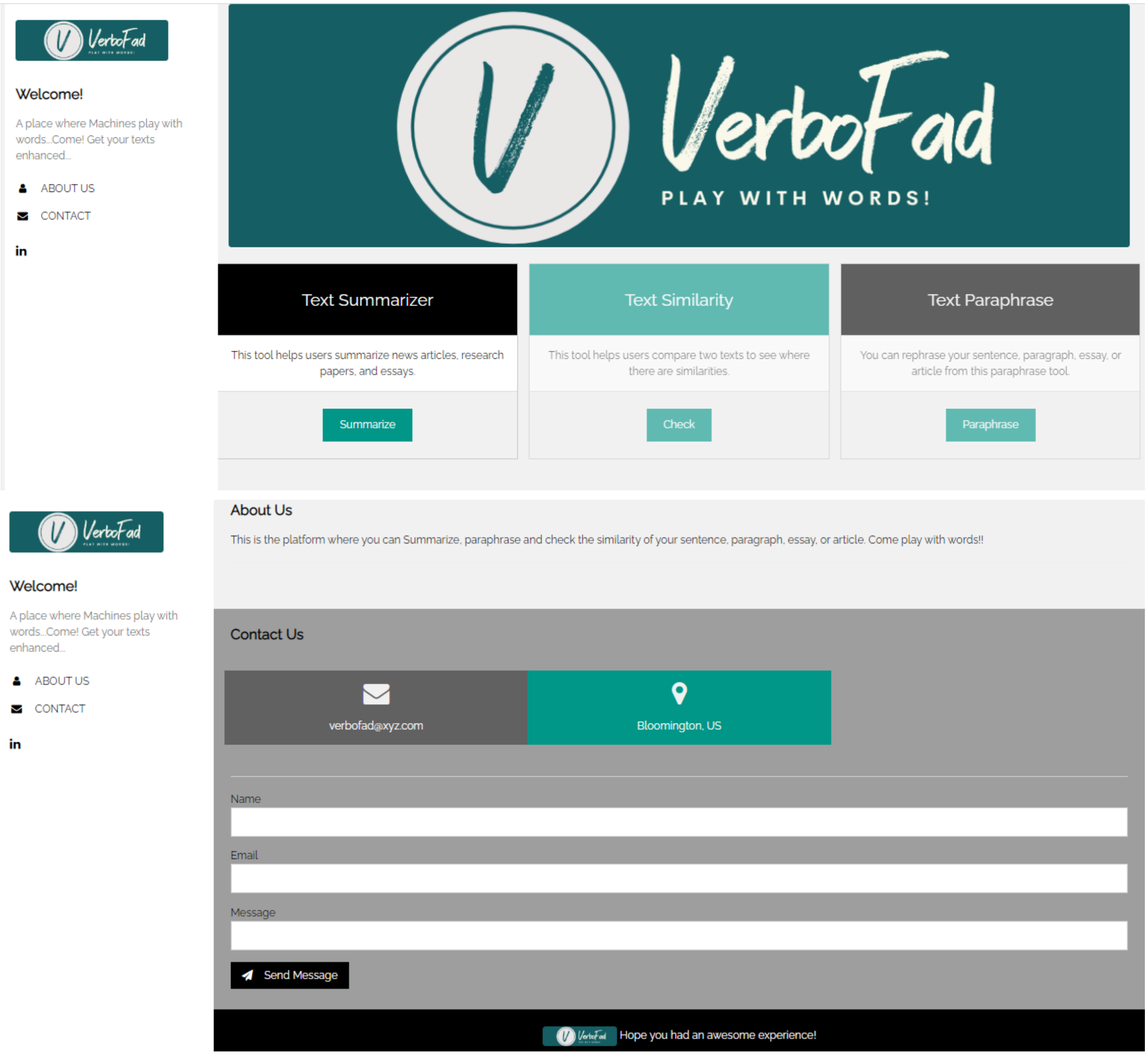
Final Project - Technical Report

Application:

Images shown below shows the Application's Home page (Image 1), Text Paraphrase (Image 2), Text Summarization (Image 3), Text Similarity (Image 4)



Welcome!

A place where Machines play with words..Come! Get your texts enhanced..

This tool helps users summarize news articles, research papers, and essays.

This tool helps users compare two texts to see where there are similarities.

You can rephrase your sentence, paragraph, essay, or article from this paraphrase tool.

