Mini-Project – 2B Web based on ML
(ITM 601)

T. E. Information Technology

By

Monik Kaole	**09**
Faustina Lazarus	**10**
Bennet Menezes	**11**
Dhruv Dave	**12**

Under Guidance of
Ms Vandana Patil
Professor

Department of Information Technology
St. Francis Institute of Technology
(Engineering College)

University of Mumbai

2021-2022

CERTIFICATE

This is to certify that the project entitled **"Castaway Corner Using ML"** is a bonafide work of **"Monik Kaole, Faustina Lazarus, Bennet Menezes, Dhruv Dave" roll no-09, roll no-10, roll no-11, roll no-12** submitted to the University of Mumbai towards completion of mini project work for the subject of **Mini Project – 2B Web Based on ML (ITM 601).**

Ms Vandana Patil
Supervisor/Guide

Dr. Joanne Gomes
HOD-IT

Examiners

1.----------

2.----------

Date:

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---	---
(Name of student and Signature)	(Name of student and Signature)
---	---
(Name of student and Signature)	(Name of student and Signature)

ABSTRACT

The world we live in today, is one which thrives on the internet. The internet has brought us closer to everything, and given us accessibility to everything, but this has a downside. Along with the increased accessibility, we also have to deal with the increased plagiarism. With numerous documents, research papers and books, available at our fingertips, plagiarism has overrun the book industry. Authors, researchers and scientists suffer heavily from this, as their own documents and books become unappealing and illegal. To avoid this, we have created a website which will allow writers and researchers to check their work and make sure its original, and not copied. Our website allows authors to upload their document for checking against a database. The database is first preprocessed using the algorithms, and then cross checked with the document uploaded. The result is shown in the form of a heat map, with different colors indicating the levels of plagiarism. The user can therefore refer to the heat map and make the appropriate changes and also download it for future use.

INDEX

List of Abbreviations

Sr. No.	Abbreviation	Full Form
1	N-Gram	A sequence of N words

List of Figures

Chapter 1
Introduction

1.1 Background

When it comes to writing books or documents, plagiarism is an increasingly widespread and growing problem in various fields. Several plagiarism techniques are used by fraudsters, ranging from a simple synonym replacement, sentence structure modification, to more complex methods involving several types of transformation. Human based plagiarism detection is difficult, not accurate, and a time-consuming process.

Our project provides authors and writers with an instrument, to find whether their content is original or plagiarized.

Literal plagiarism includes copy–paste operations and is usually easy to detect. More sophisticated forms of plagiarism may involve translation, summarization, and paraphrasing and are more difficult to recognize. One of the most difficult to detect and relatively less addressed forms is paraphrase plagiarism in which the original content may be completely reworded and altered considerably.

1.2 Scope of the project

We see Castaway Corner as an opportunity to attract a niche audience of authors by providing them with necessary tools at one place. We look to upgrade our website, and make it more fundamental, functional and convenient for authors and developers, by introducing features that allow our target audience to have a better, and a more fulfilling writing experience, by not just cross-checking and proofreading their work, but also by getting plagiarism checks. The widespread usage of paraphrasing techniques for plagiarizing text has motivated the current work.

The objective of this work is to investigate the suitability of utilizing a machine learning-based paraphrase recognition system for plagiarism detection. Various lexical, syntactic, and semantic features, which reflect the degree of similarity between the source and suspicious text, are extracted. These are used as input to a support vector machine classifier, which determines if the source text has been plagiarized.

1.3 Objectives

Our objective is to give authors a place where they can manage all their work, with the utmost efficiency. We want to create a safe place for authors, to manage and keep track of their work, and at the same time proofread and cross check their work, with the works of other authors, working in the same direction. We wish to create a community of authors, who can publish their stories and also use the platform's tools at the same time.The objectives of any plagiarism checker include to find the similarities in the text and ensure that the document is original.

It further implies that no part of the document is copied from another writer's work. This becomes critical in academic assignments, where it is very convenient to copy the work of other writers from online sources and present it as their own work. Plagiarism checkers have evolved as a result and most of the academic institutes and many business organizations use the plagiarism checker as a prerequisite to finding out if the submitted work is original. A plagiarism checker is a nice way for checking the originality of a document. However, the results of each plagiarism checker should be read and understood in order to address the shortcomings in the document.

Problem Statement

To keep perfect track of every sentence, statement, and error sometimes can prove to be a real burden. This in turn makes the story seem unappealing and incomplete and leads to the author losing out on what could be a great chance at success. A problem related to plagiarism is the misuse of sources. When using and acknowledging sources' ideas in their essays, writers should take care not to distort or misrepresent the original text's information in any way.

Chapter 2
Literature Review

Antiplag: AntiPlag is developed using the tri-gram sequence matching technique. Three sets of text based assignments were tested by AntiPlag and the results were compared against an existing commercial plagiarism detection tool. AntiPlag showed better results in terms of false positives compared to the commercial tool due to the pre-processing steps performed in AntiPlag. In addition, to improve the detection latency, AntiPlag applies a data clustering technique making it four times faster than the commercial tool considered. AntiPlag could be used to isolate plagiarized text based assignments from non-plagiarised assignments easily.

Plagiarism Checker: The plagiarism checker free utility offered on SmallSEOTools' is second to none due to the advantages it provides to its users. From students to teachers, researchers, writers, publishers, and bloggers, everyone can gain the top benefits of SmallSEOTools' plagiarism detector. The plagiarism checker online tool available on this platform is a super-fast utility that generates results within a matter of seconds. For using this plagiarism detector, the users won't have to follow any convoluted procedure. The user-friendly interface of this facility makes the process to check plagiarism free from all kinds of intricacies.

Turnitin: Turnitin is an Internet-based plagiarism detection service run by the American company Turnitin. The Turnitin software checks for potentially unoriginal content by comparing submitted papers to several databases using a proprietary algorithm. It scans its own databases and also has licensing agreements with large academic proprietary databases. The essays submitted by students are stored in a database used to check for plagiarism. This prevents one student from using another student's paper, by identifying matching text between papers. In addition to student papers, the database contains a copy of the publicly accessible Internet, with the company using a web crawler to continually add content to Turnitin's archive. It also contains commercial and/or copyrighted pages from books, newspapers, and journals.

