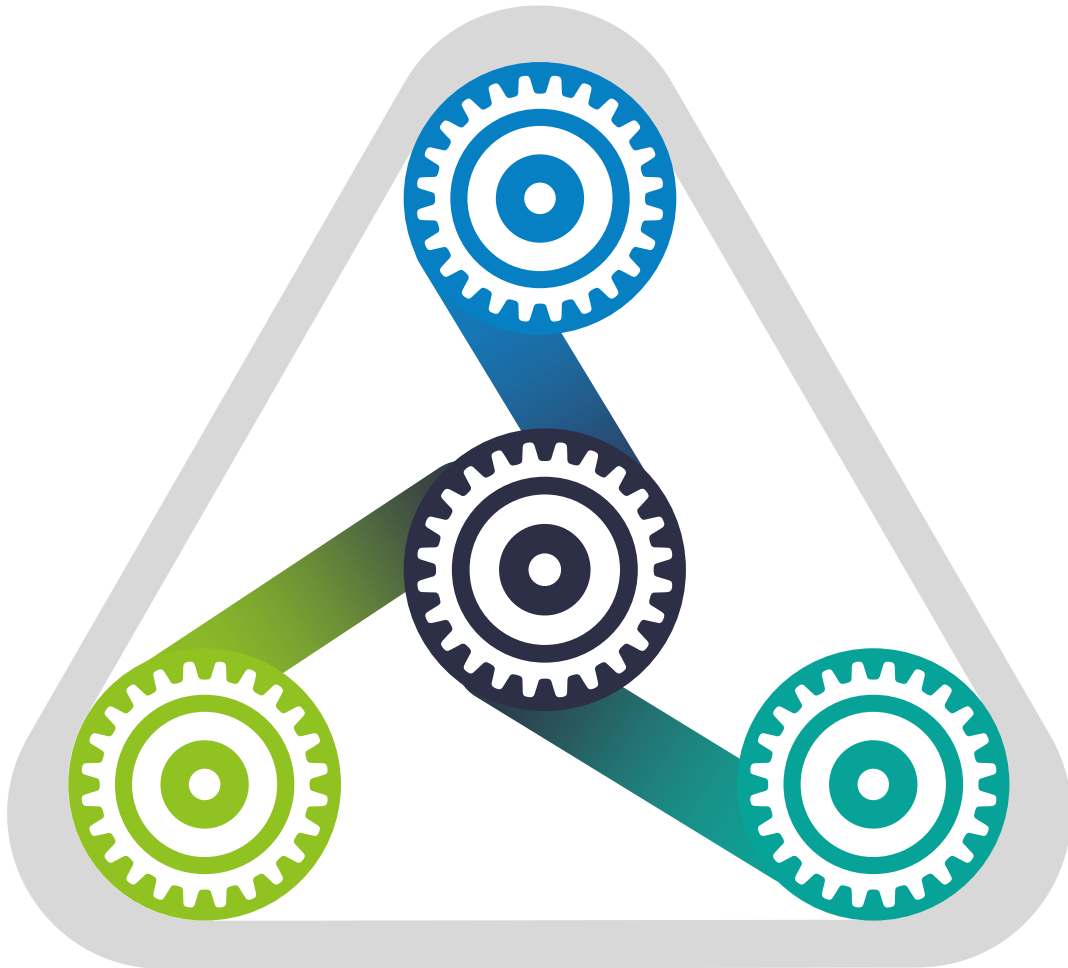


Project Completion Report



Apollo : Text
summarization, keyword
extraction from given
material

Members and roles



Text summarization

Akobir Ismatov

Keyword extraction

Nodir Makhtumov

Web interface

Kamol Umariy

Project Summary

Background and necessity of project selection

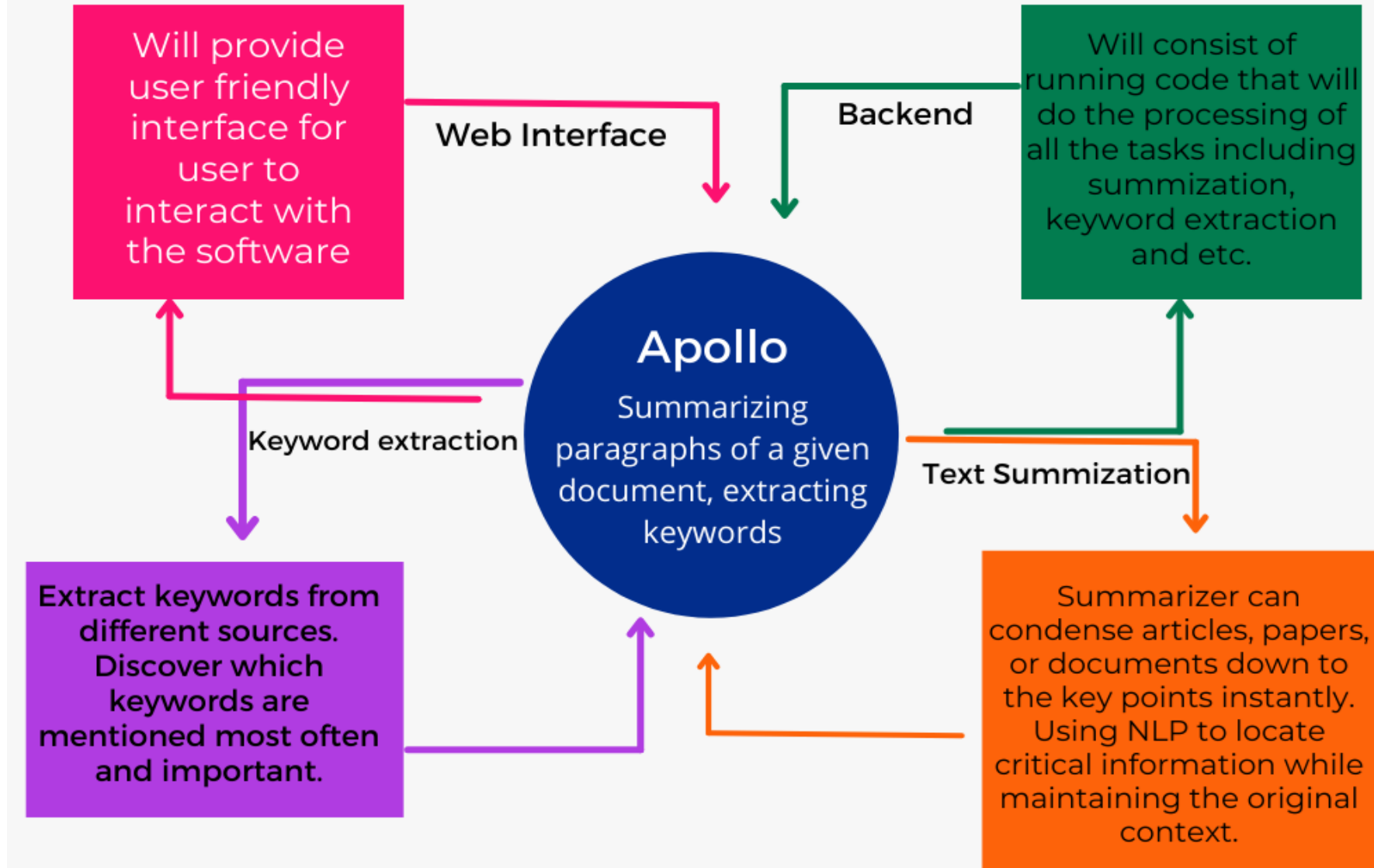
- To provide a tool for users to be able to summarize their text, article, books, etc. (also will be able to pool text from audio and video files).
- Keyword extraction will provide key ideas from different points of text that will allow users to understand