About Us

This is the platform where you can Summarize, paraphrase and check the similarity of your sentence, paragraph, essay, or article. Come play with words!!

Contact Us

verbofad@xyz.com

Bloomington, US

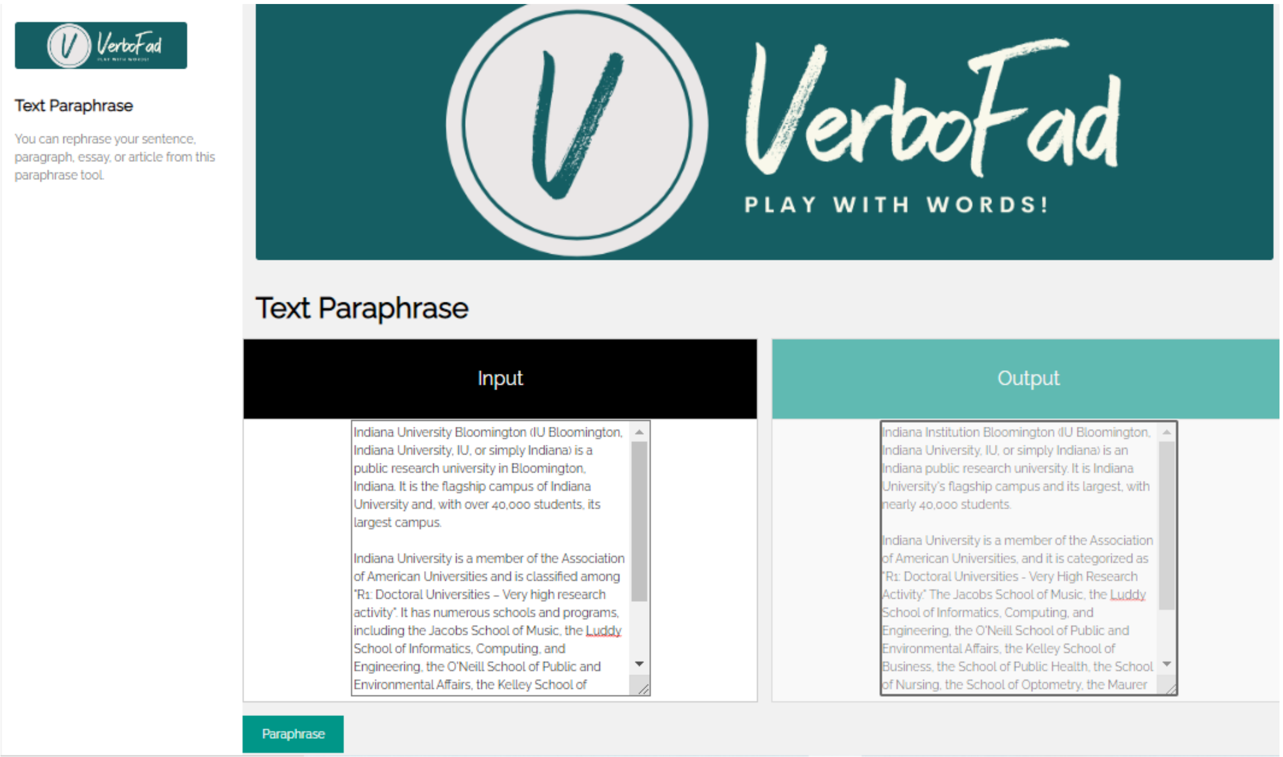
Name

Email

Message

Hope you had an awesome experience!

Image 1: *The above image is of VerboFad's (Web Application's) Home page, displaying 3 fields on the top for Text Similarity, Summarization, and Paraphrase, with a section describing About Us, and Contact Us form at the bot*



Text Paraphrase

You can rephrase your sentence, paragraph, essay, or article from this paraphrase tool

