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 Methods

2.1 Dataset

There are many ways to measure SST include from instruments deployed from boat, remote observations from satellites, and moored buoys. In this

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

report we will look at marine buoys are part of the Environment Canada Meteorological Service of Canada (MSC) buoy network. This data is available for download from http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/waves-vagues/data-donnees/index-eng.asp as CSV files. The format of the CSV files is documented here http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/waves-vagues/formats-eng.html. In particular, depending on the model of the marine buoy, several data field are available as shown in Table. Measuring sea surface temperature is not straightforward since there are many definition of precisely where the temperature is being measured. For this report, we will focus on `SSTP` as the variable of interest.

Variable	Description and units
`WDIR`	Direction from which the wind is blowing (° True)
`WSPD`	Horizontal wind speed (m/s)
`WSS$`	Horizontal scalar wind speed (m/s)
`GSPD`	Gust wind speed (m/s)
`ATMS`	Atmospheric pressure at sea level (mbar)
`DRYT`	Dry bulb temperature (° C)
`SSTP`	Sea surface temperature (° C)
`SLEV`	Observed sea level
`SST1`	Average sea temperature from the non-synoptic part of WRIPS buoy data (° C)
`HAT$`	Water temperature from high accuracy temperature sensor (° C)

Table 1: Variables in marine buoy data

Specifically we look at the data from Station 44255 - NE Burgeo Bank. This marine buoy is owned and maintained by Environment and Climate Change Canada. It is 6-meter NOMAD buoy located at 47.270 N 57.340 W. There is data available for this buoy from 1998 until 2017.

2.2 Marine Heat Waves

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

As shown in Figure, the SST has a typical annual cycle between about 0° C and 20° C. There is some missing data between the years 2002 and 2004. In 2005, 2006, and 2008 there is increased noise in the data suggesting that something is wrong with the data. It would be strange for the SST to be considerably below zero. More significantly, the data from early 2010 is very suspicious as it suggests that the temperature is getting close to 80° C. Clearly we need to clean this dataset before continuing to analyze it.

This dataset from the Marine Environmental Data Section (MEDS) of the Department of Fisheries and Oceans (DFO) has already gone a quality control (QC) process. The results of QC are given by a numerical code as describe in Table.

Code	Label	Description
0	Blank	No quality control (QC) has been performed
1	Good	QC has been performed: record appears correct
3	Doubtful	QC has been performed: record appears doubtful
4	Erroneous	QC has been performed: record appears erroneous
5	Changes	The record has been changed as a result of QC
6	Acceptable	QC has been performed: record seems inconsistent with other records
7	Off Position	There is a problem with the buoy position or mooring. Data may still be useful.

Table 2: Quality control flags used in the marine buoy dataset

4 Conclusions