the context in a more detailed view, such as what is the most frequently used words and attention points.
- This application will be wrapped in a user-friendly web framework that will provide a backend to run the code and a web interface for the users to interact with it.

Purpose of the project:

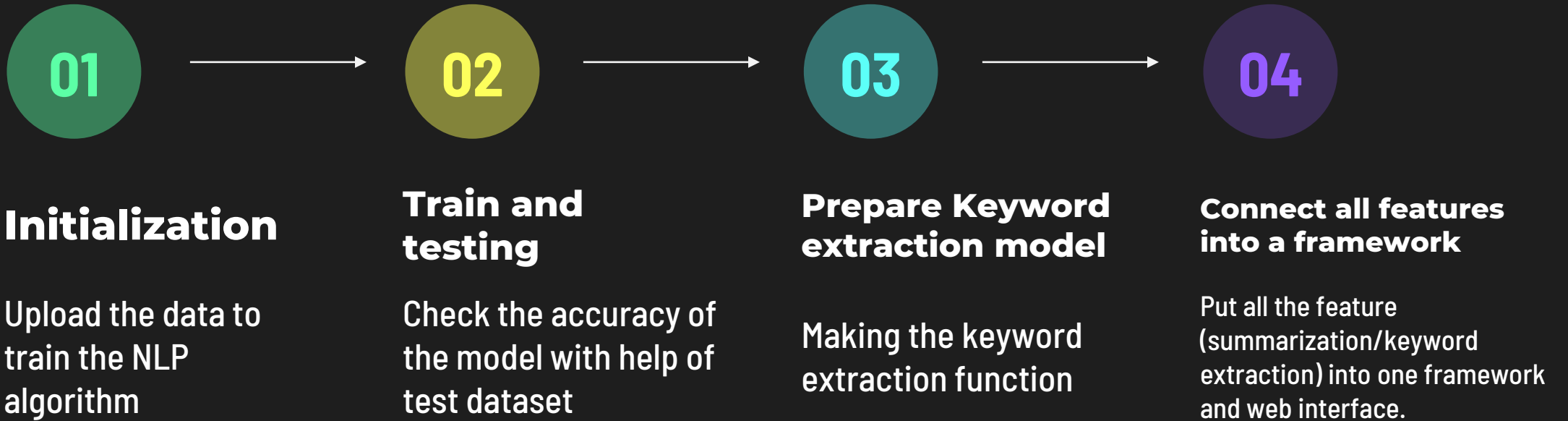
- it can help forensics to extract text and look for the keywords from video or audio files of a crime, gatherings, and such events.
- In addition, people from the education field or just anyone will be able to overview a large amount of data in shorter but clear summarized forms.

