Office for National Statistics

Measuring Sustainability Reporting using Web Scraping and Natural Language Processing

Alessandra Sozzi
alessandra.sozzi@ons.gov.uk

The sustainable development goals (SDGs) are a new, universal set of goals and targets that member states of the United Nations are expected to use to frame their agendas and political policies over the next 15 years. The Goals will be followed-up and reviewed using a set of global indicators, complemented by indicators at the regional and national levels which will be developed by each member state.

Target 12.6 is to '**Encourage companies, especially large and transnational companies, to adopt sustainable practices and to integrate sustainability information into their reporting cycle**'.

In order to measure companies response to the SDG's call for action, an indicator that quantifies the number of companies publishing sustainability reports, by turnover band, geography, national or global economy, sector and number of employees, has been proposed.

Nowadays the Web represents a medium through which corporations can effectively disseminate and demonstrate their efforts to incorporate sustainability practices into their business processes. This led to the idea of using the Web as a source of data to measure how UK companies are performing against the indicator.

DATA COLLECTION

The scraper navigates through each company website, accessing every internal link. While recursively traversing websites, the scraper flags web pages that suggest sustainability content through a keyword-based search.

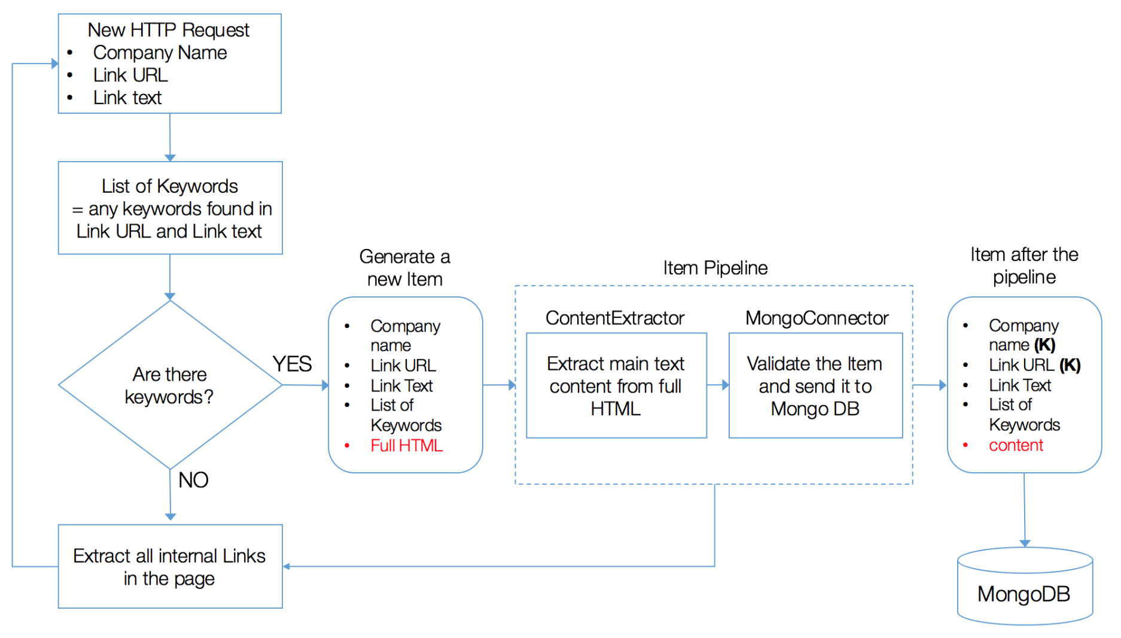
Once a page is found, it needs to be cleaned before it is saved in a MongoDB database: navigation panels, pop-up ads and advertisements, around the main textual content, tend to negatively affect the performances of NLP tasks. This is done in the Item pipeline (Figure 1).

Content extraction methods exist to analyse the HTML behind a webpage and extract what is considered to be the important text.

The Python Dragnet library is used for this, after comparing and benchmarking 4 methods against a true manually extracted content (Figure 2):

- **Dragnet**: uses machine learning models [1] to extract the main article content.
- **Readability**: was originally thought as a browser add-on and gives a score to each part of the HTML page based on a series of deterministic rules
- **BeautifulSoup get_text()**: BeautifulSoup is a Python library for parsing HTML files. The get_text() method of this library returns all the visible text of an HTML document.
- **<p> Tags**: this method simply extracts all the text enclosed within tags on a web page.

Out of a sample of 100 UK largest companies, a total of 563 sustainability-related web pages were collected from 59 of those. 35 companies did not have any sustainability pages published on their websites and thus no pages were found by the scraper. The scraper could not access or did not find the sustainability content for 6 companies.
Industry type as well as size of the companies are noticed to have an effect on whether a sustainability related content was found in their websites.


New HTTP Request
- Company Name
- Link URL
- Link text

List of Keywords = any keywords found in Link URL and Link text

Are there keywords?

YES

NO

Extract all internal Links in the page

Generate a new Item
- Company name
- Link URL
- Link Text
- List of Keywords
- Full HTML

Item Pipeline

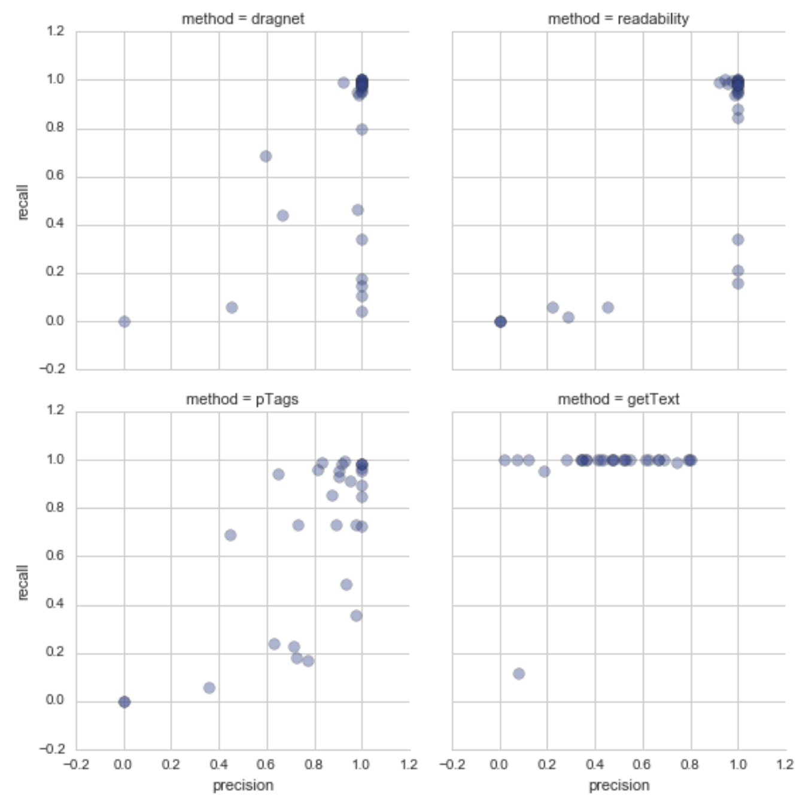
ContentExtractor: Extract main text content from full HTML

MongoConnector: Validate the Item and send it to Mongo DB

Item after the pipeline
- Company name **(K)**
- Link URL **(K)**
- Link Text
- List of Keywords
- content

MongoDB

Figure 1. High level overview of the web scraping program



	Dragnet	Readability	<p> Tags	get_text()
Precision	0.92	0.82	0.80	0.46
Recall	0.73	0.71	0.68	0.97
F1	0.76	0.73	0.71	0.59

Figure 2. Precision vs. Recall of four Content Extraction methods

TOPIC MODELLING

Latent Dirichlet Allocation (LDA) [2] is used to identify topics on the text extracted from scraped web pages. Topics allow us to understand on what areas of sustainability companies are focusing their action:

Given a collection of documents, LDA assigns to each topic a distribution over the words of the entire corpus (topic-words distributions) and to each document a distribution over topics (document-topic distributions) in an entirely unsupervised way.

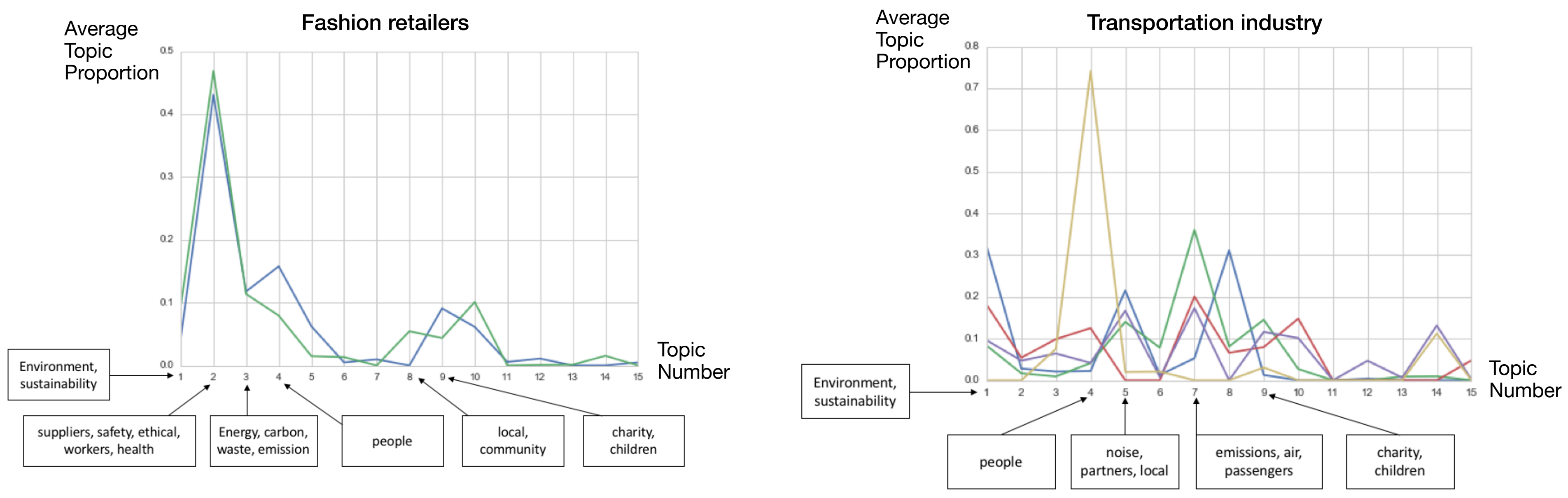


Figure 4. Transportation and fashion retail companies topic average proportions

Topic-word distributions

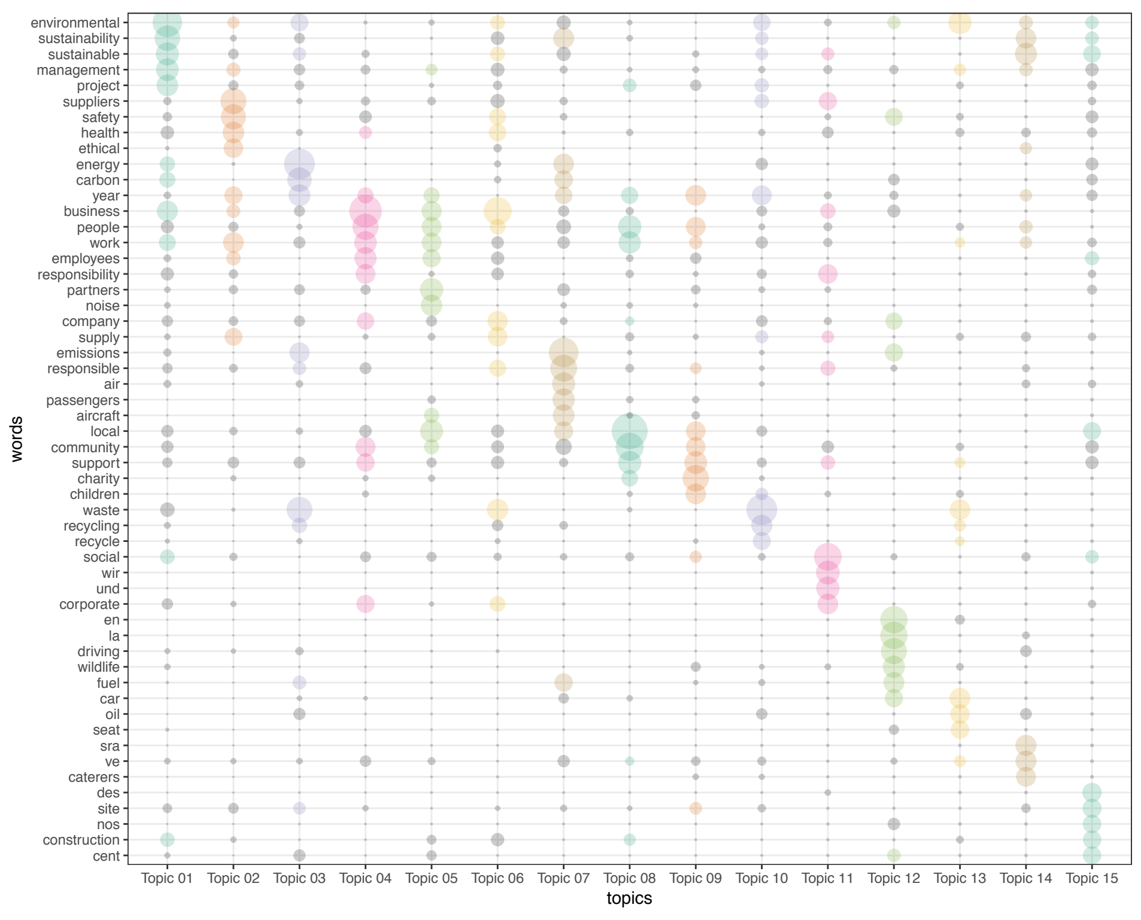


Figure 3. Termite plot

The topics are here presented in a tabular layout [3] to promote comparison of words both within and across the latent topics. For each topic, the 5 most relevant [4] words are listed on the left, based on how much discriminatory the terms are for the specific topic.

CONCLUSIONS

A gap was identified in producing data for indicator 12.6 for the SDG's in the UK and this paper discusses the development of a prototype to fill this gap. The results show that it is possible to discern the number of companies publishing sustainability information via scraping of their websites. LDA was then used to determine topics in the web pages, which shows the subject of sustainability to be much more nuanced than what might result from a mere keyword analysis.

[1] M. E. Peters and D. Lecocq, Content Extraction Using Diverse Feature Sets
[2] D. M. Blei, A. Y. Ng and M. I. Jordan, 2003, Latent Dirichlet Allocation, Journal of machine Learning research, 3, 993-1022
[3] J. Chuang, C. D. Manning and J. Heer, 2012, Termite: Visualization Techniques for Assessing Textual Topic Models, Advanced Visual Interfaces, 74–77
[4] C. Sievert and K. E. Shirley, 2014, LDAvis: A method for visualizing and interpreting topics, Proceedings of the workshop on interactive language learning, visualization, and interfaces, 63-70