USER MANUAL FOR SIMILABS 2022

A GUIDE TO USE SIMILABS

DEVELOPMENT TEAM:

HANNO VISAGIE	31594883
HANO STRYDOM	31597793
MICHAEL ROSIN	31704948
LLEWELLYN ANTHONY	32969694
ANNIKA DU TOIT	31842534
SHENÉ BOSHOFF	31775357

23/11/2022

Version 1.0.0

Table of Contents

1. Introduction

Thank you for choosing SimiLabs, a system designed to identify academic misconduct committed by students at North-West University.

1.1. Background of the application

The NWU Registrar must address plagiarism by evaluating each case individually and appointing experts to prepare technical reports. External subject matter experts (SMEs) are requested to examine the technical reports with an additional report that provides a deeper insight into the alleged plagiarism if the technical reports do not self-evidently emphasize the severity of the plagiarism. Manually comparing the allegedly plagiarized text in issue with the original text as evidence text is a requirement for the technical report, which can become difficult and lead to certain similarities being overlooked.

1.2. Purpose of the application

Through the use of a similarity metric, the software should reduce the time spent manually comparing two texts and generalise the assessment of how severe the conjectured copying is. The software must combine text-matching skills with stylometric analytics to provide more accurate reports, better explain academic misbehaviour, and enable improved decision-making.

1.3. System Capabilities

The system developed and shown in this user manual has the following capabilities:

- The system is capable of being used by multiple users at any given time.
- The user can upload documents to do a quick text analysis, extensive text analysis, and a stylometric analysis.
- The user can download reports of the results of the various analyses.

The following sections will explain how each of these functionalities works and will also assist with accessing the application on a web browser.

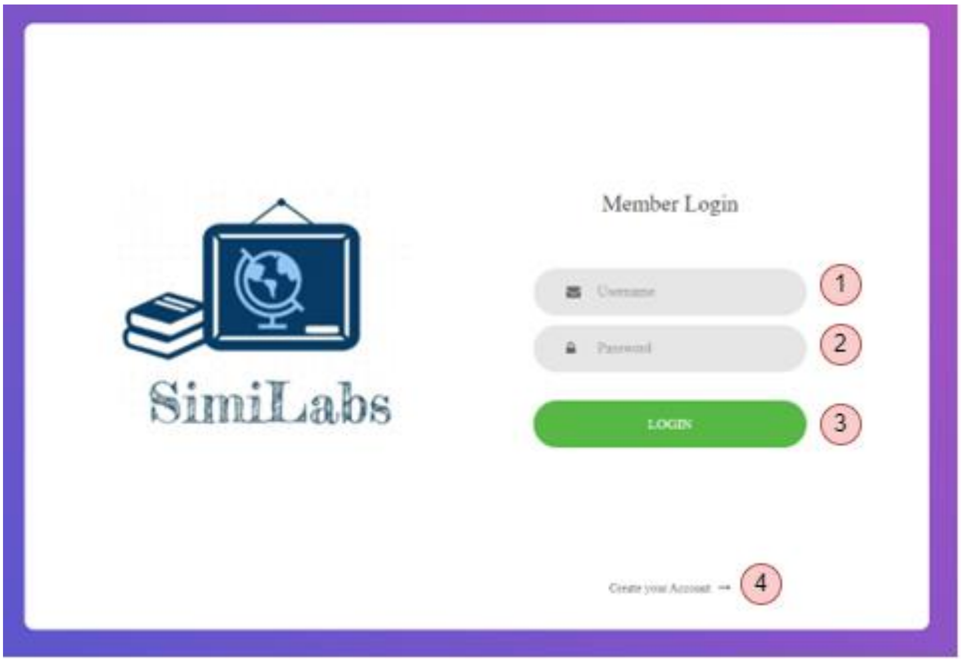
2. Using the system

2.1. Accessing the application via a web browser

The application can be accessed through any web browser on a desktop device, including a desktop PC and a laptop, by entering the following URL in the search bar:

2.2. Logging in

The first page the user will see is the login page:

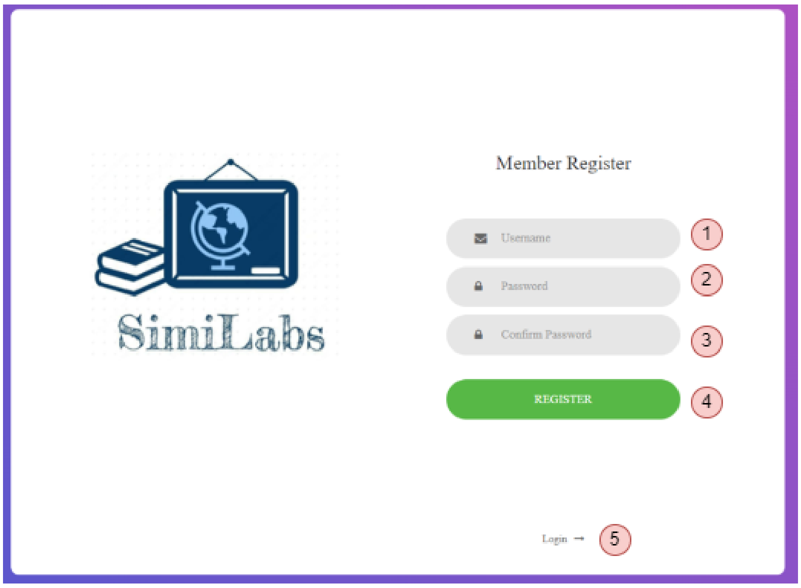


1) The user inputs their username in this box.
2) The user inputs their password in this box.
3) The user clicks on this button to log in with their credentials.
4) The user clicks on this to be redirected to the registration page if they do not have an account.

The user must log in before the application can be used.

2.3. Creating an Account

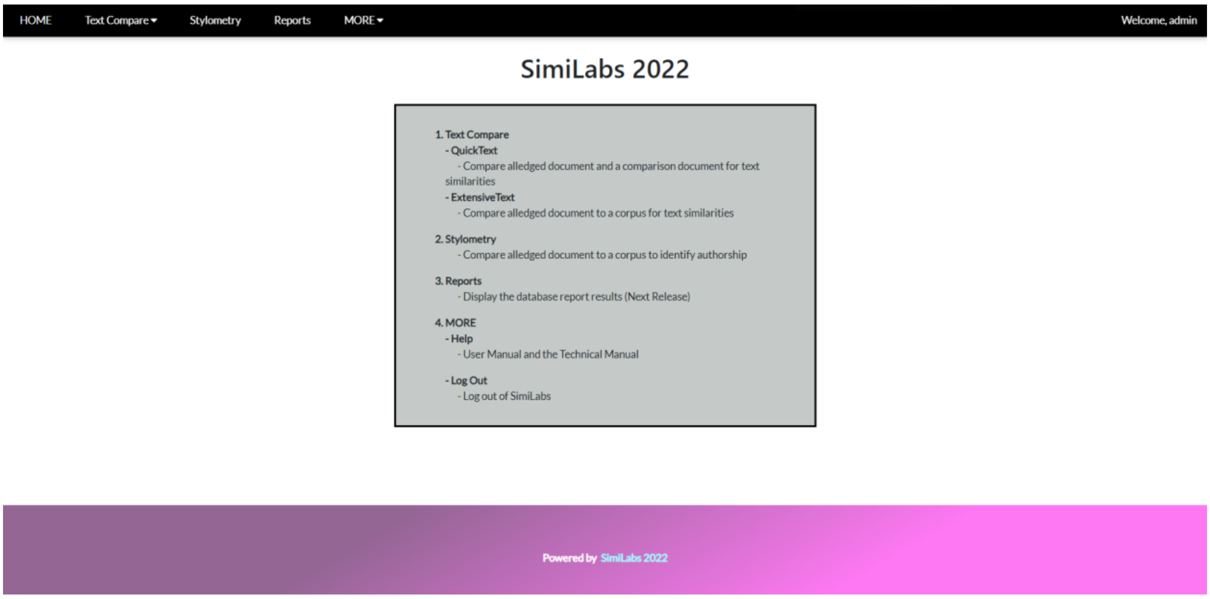
If the user does not already have an account, they can create one by clicking on Create your Account at the bottom of the screen. This will redirect the user to the registration page:



1) The user chooses a username. The username must be a valid email address.
2) The user chooses a strong password.
3) The user enters the password again to make sure they did make any errors.
4) The user clicks on this button to register a new account with their credentials.
5) The user clicks on this to go back to the login page to log into the application.

