Visualization of website publish date frequency using htmldate python package

Raufir Ahmed Shanto

August 6, 2020

1 Introduction

Metadata extraction is part of data mining and knowledge extraction. Being able to better qualify content allows for insights based on descriptive or typological information (e.g., content type, authors, categories), better bandwidth control (e.g., by knowing when webpages have been updated), or optimization of indexing (e.g., caches, language-based heuristics). It is useful for applications including database management, business intelligence, or data visualization. This particular effort is part of a methodological approach to derive information from web documents in order to build text databases for research, chiefly linguistics and natural language processing. Dates are critical components since they are relevant both from a philological standpoint and in the context of information technology.

In this project, I have used this htmldate [1] package to extract publication date for most popular 500 websites. Then I have written a python program which creates two histogram plots.

2 Dataset

I have collected the dataset from https://moz.com/top-500/download/?table=top500Domains which is a CSV file and saved as top500Domains.CSV. Upon opening the dataset, I have got to know that this dataset has following variable:

[1] https://github.com/adbar/htmldate

Variable	Description and units
`Rank`	Rank of the website according to the DA ranking
`Root Domain`	Domain Name
`Linking Root Domains`	The number of other sites that link to that page
`Domain Authority`	Search Engine Ranking

Table 1: Variables in Moz data

My point of interest in the dataset was Domain Name column. The prefix "https://" was required to call the find_date() function. Therefore, I have written a program named web_publish_date.py which takes the dataset as input, sanitizes it, finds publish date of the domain's and puts it's it into a new column in the dataframe.

To quantify whether SST for a particular area is indeed abnormally warm, we will use the Marine Heatwave (MHW) definition of Hobday et al. (manuscript submitted to Progress in Oceanography). An algorithm for detecting a MHW has been implemented in the Python package 'marineHeatWave'.

3 Results

I have visualized the publish date of domains in two plots. One of them is publish date vs year timeline another is publish date vs month timeline. The results were generated using matplotlib library of python.

4 Conclusions

From the results we can observe that, most of the popular websites are published or updated in year 2020. On a different perspective we can see that, most of the websites are published or updated in January or August.

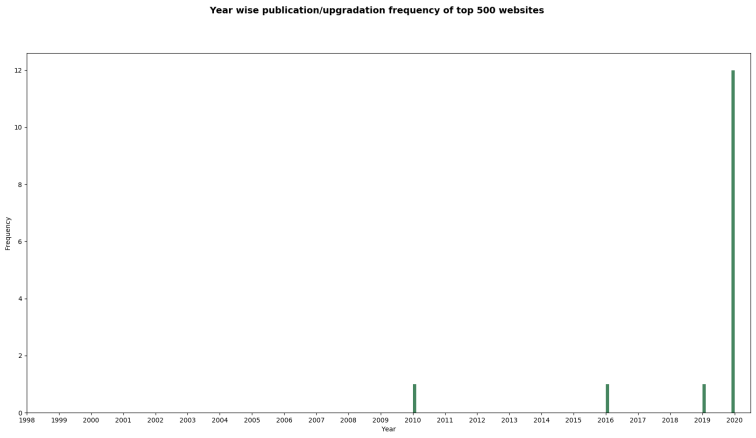

Figure 1: Website publish/upgrade date in yearly timeline

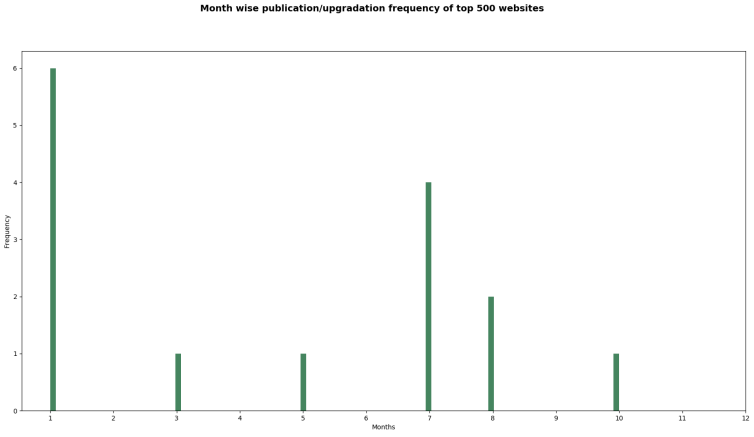

Figure 2: Website publish/upgrade date in monthly timeline