“REVIEWER FINDER”

A Machine Learning Project

Creator: D. Stamatakis (MscRes Student)

Supervisor: Professor C. Patrikakis

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA

TABLE OF CONTENTS

1.1 HIGH-LEVEL OVERVIEW (*PREDICTION*)

Main Purpose of the Application:

1. Receive a ***PDF document*** as an ***Input***.
2. ***Predict*** the most suitable ***Reviewer*** for the document.

How is this achieved:

1. By ***Extracting*** the document's ***Keywords*** using the ***RAKE*** (Rapid Keyword Extraction) Algorithm.
2. Then, these keywords are ***Filtered*** to increase their quality (e.g., duplicate words, verbs and single words > 60 characters are removed).
3. Next, the filtered keywords are inserted into the ***Trained Model***.
4. Finally, the Model predicts the best Author based on its trained datasheet.

1.2 HIGH-LEVEL OVERVIEW (*NEW MODEL*)

Main Purpose of the Application:

1. Receive a ***Directory Path*** as an ***Input.*** *(Must be structured in a specific way)*
2. ***Train & Save*** a new ***Model*** based on that ***Data*** contained in the provided ***Directory***.

How is this achieved:

1. By ***Extracting*** the document's ***Keywords*** using the ***RAKE*** (Rapid Keyword Extraction) Algorithm for each Author, for each of their Documents.
2. External *.txt* and *.json* files are created to organize and store the outputted information.
3. Then, for each Author: ***Filtering*** is performed and all keywords (from every document) are bundled together into a single *.txt* file.
4. Next, ***TF-IDF*** Algorithm ***Ranks*** these keywords based on their frequency across all Author's Documents (corpus).
5. Afterwards, a ***Multi-Class Experiment*** is performed in order for the ***AutoML*** (ML.NET) to ***discover*** which Machine Learning Algorithm is best suited for the current task.
6. Finally, ***AutoML*** returns a ***Trained model*** which is capable of predicting the ***Best Reviewer.***

2.1 MAIN COMPONENTS

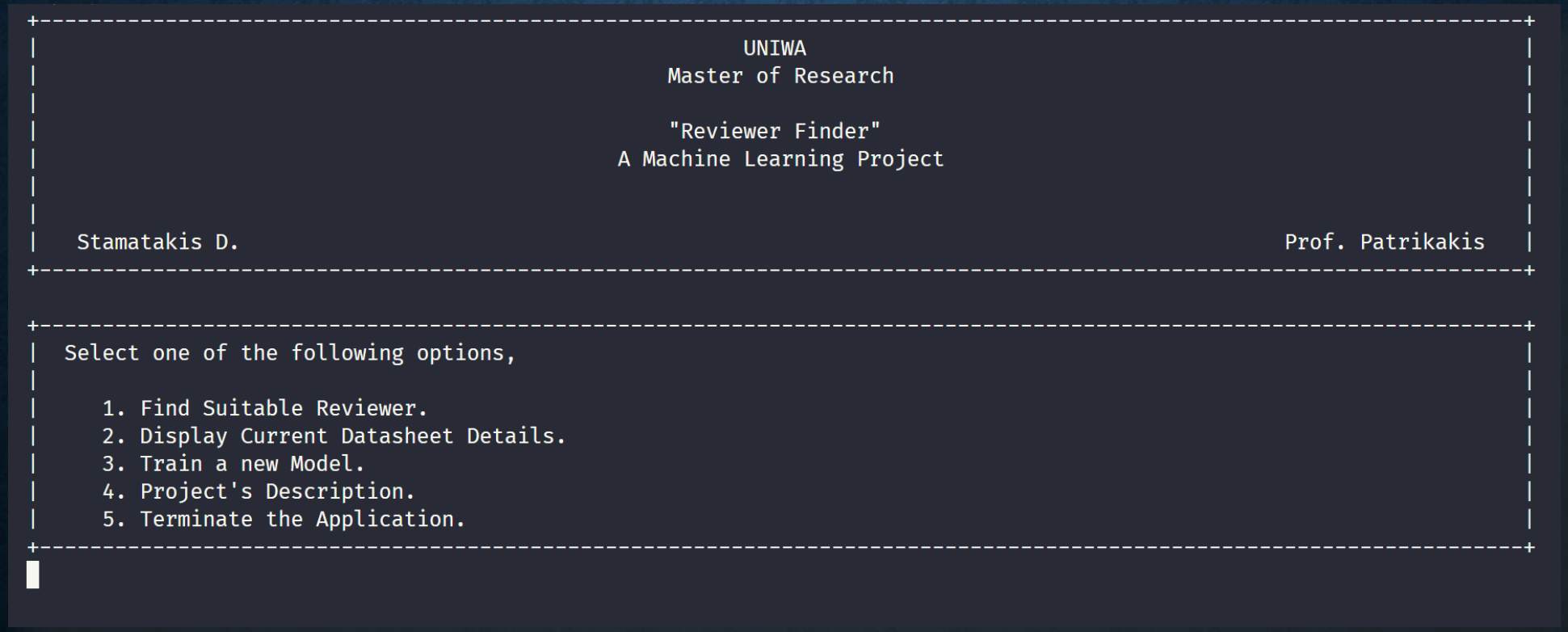
Foundation Elements:

- Programming Language: *C# or CSharp*
- Framework: .NET 6.0 — The Microsoft's Cross-Platform Runtime.

- External Modules (Packages):
  - .NET Ecosystem:
    - ML.NET (Microsoft.ML) — Microsoft's free and open-sourced Machine Learning Framework
    - ML.AutoML — Generates Trained Model with minimal configurations
    - Newtonsoft.Json — Coverts C# Object into JSON (Serialization)
    - PdfPig — Free and open-sourced Library for processing PDF files
    - Python.NET — Enables the execution of Python scripts and modules inside C#
  - Python
    - rake_nltk — Python Version of the RAKE Algorithm infused with NLTK.

3.1 APPLICATION PRESENTATION

By Running the Application the user is greeted with the Welcome and Options Menus



3.2 APPLICATION PRESENTATION

By Pressing "**1**" and hitting "Enter", the Application will start the ***Process of Finding a Reviewer***

```
+-------------------------------------------------------------------------------------------------------------------------+
|  You have selected the: [Find Suitable Reviewer]                                                                        |
|                                                                                                                         |
|    NOTE_1: The Path must contain only ASCII characters.                                                                 |
|    NOTE_2: Running as 'Administrator' might fix some issues                                                             |
|                                                                                                                         |
+-------------------------------------------------------------------------------------------------------------------------+
|  Provide the Absolute Path of you PDF file:                                                                             |
|                                                                                                                         |
C:\Users\Jimzord12\Desktop\MscRes\Thesys_Final_v3.0.pdf
```

The App will request of the User to insert a ***PDF File Path***.

For this demonstration the author's thesis PDF file is provided which is about "***Blockchain & Gaming***".

3.3 APPLICATION PRESENTATION

Afterwards, the Application will perform the: ***Text and Keywords Extraction***

```
+-------------------------------------------------------------------------------------------------------------------+
|  Provide the Absolute Path of you PDF file:                                                                       |
|                                                                                                                   |
C:\Users\Jimzord12\Desktop\MscRes\Thesys_Final_v3.0.pdf
| ->   Selected File: [Thesys_Final_v3.0]                                                                           |
|                                                                                                                   |
| ->   1) Extracting Text from PDF...                                                                               |
|                                                                                                                   |
| ->   2) Extracting Keywords using RAKE...                                                                         |
|                                                                                                                   |
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Jimzord12\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Jimzord12\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\Jimzord12\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
| ->   Extracted : (2950) keywords                                                                                  |
| ->   3) Converting Keywords Object into a Single string...                                                        |
|                                                                                                                   |
| ->   4) Transforming Data to AutoML friendly format...                                                            |
|                                                                                                                   |
| ->  Select 1 of 3 options:                                                                                        |
| ->     1) Provide the [Path] that contains your models                                                            |
| ->     2) Enter the word [def] to use the Default one                                                             |
| ->     3) Enter the word [local] to see the locally stored models                                                 |
|                                                                                                                   |
```

For this Presentation ***the 3rd Option will be selected***.