Viper: Though the process is simple, and familiar to anyone who has used a plagiarism checker in the past, it does have a few interesting features. One of the biggest being its ability to match against a local database, the Web or both. This means that, if you have a pool of content you want to test against, you can do that with or without also checking the broader

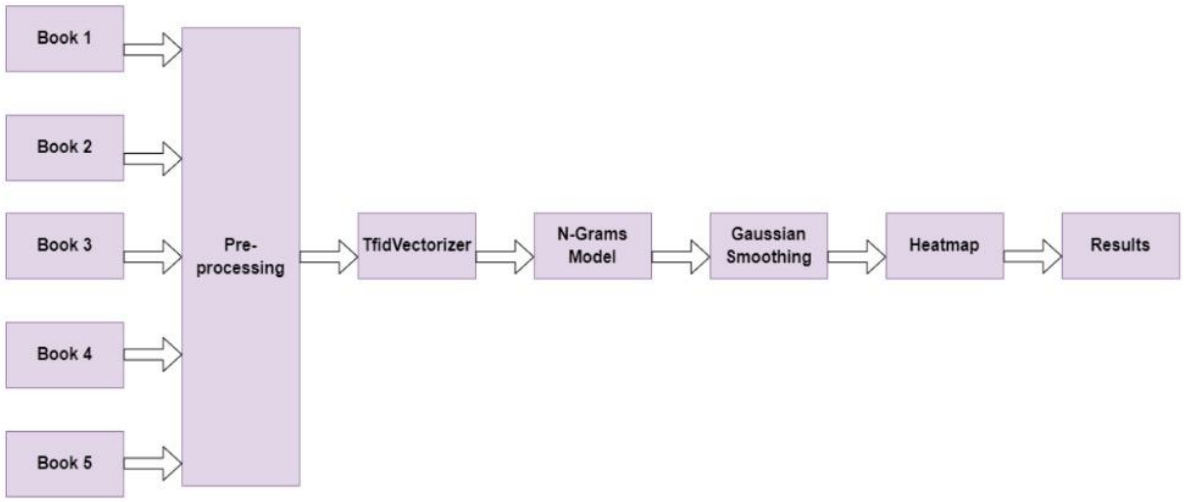
Web. Beyond those two features, both of which can actually be found in other applications or services, the rest of the application is fairly straightforward. While that is not a bad thing in and of itself, the problem is that it doesn't seem to do the job it set out to.

Sr No.	Methodology	Advantages	Disadvantages
1	Antiplag: An innovative plagiarism detection tool	This tool won first prize in plagiarism detection competition of 2012 as best plagiarism detection tool. This tool supports R&D ,creativity in science ,and increasing transport work in research. Diverse in text plagiarism was found large and no best tool existed which led to research.	Essays receiving high plagiarism rating, with proper citation, due to long quotes. It cannot prompt students to cite information that needs credibility It cannot check citation styles (APA, MLA, etc.)
2.	Plagiarism checker: A web tool for text plagiarism detection Motivation of Tool:	Text plagiarism detection involves a huge search task with comparative matching building complexity in software. As such graph- based search methodology has been adopted here. Trial experiments done on tools show marvelous results with level of similarity score and sources of plagiarism	plagiarism checkers will make an attempt to separate out attributed use, given the variety of attribution styles it isn't always possible. Given how common some phrases are in the English language, plagiarism

		found.	checker will report matches that are actually just coincidence.
3.	Turnitin: Technology to Improve student writing and research Motivation of Tool:	The tool learns overtime and as such would be the best tool in coming years. it also detects image-based plagiarism and is constantly under upgrade	Marks citations as names as plagiarized text
4.	Viper: Anti plagiarism scanner Motivation of Tool:	Viper is quickly turning into the copyright infringement checker of decision, ascending well beyond other written falsification checkers, with more than 10 billion assets examined and a simple interface which features potential ranges of literary theft in your work. An awesome device for understudies, educators, speakers and scholastics, Viper will examine billions of assets to check for occurrences of written falsification in expositions, articles, theses, bits of coursework, sites and that's only the tip of the iceberg.	Writeups receiving high plagiarism rating, with proper citation, due to long quotes. Citations marked as plagiarized text.

Chapter 3
Proposed Work

3.1 Architecture Details (module description)



The dataset contains 4 different books written by different authors. We will then convert those datasets, into txt format and will feed into the code or model which we have trained. After applying the vectorize and similarity algorithms in the dataset, we get back results in float format, which if we multiply by 100, we get it into the percentage the dataset has been plagiarized.

3.1.1 Books

Dataset:

Book1: Famous Composers

Book2: Living in the light

Book3: Start new

Book4: Emma

Book5: The Adventures of Dr. Thorndyke

Books dataset: https://www.kaggle.com/bilalyussef/google-books-dataset

The above datasets were downloaded from various sources.

3.1.2 Pre-Processing

We removed special characters from our books because the algorithm won't process these characters. We also removed it from the database we are using.

3.1.3 TfidVectorizer

TfidVectorizer /Vector embeddings are one of the most fascinating and useful concepts in machine learning. They are central to many NLP (Natural Language Processing), recommendation, and search algorithms.