2.4. Navigating the Home page

Upon successful log-in, the user will be redirected to the home page:



SimiLabs 2022

1. Text Compare
 - QuickText
  - Compare alledged document and a comparison document for text similarities
 - ExtensiveText
  - Compare alledged document to a corpus for text similarities

2. Stylometry
  - Compare alledged document to a corpus to identify authorship

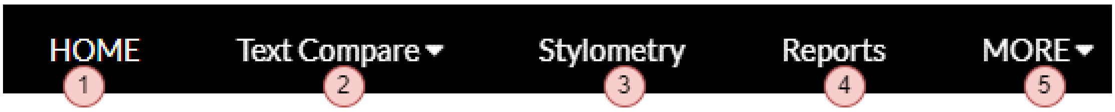
3. Reports
  - Display the database report results (Next Release)

4. MORE
 - Help
  - User Manual and the Technical Manual

 - Log Out
  - Log out of SimiLabs

Powered by SimiLabs 2022

The toolbar at the top of the screen shows the different options available to choose from:

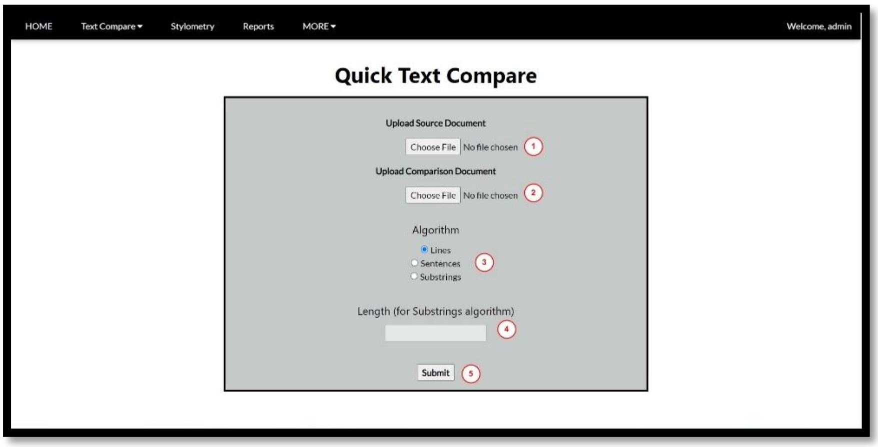


1) This redirects the user to the home page.
2) The user has an option of a quick text comparison or an extensive text comparison.
3) Redirects the user to the stylometry page.
4) Reports
5) The user can choose between a help option and logging out of the application.

3. Using the Text Comparison Feature

3.1. Quick text comparison

QuickText landing page



Quick Text Compare

Upload Source Document

Upload Comparison Document

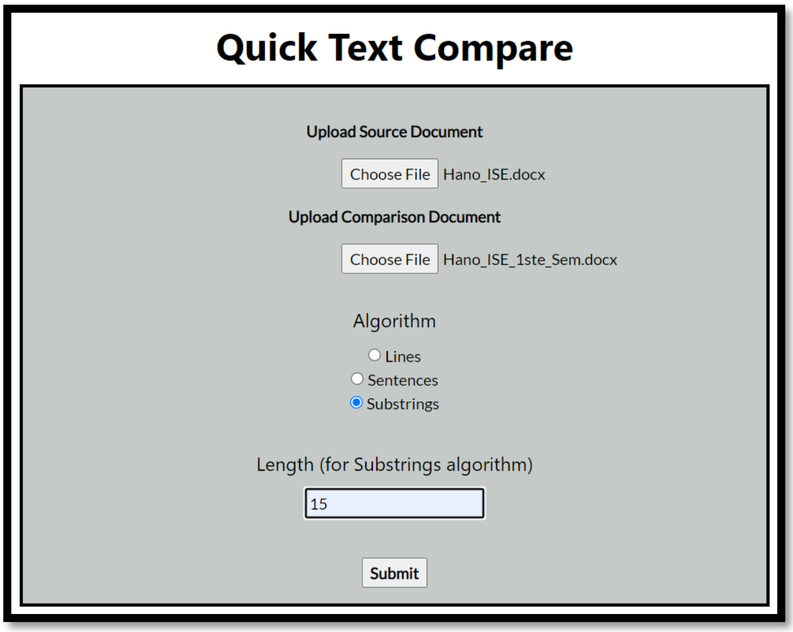
Algorithm

● Lines
○ Sentences
○ Substrings

Length (for Substrings algorithm)

1. Upload an alleged document
   - This is a document of the student you suspect of plagiarism.
2. Upload a comparison document
   - This is a document of a student you suspect the above-mentioned student plagiarised from.
3. Select a comparison algorithm
   - Lines
     - This compares the document line for line and highlights the matches.
   - Sentences
     - This compares every sentence in the alleged document, with every sentence in the comparison document and highlights the matches.
   - Substrings
     - Returns substrings with a specified length that are the same
4. Substring Length
   - Give a substring length you want to use to compare the documents.
5. Submit
   - Submit the files, values, and options, that will redirect you to the report page.

Example of QuickText input



Quick Text Compare

Upload Source Document

Choose File Hano_ISE.docx

Upload Comparison Document

Choose File Hano_ISE_1ste_Sem.docx

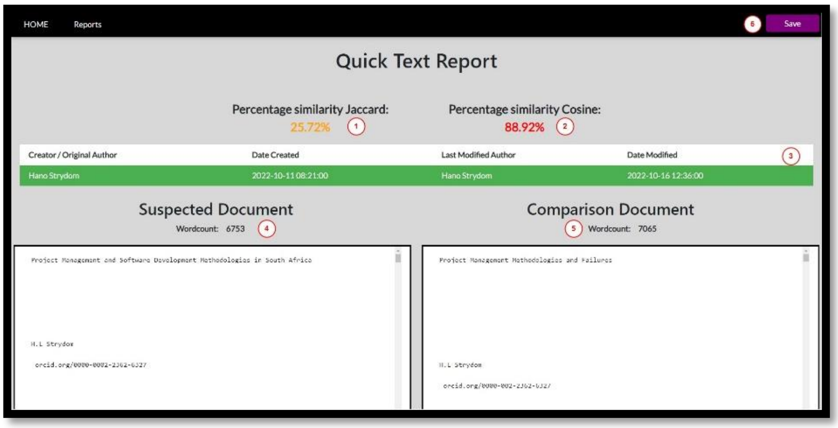
Algorithm

○ Lines
○ Sentences
● Substrings

Length (for Substrings algorithm)

15

QuickText Report Page



Quick Text Report

Percentage similarity Jaccard: 25.72%

Percentage similarity Cosine: 88.92%

Creator / Original Author	Date Created	Last Modified Author	Date Modified
Hano Strydom	2022-10-11 08:21:00	Hano Strydom	2022-10-16 12:36:00

Suspected Document

Wordcount: 6753

```
Project Management and Software Development Methodologies in South Africa



H.L Strydom

 orcid.org/0000-0002-2362-6327
```

Comparison Document

Wordcount: 7065

```
Project Management Methodologies and Failures




H.L Strydom

 orcid.org/0000-002-2362-6327
```

