htmldate

Abdulqudus Olayinka Awesu

2020 August 06

Chapter 1

Introduction

Nowadays, we have lots of data moving around the globe, making data extraction clumsy, but with the inclusion of dates, web content is analyzed easily. Having an adequate and updated report is crucial in the area of research and publication, which can minimize the issues caused by the information explosion. When researchers have high levels of expertise, social interactions provide them with more precious insights than reading journal papers, and they value discussions with peers for keeping up to date. htmldate is a python package that scans throgh the web for original and updated web articles and blogs date. htmldate make use of module call finddate to scan through for dates on web pages

1.1 What is the problem?

We would like to know how often information are been updated on *MIT-News* concerning machine learning article or blogs, for better clarification htmldate was used.

1.2 Data source

we have used the *MIT-News* data for a clarification of our trend on machine learning

`http://news.mit.edu/topic/machine-learning?page=1`

1.3 Description of workflow

- we Imported htmldate modules, numpy, pandas, matplotlib.pyplot, ipywidget, interact, csv,
- We used ipywidget to beautify the notebook
- The list was saved into a csv file.
- we used widget to design the drop down list.
- we used widget and wrote a function to select each web link to display the publication date
- We used matplotlib to plot the trends of how often articles are posted.
- Used LaTex to create report PDF.
- We Created a Makefile for the automation of this project.

Chapter 2

Graphs

2.1 matplotlib Garaph

PLOT SHOWING PUBLICATION DATES OF DIFFERENT WEB ARTICLES

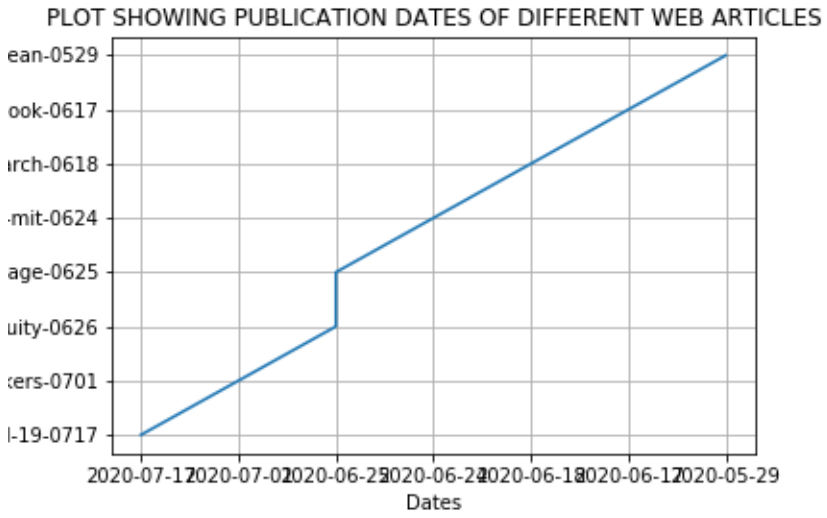

Figure 2.1: Publication dates of different web articles

Chapter 3

htmldate Reports

htmldate has not been able to perfect the extraction of publication dates from web links without a guessing. The table below shows the documentation of Python packages similar to htmldate

3.1 Analysis of htmldate

Python Package	Precision	Recall	Accuracy	F-Score	Time
newspaper 0.2.8	0.888	0.407	0.387	0.558	81.6
goose3 3.1.6	0.887	0.441	0.418	0.589	15.5
date_guesser 2.1.4	0.809	0.553	0.489	0.657	40.0
news-please 1.5.3	0.823	0.660	0.578	0.732	69.6
articleDateExtractor 0.20	0.817	0.635	0.556	0.714	6.8
htmldate 0.7.0 *(fast)*	**0.903**	0.907	0.827	0.905	**2.4**
htmldate[all] 0.7.0 *(extensive)*	0.889	**1.000**	**0.889**	**0.941**	3.8

Figure 3.1: Result of htmldate in comparison with its alternative

@articleBarbaresi2020, DOI = 10.21105/joss.02439, url = `https://doi.org/10.21105/joss.02439`, year = 2020, publisher = The Open Journal, volume = 5, number = 51, pages = 2439, author = Adrien Barbaresi, title = htmldate: A Python package to extract publication dates from web pages, journal = Journal of Open Source Software