Project Content

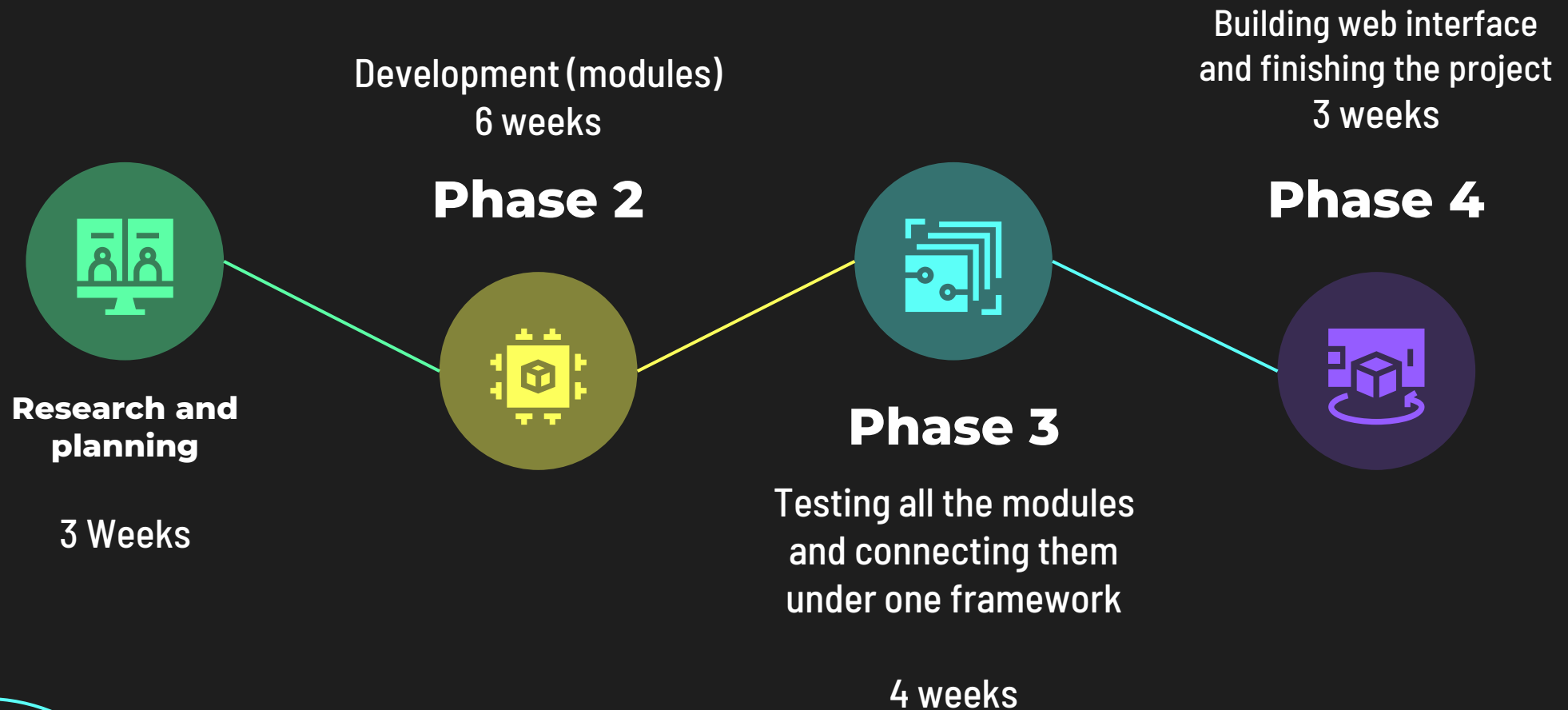
System Concept Map



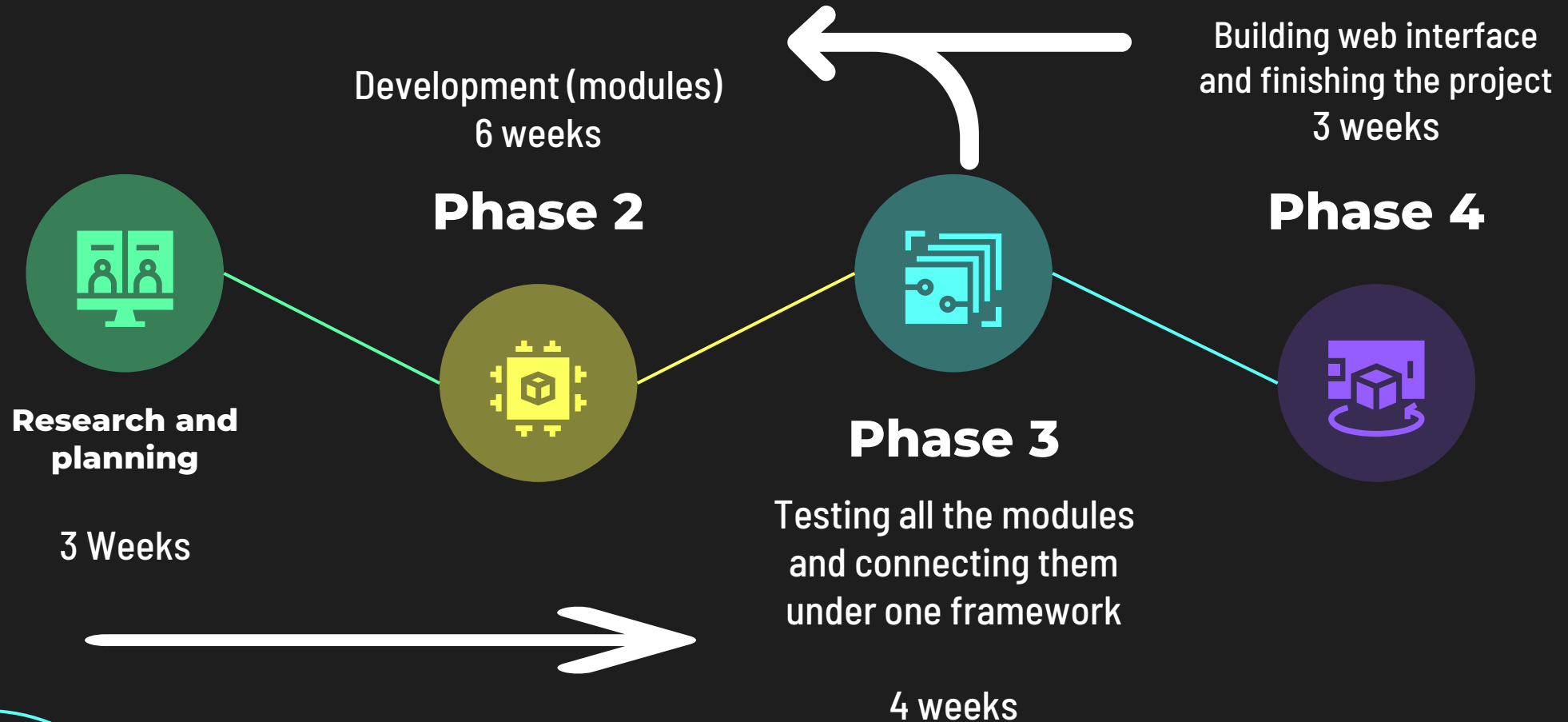
Flowchart



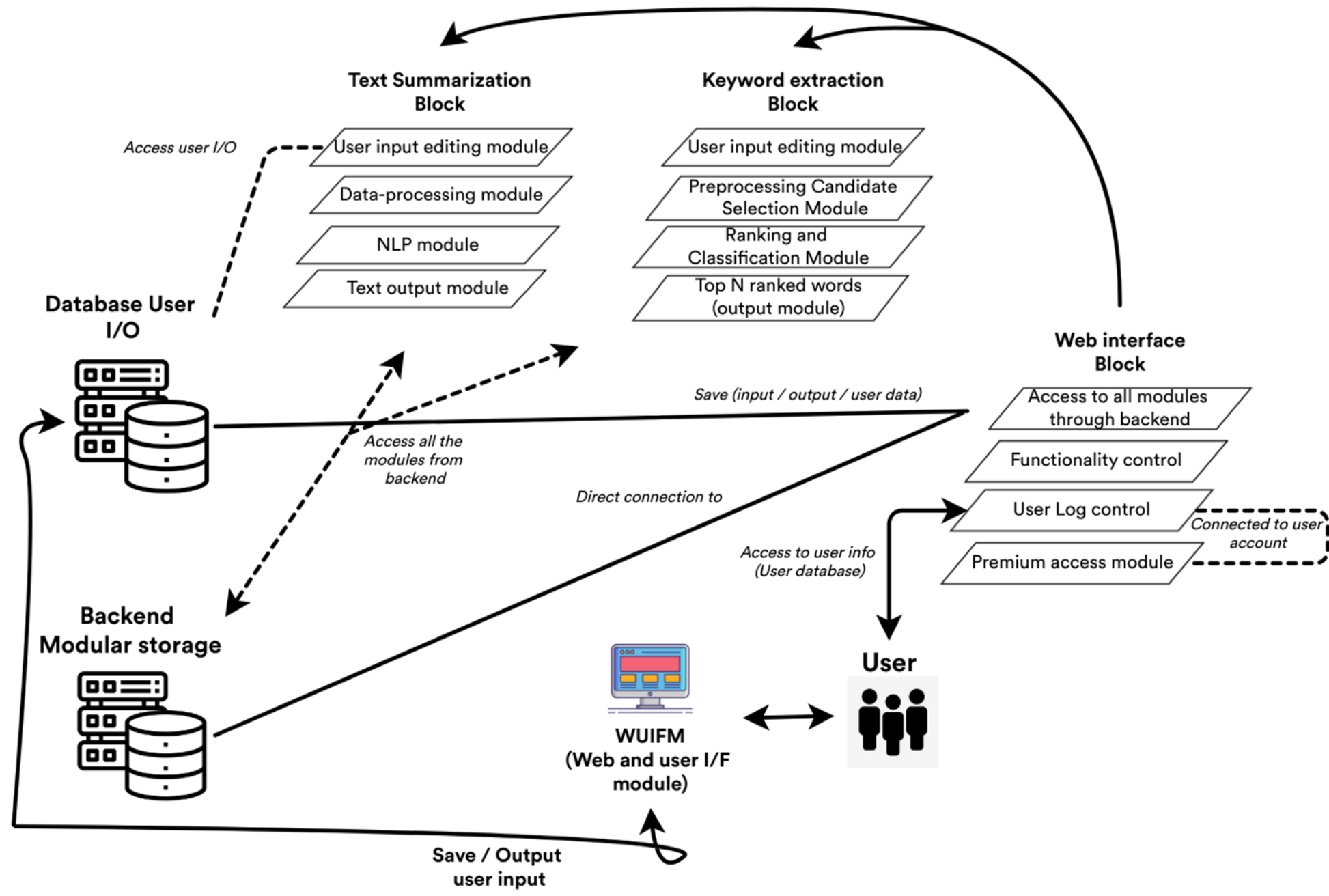
Project timing (Planned)



Project timing (actual)



HLD



Project Result

With help of newspaper3k we extract text from URL: (Old skins cells reprogrammed to regain youthful function: 873 words)

```
from newspaper import Article
```

```
url = 'https://www.sciencedaily.com/releases/2022/04/220408083901.htm'
```

```
article = Article(url)
```

```
article.download()
```

```
article.parse()
```

```
article.text
```

```
'Research from the Babraham Institute has developed a method to \'time jump\' human skin cells by 30 years turning back the ageing clock for cells without losing their specialised function. Work by researchers in the Institute\'s Epigenetics research programme has been able to partly restore the function of older cells, as well as rejuvenating the molecular measures of biological age. The research is published today in the journal eLife and whilst at an early stage of exploration, it could revolutionise regenerative medicine.\n\nWhat is regenerative medicine?\n\nAs we age, our cells\' ability to function declines and the genome accumulates marks of ageing. Regenerative biology aims to repair or replace cells including old ones. One of the most important tools in regenerative biology is our ability to create \'induced\' stem cells. The process is a result of several steps, each erasing some of the marks that make cells specialised. In theory, these stem cells have the potential to become...'
```

