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 paper 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 the month of 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 for which 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 were 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

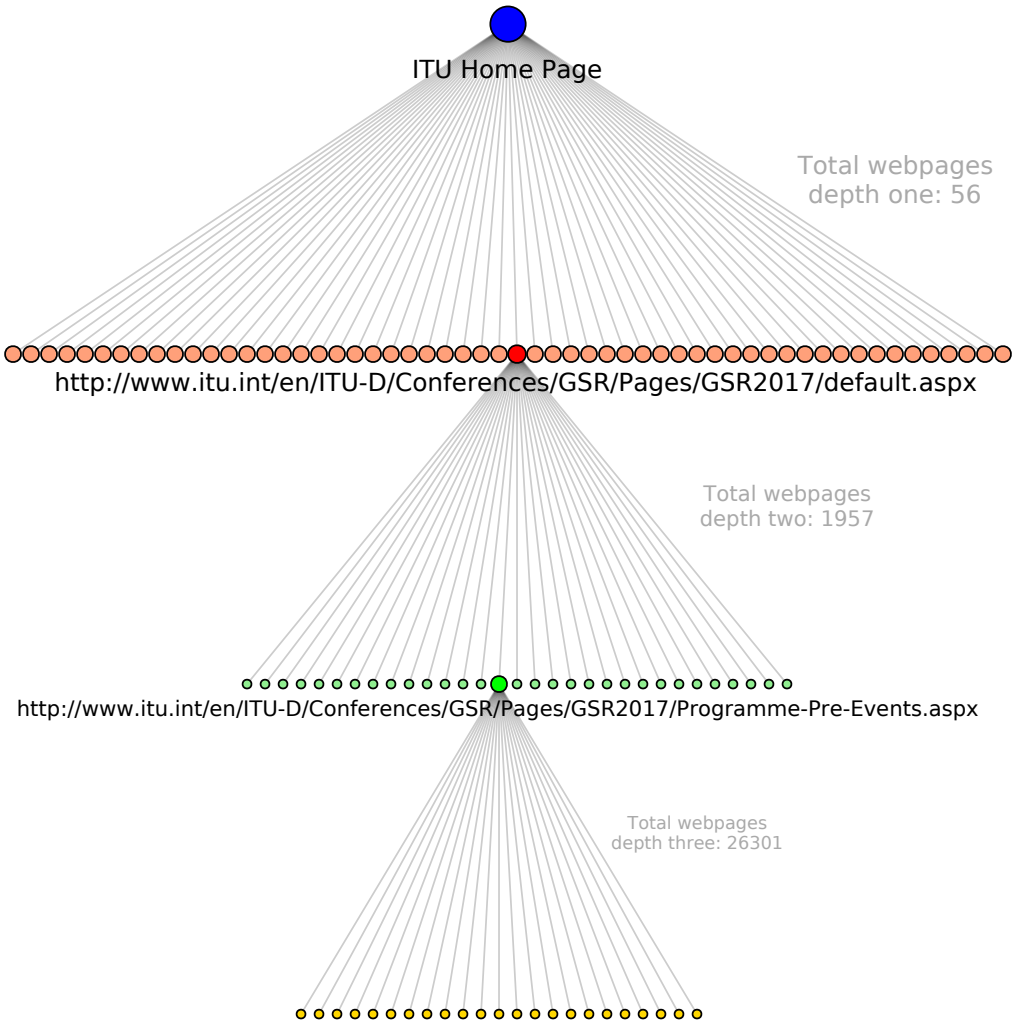

FIGURE 2: ITU website – Diagram of layers

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 answer this question in an objective and automated 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.