Once the Keywords are obtained, the App will ask the User to select one of the provided options.

1) In case the User possesses trained models (compatible with ML.NET) and wishes to use those, their Directory Path should be entered.
2) If the User simple wishes for quick demonstration of the App's functionality and features, the word "def" should be typed.
3) Finally, in case the User has trained a couple of Models (using this App) but did not specify a Path for them to be saved, they will be saved in a local folder. By entering "local", those Models can be accessed.

3.4 APPLICATION PRESENTATION

Next, the Application will ***Display all available Trained Models***

```
| ->  Select 1 of 3 options:
| ->     1) Provide the [Path] that contains your models
| ->     2) Enter the word [def] to use the Default one
| ->     3) Enter the word [local] to see the locally stored models
|
local
| ->      > 1) 'superSayan_01.zip'
| ->      > 2) 'TestModel_01.zip'
| ->      > 3) 'TestModel_02.zip'
| ->      > 4) 'TestModel_1_01.zip'
| ->      > 5) 'tfidfModel.zip'
|
| ->  Select a Model by Entering its Number
```

For this Presentation ***the 1rd Model will be selected***.

3.5 APPLICATION PRESENTATION

Finally, the Application will ***Predict*** the most ***Suitable Reviewer***

```
|
| ->  You have selected the: (superSayan_01)
| ->   5) Loading the Trained Model...
|
| ->   6) The Most Suitable Reviewer is: [Vitalik Buterin]
|
| ->   7) The Scientific Field should be: [Blockchain]
|
|
|  Press 'q' for (Main Menu) or 'Enter' to (repeat the process)
+-------------------------------------------------------------------
```

Note: Vitalik Buterin is the Founder of Ethereum.

Therefore, the ***Prediction*** is correct.

3.6 APPLICATION PRESENTATION

Let's quickly explore the rest of the available Features

Detailed Description is ***skipped*** as the meaning of each Process can be inferred by the Screen Shots.

```
|  NOTE: If you are working with your own data,
|  first [Train a Model]. By doing this, a proccess
|  of writing external files is performed
|  which is necessary for this method to work.
|
|  Provide the [Path] of your data. Ex. C:\Journals\Authors
C:\FastPaths\AuthorsAndPapers
|
|  You are probably qualified to run this method
|
|    ==> Author's Name             | Scientific Field          | Documents  | Total Keywords | Score
|    ==> Adi Shamir                | Cybersecurity             | 5          | 2257           | 0.90280
|    ==> Andrew Ng                 | Artificial_Intelligence   | 5          | 5              | 0.00200
|    ==> Angelos Stavrou           | Cybersecurity             | 5          | 19             | 0.00760
|    ==> Bhavani Thuraisingham     | Cybersecurity             | 5          | 112            | 0.04480
|    ==> Elaine Shi                | Blockchain                | 3          | 1              | 0.00067
|    ==> Emin Gün Sirer            | Blockchain                | 5          | 14             | 0.00560
|    ==> Geoffrey Hinton           | Artificial_Intelligence   | 5          | 6              | 0.00240
|    ==> Jeffrey Ullman            | DataScience_and_BigData   | 5          | 27622          | 11.04880
|    ==> Jiawei Han                | DataScience_and_BigData   | 5          | 62             | 0.02480
|    ==> Joseph Bonneau            | Blockchain                | 3          | 263            | 0.17533
|    ==> Josh Pauli                | Cybersecurity             | 4          | 2203           | 1.10150
|    ==> Judea Pearl               | Artificial_Intelligence   | 5          | 9              | 0.00360
|    ==> Jure Leskovec             | DataScience_and_BigData   | 5          | 1220           | 0.48800
|    ==> Madhusanka Liyanage       | Blockchain                | 5          | 58             | 0.02320
|    ==> Trevor Hastie             | DataScience_and_BigData   | 5          | 784            | 0.31360
|    ==> Vipin Kumar               | DataScience_and_BigData   | 5          | 5207           | 2.08280
|    ==> Vitalik Buterin           | Blockchain                | 5          | 4295           | 1.71800
|    ==> Wenjing Lou               | Cybersecurity             | 5          | 165            | 0.06600
|    ==> Yann Lecun                | Artificial_Intelligence   | 5          | 19             | 0.00760
|    ==> Yoshua Bengio             | Artificial_Intelligence   | 5          | 623            | 0.24920
|  Press any key to go Main Menu...
```

