**SDG Corpus**

**Update Report**

PREPARED FOR

# Introduction

The purpose of this document is to provide the reader with an insight about the SDG corpus update conducted on 30-08-2020, as well as the testing process used to determine the accuracy results.

# Corpus Versions

In this section, we present the three versions of the SDG corpus used in the testing process. They are as follows:

* **Original Corpus:** This is the version currently running on ISTISHARA consisting of the previous version of the SDG metadata documents augmented with the wiki articles concerning each goal.

* **New Metadata:** This version consists only of the new version of the SDG metadata documents found on the [UN SDG Indicators Page](https://unstats.un.org/sdgs/metadata/) .

* **New Metadata with Wikipedia Augmentation:** This version consists of the new SDG metadata documents found on the [UN SDG Indicators Page](https://unstats.un.org/sdgs/metadata/) . In addition to the metadata, the Wikipedia articles have been added in this version as well.

# Testing Process

In this section, we describe our test documents set as well as the scoring process we used in order to determine the accuracy of each corpus.

## **Test Documents**

The test documents are ESCWA publications annotated by human experts. For each document, we have a list of related SDGs according to the labeling done by the ESCWA staff.

We started off with 170 documents but after removing non-english documents as well as the duplicates, we were left with 150 documents that we used for testing.

## **Scoring Process**

The scoring process is the same as ISTISHARA’s: An SDG is considered relevant if its similarity score is equal or greater to a certain threshold.

Based on some previous tests, we deduced that an adequate threshold for the SDG corpus would be around 0.07. This value is considered to be the threshold in the current tests.

A document is considered as *relevant* with respect to an SDG if one of the SDGs in the related SDGs list is considered as a relevant SDG by ISTISHARA.

The final corpus accuracy will be calculated by dividing *the number of relevant documents* over the total *number of test documents*.

# Testing Results

In this section, we showcase a comparative table presenting the results of each version of the SDG corpus

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Original Corpus** | **New Metadata** | **New Metadata with Wikipedia Augmentation** |
| Number of relevant documents | 138 | 133 | 134 |
| Number of testing documents | 150 | 150 | 150 |
| Accuracy (%) | **92** | **88.6** | **89.3** |

Results show a small drop in accuracy, with the new corpora. This can highlight the fact that the test documents were manually labelled by humans having in mind the original descriptions of the SDGs and not the new ones. Now that the SDG corpus has been updated, the manual annotations might need to be manually re-evaluated accordingly.

# Non-Relevant Documents

Among the non-relevant documents, a set of eight documents was constantly showing regardless of the corpus version used. This is due to multiple reasons:

* One document (*Review of Innovative and Appropriate Technologies for Waste Management in Morocco and the Arab Region*) talks about waste management techniques and was labeled as SDG 17 by the human expert, while ISTISHARA classified it under SDGs 7, 11 and 12. These SDGs might be closer to the waste management topic than SDG 17. The mentioned document may need to be re-assessed by a human expert in light of the above results.
* Two documents were either infographic documents or numerical chart documents.

* One document had a large Arabic section within it.