Checking the alignment of global actions and development targets

Text mining project

Ivan Vallejo

June 19, 2017

Introduction

In September 2015, the United Nations (UN) adopted the 2030 Agenda for Sustainable Development (General Assembly resolution 70/1, 2015). This document identifies a set of priority areas in the global development agenda which supersede those set by the Millennium Development Goals (General Assembly resolution 55/2, 2000).

In particular, the 2030 Agenda for Sustainable Development identifies 17 Sustainable Development Goals (SDGs), each corresponding to a given area of action (Figure 1).

FIGURE 1: List of United Nations Sustainable Development Goals. Source: Sustainable Development Knowledge Platform, UN Department of Economic and Social Affairs.

Following the adoption of the SDGs, each agency of the United Nations development system (i.e. the diverse United Nations funds, programmes and specialized agencies) has been required to align its actions to those areas within their mandate included in the SDGs.

This project uses text mining methods to assess whether the activities of a UN specialized agency, the International Telecommunication Union (ITU), are aligned with the Sustainable Development Goals. ITU is the UN specialized agency for information and communication technologies (ICTs).[1]

The procedures and methods proposed in this project can be easily generalized and applied to assess the degree of alignment of other agencies, provided that the thematic dictionaries are adapted to the area of action of each agency.

Data

Data were obtained by scraping the ITU website. A hierarchical procedure was followed: starting at the ITU home page, each link to another ITU webpage was recorded and scraped. The process was iterated three times, thus collecting data for all ITU webpages at a distance of three clicks from the home page. Figure 2 shows an example of a hierarchical path for three connecting webpages, as well as the total number of pages per layer.

Each webpage was scraped just once, at its first occurrence. That is, any further links to a webpage already scraped were discarded, to avoid having repeated content in the text analysis. Data were scraped in one run in June 2017.

Given the large number of webpages at depth three – more than 25'000 webpages three clicks away from the home page – the text scraping was limited to depth one and depth two. Webpages containing less than 10 terms were discarded, because their classification would be unreliable. In total, the text of about 2'000 pages was retained. The full text of each webpage was considered as a document unit.

Figure 3 shows a graph representation of the webpages whose text was scraped.

In addition to the data collected from the ITU website, the official text description of SDGs 4, 5, 9, 14 and 17 was obtained from the Sustainable Development Knowledge Platform.

In particular, the description contained in the tabs "Progress & Info" and "Targets & Indicators" of each concerned indicator was collated, thus producing an individual document with the text relevant for each SDG. These documents were used to build the dictionary of each concerned SDG.

