Amrita Vishwa Vidyapeetham

Text Summarizer

{ https://github.com/vkmanojk/Text-Summarizer }

- Manojkumar V K
CB.EN.U4CSE17040
ASE, Coimbatore

Introduction:

Automatic Text Summarization is one of the most challenging and interesting problems in the field of Natural Language Processing (NLP). It is a process of generating a concise and meaningful summary of text from multiple text resources such as books, news articles, blog posts, research papers, emails, and tweets. The demand for automatic text summarization systems is spiking these days thanks to the availability of large amounts of textual data. The news app *Inshorts* is a big hit because of the ability to produce a concise summary in 60 words.

The web application *Text Summarizer* can be used to summarize content from a website or directly from the user. It can also be used to compare and contrast the summaries provided by various algorithms.

Tools Used:

- HTML5
- CSS
- JavaScript
- Python3

Packages Used:

For summarization:

- Spacy
- Gensim
- Sumy
- NLTK

For Web application:

- Flask
- BeautifulSoup
- Urllib

Spacy:

spaCy is an open-source software library for advanced natural language processing, written in the programming languages Python and Cython. The library is published under the MIT license. It features NER, POS tagging, dependency parsing, word vectors and more.

Gensim:

Gensim is an open-source library for unsupervised topic modelling and natural language processing, using modern statistical machine learning. Gensim is implemented in Python and Cython. Gensim is designed to handle large text collections using data streaming and incremental online algorithms, which differentiates it from most other machine learning software packages that target only in-memory processing.

Sumy:

Module for automatic summarization of text documents and HTML pages. Simple library and command-line utility for extracting summary from HTML pages or plain texts.

NLTK:

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. NLTK includes graphical demonstrations and sample data.

Flask:

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

Beautifulsoup:

Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. It is available for Python 2.7 and Python 3.

Urllib:

Urllib is a Python module that can be used for opening URLs. It defines functions and classes to help in URL actions. With Python, you can also access and retrieve data from the internet like XML, HTML, JSON, etc. You can also use Python to work with this data directly.

Features:

Apart from summarizing texts, the application can also compare and contrast the outputs from different summarizing algorithms. It uses a ***session*** variable to calculate the number of computations required for the user's request. It also keeps track of the date the user last visited by using ***cookies***. The application provides ***error handling*** especially in cases, where websites do not allow texts to be extracted or do not contain texts to be extracted.

To check these features, run the application and:

1. Note the statement "Your requests required [num] computations in the current session" (session variable)
2. The sentence "You already visited this page on [date]" uses cookies to store and retrieve dates previously visited by the user.
3. In the enter URL section, give the URL of a website with no text. Example URL: https://wallup.net/wp-content/uploads/2017/11/23/526001-computer-keyboards.jpg
4. Enter text in the text column or URL in the Enter URL column and witness the automatic summarization.
5. Click Compare Summarizers in the navbar. Then enter the text to be summarized. Different outputs are produced for different algorithms (packages).

Environmental Requirements:

- Python (64 bit; the application will not run in a 32-bit version)
- Verified on Ubuntu and Windows 7, 8 and 10
- Pip install the following:
    - beautifulsoup4
    - bs4
    - Flask
    - nltk
    - spacy
    - sumy
    - thinc
    - urllib3
    - gensim
    - gensim-sum-ext
    - numpy
    - cython
    - https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-2.2.5/en_core_web_sm-2.2.5.tar.gz#egg=en_core_web_sm==2.2.5

It is necessary that the command line is ***run as administrator*** to link en_core_web to en. If an error occurs, install every package specified above except the last one. Instead, type:

python −m spacy download en

The above packages are mentioned in requirements.txt file and therefore can be directly installed using the command:

