Improving readability and searchability of documents provided by the dutch government under the WOB

Exploratory data analysis

Author: Justin Bon
Information Studies - Data Science
University of Amsterdam
The Netherlands
justin.bon@student.uva.nl

Supervisor: Maarten Marx
University of Amsterdam
The Netherlands
M.J.Marx@uva.nl

KEYWORDS

Named Entity Recognition, Meta Data Extraction, Co-occurrence Network

GitHub

https://github.com/JustinBon/thesis/

1 INTRODUCTION

The goal of this thesis is to create a tool that will greatly improve the machine-readability of Freedom of Information act (Wob) documents. This will be done with off-the-shelf tools so the only data that will be used in the project are the documents themselves for testing. The documents that will be used in creating this tool come from 119 separate Wob request (in a Wob request a civilian request documents from the government about a topic and the government giving those documents they deem relevant) about the governments handling of the coronavirus pandemic. Because the content of the data is not important to the project, just that it comes from a Wob request, any other set of Wob documents of sufficient size could also be used.

This thesis is part of a bigger project done in cooperation with other parties that consists of three main parts. First is the text extraction from the documents. Second is splitting the documents. These two parts will be done by other parties. The third part, and the focus of this thesis is the knowledge extraction. This means that the data that will be used for this thesis is the result of the work done by the other parties and that the data will not be available until that work is done. However, knowledge extraction can still be done by using the documents that do not need special text extraction or cleaning. The analysis of the data will be done using python [1], and the python package pandas [2].

2 OVERVIEW

As stated in the introduction, the data consists of documents from 119 Wob request about the coronavirus pandemic. The 119 requests themselves consist of a total of 368 documents. The contents of these documents are varied but generally fall into one of two categories: decisions and appendices. The decision documents are the decisions of the relevant government ministries about whether to release the requested documents. The decision also states which documents will fall within the bounds of the request, reasoning and motivation for why some documents don't fall within the bound of the request, and motivation for limited censorship in the released documents. The censorship is done for the sake of privacy, so names, email addresses, and personal views are subject to censoring. There is one decision document per Wob request. The appendices are the actual documents that the government released. Usually this is just one or multiple PDF files that consists of multiple smaller documents. These smaller documents can again be categorized:

- Information presentations of different ministries or companies
- Official government documents
- Reports
- lists of emails or other messages send or received by government officials or other smaller documents like memo's

The appendices are a compilation of some or all of these types of documents. Besides the decision and appendix, every Wob request also comes with an inventory list. An inventory list states for every document if it will be made public, partly public or not public at all. These list will not be used for this thesis.

There are at total of 367 documents from 119 Wob requests. This excludes the 119 inventory lists. The files have a wide range in sizes. Most of the files though are less than 1 MB. The average file size is 21 MB with a standard deviation of 103 MB. The standard deviation is so high because there are some very large outliers. The largest file for example is 1241 MB. There are also 28 files that have a size of 0. These files are corrupted and cannot be opened. All documents combined are 7.8 GB.

3 ANALYSIS

Currently, all the data is in the form of PDF files. As stated above, these PDF's can contain a lot of other, smaller files. In a perfect world, all of these smaller files would have the original documents so that the text can be easily extracted. However, this is not the case for most of the PDF's. A lot of the documents are scans of paper documents or screenshots of digital ones. When this happens, text cannot be extracted in the normal way and the documents are not machine readable. Some of the PDF's contain a combination of text that can be extracted and text that can not be extracted. To test how much of the text is readable, every PDF documents was put through the PyPDF2 pdf file reader to extract all possible text. Then, the number of pages, text, and characters was calculated, the results of which are can be seen in figure 1.

