

US State Tourism

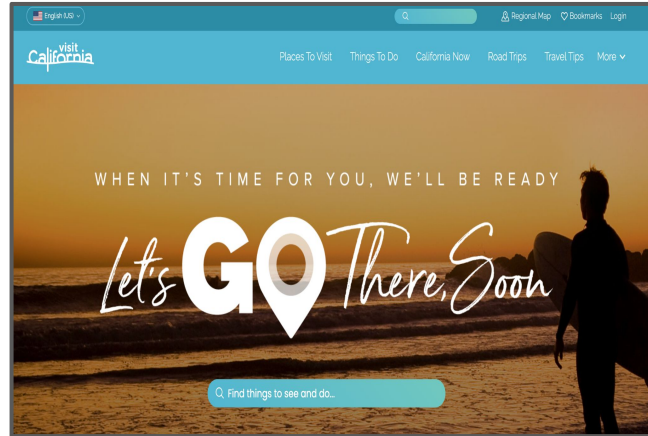
By Ashutosh Bhattarai, Shane Grishaw, Abhinav Verelly, David Gruhn

Outline

- Introduction
- ER Diagram
- Data Extraction
- Data Visualization
- Lessons Learned
- Future Work
- Acknowledgements
- References

Introduction

- Destination Management Organization
 - Funds future tourist attractions
 - Provides understanding of tourist habits



