Info@rte.ie>
Subject: Copyright Request

You don't often get email from x19404024@student.ncirl.ie. [Learn why this is important](#)

Dear RTE,

Hello, my name is Aaron Flynn, and I am a student at the National College of Ireland. I am currently developing a news curation site as part of the course work to be submitted for my final year computing project and I would like to use the news articles published on <https://www.rte.ie/> as a data source to include in the curation.

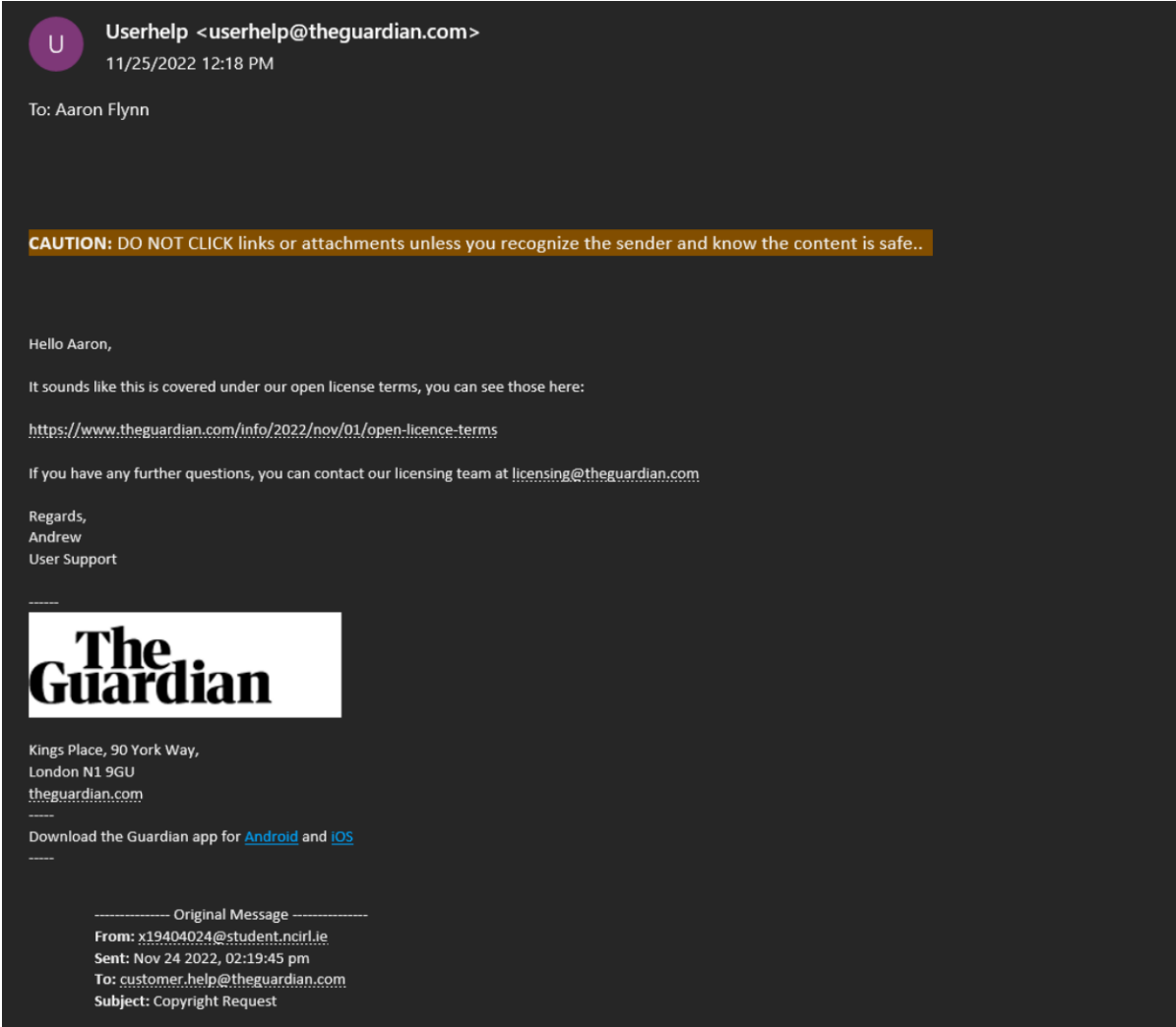
The project will not be monetized, and I will follow the normal procedure of using small text snippets of your articles followed by links to the article. Credits to the original author and source/site will also be appended, so I will not be republishing any material in full.