Exploring: ***Option #2***

```
|  Select one of the following options,
|
|     1. Find Suitable Reviewer.
|     2. Display Current Datasheet Details.
|     3. Train a new Model.
|     4. Project's Description.
|     5. Terminate the Application.
```

Comments:

1. The Directory's Structure must have specific form (an example is given in the "Train a new Model" Feature)
2. The Score is a very simple Metric though on the spot. This is the formula:

$$f(\text{documentAmount}, \text{keywords}) = \begin{cases} -1 & \text{if keywords} = 0 \\ \frac{\text{keywords}}{\text{documentAmount} \times 500} & \text{otherwise} \end{cases}$$

More research should be conducted for the ***Coefficient*** (500)

3. As you observe most Documents have a score < 1, which is not ideal. This is probably related to the ***PDF Library*** used.

3.7 APPLICATION PRESENTATION

Let's quickly explore the rest of the available Features

Detailed Description is ***skipped*** as the meaning of each Process can be inferred by the Screen Shots.

Exploring: ***Option #3***

```
+-----------------------------------------------
|  Select one of the following options,
|
|     1. Find Suitable Reviewer.
|     2. Display Current Datasheet Details.
|     3. Train a new Model.
|     4. Project's Description.
|     5. Terminate the Application.
+-----------------------------------------------
```

```
+-------------------------------------------------------------------------------------------------------+
|  You have selected the: [Train a new Model]                                                           |
|                                                                                                       |
|  Provide the [Absolute Path] of your Authors Directory.                                               |
|  The Directory should much the structure below:                                                       |
|                                                                                                       |
|   (rootDir) [This is the requested Path]                                                              |
|   |--> Author_1 (Folder)                                                                             |
|     |-----> Document_1.pdf                                                                            |
|     |-----> Document_2.pdf                                                                            |
|     |-----> Document_N.pdf                                                                            |
|   |--> Author_N (Folder)                                                                             |
|     |-----> Document_1.pdf                                                                            |
|     |-----> Document_2.pdf                                                                            |
|     |-----> Document_N.pdf                                                                            |
|                                                                                                       |
C:\FastPaths\AuthorsAndPapers
```

In this structure a folder (root) must contain many folders. The names of those folders must be Author names. Inside every Author folder, a collection of their Documents (in pdf format) should exist.

Comments:

No comments.

3.8 APPLICATION PRESENTATION

Let's quickly explore the rest of the available Features

```
|
C:\FastPaths\AuthorsAndPapers

The Number of Available Authors: 20

 ===> Adi Shamir                | Cybersecurity
 ===> Andrew Ng                 | Artificial_Intelligence
 ===> Angelos Stavrou           | Cybersecurity
 ===> Bhavani Thuraisingham     | Cybersecurity
 ===> Elaine Shi                | Blockchain
 ===> Emin Gün Sirer            | Blockchain
 ===> Geoffrey Hinton           | Artificial_Intelligence
 ===> Jeffrey Ullman            | DataScience_and_BigData
 ===> Jiawei Han                | DataScience_and_BigData
 ===> Joseph Bonneau            | Blockchain
 ===> Josh Pauli                | Cybersecurity
 ===> Judea Pearl               | Artificial_Intelligence
 ===> Jure Leskovec             | DataScience_and_BigData
 ===> Madhusanka Liyanage       | Blockchain
 ===> Trevor Hastie             | DataScience_and_BigData
 ===> Vipin Kumar               | DataScience_and_BigData
 ===> Vitalik Buterin           | Blockchain
 ===> Wenjing Lou               | Cybersecurity
 ===> Yann Lecun                | Artificial_Intelligence
 ===> Yoshua Bengio             | Artificial_Intelligence

- Received Authors: 20
- Selected Author: [Adi Shamir]
- Author Papers Amount: [5]
*** Author: [Adi Shamir] Paper Name: [doc_1]
*** Data Length: [1]
*** Author: [Adi Shamir] Paper Name: [doc_2]
*** Data Length: [2]
*** Author: [Adi Shamir] Paper Name: [doc_3]
*** Data Length: [3]
*** Author: [Adi Shamir] Paper Name: [doc_4]
*** Data Length: [4]
*** Author: [Adi Shamir] Paper Name: [doc_5]
*** Data Length: [5]
```