[1] For more information, see ITU's website

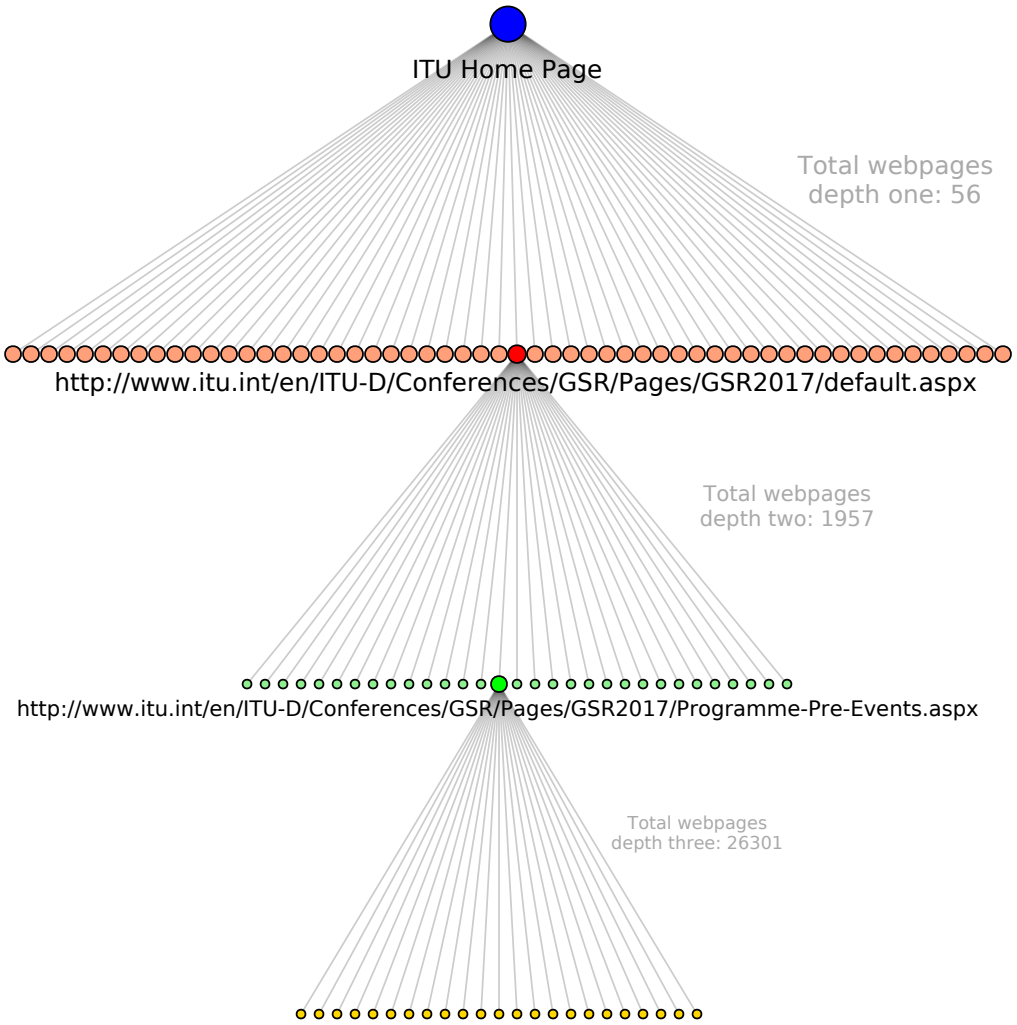


FIGURE 2: ITU website – Diagram of layers, June 2017

Research question

> Are the activities of the United Nations agencies aligned with the Sustainable Development Goals?

This is a crucial question when evaluating the efficiency of the activities undertaken by the United Nations development system. Furthermore, it is relevant for accountability purposes, more so considering the large amount of resources channeled to the UN system[2] and the critical

[2]For instance, USD 5.4 billion for the 2016-2017 Biennium. Source: UN Administrative and Budgetary Committee

actions expected from the UN agencies by the Member States.[3]

Indeed, the importance of aligning the plans and activities of each UN agency with the SDGs has been highlighted in several UN documents. For instance, in paragraph 88 in General Assembly resolution 70/1 (2015) and paragraphs 17-20 and 78-79 in General Assembly resolution 71/243 (2016).

This project aims to be an initial proof of concept on whether text mining and, in particular, bag-of-words approaches can be used to assess the alignment with the SDGs in an objective and systematic way based on data publicly available on the UN websites.

This project focuses on the use case of the International Telecommunication Union. Therefore, it tries to answer the research question for this particular UN agency.

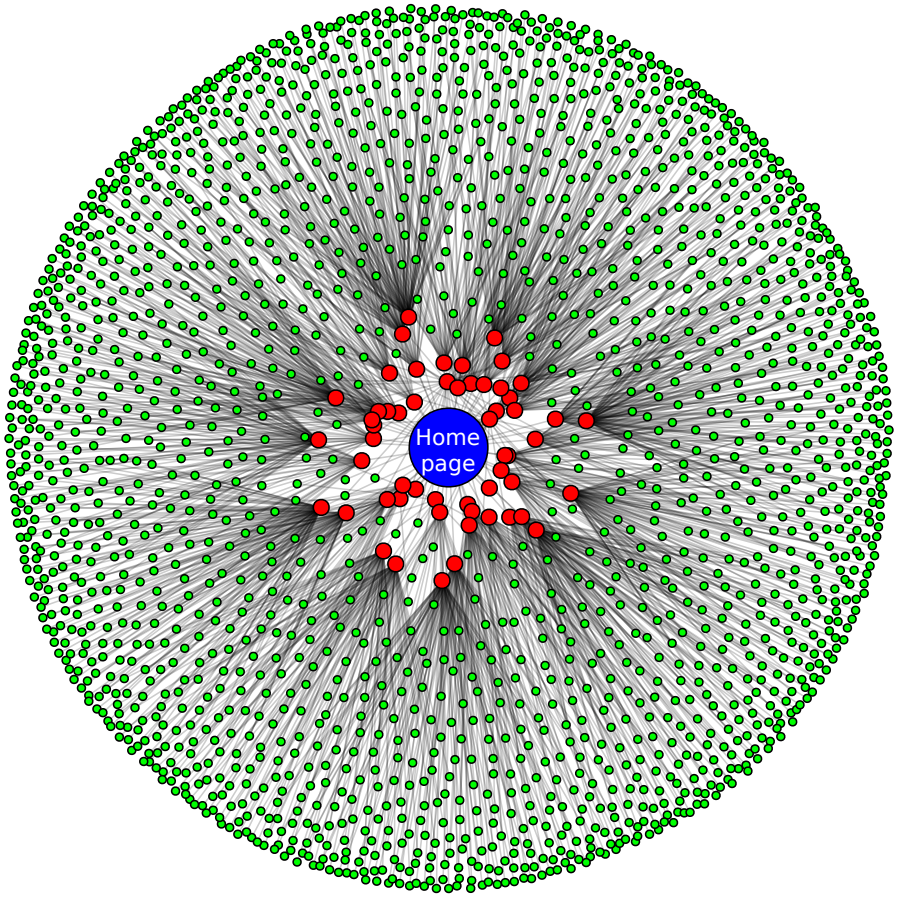


FIGURE 3: Network diagram scraped webpages, June 2017.