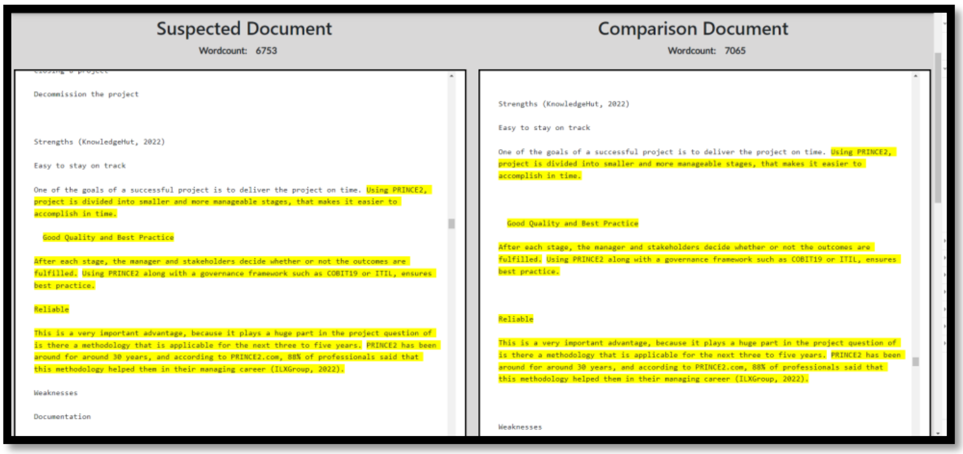
Components

1. Jaccard Similarity

    - The Jaccard similarity measures the text similarity between two documents to determine which members sentences/words/phrases are the most similar. The Jaccard similarity is calculated by dividing the number of similarities in both documents by the number of similarities in either document.
    - *The Jaccard similarity is preferred when comparing documents with 500 words or more.*

2. Cosine Similarity
   - Cosine Similarity is a measurement that quantifies the similarity between two or more vectors.
   - *The Cosine similarity is preferred when comparing document with 500 words or less.*
3. Metadata Table
   - This table displays the document metadata that can be used to verify the authors.
     - **Creator / Original Author** is the user that created the document
     - **Date Created** is the date the document was created
     - **Last Modified Author**: the author that made the last modifications
     - **Date Modified** is the date the document was last modified.
   - If the **Creator / Original Author** happens to be “Windows User”, it most likely indicates the user edited content on a document template.
   - If the **Creator / Original Author** and the **Last Modified Author** matches, the bottom table row will become green when you hover over it, otherwise the table row will be highlighted in red.
4. Suspected Document
   - The “Suspected Document” content is displayed and depending on the algorithm previously specified, the plagiarised text is highlighted in yellow.
   - A wordcount of the document is also displayed, allowing the user to know which Similarity score to look at.
5. Comparison Document
   - The “Comparison Document” is displayed and depending on the algorithm previously specified, the plagiarised text is highlighted in yellow.
   - A total wordcount of the document is also displayed, allowing the user to know which Similarity score to look at.
6. Save
   - When clicking this button, the results will be saved to a PDF and a prompt will allow the user to specify a location to save the report.

Example of plagiarised text highlights



Suspected Document

Wordcount: 6753

Decommission the project

Strengths (KnowledgeHut, 2022)

Easy to stay on track

One of the goals of a successful project is to deliver the project on time. Using PRINCE2, project is divided into smaller and more manageable stages, that makes it easier to accomplish in time.

Good Quality and Best Practice

After each stage, the manager and stakeholders decide whether or not the outcomes are fulfilled. Using PRINCE2 along with a governance framework such as COBIT19 or ITIL, ensures best practice.

Reliable

This is a very important advantage, because it plays a huge part in the project question of is there a methodology that is applicable for the next three to five years. PRINCE2 has been around for around 30 years, and according to PRINCE2.com, 88% of professionals said that this methodology helped them in their managing career (ILXGroup, 2022).

Weaknesses

Documentation

Comparison Document

Wordcount: 7065

Strengths (KnowledgeHut, 2022)

Easy to stay on track

One of the goals of a successful project is to deliver the project on time. Using PRINCE2, project is divided into smaller and more manageable stages, that makes it easier to accomplish in time.

Good Quality and Best Practice

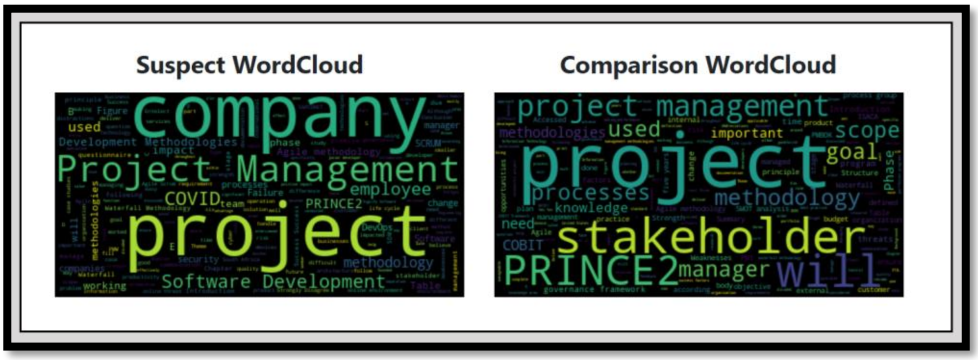
After each stage, the manager and stakeholders decide whether or not the outcomes are fulfilled. Using PRINCE2 along with a governance framework such as COBIT19 or ITIL, ensures best practice.

Reliable

This is a very important advantage, because it plays a huge part in the project question of is there a methodology that is applicable for the next three to five years. PRINCE2 has been around for around 30 years, and according to PRINCE2.com, 88% of professionals said that this methodology helped them in their managing career (ILXGroup, 2022).

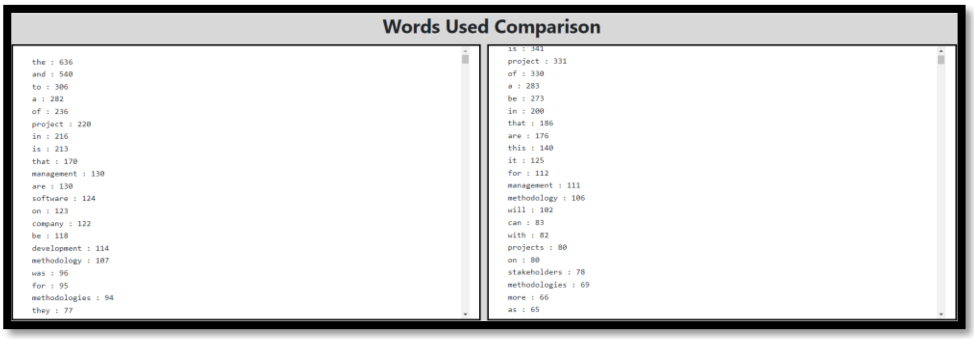
Weaknesses

Wordcloud



- WordCloud is a function that generates an image of either text, displaying the most used words within the "Suspect Document" and the "Comparison Document".
- Stopwords such as "a, the, is, are" are ignored. The larger the word in the WordCloud, the bigger the word will appear in the image.

Word Count



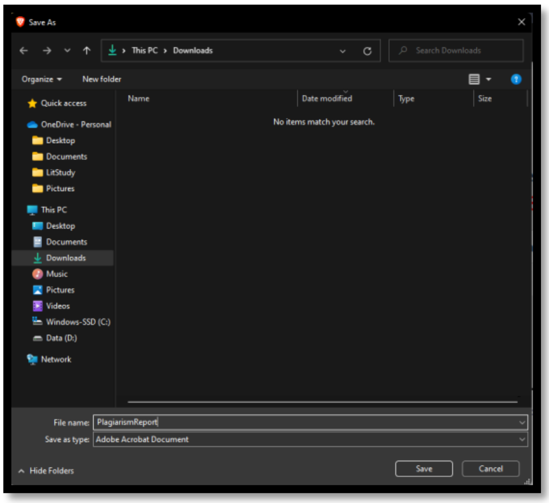
Words Used Comparison

```
the : 636
and : 540
to : 306
a : 282
of : 236
project : 220
in : 216
is : 213
that : 170
management : 130
are : 130
software : 124
on : 123
company : 122
be : 118
development : 114
methodology : 107
was : 96
for : 95
methodologies : 94
they : 77
```

```
is : 341
project : 331
of : 330
a : 283
be : 273
in : 200
that : 186
are : 176
this : 140
it : 125
for : 112
management : 111
methodology : 106
will : 102
can : 83
with : 82
projects : 80
on : 80
stakeholders : 78
methodologies : 69
more : 66
as : 65
```

- Word Count is a function that displays all the words used in both the “Suspect Document” and the “Comparison Document”.
- The function will also count how many times each word appears in the documents.