If you agree to provide me with permission to use your material in my project, I would need a letter of agreement that confirms that I have been granted use for processing, analysis and public dissemination of the material.

Thank you for the consideration of my permission request and if there's any questions related to the use of the material or any other concerns that may come up, feel free to email me.

Kind Regards,
Aaron Flynn

For use of The Guardian's articles to be used in the same capacity as RTE's articles, I got the following letter of agreement as long as I stick to their open license agreement.



U Userhelp <userhelp@theguardian.com>
11/25/2022 12:18 PM

To: Aaron Flynn

CAUTION: DO NOT CLICK links or attachments unless you recognize the sender and know the content is safe..

Hello Aaron,

It sounds like this is covered under our open license terms, you can see those here:

<https://www.theguardian.com/info/2022/nov/01/open-licence-terms>

If you have any further questions, you can contact our licensing team at licensing@theguardian.com

Regards,
Andrew
User Support

The Guardian

Kings Place, 90 York Way,
London N1 9GU
theguardian.com

Download the Guardian app for [Android](#) and [iOS](#)

----- Original Message -----
From: x19404024@student.ncirl.ie
Sent: Nov 24 2022, 02:19:45 pm
To: customer.help@theguardian.com
Subject: Copyright Request

Their open license agreement is described here. The open licence terms will be used for every publication, regardless of if they require it or not.

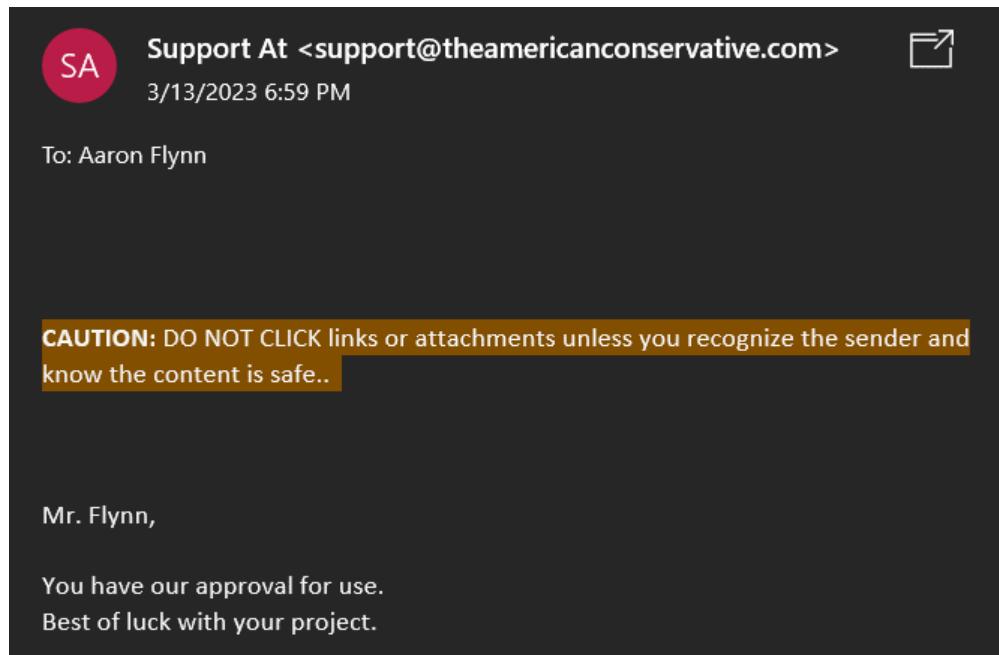
<https://www.theguardian.com/info/2022/nov/01/open-licence-terms>

In summary, you are permitted to use our content without payment of a fee for the following purposes:

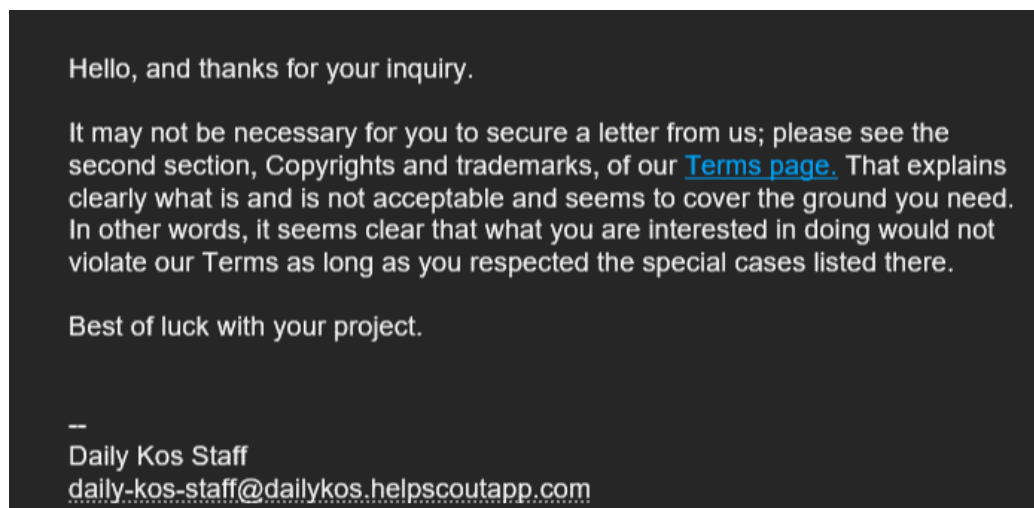
- In a short quotation of a maximum of 100 words in length excluding advertising and endorsement purposes; or
- By students in coursework or dissertations created in the course of full- or part-time study at a recognised school or university or other place of learning provided that such coursework or dissertation shall not be released for external publication in any form; or
- In church or village or parish newsletters provided that in the case of articles you shall not exceed 500 words; or
- Use of articles in personal, non-commercial blog- and websites subject to:
i) a link back to theguardian.com website; and ii) a limit of 500 words.

You may edit our articles on the condition that their integrity is maintained

For use of The American Conservative's articles to be used in the news curation application, I got the following letter of agreement



For use of The Daily Kos' articles to be used in the news curation application, I got the following agreement. My use of their content will be following their terms, which can be found here: <https://www.dailykos.com/terms>



CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form	Yes / No
OR	
Citation and link to the web site where permission is granted – provided in this form	Yes / No

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

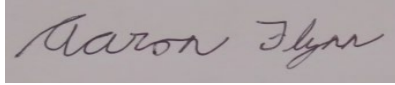
- Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study
- Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation
- No confidential data will be used for personal advantage or that of a third party

Informed consent

- Consent to participate in the study has been given freely by the participants
- participants have the capacity to understand the project goals.
- Participants have been given information sheets that are understandable
- Likely benefits of the project itself have been explained to potential participants
- Risks and benefits of the project have been explained to potential participants
- Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress
- The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty
- Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)
- Participants have been informed of potential conflict of interest issues
- The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature:

A rectangular box containing a handwritten signature in cursive script that reads "Aaron Flynn".

Date:

31/03/2023

Appendix I

1) Fingal Open Data: <http://data.fingal.ie/About>

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

2) Eurostat: <https://ec.europa.eu/eurostat/about/policies/copyright>

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

- the source is indicated as Eurostat

- when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Appendix II

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

13.2. Reflective Journals

Student Name	Aaron Flynn
Student Number	X19404024
Course	BSHCSD4
Supervisor	Adriana Chis

Month: October

What?

Reflect on what has happened in your project this month?

This month I completed the project pitch and the draft for the project proposal. To do these, I spent a lot of time thinking about potential project ideas that would fulfil the requirements and would also be an engaging piece of work that would fill up the many months until May. The Proposal included the focusing in on the different technologies I would need to focus in on to build the news curation site and the many different elements comprised in it.

So What?

Consider what that meant for your project progress. What were your successes? What challenges still remain?

These two deliverables were large components for the project as now that an idea and general roadmap for the development of the project created, I can now begin work on getting a skeleton for the project created.

The project proposal in particular answered many questions regarding what the next several months of the project will look like and how it will be completed. With a Gantt chart made and the technical approach fleshed out, there are now more avenues

The challenges that remain for the project is to clear up the few amendments needed from the reception of the project pitch and to focus in on the few chief libraries needed for the technical development. I have already completed a number of tutorials in Python, Django and Celery though now the chief challenge for the project is to get a foundation for the project down.

Now What?

What can you do to address outstanding challenges?

To address the outstanding challenges laid out above, the first step will be to get started on the UX and UI design so that I can start coding the application and the few necessary components needed for the foundation. Though, to make actual progress on that goal, I will need and am continuing to do, more research

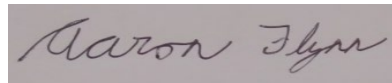
on libraries and potential approaches needed to build the application and what it will look like.

The second challenge that will need to be addressed is the amendment needed for the pitch. I plan to do this

soon by setting up a one on one meeting with my supervisor and discussing potential routes for going through

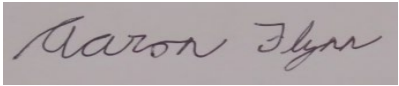
them and arriving at a conclusion.