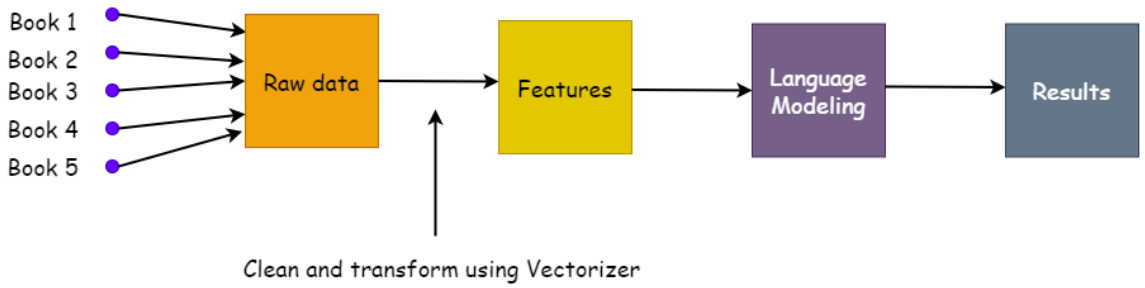


Fig 3.1.2.1 Feature Engineering

3.1.4 N-Grams Model

Language Modeling: A language model is a statistical model that captures relevant linguistic features of the corpus on which it is trained. At a basic level, it should capture the frequency distribution of letters and words. A more advanced language model should capture syntactic and grammatical dependencies, such as agreement and inflection, and semantic properties, such as which words are likely to occur in a given context. Language models are typically used for two main tasks: scoring and generation. In scoring, the language model gives a probability score to a certain word occurring in a given context. There are two different approaches for language modeling — N-gram models and RNNs.

N-gram Language Model: An N-gram language model scores words based on the preceding window of context. Although the N-gram model is not very sophisticated and fails to handle long-range dependencies and abstract semantic information, we can actually see this as a feature rather than a bug for this task. Other language models, such as those based on Recurrent Neural Networks or Transformers, are better at capturing long-range dependencies and higher levels of abstraction. For plagiarism, however, the emphasis is on copied sequences of words, not on similarities at an abstract level. A paraphrasing should not set off an alarm, but a direct copying should.

3.1.5 Gaussian Smoothing

The effect of Gaussian smoothing is to blur an image, in a similar fashion to the mean filter. The degree of smoothing is determined by the standard deviation of the Gaussian. (Larger standard deviation Gaussians, of course, require larger convolution kernels in order to be accurately represented.)

3.1.6 Heatmap

We can represent a book as a heatmap image where each pixel corresponds to the score of one word. This allows us to quickly gauge if plagiarism is likely, and which parts of a text were most likely to have been plagiarized. Visualizing information in this way is more useful than looking at an array of numerical scores or a summary statistic of all the scores.

3.1.7 Results

The results of the plagiarism detection are shown in the form of a heat map. The entered document is shown on the left, and the heat map is shown on the right. We can see the levels or amount of plagiarism in each line by hovering over the different sections of the heat map to see the percentage of plagiarism.

Chapter 4
Implementation Details

4.1 Dataset Details

We are using a dataset, which includes books written by different authors.

Book1: Famous Composers

FamousComposers - Notepad

File Edit View

On the river Leitha, in Lower Austria, and some fifteen miles south-east from Vienna, is a village so insignificant that it is not set down on the ordinary maps. It is called Rohrau, and there, during the night of March 31, 1732, and descended from a long line of humble hand-toilers, was born Franz Joseph Haydn, who was destined to make the family name immortal. His father, Mathias Haydn, was a master wheelwright, whose father, Thomas Haydn, had followed the same occupation. The mother of Franz, or Joseph, as he is now called, was Maria Koller, daughter of the market inspector of the locality, and a cook in the household of Count Harrach, the lord of the village. The ancestry of the Haydns is undistinguished as far back as it can be traced. This union of the wheelwright and the cook resulted in a family of twelve children, of whom three developed into musicians. They were Franz Joseph, the subject of this sketch, Johann Michael, the church composer, and Johann Evangelist, a singer of no special excellence. There is no record of musical talent on the side of either the Haydns or the Kollers previous to its appearance in the family of Mathias, and its sudden development in three of the offspring of this marriage is inexplicable. In addition to his occupation as a wheelwright, Mathias Haydn officiated as sexton of his parish. Both he and his wife were able to sing sufficiently well to increase their scant earnings by singing in church on Sundays and holidays, and at fairs and festivals. They also indulged in music at home, after a rude fashion, the father accompanying the voices on the harp, which he had learned to play by ear. The parents of the future composer were hard-working people who feared God, and so thoroughly did they instill their religious feelings into their children, that Haydn felt the influence of this early discipline all through his long life. Of his earliest years but little is known except that, while yet a tender child, he began to manifest the musical instinct that was in him by singing the simple tunes that his father was able to strum on the harp, and by exciting wonder at the correctness of his ear and his keen sense of rhythm. These gifts, however, are by no means rare in children, and the possession of them does not necessarily insure that their possessors shall develop into Haydns and Mozarts.

Book2: Living In the Light

Living in the Light - Notepad

File Edit View

I have always had a burning desire to understand how the universe works, what life is all about, and the meaning and purpose for which I am here. In retrospect, I can see that my entire life has been devoted to my search for truth and understanding. I was brought up in a very intellectual, well-educated, non- religious family. My parents were essentially atheists, and very early on, I remember having the attitude that a belief in God was a human fabrication, a fantasy, a superstition created to help people feel better about the totally unexplained, and unexplainable, predicament we seem to find ourselves in. Human existence, or anyother kind of existence, was simply an accident of nature and had no particularly fathomable meaning. I preferred to admit that I didn't know how we got here or why, rather than to adopt a simplistic explanation merely to gain a sense of security. I believed that truth was rational and anything that couldn't be proved scientifically didn't exist. I also felt somewhat condescending toward people who were weak enough to have to make up a god to believe in. The positive side of this upbringing was that I didn't get a lot of the rigid dogma and deeply negative messages about right and wrong, heaven and hell, and sin that so many people receive in their early religious training. On the other hand, I had no conscious concept or experience of the spiritual dimension of life, and no answers or the questions I had about the meaning and purpose of my life. My parents really wanted a child, and were very loving to me.Unfortunately they were unable to work out their own relationship and were divorced when I was two years old. Although I don't remember it clearly, I know this event had a major impact on my life and affected my later patterns in relationship. After the divorce, I lived with my mother who never remarried or had any other children. My father did remarry, and I often visited my father and his other family. My mother developed a successful career as a city planner in the days when there were few women in that field. She dealt with the usual challenges of single parenting trying to balance the needs of her child with the demands of her work. Being the only child of a working mother, I developed a strong sense of responsibility and self-sufficiency quite early. My mother is a very adventurous person. She loves to try new things, and for me, she was a great role model of fearlessness and pioneer spirit. She had been one of the first educated American women in her generation to have natural childbirth. I was the first baby her doctor had ever delivered without an anesthetic. I was blessed with a very fortunate birth. My mother loves to explore new places and we traveled a lotwhen I was a child all over the United States, to the West Indies, Mexico, Hawaii, Europe. We also moved frequently whenever my mother changed jobs. Until I was about fifteen, I had never lived in one place longer than two or three years.My mother's family had been Quakers and we still used the plain language when speaking to my grandmother saying thee rather than you for the Quakers is an acknowledgement of the god within each person. So, on a deep level, I absorbed the profound respect for spirit and concern for humanity that is woven into the fabric of the Quaker religion, which I feel had a strong influence on me later in my life. When I was fourteen years old I went through an enormous emotional crisis. Triggered initially by the collapse of my first romance, it snowballed into a deep and longlasting existential despair. I took a long hard look at life and recognized that there was really no point or meaning to it. I could see that all the things that were supposed to provide significance in life education, success, relationships, money were in themselves ephemeral, meaningless, and empty. There didn't seem to be any- thing else to fill the void. I was deeply disillusioned and depressed, and basically remained in that state for several