Save Button

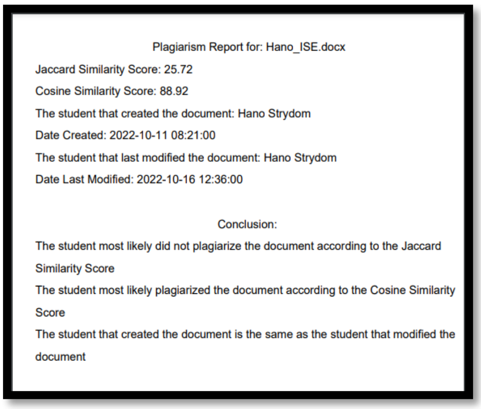


File name: PlagiarismReport

Save as type: Adobe Acrobat Document

- When the “Save” button is clicked, a prompt appear where the user can specify where they want to save the QuickText report.

QuickText Report



Plagiarism Report for: Hano_ISE.docx

Jaccard Similarity Score: 25.72

Cosine Similarity Score: 88.92

The student that created the document: Hano Strydom

Date Created: 2022-10-11 08:21:00

The student that last modified the document: Hano Strydom

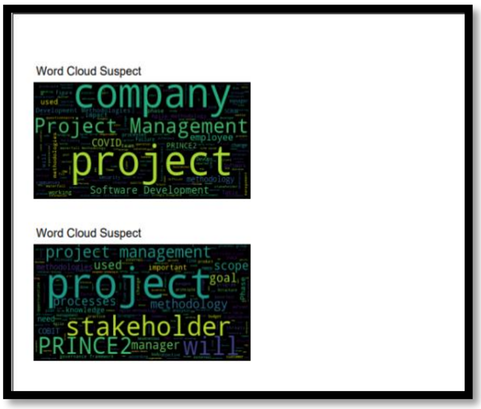
Date Last Modified: 2022-10-16 12:36:00

Conclusion:

The student most likely did not plagiarize the document according to the Jaccard Similarity Score

The student most likely plagiarized the document according to the Cosine Similarity Score

The student that created the document is the same as the student that modified the document



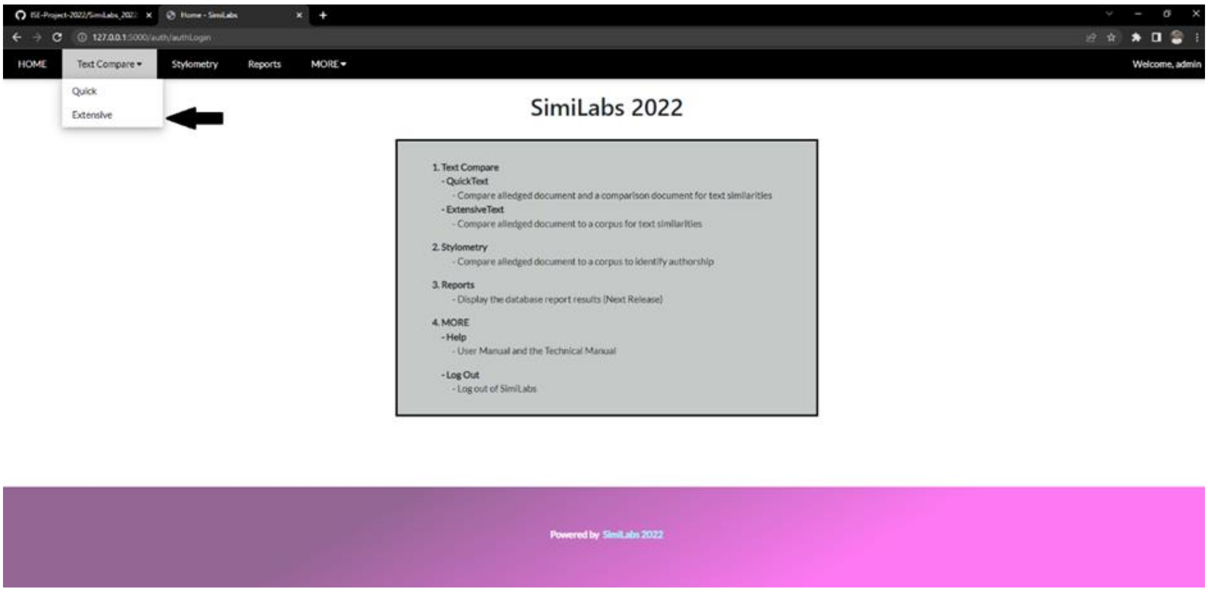
Word Cloud Suspect

Word Cloud Suspect

- Above is an example of a QuickText Report.
- The user can save this report for later analysis.

3.2. Extensive text comparison

The extensive text functionality can be accessed by navigating to the Text Compare Drop Down and selecting extensive at the top left of the screen.



SimiLabs 2022

1. Text Compare
   - QuickText
     - Compare alledged document and a comparison document for text similarities
   - ExtensiveText
     - Compare alledged document to a corpus for text similarities

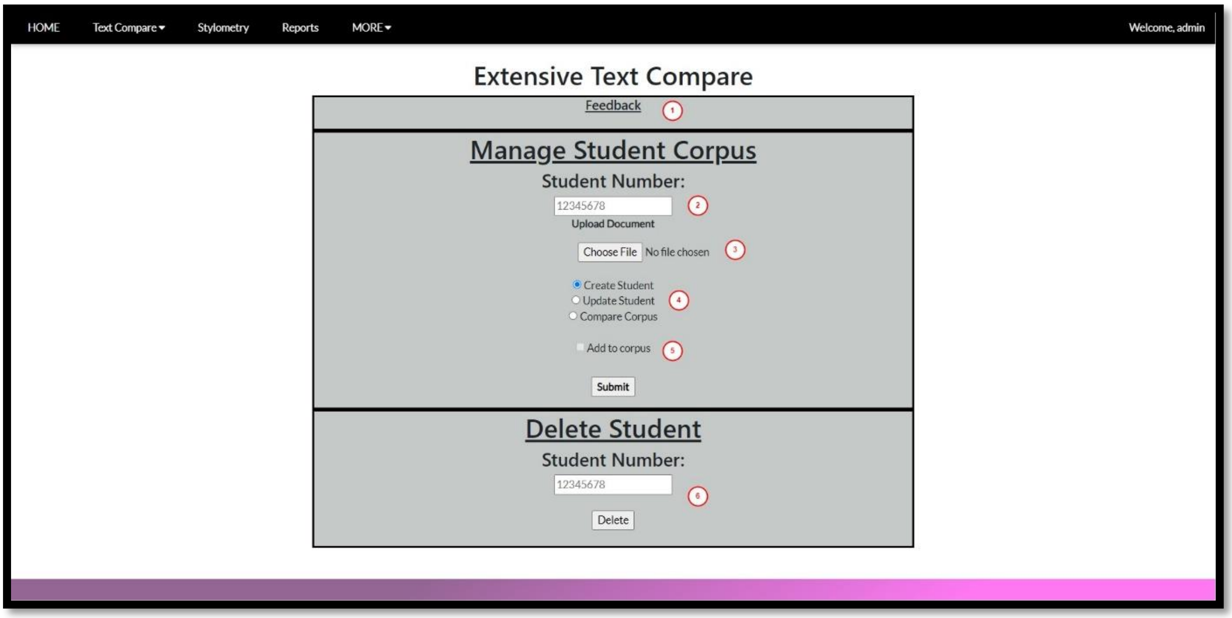
2. Stylometry
   - Compare alledged document to a corpus to identify authorship

3. Reports
   - Display the database report results (Next Release)

4. MORE
   - Help
     - User Manual and the Technical Manual
   - Log Out
     - Log out of SimiLabs

The user will be greeted with the main extensive text comparison page which will enable the user to:

- Create a corpus for the specified student number and related data files
- Update a specified student corpus
- Compare a document to a corpus of the specified student number
- Remove specified student data files