Exploring: ***Option #3***

```
+-----------------------------------------
|  Select one of the following options,
|
|     1. Find Suitable Reviewer.
|     2. Display Current Datasheet Details.
|     3. Train a new Model.
|     4. Project's Description.
|     5. Terminate the Application.
+-----------------------------------------
```

Comments:

1. A preview of the provided Directory is return (Red Box)
2. Below starts a long list of data functioning like a process bar.

3.9 APPLICATION PRESENTATION

Let's quickly explore the rest of the available Features

Exploring: ***Option #3***

```
+-------------------------------------------
|  Select one of the following options,
|
|     1. Find Suitable Reviewer.
|     2. Display Current Datasheet Details.
|     3. Train a new Model.
|     4. Project's Description.
|     5. Terminate the Application.
+-------------------------------------------
```

```
- Selected Author: [Yoshua Bengio]
- Author Papers Amount: [5]
*** Author: [Yoshua Bengio] Paper Name: [doc_1]
*** Data Length: [69]
*** Author: [Yoshua Bengio] Paper Name: [doc_2]
*** Data Length: [70]
*** Author: [Yoshua Bengio] Paper Name: [doc_3]
*** Data Length: [71]
*** Author: [Yoshua Bengio] Paper Name: [doc_4]
*** Data Length: [72]
*** Author: [Yoshua Bengio] Paper Name: [doc_5]
*** Data Length: [73]
[Yoshua Bengio]: Number of PaperData Elements: (73)
|  AutoML requires some time to figure out the suitable algorithm and train the model.
|  Please provide the time in (seconds). Recommendations:
|     > Teting: 20~30 seconds
|     > Small Datasheet (~100-150 docs): 100~120 seconds
|     > Big Datasheet (200+ docs): 240+ seconds
20
|
|  Enter a name for the Model:
PresentationTest 001
|  Commencing Training...
|
|  Training Completed!
|
|  The model can be found here: [C:\Users\Jimzord12\Documents\GitHub\MscRes-PaperReviewerFinder\PresentationTest_001.zip]|
|
|  Press 'q' for (Main Menu) or 'Enter' to (repeat the process)
+-------------------------------------------------------------------------------------------------------------------------
```

Comments:

1. As indicated in the Red Box, when creating a new model, AutoML requires a "time to train" input before training the model.
2. The provided recommendations should be taken will a grain of salt, their values derive from the minimal experience the developer gathered while creating the App.

3.10 APPLICATION PRESENTATION

Let's quickly explore the rest of the available Features

```
+-------------------------------------------------------------------------------------------------------------------------+
|  Select one of the following options,                                                                                   |
|                                                                                                                         |
+-------------------------------------------------------------------------------------------------------------------------+
|  Creator:        D. Stamatakis                                                                                          |
|  Supervisor:     Prof. Patrikakis                                                                                       |
|  Subject:        Machine Learning                                                                                       |
|                                                                                                                         |
|  Tools Used:                                                                                                            |
|  ~~~~~~~~~~~                                                                                                            |
|     ML.NET                                                                                                              |
| ->     An Free Open-souce Machine Learning Framework                                                                    |
|                                                                                                                         |
|     Python.NET                                                                                                          |
| ->     An Free Open-souce Python Runtime for C#                                                                         |
|                                                                                                                         |
|     PdfPig                                                                                                              |
| ->     An Free Open-souce PDF Manipulation Library                                                                      |
|                                                                                                                         |
|     RAKE Algorithm (rake_nltk)                                                                                          |
| ->     An Free Open-souce Python Module that implements RAKE                                                           |
|                                                                                                                         |
|     TF-IDF Algorithm (ML.NET)                                                                                           |
| ->     A Feature of Text_Featurizing_Estimator Class from ML.NET                                                        |
|                                                                                                                         |
|     AutoML (ML.NET)                                                                                                     |
| ->     Provides methods and processes for automatically selecting the best algorithm given the task.                    |
|                                                                                                                         |
|  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  |
|                                                                                                                         |
|  Getting Started:                                                                                                       |
|  ~~~~~~~~~~~~~~~~                                                                                                       |
|  Sadly, there are a few things that have to be done manually:                                                           |
|                                                                                                                         |
| ->  1. Search for ('Cntl+F') [Runtime.PythonDLL =] which is located at the Rake.cs file                                 |
|                                                                                                                         |
| ->  2. Once found, insert the ABS Path of your python3.XX.dll                                                           |
|                                                                                                                         |
| ->  3. You will have to download and install .NET 7.0 or 6.0, just Google it                                            |
|                                                                                                                         |
| ->  4. Add more...                                                                                                      |
|                                                                                                                         |
|  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  |
|                                                                                                                         |
|  For more Details visit the GitHub Repository:                                                                          |
|                                                                                                                         |
|  Thank you for reading! Happy Coding!                                                                                   |
|                                                                                                                         |
|  Press any key to go Main Menu...                                                                                       |
```