Book3: Start New

StartNew - Notepad

File Edit View

We already have everything we need. There is no need for self-improvement. All these trips that we lay on ourselves the heavy-duty fearing that we're bad and hoping that we're good, the identities that we so dearly cling to, the rage, the jealousy and the addictions of all kinds never touch our basic wealth. They are like clouds that temporarily block the sun. But all the time our warmth and brilliance are right here. This is who we really are.We are one blink of an eye away from being fully awake. Looking at ourselves this way is very different from our usual habit. From this perspective we don't need to change: you can feel as wretched as you like, and you're still a good candidate for enlightenment. You can feel like the world's most hopeless basket case, but that feeling is your wealth, not something to be thrown out or improved upon. There's a richness to all of the smelly stuff that we so dislike and so little desire. The delightful things what we love so dearly about ourselves, the places in which we feel some sense of pride or inspiration these also are our wealth. With the practices presented in this book, you can start just where you are. If you're feeling angry, poverty-stricken, or depressed, the practices described here were designed for you, because they will encourage you to use all the unwanted things in your life as the means for awakening compassion for yourself and others. These practices show us how to accept ourselves, how to relate directly with suffering, how to stop running away from the painful aspects of our lives. They show us how to work openheartedly with life just as it is. When we hear about compassion, it naturally brings up working with others, caring for others. The reason we're often not there for others whether for our child or our mother or someone who is insulting us or someone who frightens us is that we're not there for ourselves. There are whole parts of ourselves that are so unwanted that whenever they begin to come up we run away.

Book 4: Emma

Emma - Notepad

File Edit View

Emma Woodhouse handsome clever and rich with a comfortable home and happy disposition seemed to unite some of the best blessings of existence and had lived nearly twentyone years in the world with very little to distress or vex her

She was the youngest of the two daughters of a most affectionate indulgent father and had in consequence of her sisters marriage been mistress of his house from a very early period Her mother had died too long ago for her to have more than an indistinct remembrance of her caresses and her place had been supplied by an excellent woman as governess who had fallen little short of a mother in affection

Sixteen years had Miss Taylor been in Mr Woodhouses family less as a governess than a friend very fond of both daughters but particularly of Emma Between them it was more the intimacy of sisters Even before Miss Taylor had ceased to hold the nominal office of governess the mildness of her temper had hardly allowed her to impose any restraint and the shadow of authority being now long passed away they had been living together as friend and friend very mutually attached and Emma doing just what she liked highly esteeming Miss Taylors judgment but directed chiefly by her own

The real evils indeed of Emmas situation were the power of having rather too much her own way and a disposition to think a little too well of herself these were the disadvantages which threatened alloy to her many enjoyments The danger however was at present so unperceived that they did not by any means rank as misfortunes with her

Sorrow camea gentle sorrowbut not at all in the shape of any disagreeable consciousnessMiss Taylor married It was Miss Taylors loss which first brought grief It was on the weddingday of this beloved friend that Emma first sat in mournful thought of any continuance The wedding over and the bridepeople gone her father and herself were left to dine together with no prospect of a third to cheer a long evening Her father composed himself to sleep after dinner as usual and she had then only to sit and think of what she had lost

The event had every promise of happiness for her friend Mr Weston was a man of unexceptionable character easy fortune suitable age and pleasant manners and there was some satisfaction in considering with what selfdenying generous friendship she had always wished and promoted the match but it was a black mornings work for her The want of Miss Taylor would be felt every hour of every day She recalled her past kindnessthe kindness the affection of sixteen yearshow she had taught and how she had played with her from five years oldhow she had devoted all her powers to attach and amuse her in healthand how nursed her through the various illnesses of childhood A large debt of gratitude was owing here but the intercourse of the last seven years the equal footing

Ln 1, Col 1 100% Unix (LF) UTF-8

Book 5: The Adventures of Dr. Thorndyke

The Adventures of Dr. Thorndyke - Notepad

File Edit View

THE MECHANISM OF CRIME

A surprising amount of nonsense has been talked about conscience On the one hand remorse or the againbite as certain scholars of ultraTeutonic leanings would prefer to call it on the other hand an easy conscience these have been accepted as the determining factors of happiness or the reverse

Of course there is an element of truth in the easy conscience view but it begs the whole question A particularly hardy conscience may be quite easy under the most unfavourable conditionsconditions in which the more feeble conscience might be severely afflicted with the againbite And then it seems to be the fact that some fortunate persons have no conscience at all a negative gift that raises them above the mental vicissitudes of the common herd of humanity

Now Silas Hickler was a case in point No one looking into his cheerful round face beaming with benevolence and wreathed in perpetual smiles would have imagined him to be a criminal Least of all his worthy highchurch house keeper who was a witness to his unvarying amiability who constantly heard him carolling lightheartedly about the house and noted his appreciative zest at mealtimes