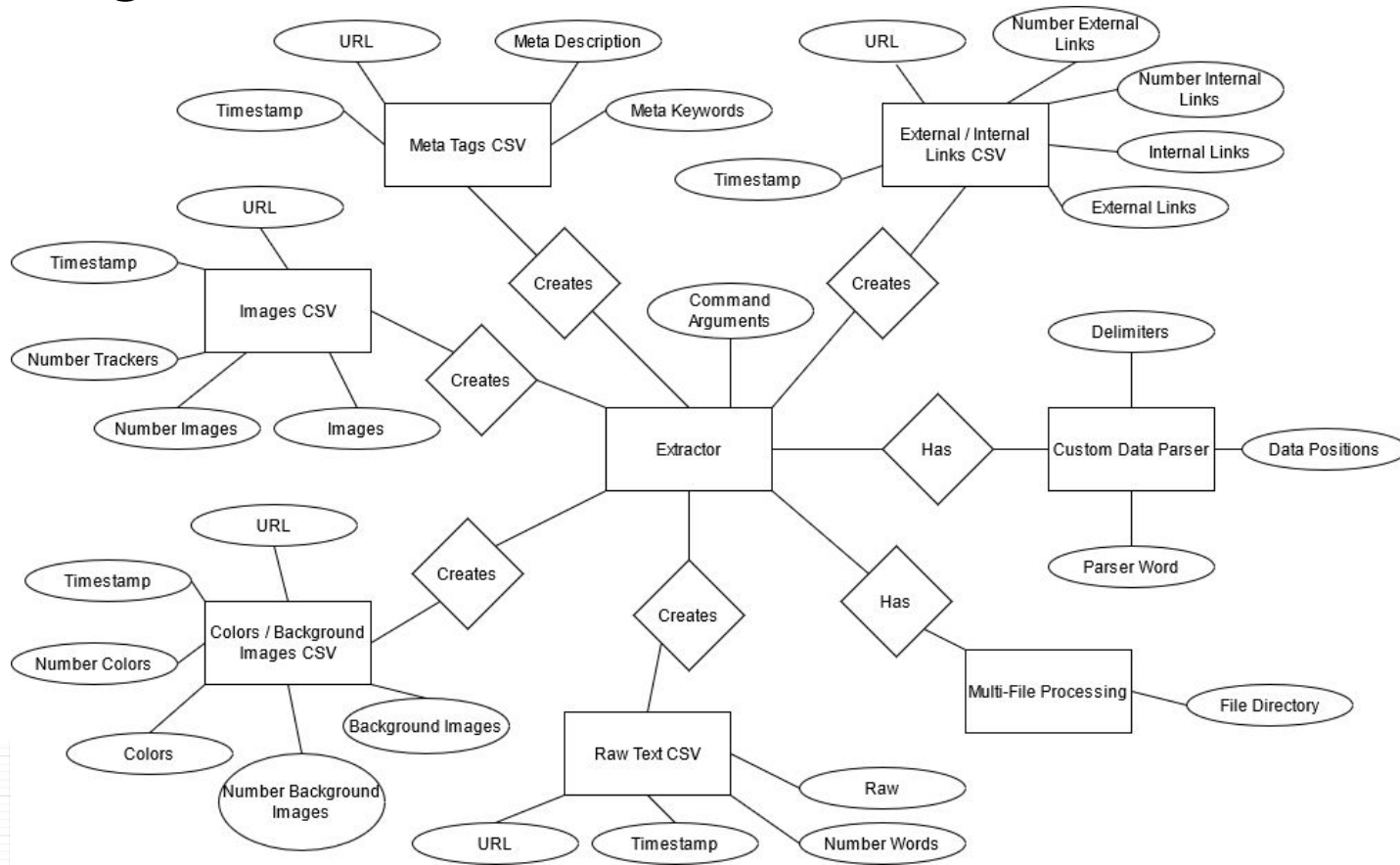
Project Goals

- Parse through snapshots of the states of California, Colorado and Virginia that are stored by the Internet Archives
- Perform data extraction on desired data
- Visualize the data in an easy to read format



BeautifulSoup

ER Diagram



Data Extraction

Extraction Processes Completed:

- External and Internal Links
- Meta Tags
- Images and Trackers
- Colors and Background Images
- Raw Text and Word Count
- Log File
- *Configuration Arguments Created*
- *Multi-CSV Parsing*
- *Multi-File Processing*
- *Custom Data Parser*

Data Extraction - Images

Timestamp	URL	Number_Trackers	Number_Images	Images
20121117	http://www.colorado.com:80/things-to-do?width=640&height=500&inline=true	1	6	mock.jpg, mock_two.jpg, alsomock.jpg, moremock.jpg, evenmore.jpg, finally.jpg
20130117	http://www.colorado.com:80/things-to-do?width=640&height=500&inline=true	1	6	mock.jpg, mock_two.jpg, alsomock.jpg, moremock.jpg, evenmore.jpg, finally.jpg
20130319	http://www.colorado.com:80/things-to-do?width=640&height=500&inline=true	1	6	mock.jpg, mock_two.jpg, alsomock.jpg, moremock.jpg, evenmore.jpg, finally.jpg
20130519	http://www.colorado.com:80/things-to-do?width=640&height=500&inline=true	2	6	mock.jpg, mock_two.jpg, alsomock.jpg, moremock.jpg, evenmore.jpg, finally.jpg
20120613	http://www.colorado.com:80/things-to-do?listing=non-listing	1	12	mock.jpg, mock_two.jpg, alsomock.jpg, moremock.jpg, evenmore.jpg, finally.jpg, mock.jpg, mock_two.j
20141116	http://www.colorado.com:80/things-to-do?listing=non-listing	4	15	mock.jpg, mock_two.jpg, alsomock.jpg, moremock.jpg, evenmore.jpg, finally.jpg, mock.jpg, mock_two.j
20150116	http://www.colorado.com:80/things-to-do?listing=non-listing	4	3	mock.jpg, mock_two.jpg, alsomock.jpg
20150215	http://www.colorado.com:80/things-to-do?listing=non-listing	4	6	mock.jpg, mock_two.jpg, alsomock.jpg, moremock.jpg, evenmore.jpg, finally.jpg

Note: Most CSVs will have a counter before lists to allow for easier parsing of the list