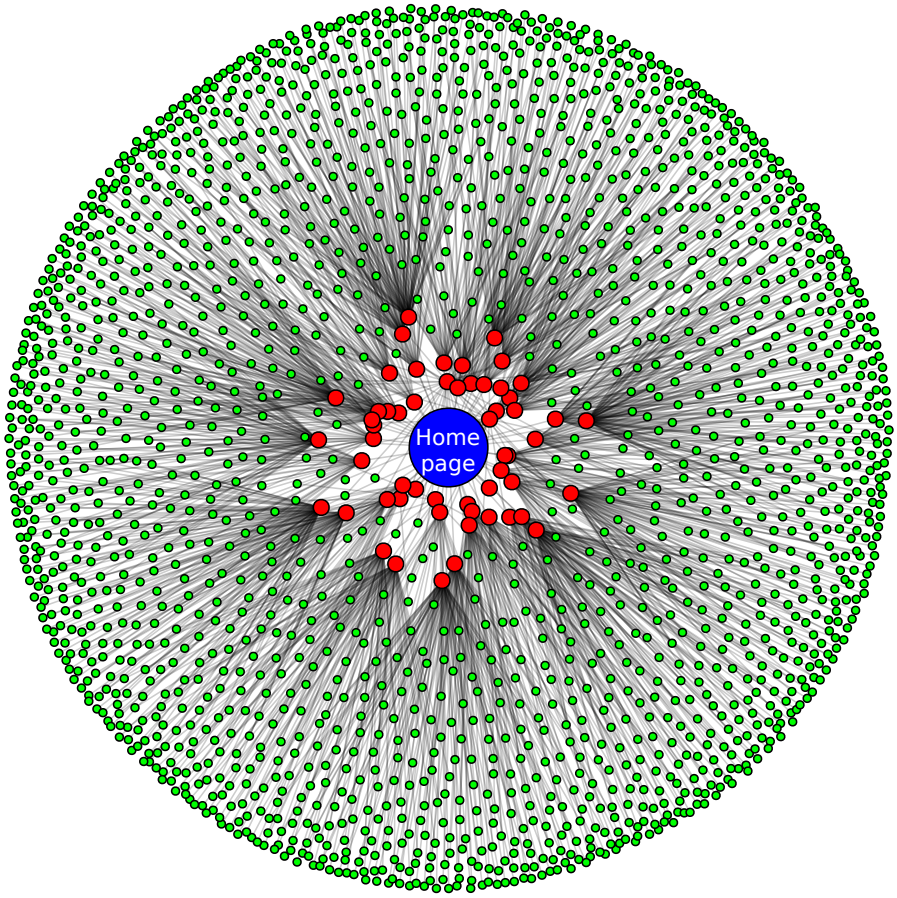

FIGURE 3: Network diagram scraped webpages.

[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 some dictionaries 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) for ITU.

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

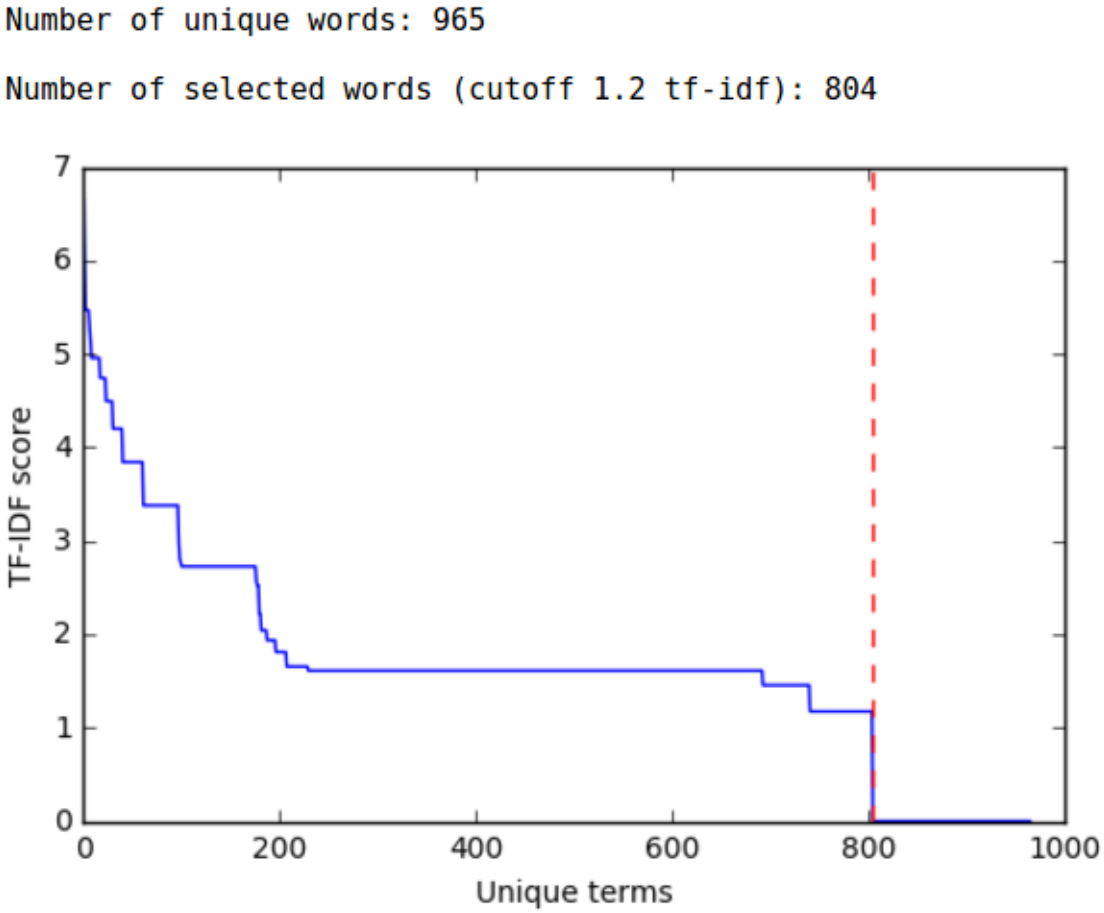

FIGURE 4: Number of terms and TF-IDF of the corpus used to create the dictionaries.

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 retain 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.

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 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 total TF-IDF of that document. The results confirm that each dictionary identifies a unique SDG (6).

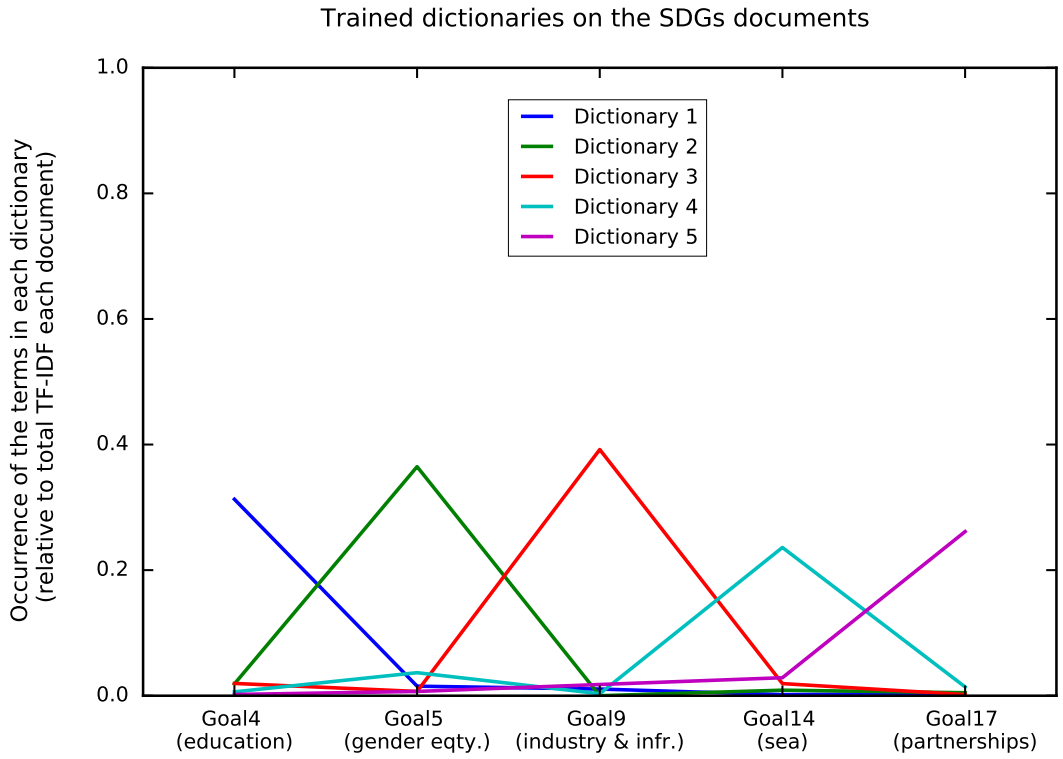

FIGURE 6: Results of applying the dictionaries to the SDG corpus.

ITU website

We proceed...

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).

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).