Yet it is a fact that Silas earned his modest though comfortable income by the gentle art of burglary A precarious trade and risky withal yet not so very hazardous if pursued with judgment and moderation And Silas was eminently a man of judgment He worked invariably alone He kept his own counsel No confederate had he to turn Kings Evidence at a pinch no one he knew would bounce off in a fit of temper to Scotland Yard Nor was he greedy and thriftless as most criminals are His scoops were few and far between carefully planned secretly executed and the proceeds judiciously invested in weekly property

In early life Silas had been connected with the diamond industry and he still did a little rather irregular dealing In the trade he was suspected of transactions with IDBs and one or two indiscreet dealers had gone so far as to whisper the ominous word fence But Silas Smiled a benevolent smile and went his way He knew what he knew and his clients in Amsterdam were not inquisitive

Such was Silas Hickler As he strolled round his garden in the dusk of an October evening he seemed the very type of modest middleclass prosperity He was dressed in the travelling suit that he wore on his little continental trips his bag was packed and stood in readiness on the sitting room sofa A parcel of diamonds purchased honestly though without impertinent questions at Southampton was in the inside pocket of his waistcoat and another more valuable parcel was stowed in a cavity in the heel of his right boot In an hour and a half it would be time for him to set out to catch the boat train at the junction

Ln 1, Col 1 100% Unix (LF) UTF-8

In this dataset, there are no missing values nor any numerical or nominal values. The datasets are books of different authors. The reason why this dataset was chosen was regarding our last semester's project Castaway Corner which is a book/story/game writing website.

The database we are using for verifying whether the dataset is plagiarized is database.txt where it contains 10 different books.

4.2 Algorithm Details

The algorithm we are using are:

Language Modeling:

A language model is a statistical model that captures relevant linguistic features of the corpus on which it is trained. At a basic level, it should capture the frequency distribution of letters and words. A more advanced language model should capture syntactic and grammatical dependencies, such as agreement and inflection, and semantic properties, such as which words are likely to occur in a given context. Language models are typically used for two main tasks: scoring and generation. In scoring, the language model gives a probability score to a certain word occurring in a given context.

To apply language modeling to plagiarism detection, we have to train a language model on a bunch of text that you think people may copy from. This aggregated dataset will be our training data that we use to build a language model, which captures the statistical features of the text. Once we have this language model, we can run student work through the language model to assign scores. A higher score means the work is more predictable from the

training data, and represents a higher likelihood of plagiarism. There are two different approaches for language modeling — N-gram models and RNNs.

N-gram Language Model:

An N-gram language model scores words based on the preceding window of context. Although the N-gram model is not very sophisticated and fails to handle long-range dependencies and abstract semantic information, we can actually see this as a feature rather than a bug for this task. Other language models, such as those based on Recurrent Neural Networks or Transformers, are better at capturing long-range dependencies and higher levels of abstraction. For plagiarism, however, the emphasis is on copied sequences of words, not on similarities at an abstract level. A paraphrasing should not set off an alarm, but a direct copying should.

To implement an N-gram language model in Python, we can use the NLTK library (one option among many). The basic steps of training a language model are the following:

Read in and pre-process a training data file (e.g. remove punctuation, casing, and formatting). We would be left with something like this is an example sentence

Tokenize the training data (i.e. separate into individual words) and add padding at the beginning. This would leave us with ['<s>', '<s>', 'this', 'is', 'an', 'example', 'sentence'] .

Generate N-grams from the training data using the nltk.ngrams or nltk.everygrams methods. For an N-gram size of 3, this would give us something like [('<s>', '<s>', 'this'), ('<s>', 'this', 'is'), ('this', 'is', 'an'), ('is', 'an', 'example'), ('an', 'example', 'sentence')] . Note that everygrams would also give us the unigrams and bigrams, in addition to trigrams.

Fit a model using these N-grams. NLTK has various models that can be used, ranging from a basic MLE (Maximum Likelihood Estimator) to more advanced models like WittenBellInterpolated that use interpolation to deal with unseen N-grams.

Once we have the trained model, it supports various operations such as scoring a word given a context, or generating a word from the learned probability distribution.

- Read in and pre-process the testing data (the “books”).
- Tokenize the testing data.
- For each word in the text, call model.score() on that word, with the previous N-1 words as the context argument.

This gives us a list of scores between 0 and 1, one per word, where a larger score represents a higher probability that the given word was plagiarized.

$$P(the \mid its\ water\ is\ so\ transperant\ that) = C(\ its\ water\ is\ so\ transperant\ that\ the) \,/\, C(\ its\ water\ is\ so\ transperant\ that)$$