Text Paraphrase

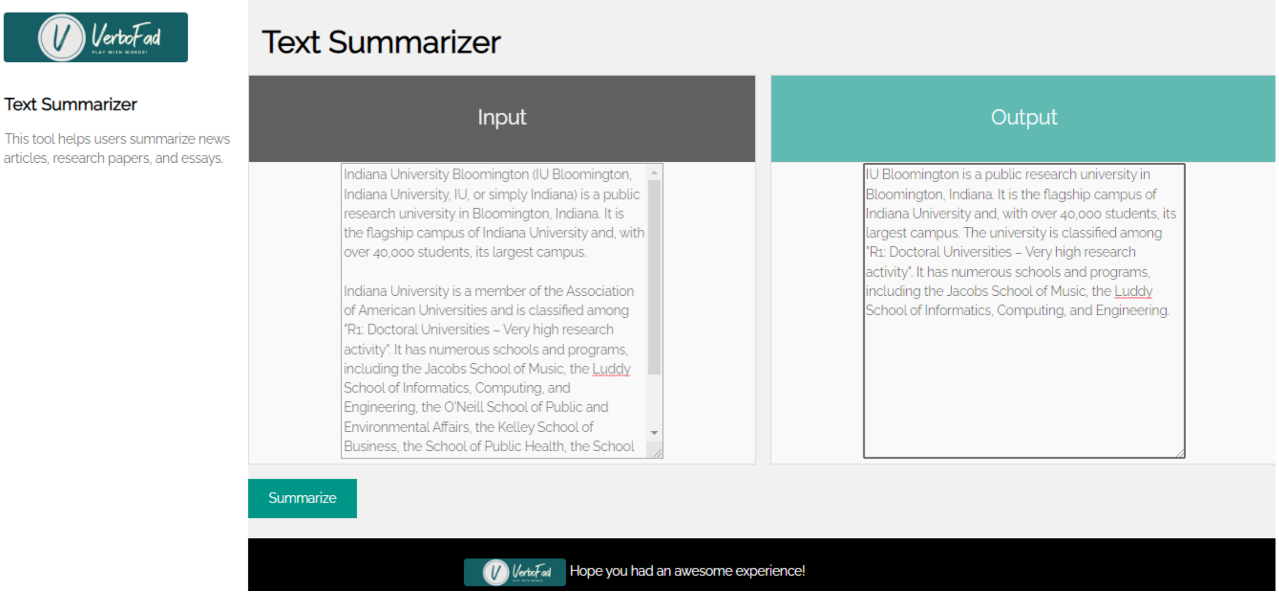
Indiana University Bloomington (IU Bloomington, Indiana University, IU, or simply Indiana) is a public research university in Bloomington, Indiana. It is the flagship campus of Indiana University and, with over 40,000 students, its largest campus.

Indiana University is a member of the Association of American Universities and is classified among "R1: Doctoral Universities – Very high research activity". It has numerous schools and programs, including the Jacobs School of Music, the Luddy School of Informatics, Computing, and Engineering, the O'Neill School of Public and Environmental Affairs, the Kelley School of

Indiana Institution Bloomington (IU Bloomington, Indiana University, IU, or simply Indiana) is an Indiana public research university. It is Indiana University's flagship campus and its largest, with nearly 40,000 students.

Indiana University is a member of the Association of American Universities, and it is categorized as "R1: Doctoral Universities - Very High Research Activity." The Jacobs School of Music, the Luddy School of Informatics, Computing, and Engineering, the O'Neill School of Public and Environmental Affairs, the Kelley School of Business, the School of Public Health, the School of Nursing, the School of Optometry, the Maurer

Image 2: *The above image is of Text Paraphrase Web page where the user can enter their text in the Input field on the left side and get it paraphrased (in the output field on the right side) with a single click (Paraphrase button present at the bottom)*



Text Summarizer

This tool helps users summarize news articles, research papers, and essays.