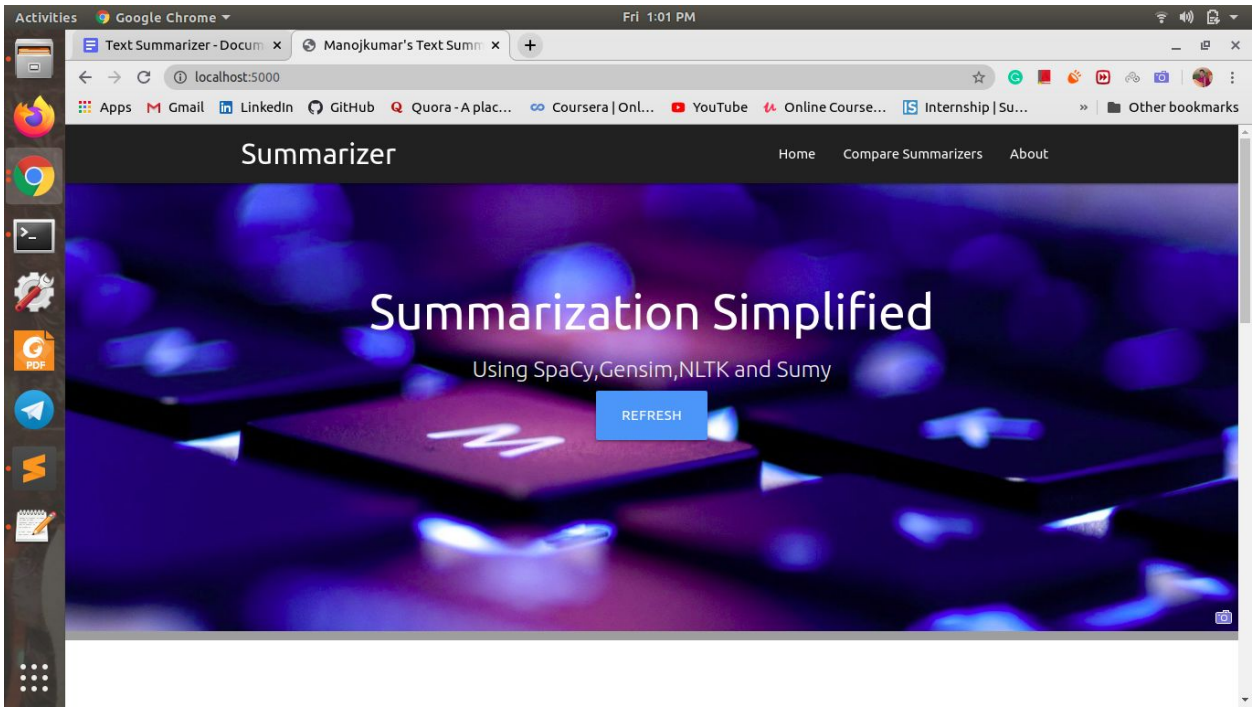
pip install −r requirements.txt

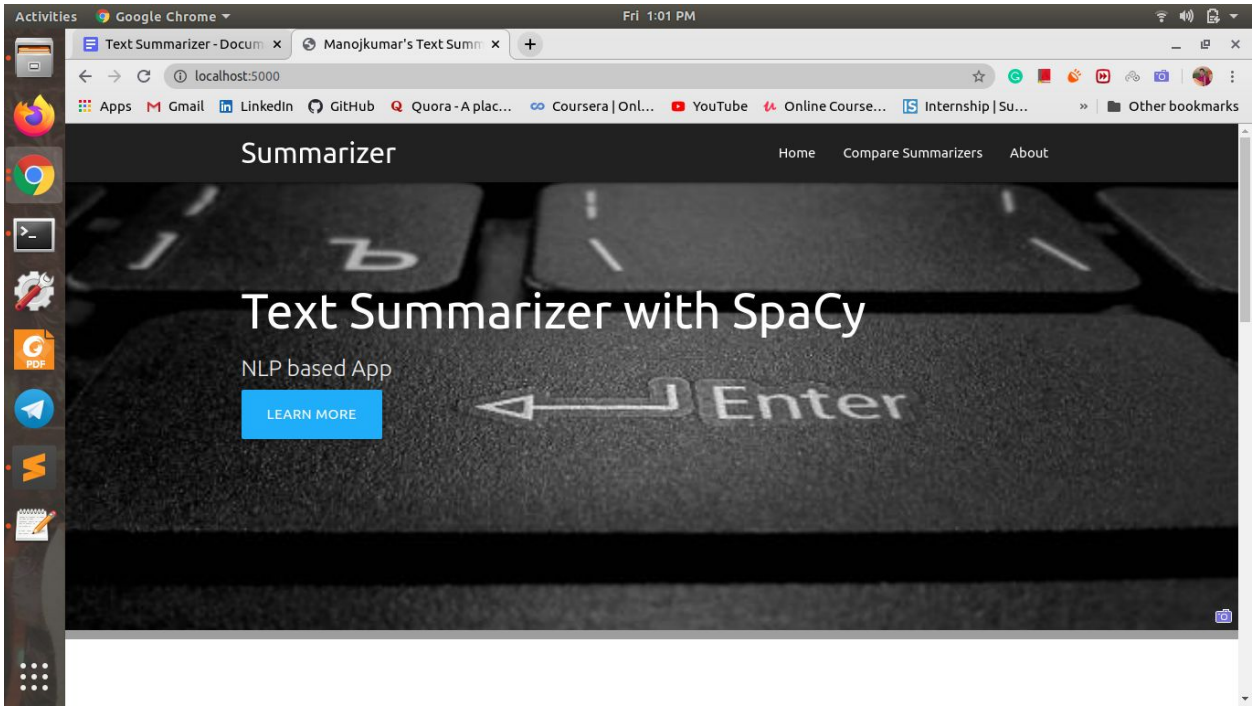
Then run the app using:

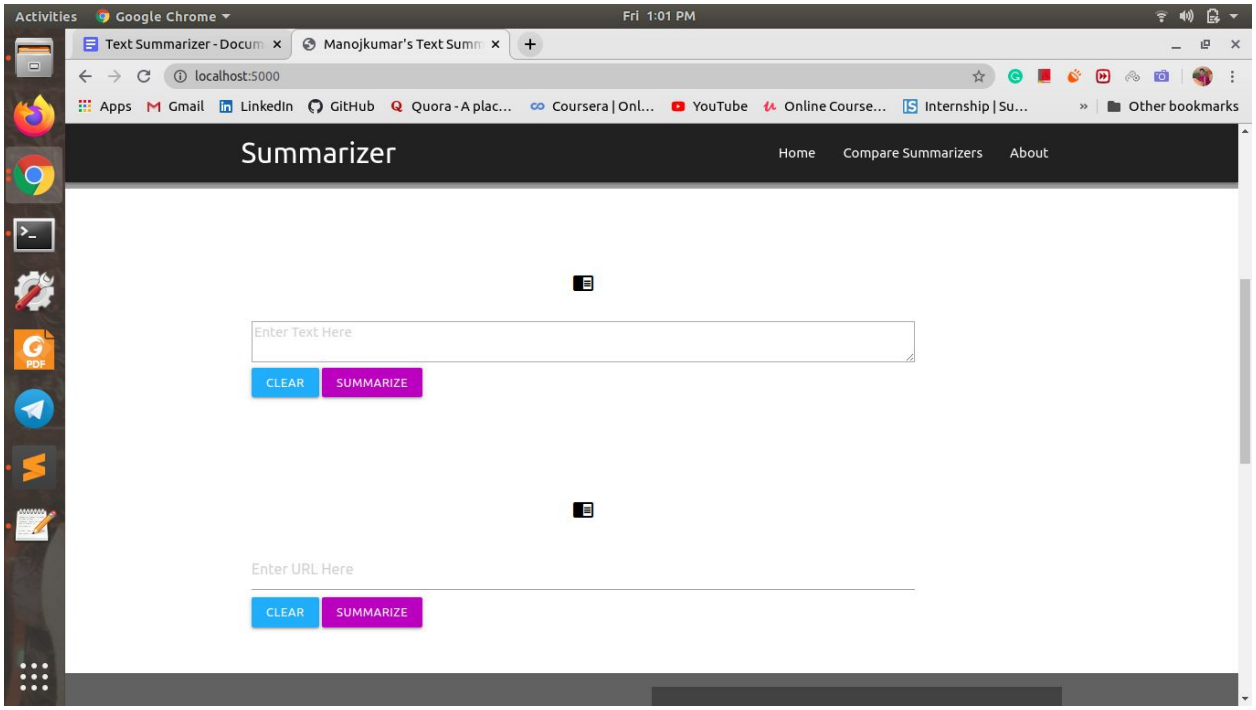
python app.py

Screenshot:

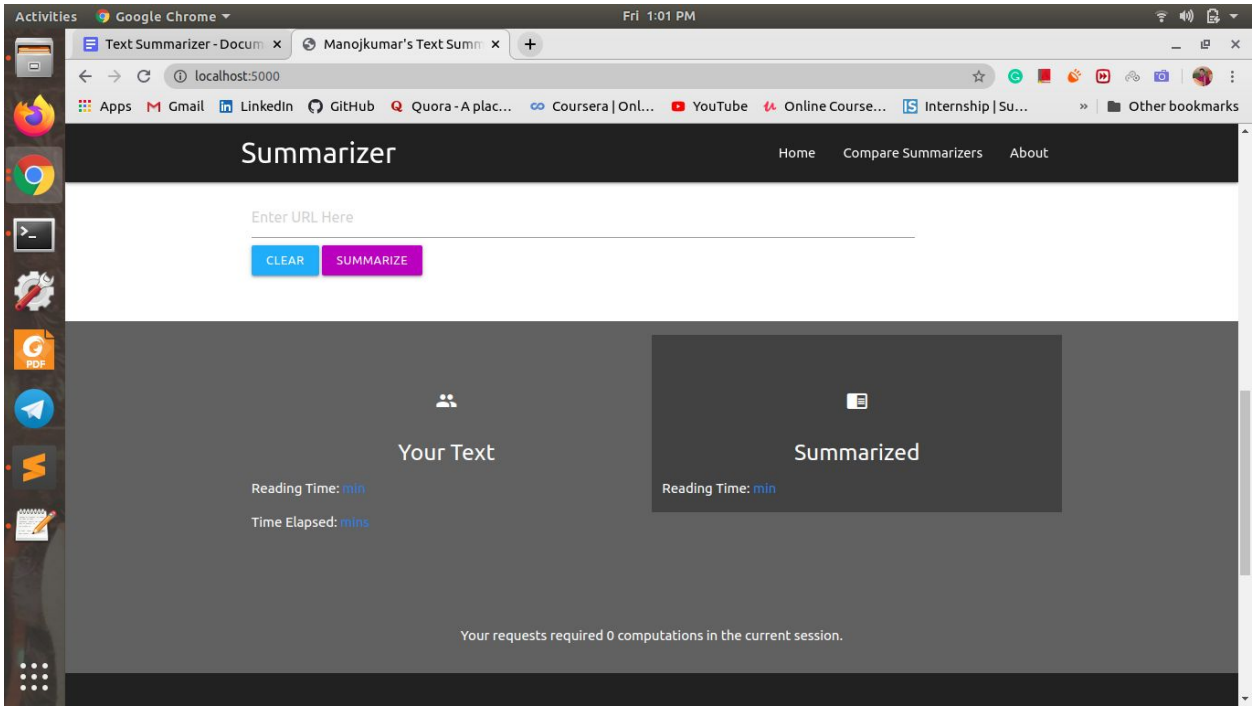
UI:



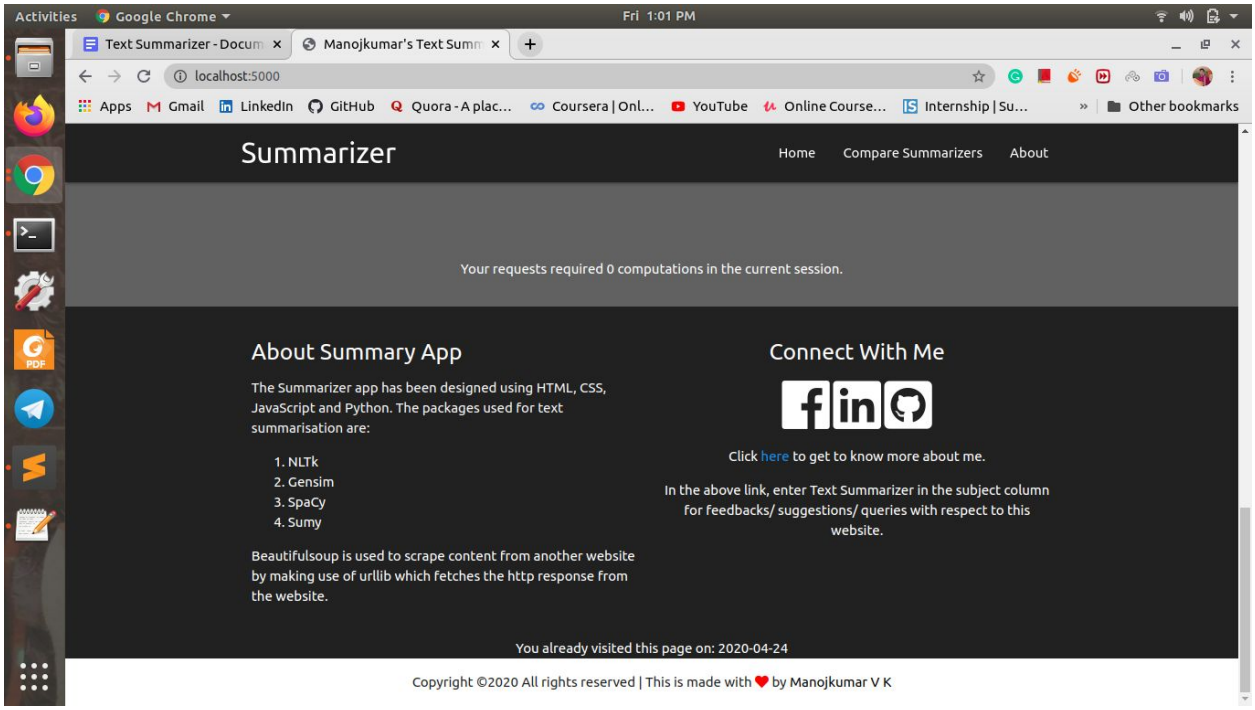

Activities
Google Chrome
Fri 1:01 PM
Text Summarizer - Docum
Manojkumar's Text Summ
localhost:5000
Apps
Gmail
LinkedIn
GitHub
Quora - A plac...
Coursera | Onl...
YouTube
Online Course...
Internship | Su...
Other bookmarks
Summarizer
Home
Compare Summarizers
About
Text Summarizer with SpaCy
NLP based App
LEARN MORE
Enter


Activities
Google Chrome
Fri 1:01 PM
Text Summarizer - Docum
Manojkumar's Text Summ
localhost:5000
Apps
Gmail
LinkedIn
GitHub
Quora - A plac...
Coursera | Onl...
YouTube
Online Course...
Internship | Su...
Other bookmarks
Summarizer
Home
Compare Summarizers
About
Enter Text Here
CLEAR
SUMMARIZE
Enter URL Here
CLEAR
SUMMARIZE