Extensive Text Compare

Feedback

Manage Student Corpus

Student Number: 12345678

Upload Document: No file chosen

● Create Student
○ Update Student
○ Compare Corpus

□ Add to corpus

Delete Student

Student Number: 12345678

1. Feedback
   - This is where feedback will be given after a student is created, updated, or deleted.
2. Student number
   - The user must enter a student number every time when utilising the functionality of the extensive text comparison. The student number will be utilised to identify the correct corpus of which to either create, update or compare to.
3. Choose file
   - The user must specify a file which will be used to either create or update a corpus depending on the selected radio buttons. If the compare corpus button is selected, the file which was uploaded will be used for comparison to the corpus.
4. Select action
   - The user has 3 options, of which one option should be selected, in the form of radio buttons to choose from, namely:
     a) Create a student
        - The create student functionality enables the user to specify a student number and provides the ability to upload either a PDF or Microsoft Word document which will be added to the corpus

of the specified student when the user clicks on the submit button.

b) Update a student

   - The update student functionality allows the user to update a corpus by specifying a student number and selecting either a PDF or Microsoft Word document to be added to the student corpus when the user clicks on the submit button.

c) Compare a student corpus to a document

   - The compare student functionality provides the user with the ability to compare a specified student corpus to a document uploaded by the user in order to determine the similarity between the document and the specified student corpus. The user enters the student number, uploads the documents to which the user wants to compare the corpus to, and selects the compare corpus radio button. The document can either be a PDF or Microsoft Word document and the user should click on the submit button when all the required information is specified.

5. Add to Corpus checkbox
   - Before selecting submit, the user has the ability to check the Add to Corpus checkbox. This will automatically add the uploaded document to the specified user’s corpus after it has compared the document to the corpus.
6. Delete student
   - The delete student functionality provides the user with the ability to remove the related data files of the specified student number.

Extensive text comparison algorithm

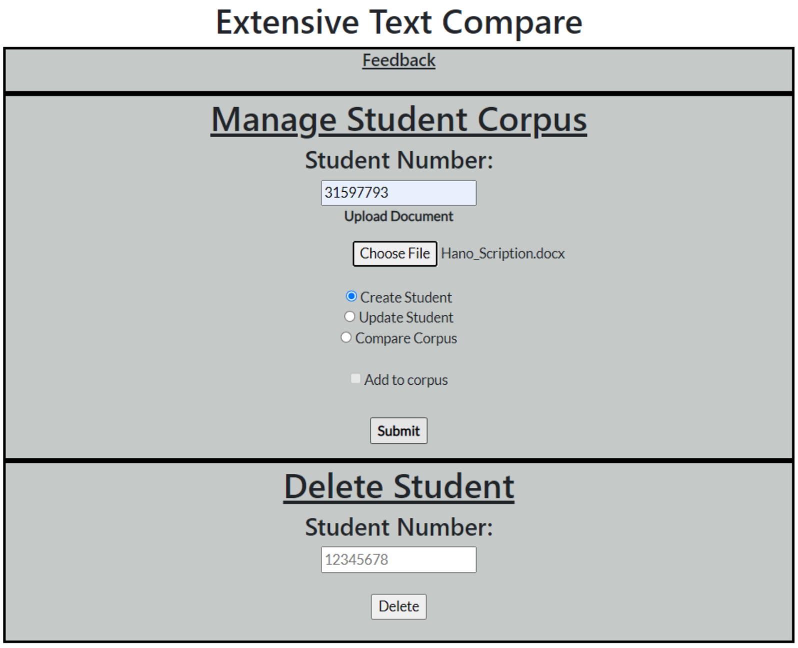
LSI

By creating a collection of ideas associated with the documents and terms, latent semantic analysis (LSA), a method in natural language processing, specifically distributional semantics, analyses relationships between a set of documents and the terms they contain. LSA believes that words with similar meanings will appear in texts with a similar structure (the distributional hypothesis). Singular value decomposition (SVD), a mathematical method, is used to condense a large piece of text into a matrix with word counts per document (rows represent unique words and columns represent each document). This technique reduces the number of rows while maintaining the similarity structure among columns. The cosine of the angle produced by any two vectors formed by columns is then used to compare documents, as is the dot product formed by the normalisation of the two vectors. Values near 1 reflect documents that are extremely similar, while values near 0 describe documents that are quite different. In the context of its application to information retrieval, it is sometimes called latent semantic indexing (LSI)

TF-IDF

TF-IDF, short for term frequency-inverse document frequency, is a numerical statistic used in information retrieval that aims to capture the significance of a word to a document in a collection or corpus. In information retrieval, text mining, and user modelling searches, it is frequently employed as a weighting factor. To account for the fact that some words are used more frequently than others overall, the TF-IDF value rises according to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the term. One of the most common term-weighting techniques used nowadays is TF-IDF. According to a 2015 survey, TF-IDF is used by 83% of text-based recommender systems in digital libraries.

Student creation example



Extensive Text Compare

Feedback

Manage Student Corpus

Student Number:

31597793

Upload Document

Choose File Hano_Scription.docx

● Create Student
○ Update Student
○ Compare Corpus

□ Add to corpus

Submit

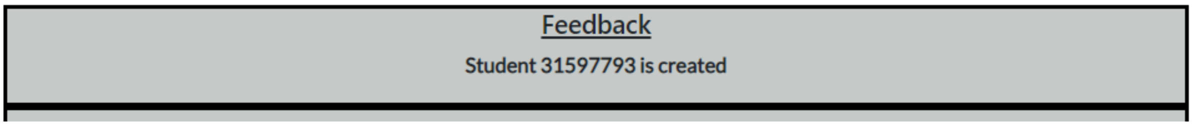
Delete Student

Student Number:

12345678

Delete

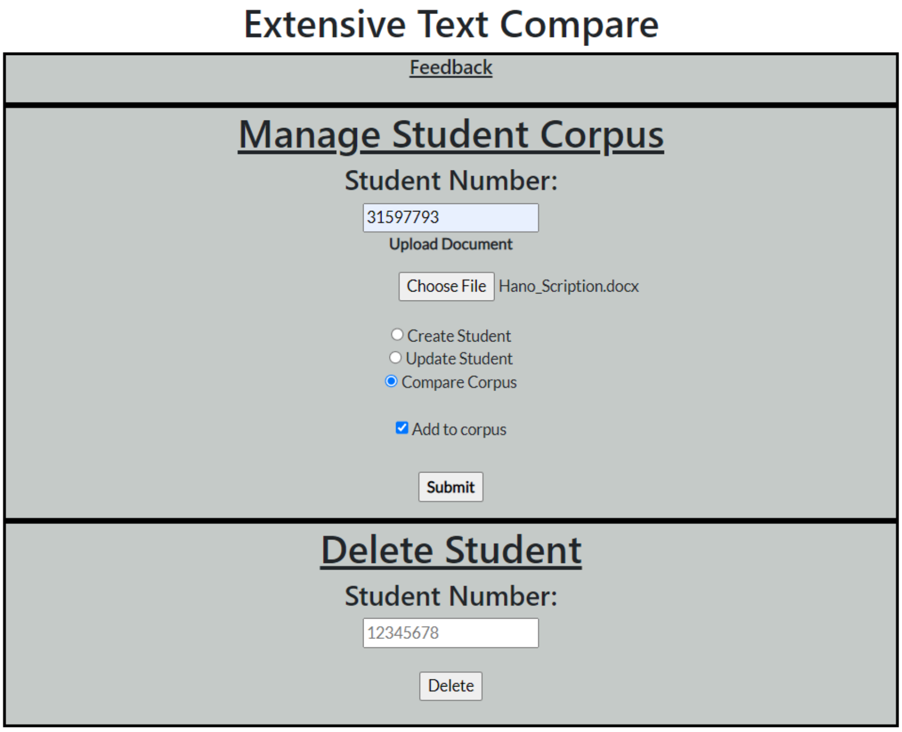
Feedback example



Feedback

Student 31597793 is created

Student Comparison example



Extensive Text Compare

Feedback

Manage Student Corpus

Student Number:

31597793

Upload Document

Choose File Hano_Scription.docx

○ Create Student
○ Update Student
● Compare Corpus

☑ Add to corpus

Submit

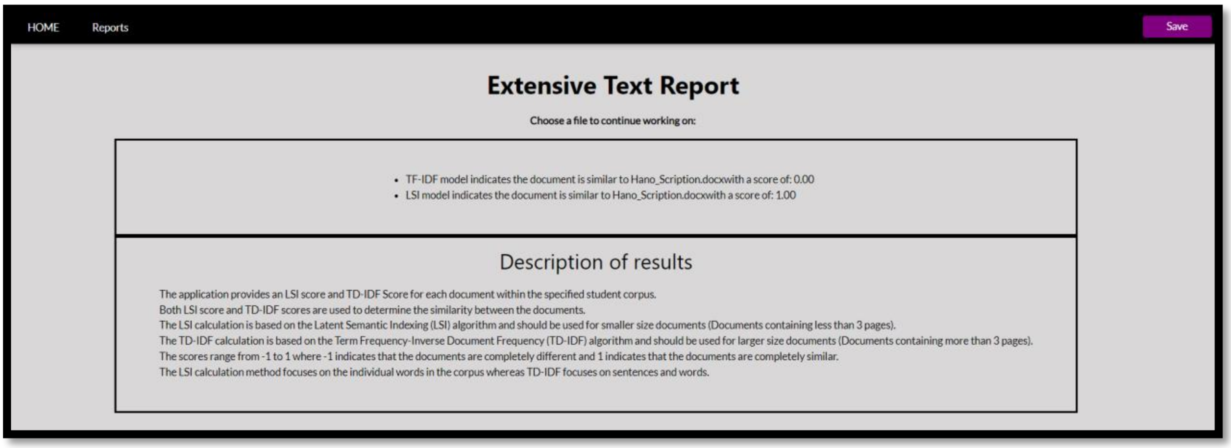
Delete Student

Student Number:

12345678

Delete

Extensive Text Report Page



Extensive Text Report

Choose a file to continue working on:

- TF-IDF model indicates the document is similar to Hano_Scription.docxwith a score of: 0.00
- LSI model indicates the document is similar to Hano_Scription.docxwith a score of: 1.00

Description of results

The application provides an LSI score and TD-IDF Score for each document within the specified student corpus.
Both LSI score and TD-IDF scores are used to determine the similarity between the documents.
The LSI calculation is based on the Latent Semantic Indexing (LSI) algorithm and should be used for smaller size documents (Documents containing less than 3 pages).
The TD-IDF calculation is based on the Term Frequency-Inverse Document Frequency (TD-IDF) algorithm and should be used for larger size documents (Documents containing more than 3 pages).
The scores range from -1 to 1 where -1 indicates that the documents are completely different and 1 indicates that the documents are completely similar.
The LSI calculation method focuses on the individual words in the corpus whereas TD-IDF focuses on sentences and words.

- Extensive Results

  The extensive text report provides the user with the results of the comparison between the uploaded document and the corpus of the specified student number. An LSI model and TD-IDF model is calculated for all the documents stored within the corpus. The uploaded document is converted to its own LSI and TD-IDF model which is used to calculate the LSI and TD-IDF similarity scores. Each line in the report page indicates the calculation type, name of the document in the corpus and its similarity score between -1 and 1.

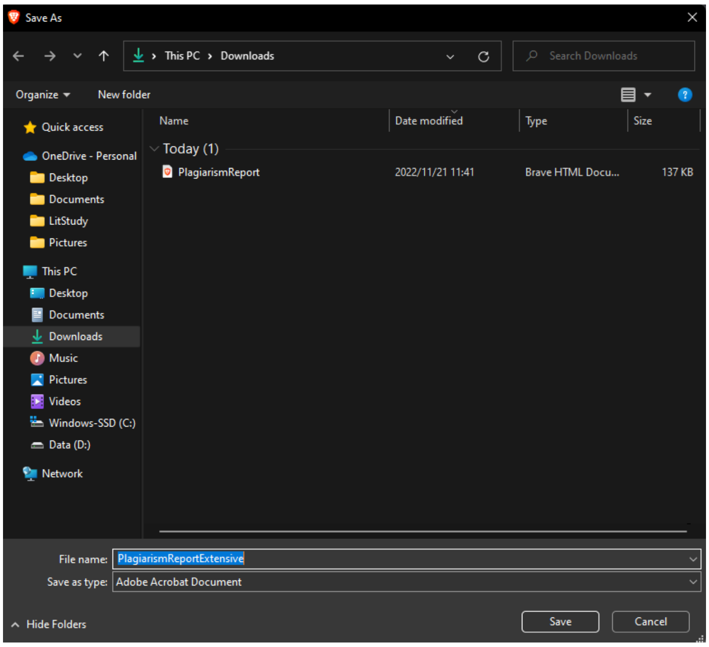
- Description of results

  The application provides an LSI score and TD-IDF Score for each document within the specified student corpus. Both LSI score and TD-IDF scores are used to determine the similarity between the documents. The LSI calculation is based on the Latent Semantic Indexing (LSI) algorithm and should be used for smaller size documents (Documents containing less than 3 pages). The TD-IDF calculation is based on the Term Frequency-Inverse Document Frequency (TD-IDF) algorithm and should be used for larger size documents (Documents containing more than 3 pages). The scores range from -1 to 1 where -1 indicates that the documents are completely different and 1 indicates that the documents are completely similar. The LSI calculation method focuses on the individual words in the corpus whereas TD-IDF focuses on sentences and words.

- Save Button

  The save button provides the user with the ability to save the generated report after a corpus comparison has occurred. This allows the user to view the results at a later stage.

Save



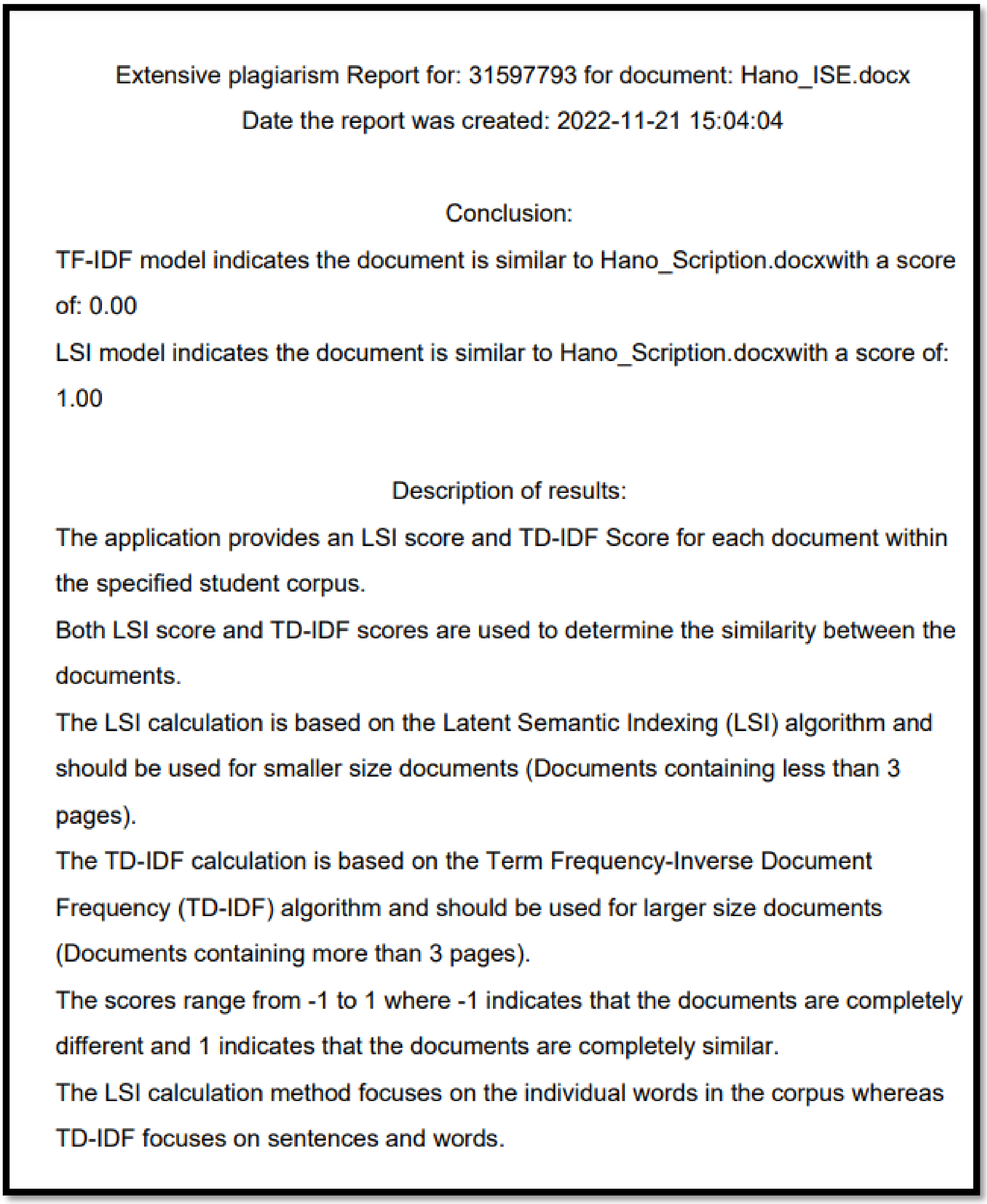
Name	Date modified	Type	Size
Today (1)			
PlagiarismReport	2022/11/21 11:41	Brave HTML Docu...	137 KB