Text Summarizer

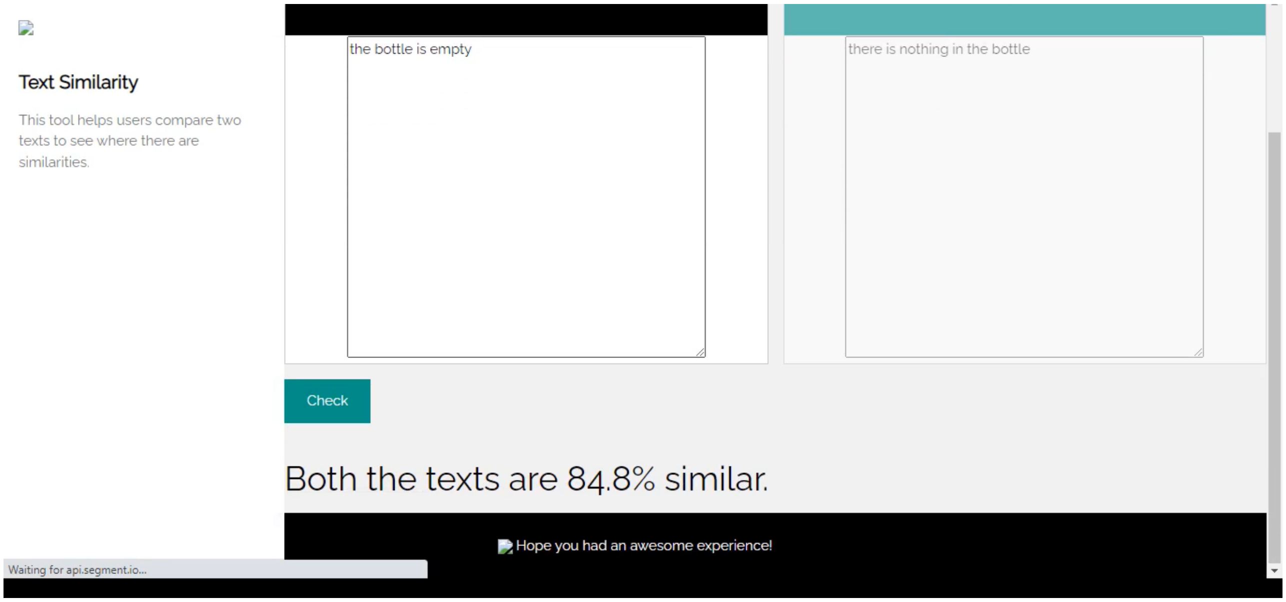
Indiana University Bloomington (IU Bloomington, Indiana University, IU, or simply Indiana) is a public research university in Bloomington, Indiana. It is the flagship campus of Indiana University and, with over 40,000 students, its largest campus.

Indiana University is a member of the Association of American Universities and is classified among "R1: Doctoral Universities – Very high research activity". It has numerous schools and programs, including the Jacobs School of Music, the Luddy School of Informatics, Computing, and Engineering, the O'Neill School of Public and Environmental Affairs, the Kelley School of Business, the School of Public Health, the School

IU Bloomington is a public research university in Bloomington, Indiana. It is the flagship campus of Indiana University and, with over 40,000 students, its largest campus. The university is classified among "R1: Doctoral Universities – Very high research activity". It has numerous schools and programs, including the Jacobs School of Music, the Luddy School of Informatics, Computing, and Engineering.

Hope you had an awesome experience!

Image 3: *The above image is of Text Summarizer Web page where the user can enter their text in the Input field on the left side and get it summarized (in the output field on the right side) with a single click (Summarize button present at the bottom)*



Text Similarity

This tool helps users compare two texts to see where there are similarities.

the bottle is empty

there is nothing in the bottle

Both the texts are 84.8% similar.

Hope you had an awesome experience!

Waiting for api.segment.io...

Image 4: *The above image is of Text Similarity Web page where the user can enter their text in the Input 1 and Input 2 field on the left side and get its similarity percentage checked (which gets displayed at the bottom of the web page) with a single click (Check button present at the bottom)*

Full GitHub URL:

We have uploaded the application source code with github.iu.edu at following URL

URL: https://github.iu.edu/ssmenon/Intro-to-NLP/tree/main/Project

Project Summary

A branch of Linguistics, Computer Science, and Artificial Intelligence called Natural Language Processing studies how computers and human language interact, with a focus on how to design computers to process and examine massive volumes of natural language data. The concept of "Natural Language Processing" has been around for a while and is now ingrained in our daily lives. In this project, we built a website “verbofad.com” that offers users a variety of language processing features, including text paraphrasing, text summarizing, and plagiarism detection between two sentences. Programming was done with Python and Flask. Python-based Flask is a microweb framework.

Project Objectives and Usefulness:

Through this project, we achieved the below mentioned primary objectives:

No other platform combines all of the functionalities in a single location. By providing all of the alternatives on one platform, we simplify things for the users and increase their accessibility. This platform “verbofad.com” can be helpful in a variety of fields, including journalism, business, and education. These are just some of the examples among many others

- The similarity feature can be useful as a Plagiarism Checker in the Education Sector.
- The summarizer feature will be useful in summarizing the notes or lectures of meetings. This can be advantageous in almost any sector today.
- Paraphrasing is also useful in the evaluation of semantic parsing and generation of new samples to expand existing corpora.

