Summify - Testing

Student 1 Name:	Benjamin Olojo	**ID Number:**	19500599
Student 2 Name:	Przemyslaw Majda	**ID Number:**	20505049

Staff Member Consulted for supervision: Dr. Jennifer Foster

Table of Contents

1. Introduction

1.1 Overview

Our testing strategy involved testing the code that we produced as well as evaluating our method for producing summaries using GPT-3. We opted for a Test-Driven Development approach, with tests for system components being developed early and used regularly throughout the development process to assess the functionality of code.

Once the pipeline for summarising video transcripts had been developed, the ROUGE-1 evaluation metric was employed to evaluate the quality of the summaries produced within our application. This was carried out using the XMediaSum40k dataset from the CLIDSUM dataset[1], containing over 67,000 samples from video dialogues of varying topics.

With the time constraints placed on the project, we opted to automate the testing of modules and approached that we developed ourselves rather than the performance of 3rd-party systems or algorithms, since they had already been tested prior to their release. Instead, we found the results of studies conducted on the summarisation capabilities[2] of GPT-3 and the fact extraction capabilities[3] of the TextRank algorithm.

2. Unit Testing

Pytest[4] was used to automate unit tests for the isolated system components as well as integration tests for the combined system components. Example test cases were curated and used to assess the functionality of the system for the various classes of inputs each component could receive.

2.1 Transcript Segmentation

The transcript segmentation component involves querying the YouTube Transcript API and dividing its response containing timestamped video transcripts into 5 minute segments. Unit tests for this component to ensure that the video transcript were correctly segmented into 5 minute segments based on the video lengths.

[1] https://github.com/krystalan/ClidSum
[2] https://arxiv.org/pdf/2209.12356.pdf
[3] https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf
[4] https://docs.pytest.org/en/7.2.x/

2.2 GPT-3 Completions

The GPT-3 completion component involves the modules for querying GPT-3 through the OpenAI API to rewrite transcript segments, create segment summaries and produce an overall summary. Unit tests were developed to ensure that each module utilising the API to GPT-3 returned the correct response type and exceeded the minimum token length.

2.3 Fact Extraction

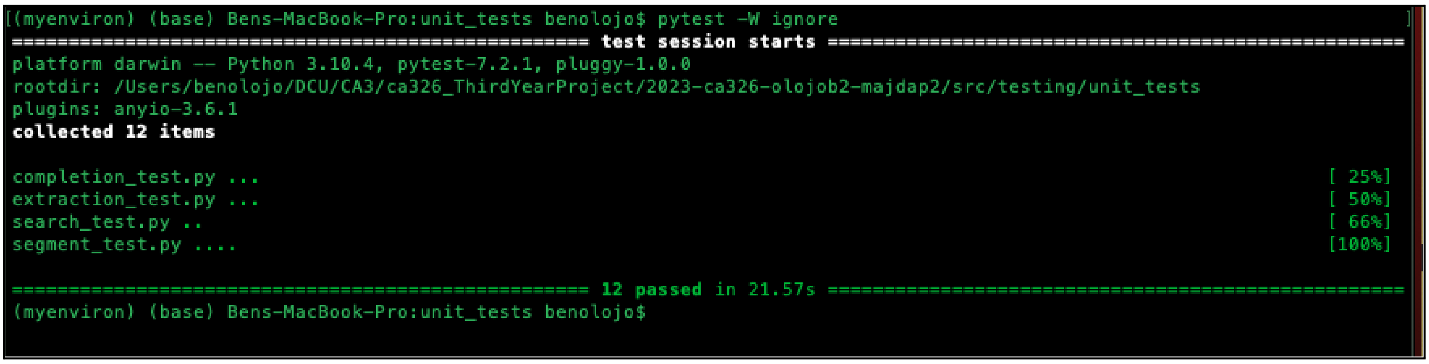
The fact extraction component involves utilising a spaCy model within an NLP pipeline containing custom components for lemmatization and stop word removal, as well as a TextRank extension for utilising the algorithm to extract key terms. Unit tests were developed for this component to ensure that the correct number of unique key terms were being returned by the component.