Student Signature

A rectangular box containing a handwritten signature in cursive script that reads "Aaron Flynn".

14.0 Student Name	Aaron Flynn
Student Number	X19404024
Course	BSHCSD4
Supervisor	Adriana Chis

Month: November

What?	
<p>This month I worked on refining the project requirements further, the chief aspect of this was working out the possible copyright infringements that would occur during the development. To do this, I contacted several news agencies I wanted to use in the news curation process and got approval to use their sites in a way that works with their policies. Lastly, there was further work started on meeting the system requirements needed for mid-point presentation, this included work on the UML diagrams and also a start on getting a prototype ready.</p>	
So What?	
<p>The letters of agreement from the copyright holders was a large positive for my project in regards to progress and addressed many concerns around the project's viability going forward. Having definitive sources of documents for the prototype is a large part of the project's concept so having those sources set in stone is a large bonus. The work started on the system requirements was also an eye-opening experience and put in perspective the multitude of work required outside the prototype and awareness of that now will allow me to put in more reasonable deadlines over the next few weeks.</p>	
Now What?	
<p>The outstanding challenges left in the project mostly concern due dates and the reasonable action to clear them up will have to do with dedicating more hours to them going forward, an action that will be simple enough to act on now that other module's CAs are closing off, allowing for more time to spend on the nitty gritty details of the project.</p> <p>Another outstanding challenge that touches on the last issue is the idea of technical knowledge in these domains. A number of the tools and frameworks I will be using in the project are not my preferred or knowledgeable areas, so time management again comes into effect at addressing these challenges</p>	
Student Signature	

Student Name	Aaron Flynn
Student Number	X19404024
Course	BSHCSD4
Supervisor	Adriana Chis

Month: January

What?

This month I worked further on developing the project more and incorporating more functionality that will get the project closer to where it needs to be by the deadline. Primarily, I worked more in the Django framework the project is built on by developing the templates more and optimizing the functions they call on. Beyond that, I added in more functions to reach the requirements I set out to achieve and developed more of a testing framework for the functionality present in the mid-point.

Another larger part I have been working on is incorporating CI/CD (Jenkins) for the project to run the unit tests and optimize my work flow going into the final few months where they will be needed for a more rapid and successful development work flow.

So What?

Completing more parts of the project at this stage of development and expanding my reach when it comes to the requirements will allow me to catch any issues that may arise early, before they become a burden where time restrictions will impact me even more. Beyond that, further work on completing the foundation of the project and expanding the reach of unit tests for them, will allow for the inclusion of the more difficult aspects of the project which will no doubt lead to a lot of bugs during their implementation.

The inclusion of work on CI/CD is also a necessary aspect to this as without it, there is the chance that a broken branch or piece of code may be included within the existing code base and lead to bugs that will take far longer to catch had there not been that integration process in place, which will allow me to catch a problem or incorrect piece of code before it makes its way onto the main branch.

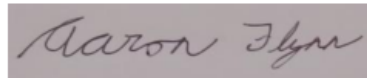
Now What?

Going into next month, my focus will be on completing the unit tests and getting a code coverage of 85%+, and in addition to that including integration tests that cover key functionality such as the web scraping and validation of data coming in, such as checks for images, the presence of a headline and synopsis etc.

I aim to get the CI/CD done very early into next month and create several test branches to make sure it works as intended. The completion of these two aspects of the project will allow me to start adding the more volatile aspects of the project, such as the synopsis generator, with more confidence.

In addition to this, I will also be sending out more requests to organizations to use their articles so I can expand the data present on the site and allow more variability for the key news categories, expanding the range and value of the project as a whole.

Student Signature

A rectangular box containing a handwritten signature in cursive script that reads "Aaron Flynn".

Student Name	Aaron Flynn
Student Number	X19404024
Course	BSHCSD4
Supervisor	Ariana Chis

Month: February

<p>What?</p> <p>This month I worked on fixing a number of issues that came up during the ongoing work on Django. The additions and fixes to the parsing functionality primarily were needed as issues came up when testing. In addition to the parsing work, I also began learning how to use pytorch and working on ways to incorporate it into my project. In the coming month, I should have a working model of it implemented.</p> <p>The incorporation of CI/CD (Jenkins) for the project is also implemented now in a rudimentary way after a number of issues with it's implementation. Now when additions to code are added, issues in other parts can be clearly seen. This will help a lot now that new functionality is being added and errors are occurring, making fixes much simpler.</p> <p>I also sent out a number of emails to procure more letters of agreements from news agencies to expand the data I am able to work with, and ultimately make the project more useful for users. Awaiting responses from those at present so will be able to gauge the data I can work with more accurately as we go into the last few months.</p> <p>In addition to this, this month I began working on the documentation of the project. With a number of pieces now done, I now have enough to write and explain why I made certain choices during development.</p>
<p>So What?</p> <p>Being able to catch the issues that came up on the Django work is a large success and speaks to the worth of putting a lot of effort into the unit tests after the midpoint submission as it will now be able to pay off in more ways, more than just this specific example. The work on the additional functionality that was not present in</p>

the mid point is also a large step in the right direction of getting a complete project, so beginning it now will make sure that these pieces arrive on time and within the deadline.