These are just some of the numerous advantages among many of our platforms.

Technical Description:

Data:

Since working with real data is the core goal of the project, we won't be utilizing a dataset to train our model. But for our testing, we'll be using the corpuses found in the nltk package. The unlabeled data will be presented as documents/sentences. The data will be transformed and normalized to eliminate superfluous alphabets and stop-words.

Link for example corpuses :- https://www.nltk.org/book/ch02.html

The live data that the user inputs will be in a string format. Our first task would be to convert that data into a set of integers for our model to work on. We will be preprocessing the data before giving it to our model. Because this is an unsupervised Project we won't be having any labeled data. We will focus more on the semantics and the relation betweens the words/documents for this project rather than giving importance to each word.

Models used:

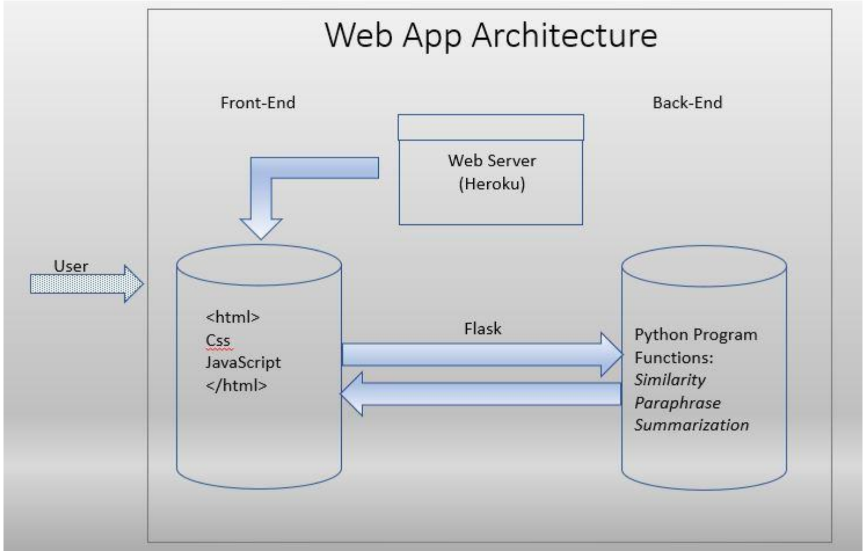


Image 4: *Web App Architecture*

We have used 3 models for the 3 different functionalities.

Similarity:- Cosine Similarity Algorithm
Paraphrasing:- Pegasus Transformer Model
Summarizer:- automatic text summarization Algorithm

View:
The views are developed using HTML, CSS, Bootstrap, JavaScript

Controller:
The controllers are developed using Flask Web Application routers & HTML hyperlinks (Navigation among web pages)

Tools Used:
Front End: HTML, CSS, Bootstrap,JavaScript
Back End: Python & Flask
Tools: Visual Studio Code

User Functionalities:

Similarity: Text similarity measures how "similar" two pieces of text are to one another in terms of their surface resemblance (lexical similarity) and their meaning (semantic similarity). The users will enter two texts into a text box, and we will compute and report the degree of similarity between the two texts. We compared Euclidean Distance, Jaccard Similarity and Cosine Similarity using several embeddings, including Word2Vec, Doc2Vec, TF-IDF, GloVe, and BERT and used the best Algorithm among them which was Cosine Similarity.

Summarization: The technique of condensing text without losing its meaningful structure is known as text summarization. We will use the model to summarize user-provided text after identifying the optimal similarity algorithm and the proper embedding technique. This model/technique is called automatic text summarization

Paraphrasing: To put it another way, paraphrasing is just another way to write, and it may be used to rephrase sentences and revise essays. We used the Parrot library which is built upon the T5 transformer to paraphrase our text. We also used the Pegasus Model and compared the results between the two. The pegasus Model performed better hence that model was used.

Teamwork:

I helped in planning the idea, features, and technical details/designing for the project. I planned the route map for the text summarizer. We created timelines for each phase of the project and I did my best to adhere to these timelines. I also helped in the integration of the models and our

front end. I am completely satisfied with my task because I learned a lot while creating the web page and developed a plan to deploy it and possibly launch it as a live website.
On a scale of 0-10 ,I'd give myself a 10.
I am also satisfied with my Team-members' work.

- Ujjwal Dubey: 10/10
- Rahul Gattu: 10/10

Everyone equally and actively participated in my team.