File name: PlagiarismReportExtensive

Save as type: Adobe Acrobat Document

- When the “Save” button is clicked, a prompt appear where the user can specify where they want to save the ExtensiveText report.

Extensive Report PDF Example



Extensive plagiarism Report for: 31597793 for document: Hano_ISE.docx

Date the report was created: 2022-11-21 15:04:04

Conclusion:

TF-IDF model indicates the document is similar to Hano_Scription.docxwith a score of: 0.00

LSI model indicates the document is similar to Hano_Scription.docxwith a score of: 1.00

Description of results:

The application provides an LSI score and TD-IDF Score for each document within the specified student corpus.

Both LSI score and TD-IDF scores are used to determine the similarity between the documents.

The LSI calculation is based on the Latent Semantic Indexing (LSI) algorithm and should be used for smaller size documents (Documents containing less than 3 pages).

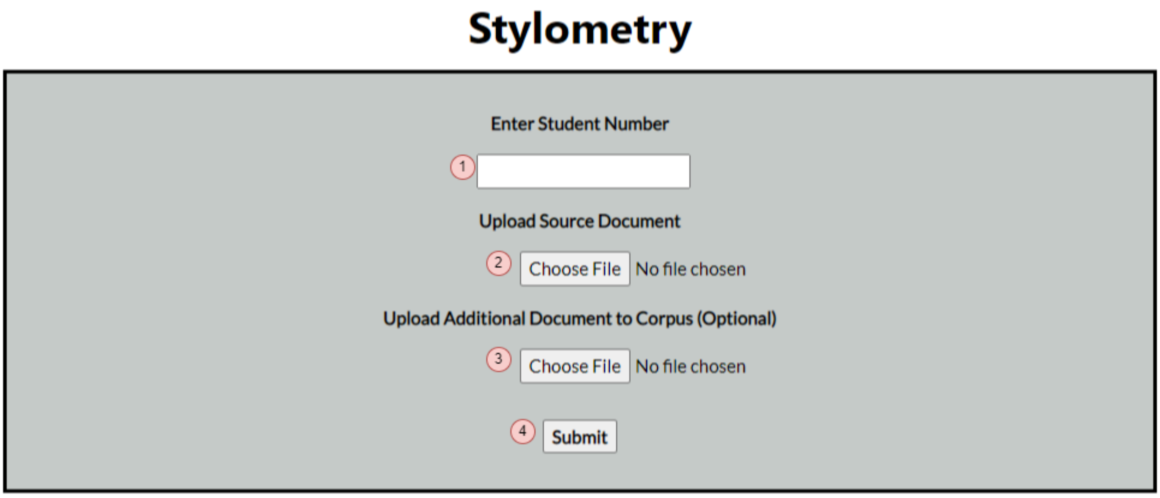
The TD-IDF calculation is based on the Term Frequency-Inverse Document Frequency (TD-IDF) algorithm and should be used for larger size documents (Documents containing more than 3 pages).

The scores range from -1 to 1 where -1 indicates that the documents are completely different and 1 indicates that the documents are completely similar.

The LSI calculation method focuses on the individual words in the corpus whereas TD-IDF focuses on sentences and words.

4. Using the Stylometry Feature

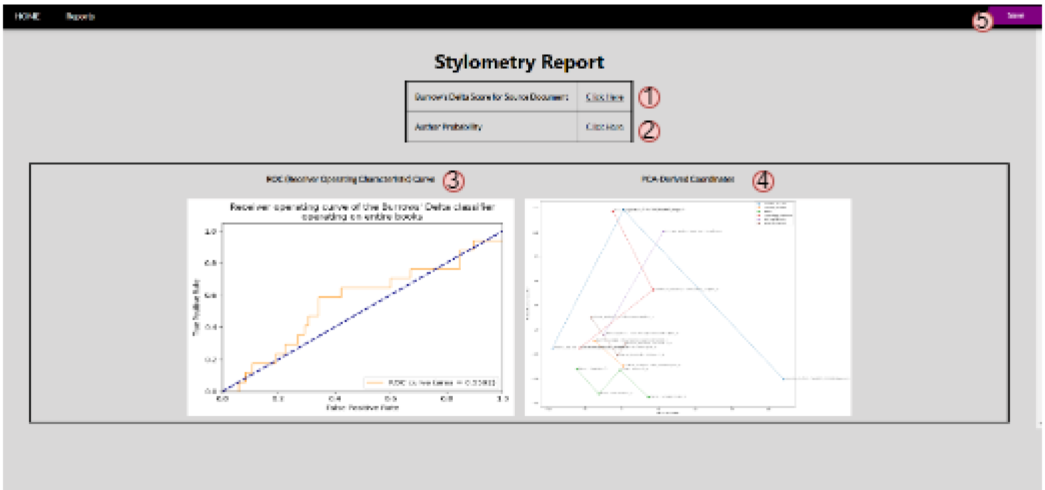
If the user chooses the stylometry option in the toolbar, they will be redirected to the stylometry page:



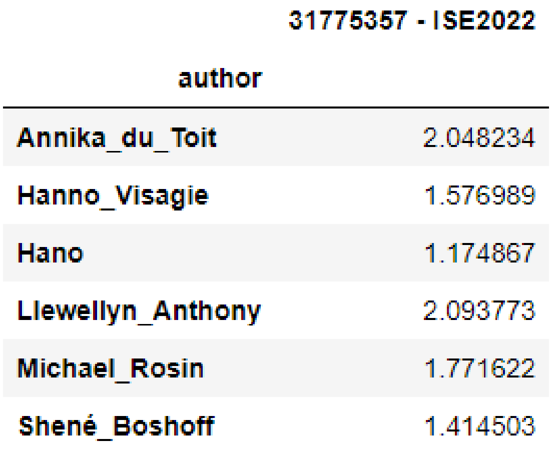
1) The user enters the student number of the student that submitted the document.
2) The user uploads the suspected document.
3) The user can choose to add additional documents to the corpus belonging to the student.
4) The user clicks on the submit button to receive the results of the stylometric analysis.