2.4 Retrieving Links

The component for performing a Google search of the most relevant Wikipedia pages through a Custom Search Engine and the Google Client API was also evaluated to assess its functionality. Unit tests for ensuring that the correct number of unique Wikipedia links were returned were developed.

In Figure 1 below, an example test run using Pytest can be seen, with all unit tests being passed in a reasonable timeframe.

Fig. 1 - Running Unit Tests



```
[(myenviron) (base) Bens-MacBook-Pro:unit_tests benolojo$ pytest -W ignore
=================================================== test session starts ===================================================
platform darwin -- Python 3.10.4, pytest-7.2.1, pluggy-1.0.0
rootdir: /Users/benolojo/DCU/CA3/ca326_ThirdYearProject/2023-ca326-olojob2-majdap2/src/testing/unit_tests
plugins: anyio-3.6.1
collected 12 items

completion_test.py ...                                                                                              [ 25%]
extraction_test.py ...                                                                                              [ 50%]
search_test.py ..                                                                                                   [ 66%]
segment_test.py ....                                                                                                [100%]

=================================================== 12 passed in 21.57s ===================================================
(myenviron) (base) Bens-MacBook-Pro:unit_tests benolojo$
```

3. Integration Testing

The overall system features two main endpoints, `/summarise` and `/links`, that involve a combination of the main components involved in producing summaries for the video transcript and retrieving the most relevant Wikipedia links respectively.

Integration tests were developed for assessing the functionality of the combined components within the running application. The Pytest module and the Requests module were used to create automated tests for sending HTTP requests to both endpoints and evaluating the responses for the different classes of inputs.

3.1 Summarise Endpoint

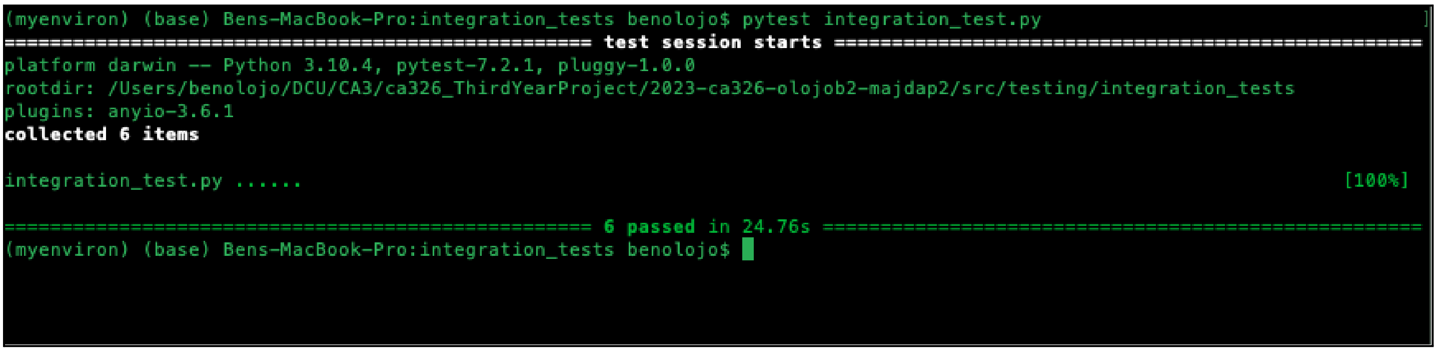
To generate summaries for YouTube videos given the Video ID, the `/summarise` endpoint involves components for retrieving and segmenting transcripts, rewriting segments, summarising segments and producing an overall summary. Integration tests were developed to ensure that the correct response code was returned for various videos and that each response exceeded a minimum length.

3.2 Links Endpoint

To retrieve the most relevant Wikipedia links given the Video ID, the `/links` endpoint involves components for extracting key terms using the TextRank algorithm and performing a google search through the custom Google Search Engine. Integration tests were developed to ensure that the correct response code was returned for various videos and that each response contained the correct number of unique responses.

In Figure 2 below, an example test run using Pytest can be seen, with all integration tests being passed in a reasonable timeframe.