After running our summarization function, we get summary of our article with 218 words

```
summarize(article.text, 0.20)
```

```
'Genome analysis showed that cells had regained markers characteristic of skin cells (fibroblasts), and this was confirmed by observing collagen production in the reprogrammed cells.\n\nIn theory, these stem cells have the potential to become any cell type, but scientists aren't yet able to reliably recreate the conditions to re-differentiate stem cells into all cell types.\n\nResearch from the Babraham Institute has developed a method to 'time jump' human skin cells by 30 years, turning back the ageing clock for cells without losing their specialised function.In 2007, Shinya Yamanaka was the first scientist to turn normal cells, which have a specific function, into stem cells which have the special ability to develop into any cell type.This allowed researchers to find the precise balance between reprogramming cells, making them biologically younger, while still being able to regain their specialised cell function.\n\nBy these two measures, the reprogrammed cells matched the profile of cells that were 30 years younger compared to reference data sets.\n\nWe have proved that cells can be rejuvenated without losing their function and that rejuvenation looks to restore some function to old cells.The new method, based on the Nobel Prize winning technique scientists use to make stem cells, overcomes the problem of entirely erasing cell identity by halting reprogramming part of the way through the process.'
```

```
import PyPDF2

reader = PyPDF2.PdfFileReader("example.pdf")
writer = PyPDF2.PdfFileWriter()
```

```
for page in reader.pages:
    writer.addPage(page)

writer.removeImages()
```

```
for page in reader.pages:
    page.compressContentStreams()
    writer.addPage(page)
```

```
with open("out.pdf", "wb") as f:
    writer.write(f)
```

We can also run module such as PyPDF2 to extract text from PDF files, its however trickier than it seems.



```
with open('out.txt', 'w') as f:
    print('', get_large_audio_transcription(path), file=f)
```

```
audio-chunks/chunk1.wav : But this week's podcast in english. com beginners podcast.
audio-chunks/chunk2.wav : We are talking about fruit and veg.
audio-chunks/chunk3.wav : No not in the supermarket not in the market.
audio-chunks/chunk4.wav : In our garden we have a large garden where we grow fruit and vegetables fruit and veg.
audio-chunks/chunk5.wav : And we grow everything from seeds in the spring we have broad beans.
audio-chunks/chunk6.wav : The beans are very long and very big.
audio-chunks/chunk7.wav : Watching the spring richard. we have lots in the summer we have two small red ones and large yellow.
audio-chunks/chunk8.wav : And we have corn.
audio-chunks/chunk9.wav : Bright yellow corn.
audio-chunks/chunk10.wav : Berry suites.
```



Split audio file into chunks => Extract text from chunks

```
And we have corn
Bright yellow corn
Berry suites
And my favorite
We also have potatoes ob jeans there
Dark purple color
We have bright red raspberries my favorites
Red strawberries and melon
```



Index sentences depending on ranking => Summarize

```
Indexes of top ranked_sentence order are [(0.09381506300819897, ['And', 'also', 'yellow', 'ones']), (0.09373264269553319, ['Bright', 'yellow', 'corn']), (0.08949551114648545, [
Summarize Text:
And also yellow ones. Bright yellow corn
```

```
▶ from multi_rake import Rake
rake = Rake()
keywords = rake.apply(text)
```

```
keywords[:10]
```

```
👤 [('machine learning algorithms', 9.0),
    ('vision chip component', 9.0),
    ('non-primary source needed', 9.0),
    ('kristen lee stated', 9.0),
    ('ceo elon musk', 8.666666666666666),
    ('minus $1 million', 8.5),
    ('driving car software', 8.5),
    ('musk offered advice', 8.166666666666666),
    ('george hotz tweeted', 8.125),
    ('ai open sourced', 7.5)]
```

```
[ ]
```

```
print("Enter your text here:\n")
text = input()
```

Enter your text here:

Hotz founded his AI startup, comma.ai, in September 2015.[45] In an interview with Bloomberg, Hotz revealed that the company was bui

```
!pip install multi_rake
```

Requirement already satisfied: multi_rake in /usr/local/lib/python3.7/dist-packages (0.0.2)
Requirement already satisfied: regex>=2018.6.6 in /usr/local/lib/python3.7/dist-packages (from multi_rake) (2019.12.20)
Requirement already satisfied: pysistent>=0.14.2 in /usr/local/lib/python3.7/dist-packages (from multi_rake) (0.18.1)
Requirement already satisfied: pycld2>=0.41 in /usr/local/lib/python3.7/dist-packages (from multi_rake) (0.41)
Requirement already satisfied: numpy>=1.14.4 in /usr/local/lib/python3.7/dist-packages (from multi_rake) (1.21.5)

```
[ ] from multi_rake import Rake
rake = Rake()
keywords = rake.apply(text)
```

keywords[:10]

```
[('machine learning algorithms', 9.0),
 ('vision chip component', 9.0),
 ('non-primary source needed', 9.0),
 ('kristen lee stated', 9.0),
 ('ceo elon musk', 8.666666666666666),
 ('minus $1 million', 8.5),
 ('driving car software', 8.5),
 ('musk offered advice', 8.166666666666666),
 ('george hotz tweeted', 8.125),
 ('ai open sourced', 7.5)]
```

```
[ ]
```