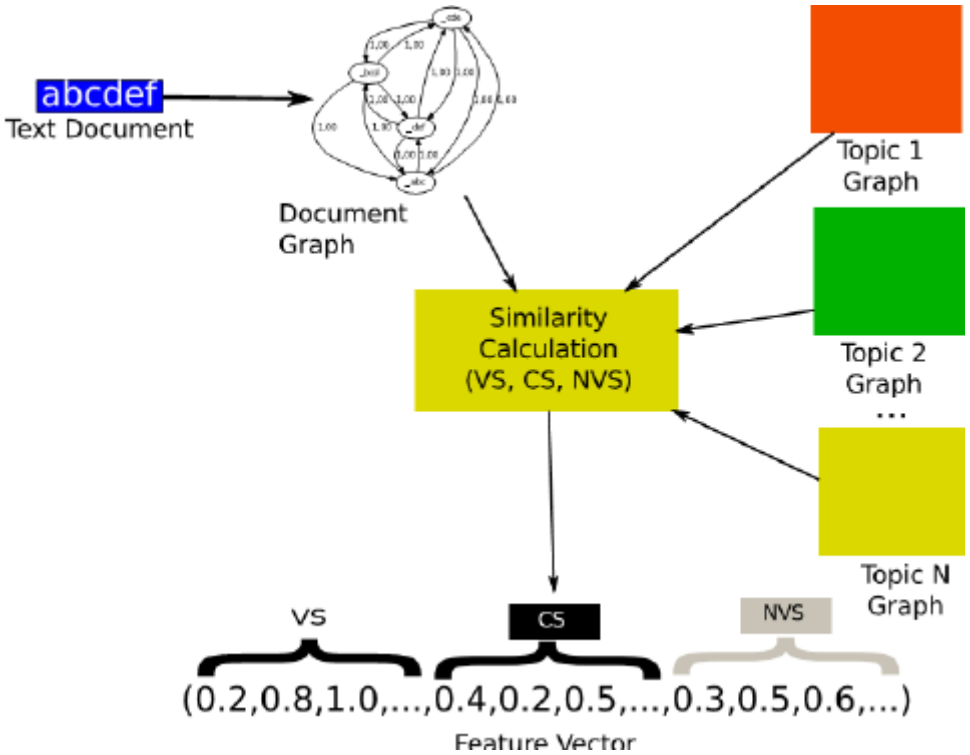


Fig 4.2.1 N-Gram

Vector embeddings

Vector Embeddings are one of the most fascinating and useful concepts in machine learning. They are central to many NLP (Natural Language Processing), recommendation, and search algorithms.

Visualization:

We can represent a book as a heatmap image where each pixel corresponds to the score of one word. This allows us to quickly gauge if plagiarism is likely, and which parts of a text were most likely to have been plagiarized. Visualizing information in this way is more useful than looking at an array of numerical scores or a summary statistic of all the scores.

4.3 Web Based Project Details

Dash is simple enough that you can bind a user interface to your code in less than 10 minutes. Dash apps are rendered in the web browser. We can deploy your apps to VMs or Kubernetes clusters and then share them through URLs. Since Dash apps are viewed in the web browser, Dash is inherently cross-platform and mobile ready. Built on top of Plotly.js, React, and Flask, Dash ties modern UI elements like dropdowns, sliders and graphs directly to analytical python code. Dash apps are composed of two parts. The first part is the "layout" of the app which basically describes how the application looks like. The second part describes the interactivity of the application. The application will run on http://127.0.0.1:8050/

A heat map is a two-dimensional representation of information with the help of colors. Heat maps can help the user visualize simple or complex information. Heat maps are used in many areas such as defense, marketing and understanding consumer behavior. Heat maps can be created with the help of software applications such as Microsoft Excel and others. We came up with the following method:

Display K words per line (K=8). This is the heatmap width in pixels. Then calculate the height (number of words in testing data divided by K).

Due to the small size of the dataset and the challenges of interpolation, there is some uncertainty in the assigned scores, so I applied Gaussian smoothing to the scores.

Reshape the array of scores into a rectangle of K columns and height rows. This requires adding zero padding to ensure the array is the correct size.

Using Plotly Heatmap to show the image using the color scale of your choice.
Show the K words of text as a y-axis tick label next to the corresponding row of the heatmap for easy side-by-side comparison. Adjust the hover data so that each pixel shows its corresponding word on hover.

4.4 Screenshots of GUI with Explanation

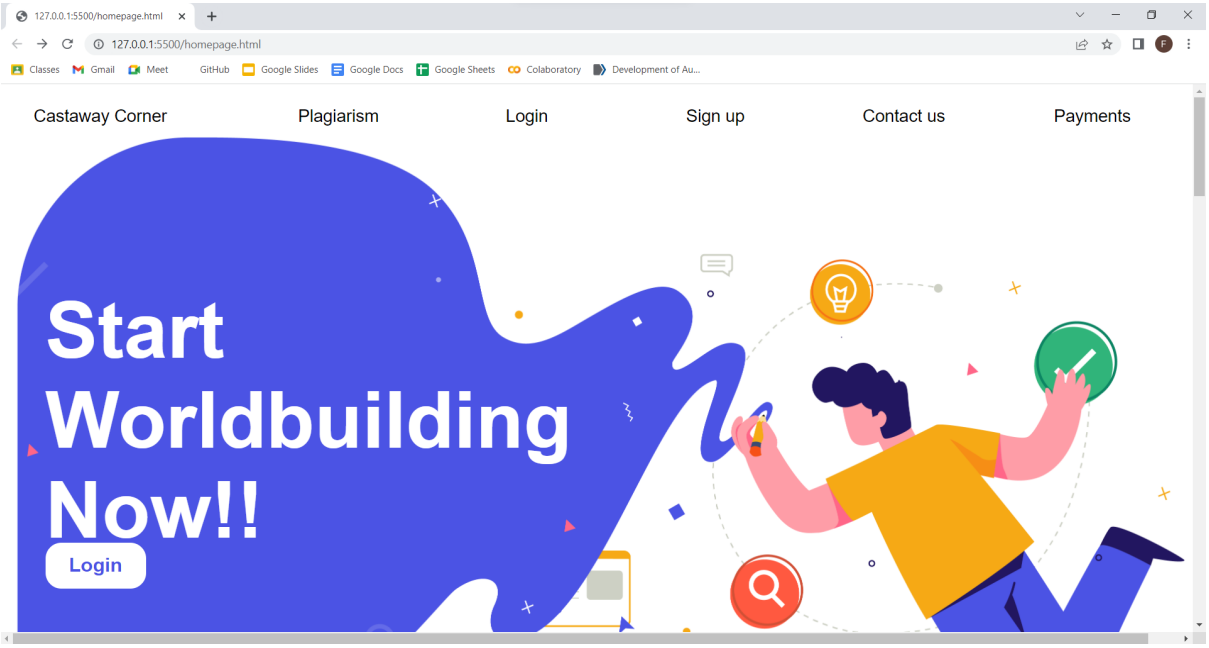


Fig 4.4.1 Home Page

This is the home page of our website. As you can see, in the navigation bar, we have the option for plagiarism. This will take us to the plagiarism page.

Fig 4.4.2 Plagiarism Page

This is the main plagiarism page. Here we will be able to see our file and the level of plagiarism in each word and line.