Exploring: ***Option #4-5***

```
+-------------------------------------------
|  Select one of the following options,
|
|     1. Find Suitable Reviewer.
|     2. Display Current Datasheet Details.
|     3. Train a new Model.
|     4. Project's Description.
|     5. Terminate the Application.
+-------------------------------------------
```

Comments:

1. The ***4th Menu*** (Project's Description) simply prints some information about the project, as it is visible on the side.
2. The ***5th Option*** (Terminate the Application) does exactly that.

4.1 INSTALLATION & SETUP

Required Software:

- ***.NET 7.0 or 6.0*** (Includes everything needed to run a C# App)
  - Link: https://dotnet.microsoft.com/en-us/download
- ***Python*** (Programming Language)
  - Link: https://www.python.org/downloads/

Getting the Code:

- GitHub Repository:
  - https://github.com/jimzord12/MscRes-PaperReviewerFinder
- (Requires ***"Git"*** to be installed on your machine)
  - Enter on your terminal:

```
git clone https://github.com/jimzord12/MscRes-PaperReviewerFinder.git
```

4.2 INSTALLATION & SETUP

Getting Dependencies :

- Using your Terminal, ***navigate*** to root folder, the one containing a ***".csproj"*** file.
- Then, enter: `dotnet restore`
- This should download all required packages/dependencies.

Code Setting:

- PythonDLL Path:
  - Search for a file named: ***"Rake.cs"*** in the project's Root Directory.
  - At the Beginning of the file (line 23), you should find this line:

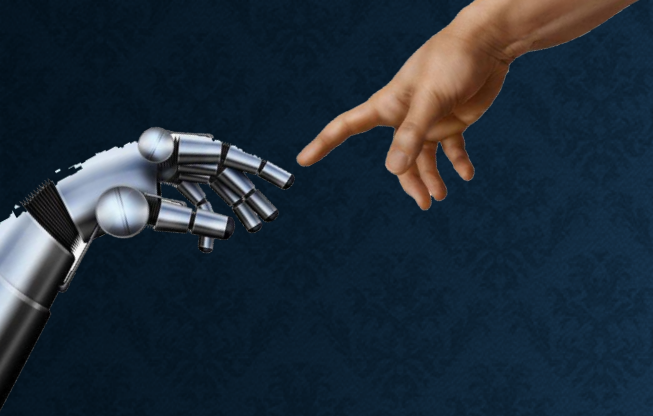
```
Runtime.PythonDLL = @"C:\Users\Jimzord12\AppData\Local\Programs\Python\Python311\python311.dll";
```

  - Replace the ***Path*** with your own. If you have ***Windows 10*** and have download the ***Python 3.11.X version***, you should only need to change the User Name as indicated below (without the {}):

```
Runtime.PythonDLL = @"C:\Users\{YourUsername}\AppData\Local\Programs\Python\Python311\python311.dll";
```

Running the App:

- Through your terminal, navigate to the project' root directory and enter: `dotnet run`



Thank you for your time!

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA