**WP4 DATA COLLECTION TASK**

This task aims to collect open-source software packages (including but not limited to npm, Python, and Ruby) that use AI/ML techniques/components. The data collection method could be:

1. Obtaining an OSS package list from existing libraries.
2. Collecting metadata of packages, such as the package description.
3. Extracting information from package description, determining whether a package uses AI/ML techniques/components. For this, the keyword matching method is used. Therefore, a keyword list has been prepared and validated.

After obtaining a list of packages using AI/ML, the source code of these packages will be collected for further analysis, such as the fairness evaluation.

*1.1 Identifying libraries*

Libraries for Open Source Software packages are collections of pre-written code that developers can use to perform specific tasks or add specific functionality to their software projects. Based on the available libraries, the following libraries has been identified for extracting and identifying OSS packages:

- Github
- Open.ai
- Libraries.io
- koha-community.org
- npmjs.com

*1.2 Keyword Selection*

There are several ways to find keywords for searching and the following methods has been used for identifying and selection appropriate keywords in Artificial Intelligence and Machine Learning.

Brainstorming: Think about the topic you want to search for and come up with a list of related words and phrases.

Analysing Articles: Analysing some journal articles and conference research papers that are published in related field and obtain keywords from it.

Through these methods, following keywords has been identified:

- Self-supervised Learning

- Zero-shot learning
- Supervised Learning
- Linear regression
- Logistic regression
- Artificial neural networks
- Unsupervised Learning
- Clustering
- Image Recognition
- Deep Learning
- Neural Networks
- Data Mining
- Natural language generation (NLG)
- Pattern recognition
- Predictive analytics

1.3 Package Collection

Based on the keywords selected, these keywords are then manually searched and obtained from the libraries identified in the step 1.1.

When packages are extracted, the following information/metadata is obtained based on the available information from the libraries. Below table shows appropriate number of packages collected per each keyword.

| **Keyword** | **Number of Packages** |
|---|---|
| Deep Learning | 15 |
| Supervised Learning | 11 |
| Zero-shot learning | 9 |
| Machine Learning | 9 |
| Artificial neural networks | 11 |
| Unsupervised Learning | 13 |
| Clustering | 13 |
| Image Recognition | 15 |
| Data Mining | 18 |
| Natural Language Processing | 14 |
| Pattern recognition | 13 |
| Predictive analytics | 13 |

Now, the packages can be retrieved for analysing the source code for further analysis and next steps of the project.