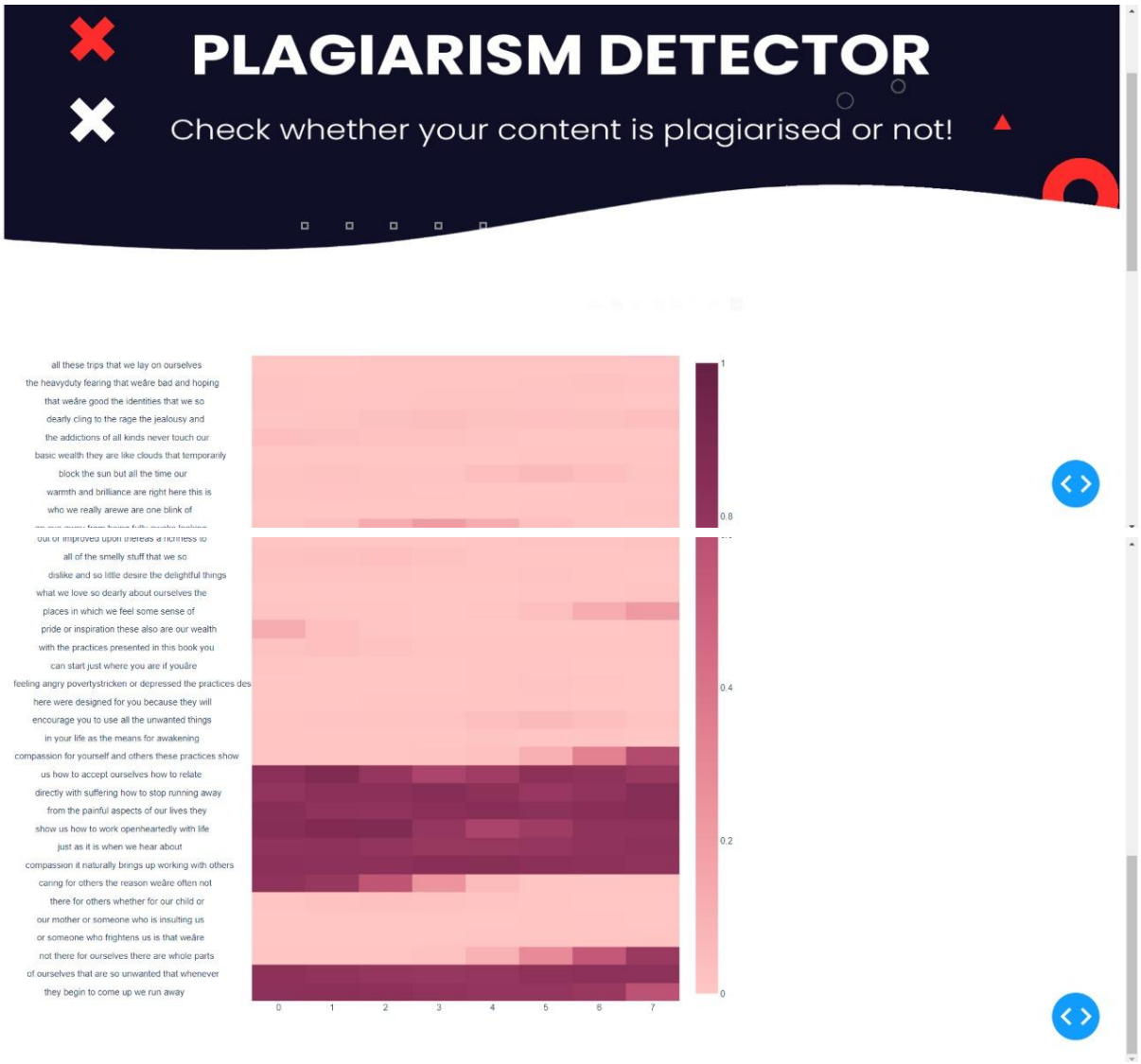


Fig 4.4.3 Plagiarism Page (Output)

We have made use of HTML and CSS to design our UI. As shown in the screenshots above, we have implemented our plagiarism detection algorithm using a heatmap. The heat map shows the plagiarized words and sentences in the book, with respect to the books in the database accessed. By hovering the mouse over the heat map, they can see the percentage The user can download a snapshot of the screen, to refer for future use. They can also enlarge the screen to see each individual word and the plagiarism percentage.

4.5 Performance Metrics

4.4.1 Accuracy

Accuracy is simply a ratio of correctly predicted observation to the total observations. If we have high accuracy then our model is best as accuracy is a great measure but only when we have symmetric datasets where values of false positives and false negatives are almost the same . For our model, we have got 0.8932 which means our model is approx. 89% accurate. So accuracy can be defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

$$Accuracy = Number\ of\ correct\ prediction/Total\ number\ of\ predictions$$

Fig 4.4.1.1 Accuracy

4.4.2 Precision

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances. Precision is a good measure to determine when the cost of False Positive is high. For our project, plagiarism detection. In this detection, a false positive means that a text that is non-plagiarised (actual negative) has been identified as plagiarized (predicted).