[3]See for instance the wording of SDG 1: "End poverty in all its forms everywhere".

Extracting content

Dictionaries

The first step towards classifying the content extracted from the ITU website is to build a series of dictionaries that characterize our topics of interest. The dictionaries should be created independently from the text on the website, because we cannot assume that our topics of interest are correctly represented on the website. Indeed, this is particularly what we want to probe: the occurrence on ITU's website of a series of topics exogenous from the website.

Our topics of interest are those relating to the SDGs which concern ITU. According to the United Nations Economic and Social Council, Statistical Commission, Forty-seventh session (2016), the following SDGs will be monitored using specific ICT indicators:[4]

Goal 4 : Quality education

Goal 5 : Gender equality

Goal 9 : Industry, innovation and infrastructure

Goal 17 : Partnership for the goals

We take as a proxy for relevance the specific mention to ICT-related metrics in the monitoring framework of an SDG. Therefore, SDGs 4, 5, 9 and 17 should be of relevance to ITU. We add to the list of topics SDG 14, which relates to life below water and should arguably not concern much ITU's activities. We will use it as a placebo against which to gauge the intensity of appearance of the other topics on the ITU website.

It is not possible to use already existing dictionaries to obtain a list of terms specific to each of these SDGs, because the content of each topic is very specific to the 2030 Agenda for Sustainable Development.

Therefore, we build our own dictionaries by applying latent Dirichlet allocation (LDA) to a set of labeled documents characteristic of these topics. We proceed as follows:

1. Create one document per SDG by collating the text contained in the tabs "Progress & Info" and "Targets & Indicators" from the Sustainable Development Knowledge Platform.

[4]See also the ITU webpage on the 2030 Agenda for Sustainable Development for a summary of the main points drawn by ITU from the United Nations Economic and Social Council, Statistical Commission, Forty-seventh session (2016).

2. Create a document-term matrix from this corpus. We remove stop words, non-alphabetic characters and stem the terms using the Porter stemmer. We drop the resulting terms containing only one character, as their interpretation is difficult for any given topic.

3. Apply a term frequency–inverse document frequency (TF-IDF) weighting and keep only those terms with a minimum TF-IDF (Figure 4).

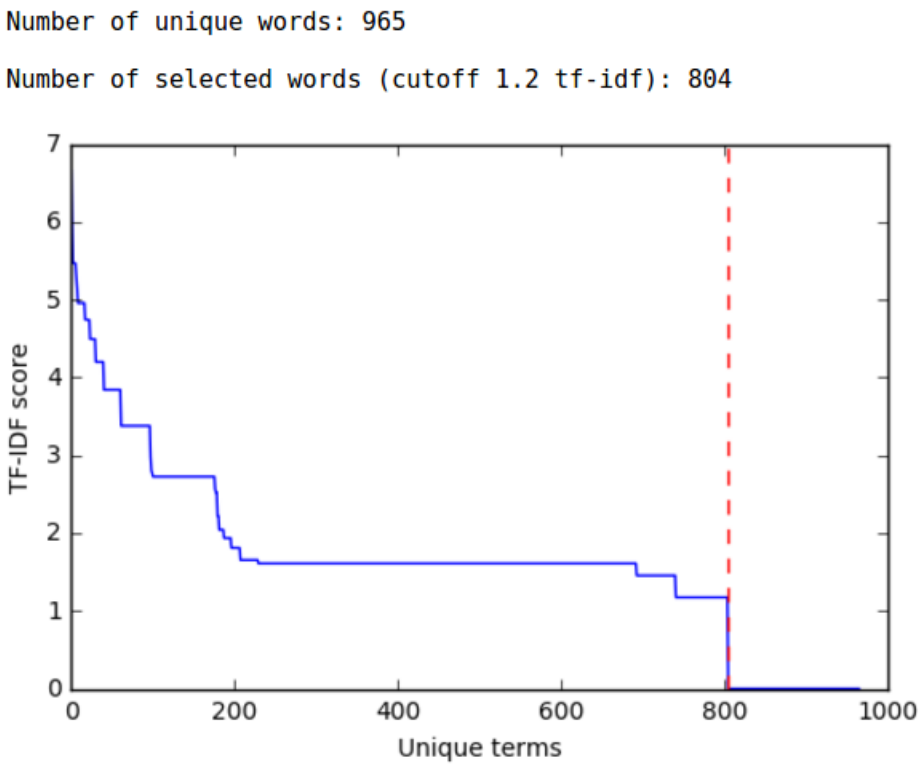
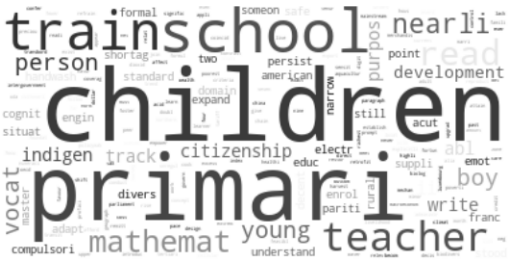


FIGURE 4: Number of terms and TF-IDF of the corpus used to create the dictionaries.

4. Run LDA on this TF-IDF document-term matrix based on 20 topics. For each document (which in turn relates to an SDG of interest), we select the topic that has the highest probability. For each of the selected topics, we build the dictionary by retaining the 50 most probable words in that topic. An illustration based on word clouds of each dictionary is shown in Figure 5.

Goal4
(education), main topic: 11



Goal5
(gender equality), main topic: 10

Goal9
(industry & infrastructure), main topic: 4

Goal14
(sea), main topic: 13

Goal17
(partnerships), main topic: 5

FIGURE 5: Word clouds of each dictionary.

5. The main terms retained seem to be in line with what would be expected of each topic. In order to formalize this judgment, we apply the dictionaries to the TF-IDF document-term matrix of the SDG corpus. That is, for each document, we add the TF-IDF of the terms included in a given dictionary and divide the result by the sum of the TF-IDF of all terms in that document. The results confirm that each dictionary identifies a unique SDG (Figure 6).

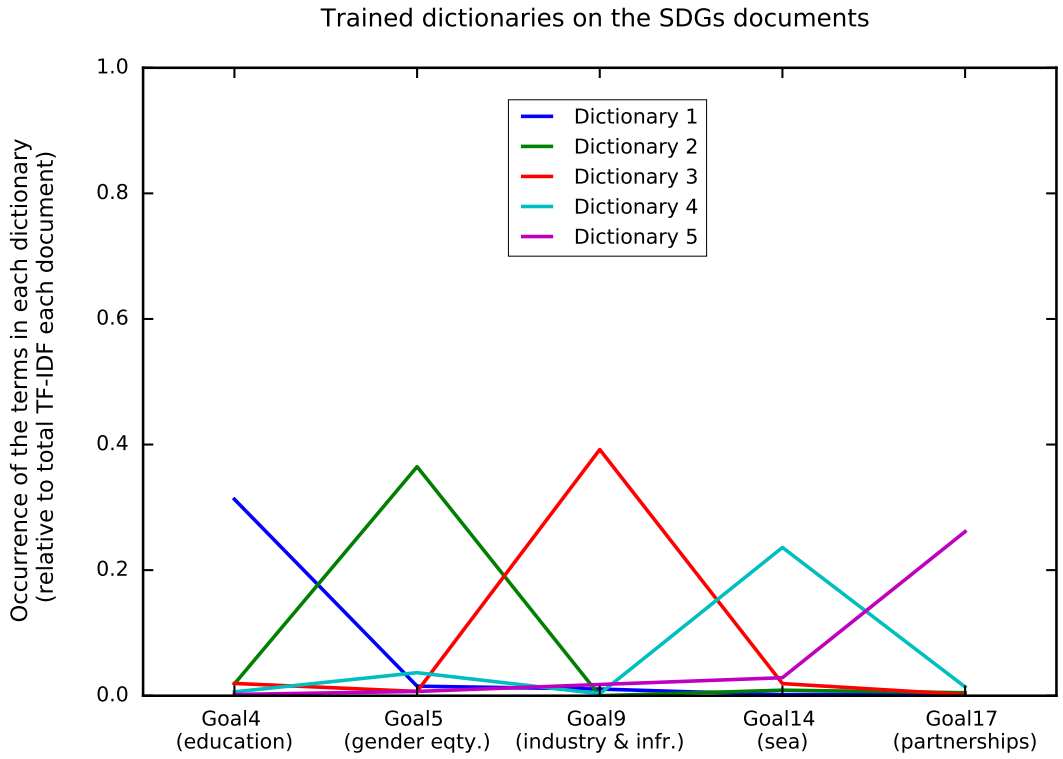


FIGURE 6: Results of applying the dictionaries to the SDG corpus.

ITU website

In order to carry out the analysis on the corpus of ITU webpages, we proceed as follows:

1. Load the data scraped from the ITU website: 1'987 documents, each corresponding to the text of an ITU webpage.

2. Create a document-term matrix from this corpus. We remove stop words, non-alphabetic characters and stem the terms using the Porter stemmer.

3. Apply the TF-IDF weighting to the document term matrix limiting the maximum number of terms to 5'000 (which is actually the final number of terms retained).

4. Apply each dictionary to the corpus of ITU websites and record the results. As mentioned in the previous section, we apply a dictionary to each document by adding the TF-IDF of the terms included in the dictionary occurring in a document and dividing the result by the sum of the TF-IDF of all terms in all documents. It is important to note that here we normalize by the sum of the TF-IDF of all documents. If we normalized by the TF-IDF result of a particular document, we would favor shorter webpages with one or

two relevant terms (i.e. we would classify them as the most intense occurrences of a given topic).

5. We record the dictionary that produces the highest result for each document: a proxy for the most likely topic among the ones characterized by the dictionaries. The actual value obtained with that dictionary is also saved as a measurement of the intensity in which the most probable topic is present in the document. Very low intensities ($< 10^{-6}$) are treated as indicating that the document does not include any of the topics represented by the dictionaries.

As a sanity check for the out-of-sample performance of the dictionaries, we inspect the webpages producing the highest results for each dictionary (Figure 7). As could be expected, each ditionary naturally identified an ITU webpage actually dedicated to explain the corresponding SDG.

The only exception is the dictionary for SDG 5 (gender), which pointed to a website listing the resolutions adopted by the last ITU plenipotentiary, several of them related to gender issues.

```
SDG4 (education): http://www.itu.int/en/sustainable-world/Pages/goal4.aspx
u'school': 1, u'engin': 1, u'person': 1, u'young': 1, u'decent': 1, u'safe': 1, u'primari': 1, u'divers': 1, u'indi
gen': 1, u'teacher': 1, u'train': 1, u'vocat': 1, u'suppli': 1, u'rural': 1, u'educ': 1, u'children': 1, u'expand':
1, u'standard': 1

SDG5 (gender): http://www.itu.int/dms_pub/itu-s/opb/conf/S-CONF-PLEN-2015-TOC-HTM-E.htm
u'harm': 1, u'telephon': 1, u'enabl': 1, u'forc': 1, u'anoth': 1, u'empower': 1, u'legal': 1, u'union': 1, u'proced
ur': 1, u'respons': 1, u'document': 1, u'constitut': 1, u'provis': 1

SDG9 (infr.): http://www.itu.int/en/sustainable-world/Pages/goal9.aspx
u'enterpris': 1, u'circumst': 1, u'chain': 1, u'credit': 1, u'alia': 1, u'drive': 1, u'landlock': 1, u'research':
1, u'capabl': 1, u'rais': 1, u'resili': 1, u'substanti': 1, u'african': 1, u'opportun': 1, u'industri': 1, u'inte
r': 1, u'line': 1, u'problem': 1, u'strive': 1

SDG14 (sea): http://www.itu.int/en/sustainable-world/Pages/goal14.aspx
u'prohibit': 1, u'unreport': 1, u'certain': 1, u'fish': 1, u'trade': 1, u'sea': 1, u'special': 1, u'subsidi': 1,
u'differenti': 1, u'intergovernment': 1, u'harvest': 1, u'activ': 1, u'legal': 1, u'treatment': 1, u'guidelin': 1,
u'negoti': 1, u'stock': 1, u'preserv': 1, u'marin': 1, u'unregul': 1, u'law': 1, u'overfish': 1, u'reflect': 1, u'r
estor': 1

SDG17 (partnership): http://www.itu.int/en/sustainable-world/Pages/goal17.aspx
u'societi': 1, u'erad': 1, u'doha': 1, u'civil': 1, u'partnership': 1, u'coher': 1, u'highli': 1, u'macroeconom':
1, u'negoti': 1, u'scienc': 1, u'statist': 1, u'financ': 1, u'coordin': 1, u'diffus': 1, u'fulli': 1, u'debt': 1,
u'net': 1, u'enabl': 1, u'invest': 1, u'rule': 1, u'issu': 1, u'paragraph': 1, u'integr': 1, u'agenda': 1, u'highli
ght': 1, u'view': 1
```

FIGURE 7: Websites with the highest intensity per topic and list of terms on the websites corresponding to the relevant dictionary.

Addressing the research question

In order to address the research question, we use the results of the classification obtained with the dictionaries.

Figure 8 shows a graphical network representation of topic assignments.

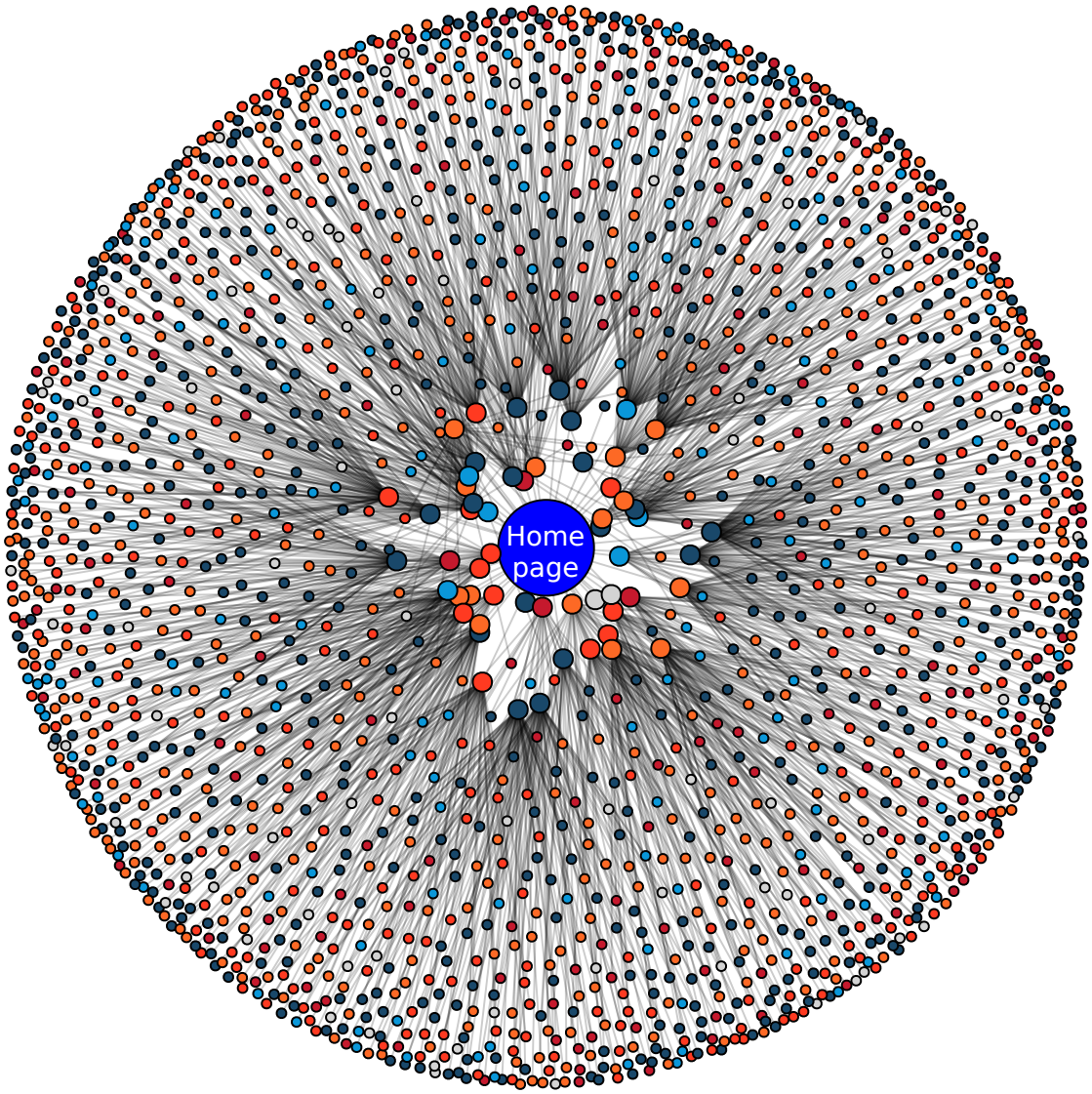


FIGURE 8: Network diagram of topic assignments – ITU website, June 2017.

Translating this diagram into numbers, Figure 9 exhibits the histograms of each category of webpages.

A majority of ITU webpages are assigned to SDG 17 (partnership) or SDG 9 (infrastructure). This is precisely what would be expected.

Indeed, SDG 17 closely corresponds to the Millennium Development Goal (MDG) 8 ("Develop a Global Partnership for Development"), which was the only MDG explicitly mentioning ICTs. Therefore, during the MDG period (2000-2015), this was the main reference in the global

agenda of relevance to ITU and therefore the organization aligned with it.

SDG 9 touches upon a core ITU activity: (ICT) infrastructure development. As a result, little extra alignment is needed to cover this topic because ITU pre-SDG activities were already heavily focused in this area.

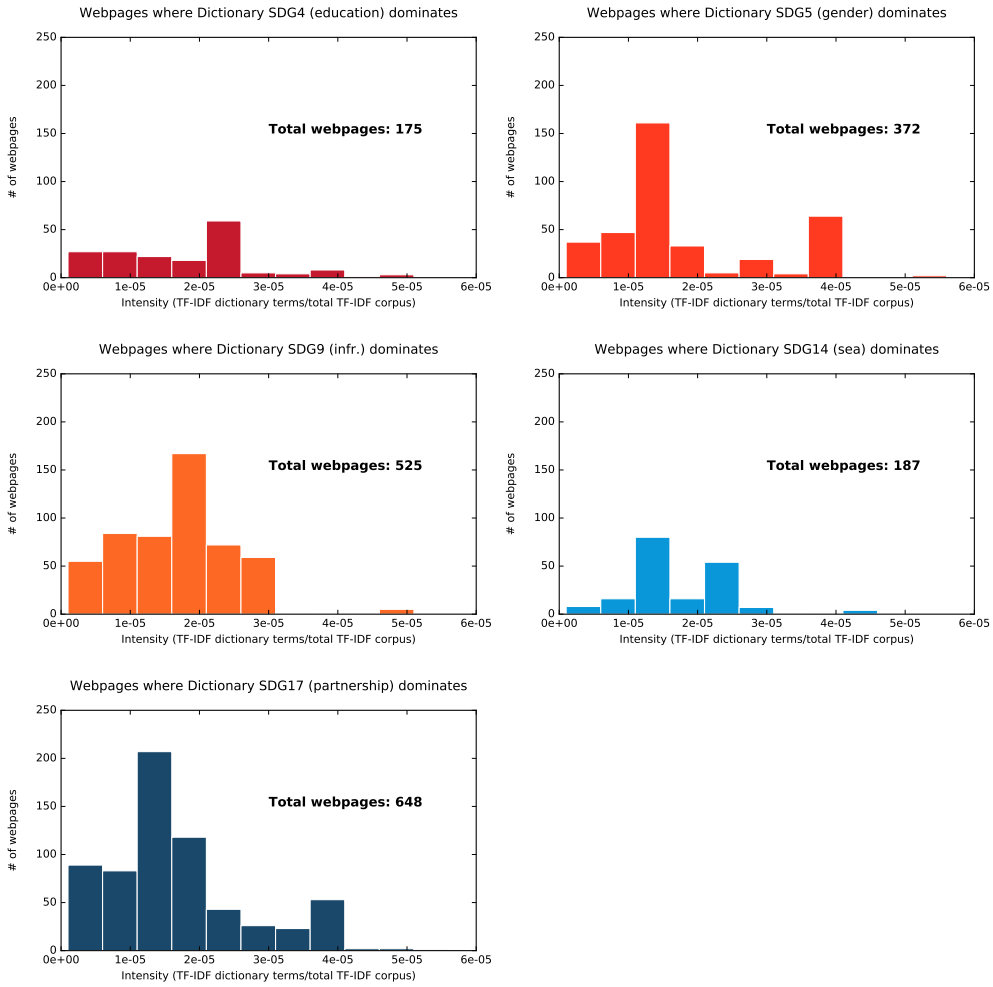


FIGURE 9: Histograms of topic assignments – ITU website, June 2017.

SDG 5 (gender) has a lower but significant number of webpages assigned to it. This suggests that the cross-sectional campaigns undertaken to mainstream gender issues in ITU have had an effect. Indeed, gender has been progressively incorporated as an area of action in the ITU

fundamental documents, starting with WTDC resolution 44 (Istanbul, 2002) and continuing with WTSA resolution 55 (Dubai, 2012), WTDC resolution 55 (Rev. Dubai, 2014) and ITU Plenipotentiary Conference resolution 70 (Rev. Busan, 2014).

SDG 4 (education) has very little presence on the ITU website. Indeed, fewer webpages are associated with this topic than with SDG 14 (sea). Since we know that the latter hardly concerns ITU's activities, we conclude that SDG 4 is also not represented on the ITU website.

This suggests that ITU's activities in the area of education are less prominent than what the 2030 Agenda for Sustainable Development would require. This may be explained by the fact that ITU's mandate in this area overlaps with that of UNESCO, which is the main UN programme in the areas of education, science, culture and communication.

It would seem that UNESCO is taking a leading role in the area of education, also when it involves ICTs, and that ITU is not intervening much. This may be a desirable course of action as long as UNESCO and ITU are coordinating their activities in this area.

Lastly, Figure 10 presents the results of the classification exercise broken down by the depth of the website (i.e. the number of clicks that separates it from ITU home page). We note that there are no major differences in the distribution of topics between depth one and depth two webpages.

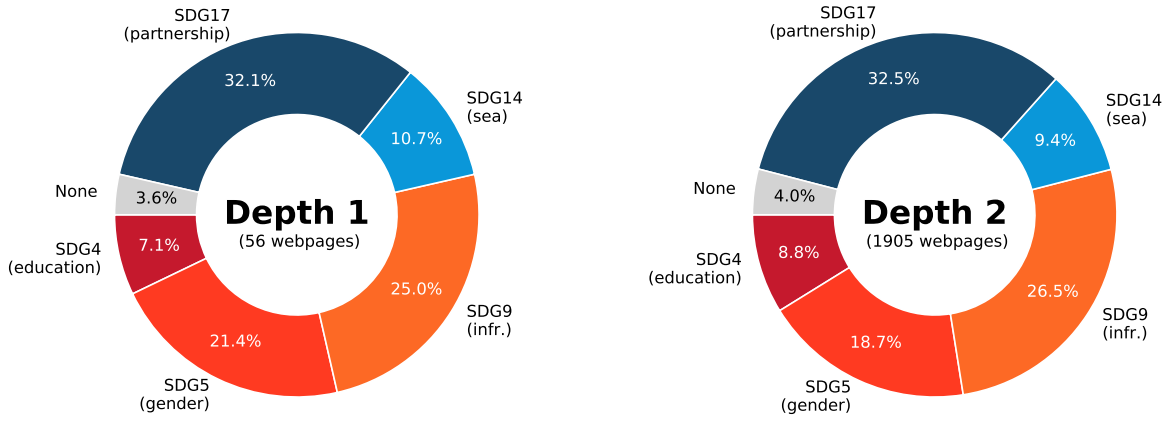


FIGURE 10: Pie charts of topic assignments, broken down by webpage depth, June 2017.