Figure 1 shows that the vast majority of the documents have almost no machine readable text. 27% of all documents do not contain a single character that a computer can read. Even more documents, 39%, do not contain a single word that the computer can read. Because of this, the figures are very unbalanced and hard

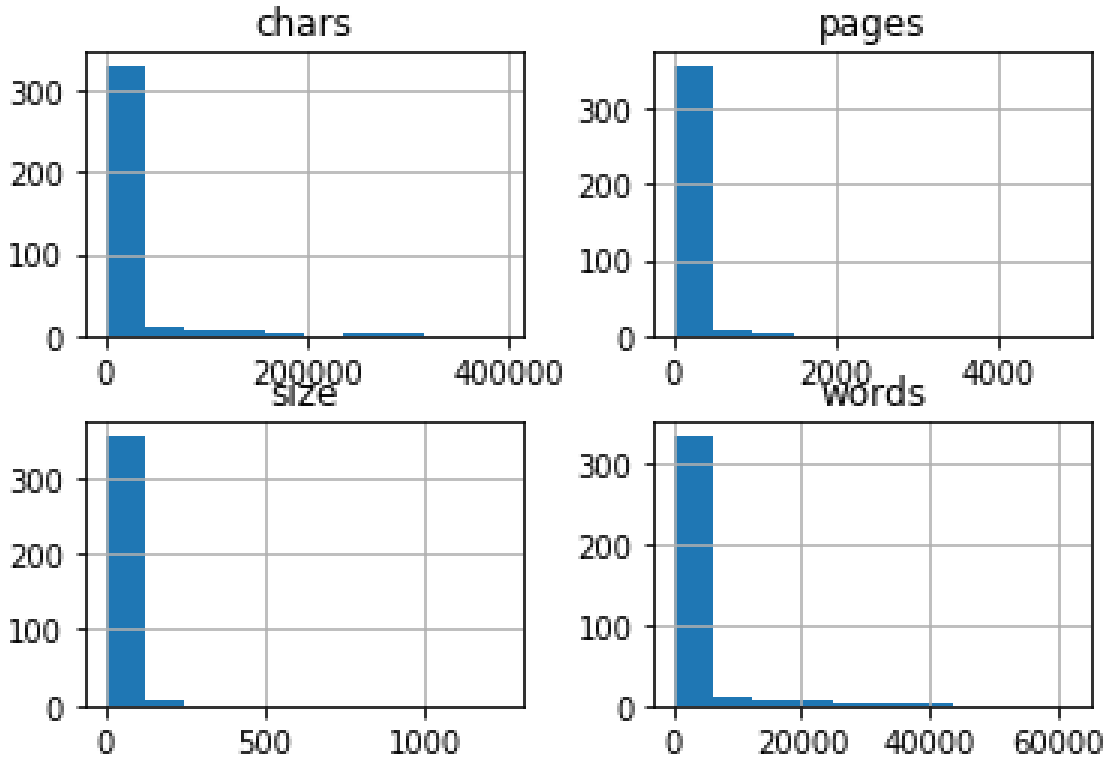

Figure 1: Histogram for the number of pages, words, and characters per file as well as the file size

	File size	nPages	nWords	nChars
Count	367	367	367	367
Mean	21.8	88.76	2272.04	18126.71
std	103.11	320.73	6804.01	51009.42
Min	0	0	0	0
25%	0.18	3	0	0
50%	2.39	12	17	568
75%	12.23	72	958	9754.5
Max	1241.34	4910	62174	394633

Table 1: Description of data

to read, so a more detailed overview of this data can be found in table 1.

Table 1 and figure 1 show the current state of the data. However, this is not the state the data will be in when it is used for this thesis. As mentioned above, this thesis is part of a bigger project and other parties will be working on extracting this text. After that is done, the data will be clean text on which to preform knowledge extraction.

Looking through the data manually (because it is not machine readable) shows a number of interesting categories that can be extracted.

- Numbers
- Dates
- Event
- Locations
- Languages
- Articles of law
- Money
- Organizations
- Percentages
- Persons
- Products
- Time
- Work of art
- government agencies

REFERENCES

[1] Van Rossum, G., and Drake Jr, F. L. *Python reference manual.* Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[2] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (2010), Stéfan van der Walt and Jarrod Millman, Eds., pp. 56 – 61.