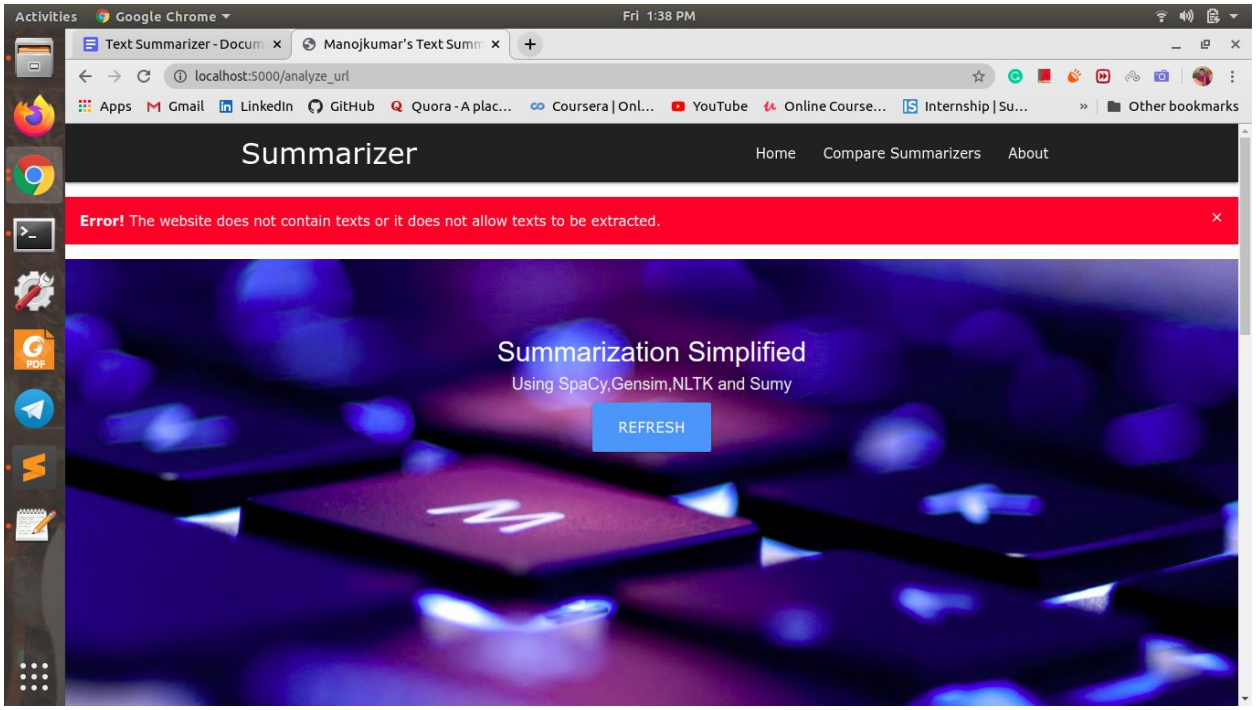
Use of session : (Your requests required 0 computations in the current session.)



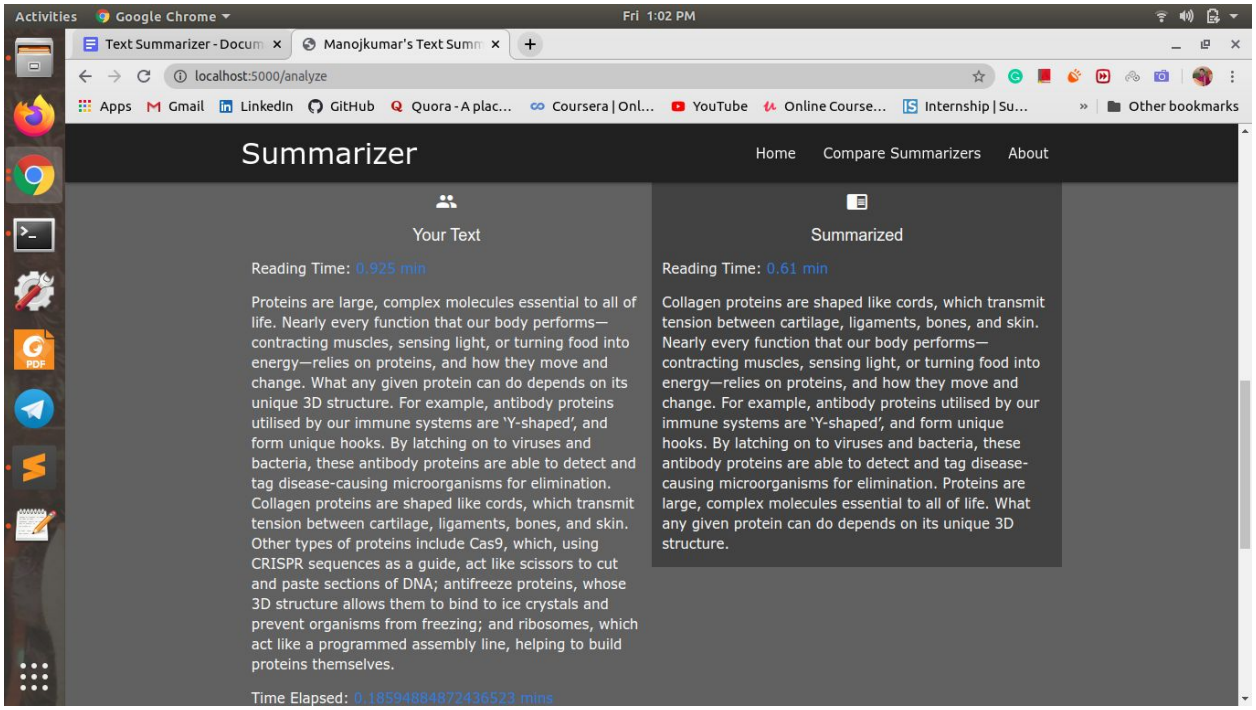
Cookie: (You already visited this page on:)