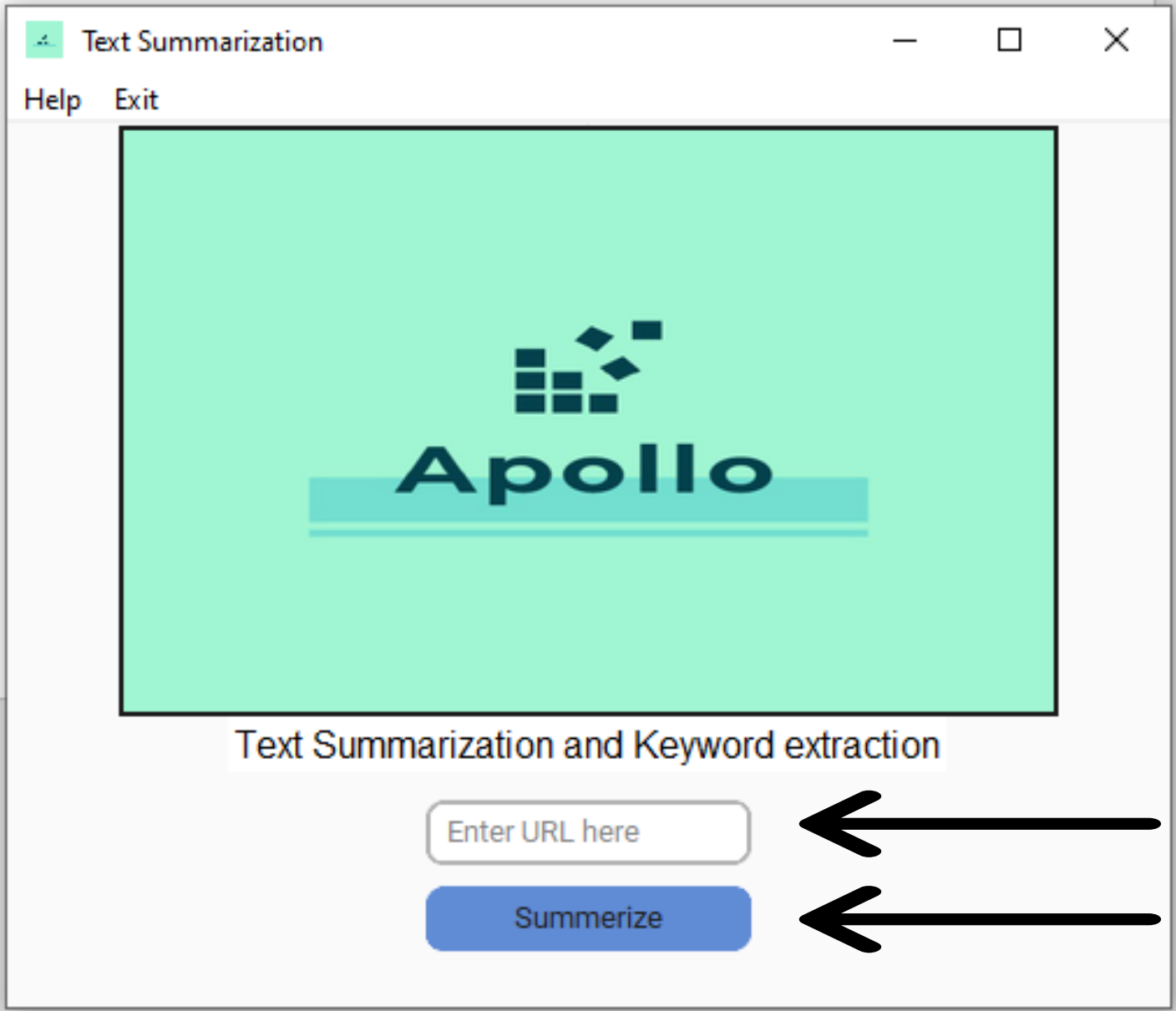


Web interface (old concept)