Fig. 2 - Running Integration Tests



```
(myenviron) (base) Bens-MacBook-Pro:integration_tests benolojo$ pytest integration_test.py
==================================================== test session starts =====================================================
platform darwin -- Python 3.10.4, pytest-7.2.1, pluggy-1.0.0
rootdir: /Users/benolojo/DCU/CA3/ca326_ThirdYearProject/2023-ca326-olojob2-majdap2/src/testing/integration_tests
plugins: anyio-3.6.1
collected 6 items

integration_test.py ......                                                                                             [100%]

===================================================== 6 passed in 24.76s =====================================================
(myenviron) (base) Bens-MacBook-Pro:integration_tests benolojo$
```

4. ROUGE Evaluation

4.1 Overview

Using the python rouge-score[5] library developed by Google, the ROUGE-1[6] metric was employed to evaluate the quality of the summaries produced using our method for creating segment summaries and an overall summary based on these segments. The XMediaSum40k dataset from the CLIDSUM dataset was used to select random samples containing dialogue text from a video and a reference (human-produced) summary.

4.2 GPT-3 and TextRank

A Jupyter notebook was created to evaluate the transcript summaries using the ROUGE-1 metric. The dataset was initially loaded and the necessary components from the application for producing the segment summaries and overall summaries were used to produce segment and overall summaries for the XMediaSum40k samples. Using the ROUGE-1 algorithm, these overall summaries were then compared to the reference summaries contained in each sample, with an average score being calculated at the end.

The sumy[7] python library for automatic text summarisation was then used with an implementation of the TextRank algorithm to produce an extractive summary of the entire dialogue sample. These extractive summaries were then compared against the reference summaries using the ROUGE-1 algorithm and averaged at the end.

Once the average scores for each summarisation method had been calculated, they were compared against each other in Table 1 that can be seen below. As can be seen below, our summarisation method outperformed the TextRank summarisation method within each category.

Table 1 - Average ROUGE-1 Scores for GPT-3 and TextRank

Summarisation Method	Precision Score	Recall Score	F-measure Score
Summify/GPT-3	**0.529146434515432**	**0.1658454205011**	**0.2451468111718**
Sumy TextRank	0.347358454032127	0.1540236399225	0.2068555662510

[5] https://pypi.org/project/rouge-score/
[6] https://aclanthology.org/W04-1013.pdf
[7] https://pypi.org/project/sumy/

4.2 GPT-3 and ChatGPT

To further assess the accuracy of our summarisation method, we decided to perform a ROUGE-1 evaluation on 50 samples of the XMediaSum40k dataset using summaries obtained by GPT-3 and ChatGPT[8], a chatbot built by OpenAI on top of OpenAI's GPT-3 family of large language models.

We aimed to evaluate our summarisation approach against another powerful large-language model capable of producing abstractive summaries on entire bodies of long-form text.

Since there isn't an API to ChatGPT available at the time of writing, the input dialogue tests had to be entered manually and the summaries produced were stored within a text file and loaded into a Jupyter notebook.

The average ROUGE-1 scores for the 50 overall summaries produced by GPT-3 were then compared to the average ROUGE-1 scores for the 50 summaries produced by ChatGPT on the entire dialogue text, which can be seen in Table 2 below.

As expected, ChatGPT outperformed our summarisation method using GPT-3 but only with regards to precision, which refers to the proportion of words suggested by the produced summary that actually appear in the reference summary.

The recall score, which our summarisation method performed better in, outlines the proportion of words in the reference summary captured altogether by the produced summary. We expect that this improved recall score is due to the option to alter model parameters provided by GPT-3, allowing for shorter summaries that are more similar to the reference summaries within the XMediaSum40k dataset. The F-measure score, based on both the precision and recall score, is likely to be influenced by this factor as well.

Table 2 - Average ROUGE-1 Scores for GPT-3 and ChatGPT

Summarisation Method	Precision Score	Recall Score	F-measure Score
Summify/GPT-3	**0.525873553222771**	**0.1627747723403**	**0.2448105516143**
ChatGPT	0.555328028968384	0.1457307663409	0.2288665574182

[8] https://openai.com/blog/chatgpt/

5. Informal Evaluation

5.1 Overview

As part of assessing the overall quality and functionality of our application, our development team conducted a series of informal evaluations. These evaluations included manually assessing the quality of transcripts summaries produced by GPT-3 on videos we had watched while taking notes and a heuristic evaluation of the UI to assess its overall usability.