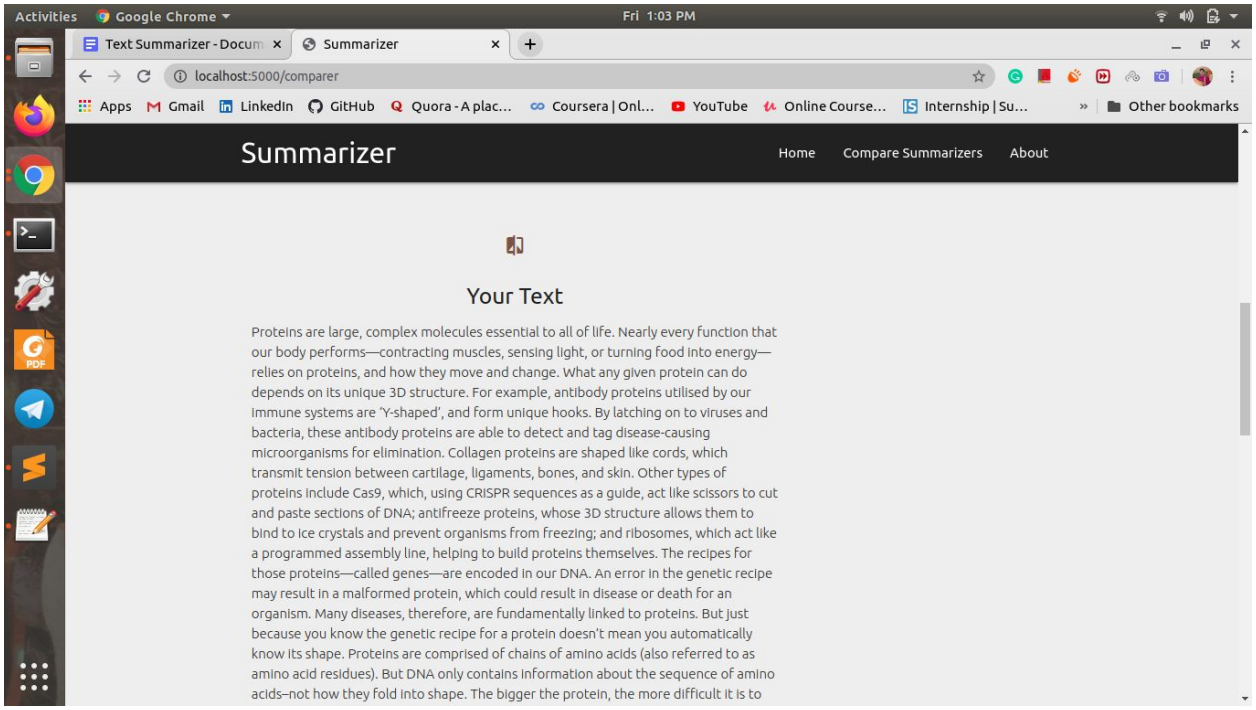
Error:

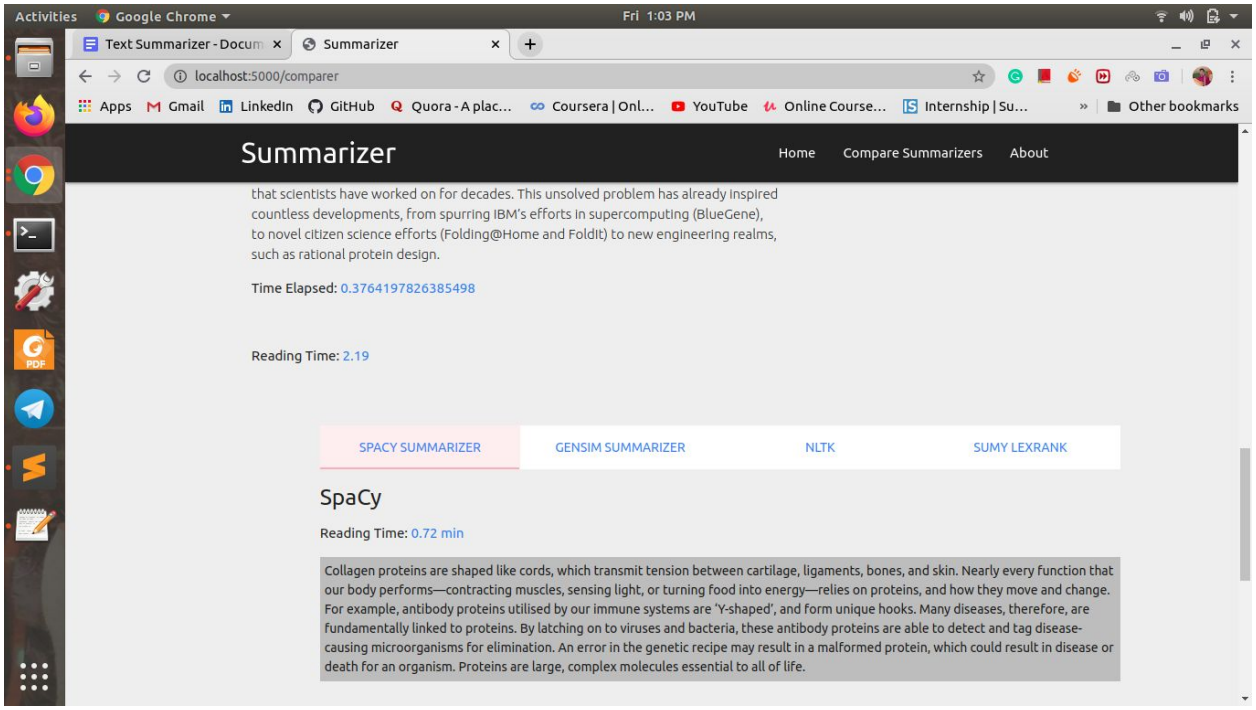


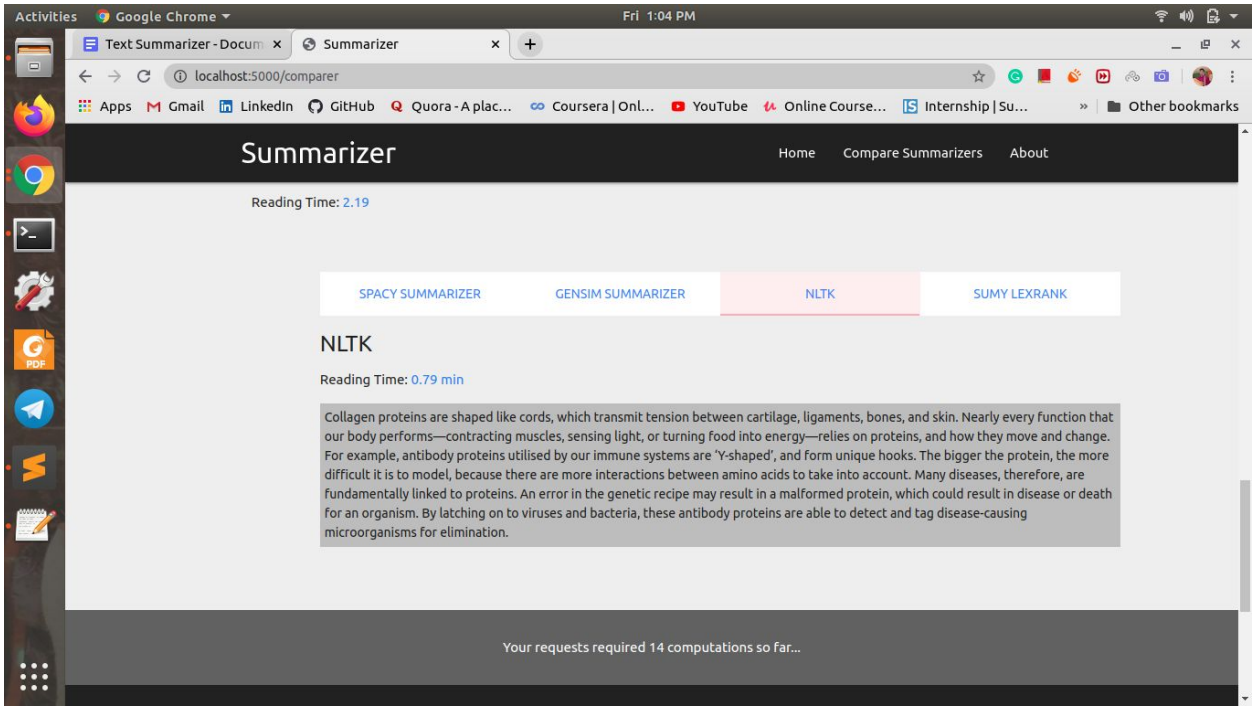
Text Summarization:

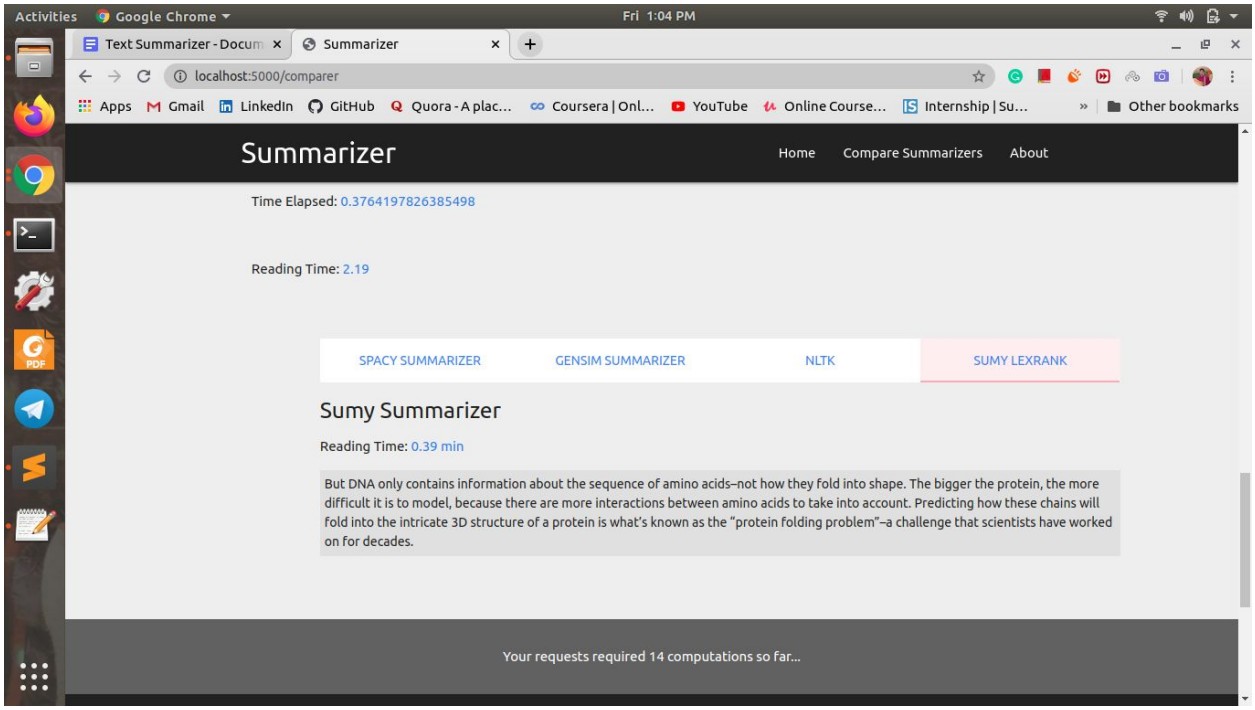


Compare Summarizers:






Activities
Google Chrome
Fri 1:04 PM
Text Summarizer - Docum
Summarizer
localhost:5000/comparer
Apps
Gmail
LinkedIn
GitHub
Quora - A plac...
Coursera | Onl...
YouTube
Online Course...
Internship | Su...
Other bookmarks
Summarizer
Home
Compare Summarizers
About
Reading Time: 2.19
SPACY SUMMARIZER
GENSIM SUMMARIZER
NLTK
SUMY LEXRANK
NLTK
Reading Time: 0.79 min
Collagen proteins are shaped like cords, which transmit tension between cartilage, ligaments, bones, and skin. Nearly every function that our body performs—contracting muscles, sensing light, or turning food into energy—relies on proteins, and how they move and change. For example, antibody proteins utilised by our immune systems are 'Y-shaped', and form unique hooks. The bigger the protein, the more difficult it is to model, because there are more interactions between amino acids to take into account. Many diseases, therefore, are fundamentally linked to proteins. An error in the genetic recipe may result in a malformed protein, which could result in disease or death for an organism. By latching on to viruses and bacteria, these antibody proteins are able to detect and tag disease-causing microorganisms for elimination.
Your requests required 14 computations so far...


Activities
Google Chrome
Fri 1:04 PM
Text Summarizer - Docum
Summarizer
localhost:5000/comparer
Apps
Gmail
LinkedIn
GitHub
Quora - A plac...
Coursera | Onl...
YouTube
Online Course...
Internship | Su...
Other bookmarks
Summarizer
Home
Compare Summarizers
About
Time Elapsed: 0.3764197826385498
Reading Time: 2.19
SPACY SUMMARIZER
GENSIM SUMMARIZER
NLTK
SUMY LEXRANK
Sumy Summarizer
Reading Time: 0.39 min
But DNA only contains information about the sequence of amino acids–not how they fold into shape. The bigger the protein, the more difficult it is to model, because there are more interactions between amino acids to take into account. Predicting how these chains will fold into the intricate 3D structure of a protein is what's known as the "protein folding problem"–a challenge that scientists have worked on for decades.
Your requests required 14 computations so far...

Steps to run:

Type the following commands in the directory containing app.py

- *pip install −r requirements.txt*
- *python −m spacy download en*
- *python app.py*

Conclusion:

In automatic text summarization, a system is confronted with real texts. The text - containing mistakes, contradictions, redundancy and incomplete information - is the concrete object that summarizers must face. Algorithms must generate the most concise summaries from a *corpus*. Writing summaries proves to be a difficult task for humans. For machines, which cannot understand the content, the challenge is even greater. Thanks to the recent developments, a machine is able to produce a summary for any given text. Although there is no perfection, it has definitely reached a great extent.