New Results

Fields for users to get help or exit the program

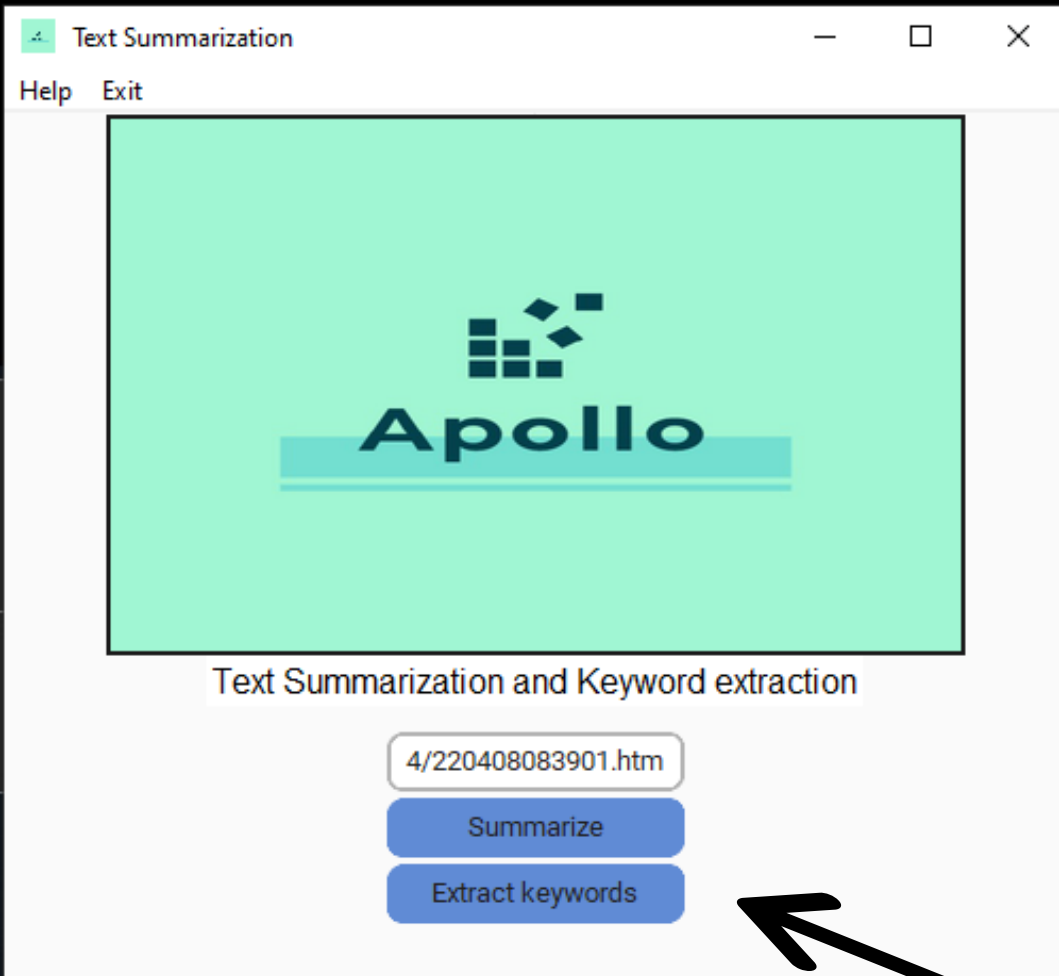


The next step is to add more features and make better design

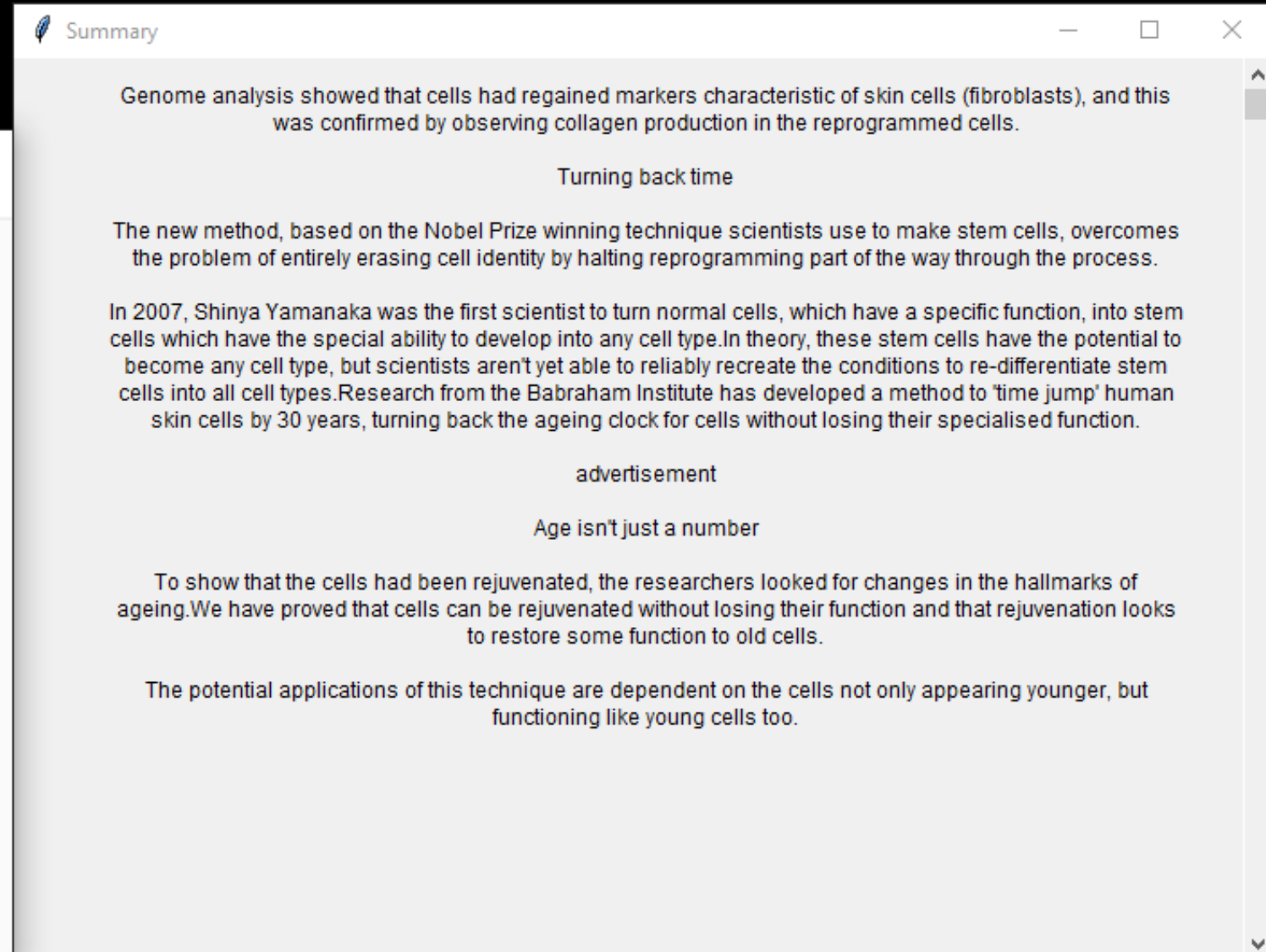
Field to input URL address for the article, news and etc.

Start process of text summarization

Text summarization Results with New interface



The screenshot shows a window titled "Text Summarization" with a menu bar containing "Help" and "Exit". The main content area features a light green background with a logo consisting of a grid of squares above the word "Apollo" in a bold, sans-serif font. Below the logo, the text "Text Summarization and Keyword extraction" is displayed. At the bottom, there is a text input field containing the URL "4/220408083901.htm", a blue "Summarize" button, and a blue "Extract keywords" button.