Conclusion

The methodology proposed in this project, based on a combination of LDA for dictionary construction and dictionary methods with TF-IDF weighting, allows to draw some meaningful conclusions on whether the activities of a given United Nations agency are aligned with the

Sustainable Development Goals.

The project has exemplified this approach using data from the International Telecommunication Union. However, the same approach could be applied to any other UN programme or agency, provided that the thematic dictionaries are adapted to the area of action of the agency/programme.

The lack of comprehensive ground truth data against which to cross-validate the classification obtained using dictionary methods makes it difficult to evaluate the accuracy of the methodology. Ideally, having a set of human-labeled webpages within the website of interest would help in the out-of-sample validation. In the absence of this, this project has enforced: (i) a sanity check based on the inspection of the most representative webpage per topic, as identified by the dictionaries; (ii) an evaluation of the results against external sources of information of ITU's activities, such as resolutions, the previous global development agenda and mandates of other agencies.

The results of these checks have been positive, but such confirming evidence has limited representativeness in quantitative terms, and relies on qualitative judgments.

A shortcoming of the methodology proposed concerns the fact that the classification of the webpages is relative. For example, dictionary 1 (education) contains 50 terms, including "someon" and "safe". A short webpage containing 20 terms, including those two, and none of the terms included in the other dictionaries would be classified as dealing with education. However, the sum of the TF-IDF of the two terms could be small and a manual inspection would not conclude that the webpage covers the topic of education.

We have enforced a rule of thumb threshold ($< 10^{-6}$) to screen the most obvious cases of webages not covering any topic. However, a more precise threshold would be needed. This threshold is difficult to derive, given the relative nature of the results of applying the dictionaries. Indeed, the numerical value produced by a dictionary depends on the number of terms in the document, the TF-IDF weighting of the terms, and the normalizing TF-IDF used (be the total sum of the corpus, which favors longer documents, or the total sum of each document, which favors shorter documents).

It is not straightforward to propose an objective threshold based on such a set of relative measures. Moreover, the identification of the threshold may require extensive empirical fine-tuning for each given corpus. This would need to be considered in future projects in this area.

References

General Assembly resolution 55/2 (2000). United nations millennium declaration. http://www.un.org/millennium/declaration/ares552e.htm. A/RES/55/2 (8 September 2000).

General Assembly resolution 70/1 (2015). Transforming our world: the 2030 agenda for sustainable development. http://www.un.org/ga/search/viewm_doc.asp?symbol=A/RES/70/1. A/RES/70/1 (25 September 2015).

General Assembly resolution 71/243 (2016). Quadrennial comprehensive policy review of operational activities for development of the united nations system. http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/71/243. A/RES/71/243 (21 December 2016).

ITU Plenipotentiary Conference resolution 70 (Rev. Busan, 2014) (2014). Mainstreaming a gender perspective in itu and promotion of gender equality and the empowerment of women through information and communication technologies. http://www.itu.int/en/ITU-D/Digital-Inclusion/Documents/Resolutions/Resolution70_PP_BUSAN_14.pdf.

United Nations Economic and Social Council, Statistical Commission, Forty-seventh session (2016). Report of the inter-agency and expert group on sustainable development goal indicators. https://unstats.un.org/unsd/statcom/47th-session/documents/2016-2-IAEG-SDGs-Rev1-E.pdf. E /CN.3/2016/2/Rev.1 (8-11 March 2016).

WTDC resolution 44 (Istanbul, 2002) (2002). Mainstreaming gender in itu-d programmes. http://www.itu.int/en/ITU-D/Digital-Inclusion/Women-and-Girls/Documents/Resolutions/WTDC%20ISTANBUL-Res-44.pdf.

WTDC resolution 55 (Rev. Dubai, 2014) (2014). Mainstreaming a gender perspective for an inclusive and egalitarian information society. http://www.itu.int/en/ITU-D/Digital-Inclusion/Documents/Resolutions/Resolution55_WTDC_DUBAI_14.pdf.

WTSA resolution 55 (Dubai, 2012) (2012). Mainstreaming a gender perspective in itu-t activities. https://www.itu.int/en/ITU-T/wtsa12/Documents/resolutions/Resolution%2055.pdf.