Stylometry results

After the analysis has been completed the user will be redirected to the Stylometry Report page:

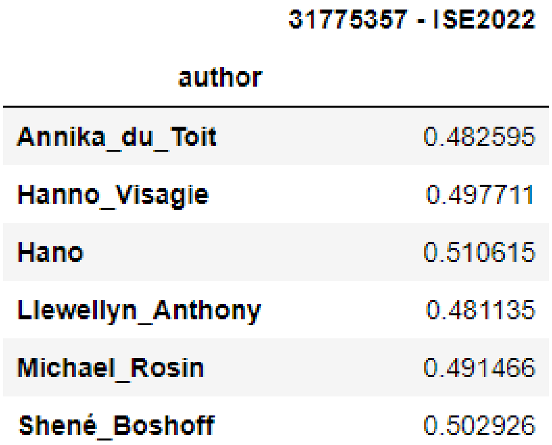


1) Contains the Burrows' data value of the source document compared to all of the students in the corpus. The user can view these values by clicking on "Click Here". The Burrows' delta is a value between 0 and 3. The higher the value, the less likely it is that the student that submitted the document is the author of the document. The lower the value, the more likely that the student is the author.



	31775357 - ISE2022
author	
Annika_du_Toit	2.048234
Hanno_Visagie	1.576989
Hano	1.174867
Llewellyn_Anthony	2.093773
Michael_Rosin	1.771622
Shené_Boshoff	1.414503

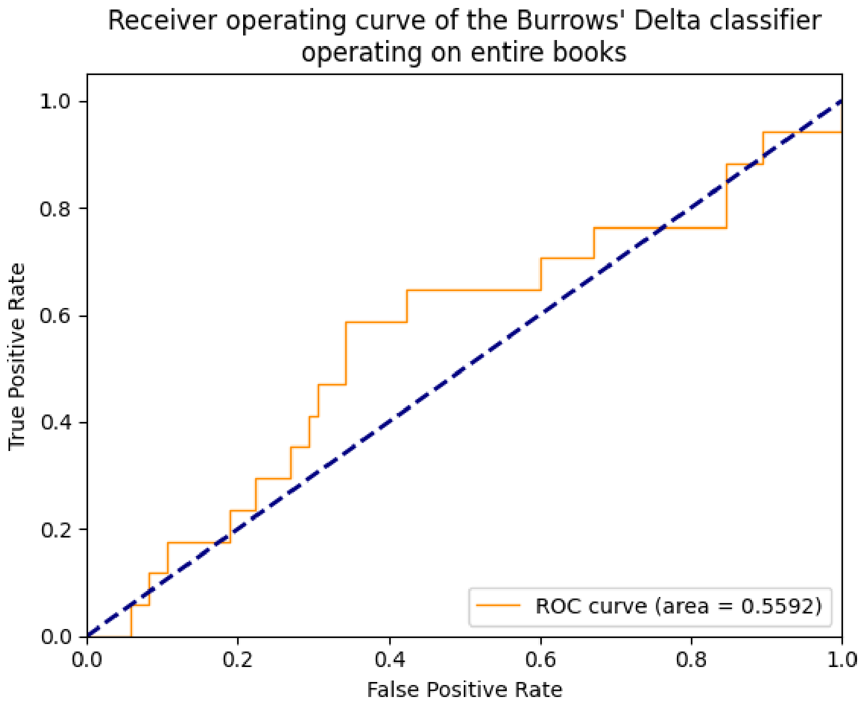
2) Contains the probability that the student that submitted the document is the author of that document compared to all other students in the corpus. The higher the probability of another student, the less likely it is that the student that submitted the document is not the sole author. The higher the probability next to the student that submitted the document, the more likely it is that the student is the author. The probability next to that student should be the highest on the list. The user can view these values by clicking on “Click Here”.



	31775357 - ISE2022
author	
Annika_du_Toit	0.482595
Hanno_Visagie	0.497711
Hano	0.510615
Llewellyn_Anthony	0.481135
Michael_Rosin	0.491466
Shené_Boshoff	0.502926

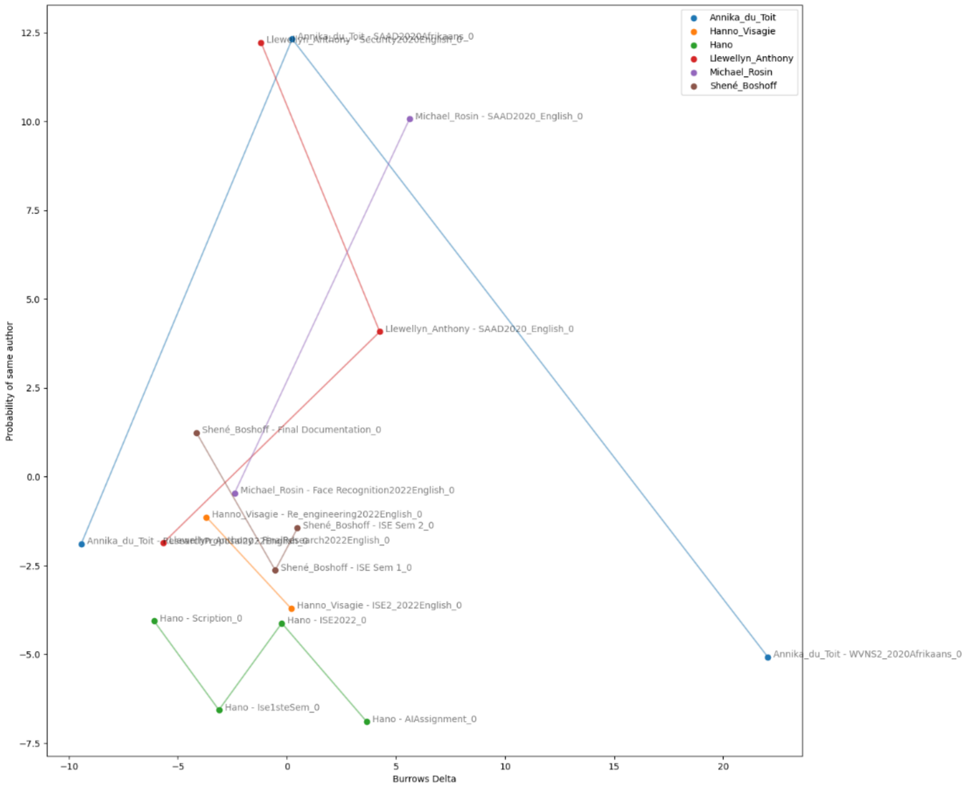
3) The Receiver Operating Characteristic (ROC) Curve is an indication of how well the analysis performed. The ROC evaluation is performed using cross-validation. Every document in the corpus is taken out and a Burrows’ model is trained on the remainder of the document, which is then tested against the document taken out. The probability scores of the results are used to calculate the ROC curve. The Area under the ROC Curve (AUC) measures the area under the ROC curve. AUC provided an aggregate measure of performance across all possible classification thresholds. It represents the probability that a random positive is positioned to the right of a random negative. The orange line represents the AUC. The closer it is to 1 the better the model performs and the more accurate it is. The model will improve as the corpus grows. The ideal AUC will have a y value of 1 and an x value of 1.

Receiver Operating Characteristic (ROC) Curve



4) The stylistic similarities between the documents in the corpus are visualized by calculating their differences and using Principle Component Analysis (PCA). Z-scores are calculated for the top 50 most common words used in every document. The z-score represents the fingerprint of each document. Ideally, the same authors should be clustered together.

Principle Component Analysis (PCA)



5) The user can generate and save a PDF report of the stylometric results. The user will be able to choose the download location of the report.

Possible errors that can be encountered

The list of all potential known faults that the user could run into when utilising the application is provided in the next part, along with some extra suggestions and guidance on how to resolve these issues.

Error Message	Reason for Error	Possible Fix
Valid username is required.	The username entered is incorrect or doesn't exist.	If the user has an account, enter the username correctly or create an account.
Password is required.	The password was not entered.	Enter the correct password.
Invalid credentials.	The username or password was entered incorrectly.	Enter the correct username and password.
Confirmation password is required.	The confirmation password was not entered.	Enter the confirmation password.
Passwords do not match.	The password and confirmation password is not the same.	Make sure both passwords are correct and the same.
Please select a file.	A file was not selected for upload.	Select a file to upload.
Something went wrong.	The length of the substring to search for was not entered when choosing the substring algorithm.	Enter the length of the substring when the substring algorithm is chosen.
Please fill out this field.	The student number was not entered.	Enter the student number.

If the user encounters any errors that are not listed above, the support team can be contacted.