The screenshot shows a window titled "Summary" with a light gray background. The text content is as follows:

Genome analysis showed that cells had regained markers characteristic of skin cells (fibroblasts), and this was confirmed by observing collagen production in the reprogrammed cells.

Turning back time

The new method, based on the Nobel Prize winning technique scientists use to make stem cells, overcomes the problem of entirely erasing cell identity by halting reprogramming part of the way through the process.

In 2007, Shinya Yamanaka was the first scientist to turn normal cells, which have a specific function, into stem cells which have the special ability to develop into any cell type. In theory, these stem cells have the potential to become any cell type, but scientists aren't yet able to reliably recreate the conditions to re-differentiate stem cells into all cell types. Research from the Babraham Institute has developed a method to 'time jump' human skin cells by 30 years, turning back the ageing clock for cells without losing their specialised function.

advertisement

Age isn't just a number

To show that the cells had been rejuvenated, the researchers looked for changes in the hallmarks of ageing. We have proved that cells can be rejuvenated without losing their function and that rejuvenation looks to restore some function to old cells.

The potential applications of this technique are dependent on the cells not only appearing younger, but functioning like young cells too.

Keyword extraction with URL

Project Effect and Next Plan

Expected effect

- Using modern transformers and NLP algorithms such as “Attention is all you need” the application needs to be capable of locating critical information while maintaining original meaning and having the smallest loss.
- In addition, the user should be able to extract text from any kind of file/format.

Next plan

- The algorithm, training process, and datasets can be increased and improved for better text summarization.
- Chapter-wise text summarization for books and large amounts of text can be applied. New features like paraphrasing, plagiarism checking, grammar checking, etc. can be added.
- Also, the user interface can be improved for better user/machine interaction.

Challenges / Benefits of current summarization model



Works best with small amount of data (not suited for long text such as books, article papers and etc.)



Has limitation on working with PDF files

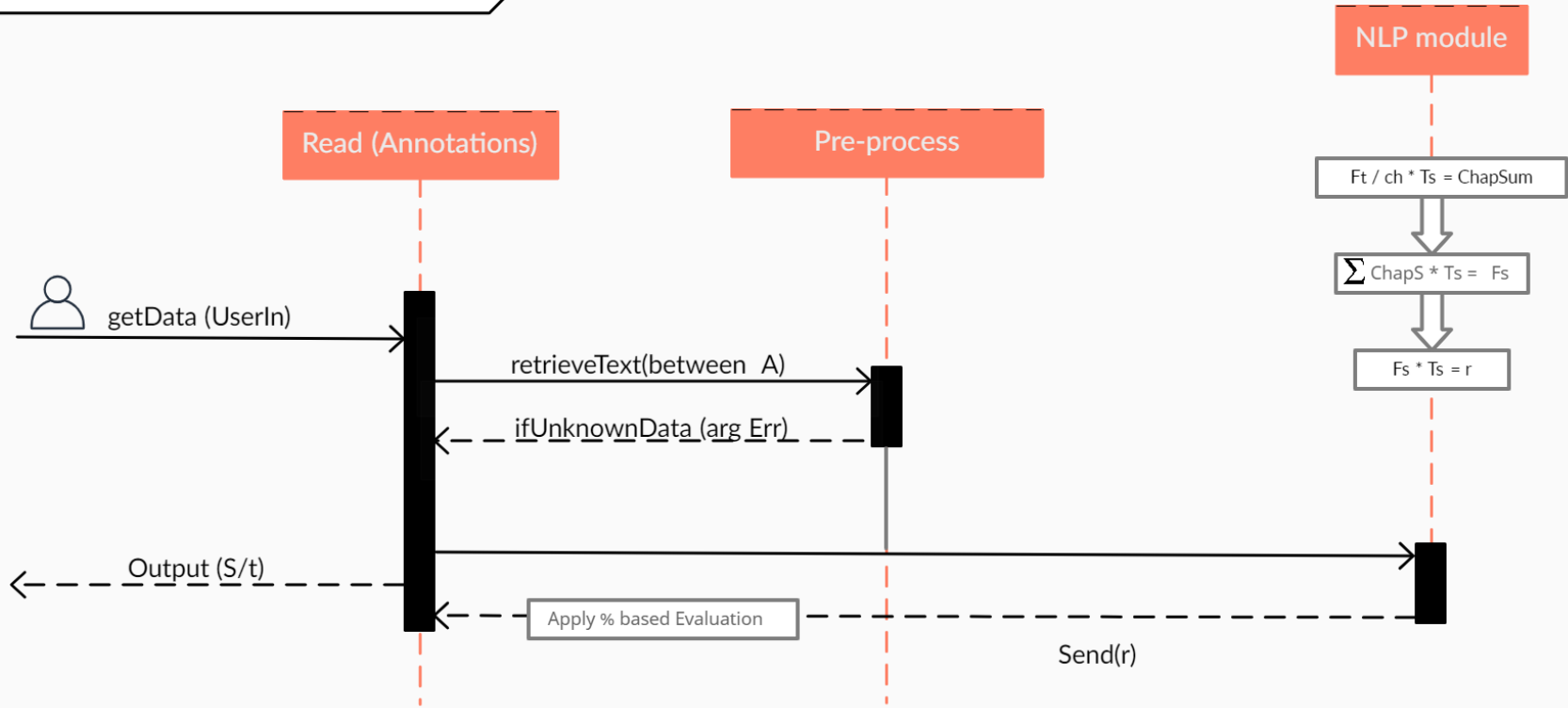


Can be improved by implementing more novel way of text summarization with help of deep / machine learning models



It suits if users needs URL, audio files , text files summarization of the limited text and provides good results.

LLD for chapter-wise concept



Possible new
concept:
Sequential
chapter-wise
summarization



THANK YOU
FOR YOUR
ATTENTION