$$Precision = True\ Positive/True\ Positive + False\ Positive$$

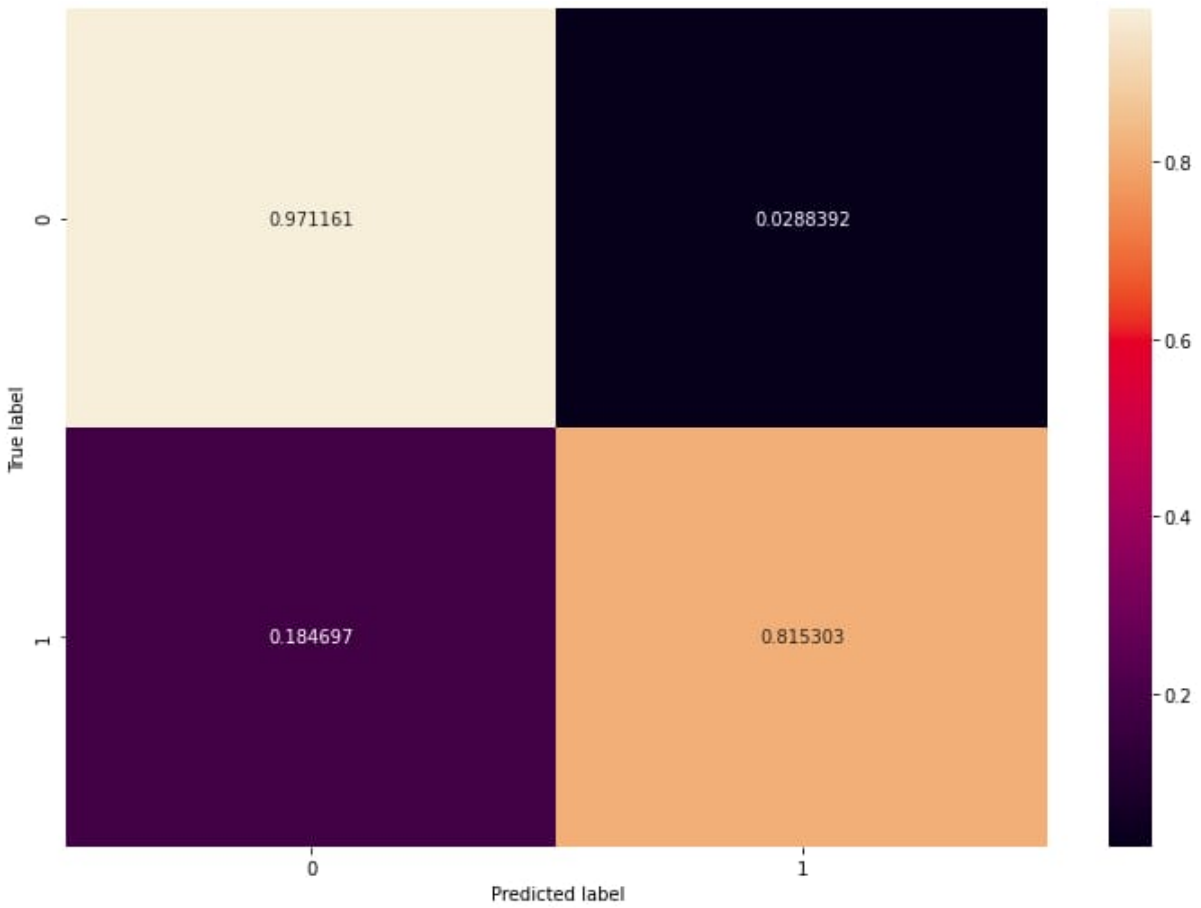
4.4.3 Recall

Recall is the opposite of Precision. It actually calculates how many of the Actual Positives our model captures through labeling it as Positive (True Positive).

$$Recall = True\ Positive/Total\ Actual\ Positive$$

Chapter 5
Results and Discussion

Confusion matrix:



Thus Accuracy our model is (TP+TN)/(TP+FP+TN+FN)= 0.8932=89%

Precision= TP/(TP+FP)= 0.971161= 97%

Recall= TP/(TP+FN)= 0.54362= 54%

As the input for the task of plagiarism detection is passage-level text, the sentence-level paraphrase recognition system has been modified to handle passages. The source and suspicious passages are split into sentences. In order to determine the closest matching source sentence for the suspicious passage sentences, the extent of unigram overlap is computed between the sentences in both the passages. For every sentence in the suspicious passage, the source sentence, which has the highest word overlap, is paired with it. The dataset contains 4 different books written by different authors. We will then convert those dataset into txt format

and will feed into the code or model which we have trained. After applying the vectorize and similarity algorithms in the dataset, we get back results in float format which if we multiply by 100 we get it into the percentage the dataset has been plagiarized.

Chapter 6
Conclusion and Future Scope

With the current state of plagiarism, there is an urgent need to design an effective mechanism for plagiarism detection. We proposed a new system for the detection of plagiarism based on machine learning methods. Its interest is in extracting the plagiarized part, without losing the sense and character of the document itself. The proposed system not only detects plagiarism, but also the probability of existence of each type of plagiarism. This shows that employing paraphrase recognition techniques, is a promising direction to explore, in the development of plagiarism detection systems.

We plan to further develop our website and add this detection system on our website like buttons, and add a grammatical checker and error checker to increase our resources to a greater extent, in order to help authors and developers to the best of our abilities. Also we will add additional features of plagiarism detection, which will detect if plagiarism exists or if the content is written from the web or not and which part is plagiarized and which part is not.

References

[1] S. Autade and A. Suryawanshi, *"A Systematic Literature survey on Plagiarism detection Tools"* International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 7, Issue 2, February 2018, Pune, India.

[2] J. Y. B. Katta, *"Machine Learning for Source-code Plagiarism Detection"*, Master of Science in Computer Science and Engineering, International Institute of Information Technology, July 2018, Hyderabad, India.

[3] E. M. Hambi and F. Benabbou, *"A New Online Plagiarism Detection System based on Deep Learning"*, (IJACSA) International Journal of Advanced Computer Science and Applications, University Hassan II, Casablanca, Morocco, Vol. 11, No. 9, 2020

[4] A. Chitra and A. Rajkumar, *"Plagiarism Detection Using Machine Learning-Based Paraphrase Recognizer"*, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India, October 8, 2014.

[5] Simple Plagiarism Detection in Python,
https://towardsdatascience.com/simple-plagiarism-detection-in-python-2314ac3aee88,
Jan 25, 2021.

Acknowledgement

We are thankful to our college **St. Francis Institute of Technology** for giving us this chance to gain exposure in solving real world problems and acquire practical knowledge and skill sets that will prove to be very crucial to our long-term career prospects. We would take this opportunity to express our sincerest gratitude to our esteemed director **Bro. Jose Thuruthiyil**, our principal **Dr. Sincy George** and our **HOD, Dr. Joanne Gomes** for their encouragement, the direction that they give to our college and us students, and the facilities provided to us.

This project, and the research that we undertook, could not have been realized without the utmost support of our project guide **Ms. Vandana Patil**, who guided us every step of the way, starting from the conception of the project, right up to the execution of the finished solution.