The functional CI/CD pipeline is another large step in the project. There is still more work and fine tuning to be done within it, the addition of docker into the mix to get a more stable local version running for testing is another possible technology I can add to the project, but that is lower on the agenda given the limited remaining time for development.

The letters of agreement are what make the project usable and without data, it loses much of it's relevance as a product. The previous data I had worked for building the project, but for making it more complete, more data sources are needed.

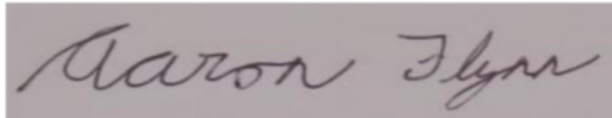
Now What?

The remaining challenges in the project still mean the last few pieces of functionality. After the work the past month getting the foundation of the project as close as I can get it to steady with the use of tests to highlight outstanding issues, I can now work on dedicating more time to learning the new technologies I need to implement them.

The problems so far for doing this remain a knowledge boundary, something I will now have more time to work on bridging as other parts of the project are reaching their close.

The start of writing the documentation is a good step but there is still a number of sections to fill out, and will be occupying more of my time in the coming months.

Student Signature

A handwritten signature in black ink on a light gray background. The signature reads "Aaron Flynn" in a cursive, slightly slanted script.

Student Name	Aaron Flynn
Student Number	X19404024
Course	BSHCSD4
Supervisor	Ariana Chis

Month: March

What?

This month I worked primarily on adding the synopsis generation functionality through pytorch and also added in more functionality for the user's and their profiles. The synopsis generation is still in an early stage of development and occasionally incorrectly generates synopsis with not ideal grammar, so more work will need to be done on that side of functionality.

The user functionality added in allows individual profiles more ability to get news stories related to their interests in their feed, along with an overhaul of the existing user, login and registration pages to fit this new functionality. In addition to this, a number of the test cases had to be updated to fit the new scheme.

I also got more letters of approval and expanded the dataset so more varied sources can be incorporated. This meant creating new web scraping scripts, which are still currently in development, and updating the existing ones due to an update in how I display the data within the Django templates.

So What?

The synopsis generation was going to be the largest undertaking in the project as it was the main technology that I had no exposure to prior to starting, so to have a working model, though still needing work, is a large step in the right direction as will be used to populate the majority of the fields a user will read.

The inclusion and editing of user functionality is also an important step as a large part of the project I wanted to implement was personality and a recommendation system based on the user's interests. This is mostly a standard in news feeds and content curation so adding this in was an important step to making it more useful to the users of the application.

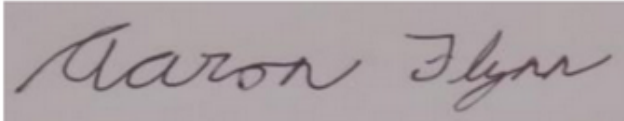
And lastly, the letters of approval and expanded dataset were important as an expanded dataset makes the curation easier and getting more varied voices into a news feed is a must. There are still email correspondences open to organize more approvals, but this is a step in the right direction.

Now What?

Coming into the last number of weeks in the project, the amount of work done so far is on track so finishing off the synopsis generation will be a need. Having it in the early stages of development will be needed as I start to implement a scheduled task so that the news is up to date for users of the application, instead of it being a daily manual process.

The user functionality too was left by the wayside as other tasks took precedent so the overhaul and revamp of that particular area of the project was needed. In addition to that, the architecture and pieces of the project are mostly in place now, so the information necessary to complete more aspects of the documentation are there. That means more accurate UML and other diagrams can be done and included in the documentation. There are still large parts of the documentation not completed so that will be another area of focus going into April.

Student Signature

A handwritten signature in cursive script that reads "Aaron Flynn". The signature is written in dark ink on a light-colored rectangular background.

14.1. Other materials used

Any other reference material used in the project for example evaluation surveys etc.