5.2 Manual Summary Analysis

Below are examples of manual analyses of summaries produced by our application on a YouTube video. Once we had watched the video, we assessed the summaries produced and ensured that they effectively conveyed the key information within the video.

Video Information

Title: Kurzgesagt - What is Intelligence?
Transcript: Manually generated
URL: https://www.youtube.com/watch?v=ck4RGeoHFko

Overall Summary

This extract from a video transcript discusses the concept of intelligence and how it is related to creativity, problem-solving, and consciousness. It highlights the differences between humans and animals in terms of their intelligence, with humans possessing a more diverse toolkit that allows them to seek out unique solutions to problems. The video was part of a series funded by the Templeton World charity foundation, which focuses on exploring the nature of intelligence, creativity, and the relationship between the two.

Segment Summaries

0:00:00 - 0:05:00: Humans value intelligence highly and view it as a trait like height or strength. It is the ability to gather knowledge, learn, be creative and form strategies, and is connected to consciousness. Intelligence is a flexible set of skills, including the ability to gather information, save it and use it to learn, and memory. It also includes learning, which is the process of putting together a sequence of thoughts or actions. It can be seen in hardwired or instinct-like reactions, and more complex animals have a wider range of problems they can solve. Creative thinking is the most impressive tool, which involves making new and unusual connections to come up with unique solutions to problems.

0:05:00 - 0:09:44: This extract from a video transcript discusses creativity and problem-solving in animals and humans. It explains that animals use physical tools, such as sticks and coconuts, and plan for the future by hoarding food. It also suggests that humans have an unusually diverse intelligence toolkit, which allows us to shape the planet but also create complex problems. The video was part of a three-part series funded by the Templeton World charity foundation.

Analysis
Within the segment summaries, GPT-3 is able to effectively capture the ideas from the video and describe the general concepts that are outlined. It could be improved by making them slightly more specific by mentioning what each animal does to convey its intelligence, eg. squirrels hoard food, apes use sticks as tools, raccoons memories how to solve problems, etc.

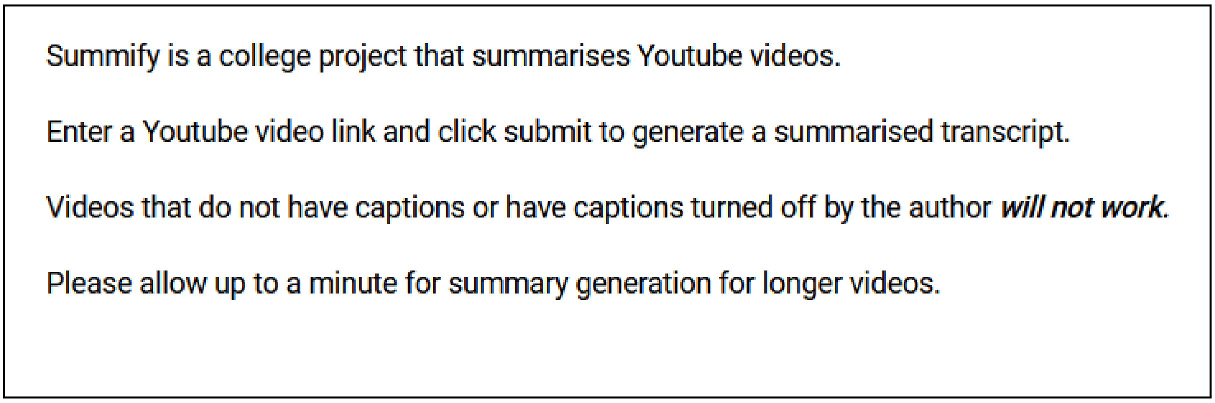
The overall summary effectively highlighted the key theme of the video while using key information from the segment summaries. However, the overall summary includes information about the video sponsor which isn’t relevant to the overall theme of the video.

5.3 Heuristic Evaluation

We conducted a heuristic inspection for producing summaries for a YouTube video with our development serving as UI experts in order to assess the quality of our UI. Through conducting a heuristic inspection, we aimed to find usability issues within our frontend UI that automated test cases would be unable to detect, as well as features that enhance the usability of the UI.

Preventing User-driven Errors
In order to prevent user driven errors, we have designed the home page in a way to maximise the user’s efficiency. There are 4 short lines of instructions on the home page that guide the user on what inputs to give the website to ensure summary generation. These instructions are at the centre of the page to ensure that they will not be missed, and they are concise to encourage the user to read them and not be bored by long instructions. The important text is bold and italicised to grab the users attention.

Fig. 3 - Home page instructions



Summify is a college project that summarises Youtube videos.

Enter a Youtube video link and click submit to generate a summarised transcript.

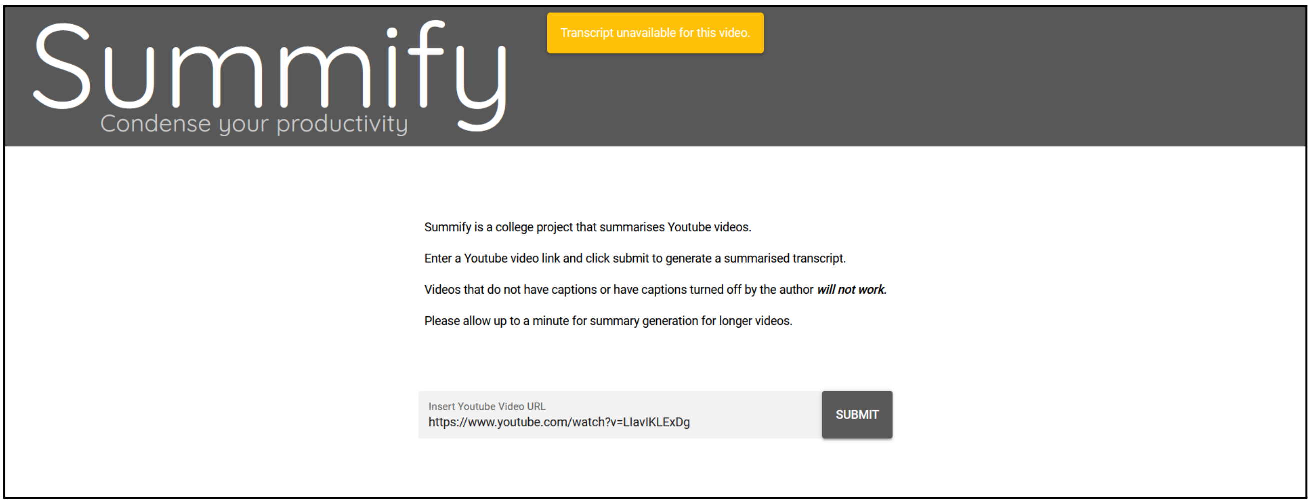
Videos that do not have captions or have captions turned off by the author ***will not work.***

Please allow up to a minute for summary generation for longer videos.

Providing Informative Feedback

Despite the explicit and simple instructions given to the user at the top of the page to enter a YouTube video URL, the input field on the home page is not restrictive and allows users to enter any text, whether it is a YouTube video URL or any other text. It does however provide messages when an incompatible string is entered, giving the user feedback as to what they should enter. If any random text is entered, a message will appear telling the user to enter in a URL. If a YouTube video URL is entered but the video does not have a transcript available, a notification message will appear informing the user of the error, and likewise if the transcript is too long.

Fig. 4 - Error messages



Transcript unavailable for this video.

Summify is a college project that summarises Youtube videos.

Enter a Youtube video link and click submit to generate a summarised transcript.

Videos that do not have captions or have captions turned off by the author ***will not work.***

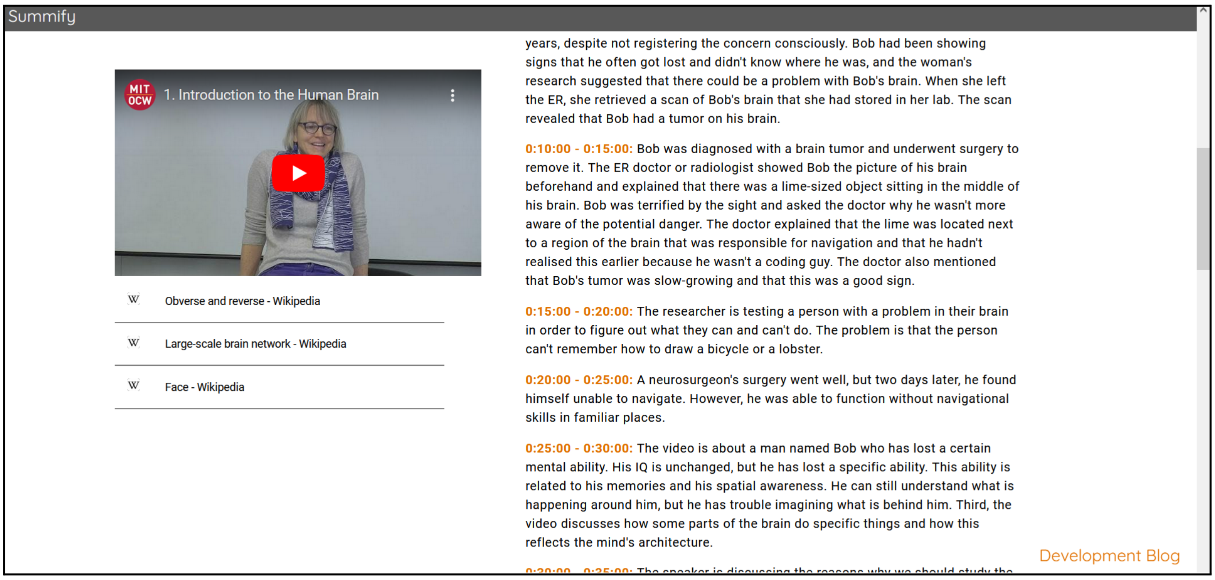
Please allow up to a minute for summary generation for longer videos.

Insert Youtube Video URL
https://www.youtube.com/watch?v=LIavIKLExDg

Providing Short-cuts for Frequent Use

When a user is conducting research and using the site to summarise information, they will likely be summarising multiple videos. To accommodate this, the website's title is always clickable and will redirect the user back to the homepage where they can immediately submit another URL for summarisation. When the user is reading through a summary, the title will shrink down to not obstruct the screen, but still provide an easily accessible link for the user if they would need it.

Fig. 5 - Home page link


Summify
MIT OCW
1. Introduction to the Human Brain
Obverse and reverse - Wikipedia
Large-scale brain network - Wikipedia
Face - Wikipedia
Development Blog

years, despite not registering the concern consciously. Bob had been showing signs that he often got lost and didn't know where he was, and the woman's research suggested that there could be a problem with Bob's brain. When she left the ER, she retrieved a scan of Bob's brain that she had stored in her lab. The scan revealed that Bob had a tumor on his brain.

0:10:00 - 0:15:00: Bob was diagnosed with a brain tumor and underwent surgery to remove it. The ER doctor or radiologist showed Bob the picture of his brain beforehand and explained that there was a lime-sized object sitting in the middle of his brain. Bob was terrified by the sight and asked the doctor why he wasn't more aware of the potential danger. The doctor explained that the lime was located next to a region of the brain that was responsible for navigation and that he hadn't realised this earlier because he wasn't a coding guy. The doctor also mentioned that Bob's tumor was slow-growing and that this was a good sign.

0:15:00 - 0:20:00: The researcher is testing a person with a problem in their brain in order to figure out what they can and can't do. The problem is that the person can't remember how to draw a bicycle or a lobster.

0:20:00 - 0:25:00: A neurosurgeon's surgery went well, but two days later, he found himself unable to navigate. However, he was able to function without navigational skills in familiar places.

0:25:00 - 0:30:00: The video is about a man named Bob who has lost a certain mental ability. His IQ is unchanged, but he has lost a specific ability. This ability is related to his memories and his spatial awareness. He can still understand what is happening around him, but he has trouble imagining what is behind him. Third, the video discusses how some parts of the brain do specific things and how this reflects the mind's architecture.

0:30:00 - 0:35:00: The speaker is discussing the reasons why we should study the