Data Extraction - Raw Text

	A	B	C	D
1	Timestamp	URL	Number_Words	Raw
2	20110207	http://www.colorado.com:80/StateParks.aspx?	4	This is Mock Data.
3	20111001	http://www.colorado.com:80/StateParks.aspx	435	Mock Data, Mock Data, Mock Data,
4	20111201	http://www.colorado.com:80/StateParks.aspx	32	Mock Data, Mock Data, Mock Data,
5	20120128	http://www.colorado.com:80/StateParks.aspx	345	Mock Data, Mock Data, Mock Data,
6	20100301	http://www.colorado.com:80/StateParks.aspx?page=2&rad=0	34	Mock Data, Mock Data, Mock Data,
7	20100302	http://www.colorado.com:80/StateParks.aspx?page=3&rad=0	888	Mock Data, Mock Data, Mock Data,
8	20100301	http://www.colorado.com:80/StateParks.aspx?page=4&rad=0	965	Mock Data, Mock Data, Mock Data,

Data Extraction - Meta Tags

	A	B	C	D
1	Timestamp	URL	Meta Description	Meta Keywords
2	20110207	http://www.colorado.com		
3	20111001	http://www.colorado.com	Colorado is cool	land, place state, mock, parks
4	20111201	http://www.colorado.com	Colorado is cool	land, place state, mock, parks
5	20120128	http://www.colorado.com	Colorado is cool	mock data, mock data, mock data, mock data, mock data
6	20100301	http://www.colorado.com	Colorado is cool	land, place state, mock, parks
7	20100302	http://www.colorado.com		pizza, coats, jackets, shoes
8	20100301	http://www.colorado.com	Mock Data	land, place state, mock, parks

Data Extraction - External/Internal Links

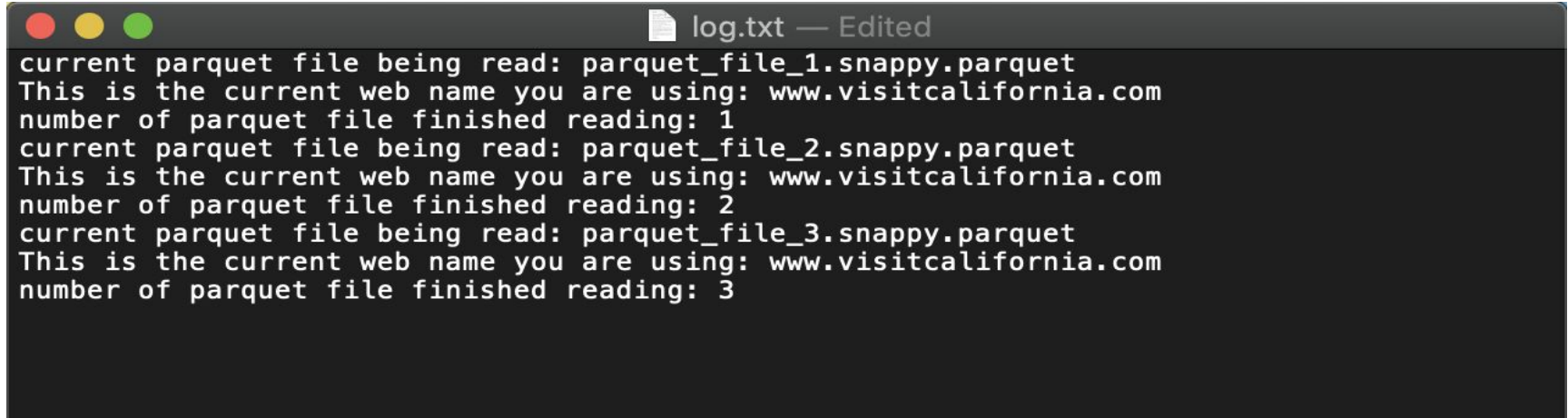
A	B	C	D	E	F
Timestamp	URL	Number_External	ExternalLinks	Number_Internal	InternalLinks
20171222	http://www.visitcalifornia.com/attraction/	2	http://www.fakewebsite.com	1	http://www.visitcalifornia.com
20181227	http://www.visitcalifornia.com/attraction/	1	http://www.fakewebsite.com	0	
20160513	http://www.visitcalifornia.com/attraction/	0		0	
20181127	http://www.visitcalifornia.com/attraction/	3	http://www.fakewebsite.com, http://www.fake	1	http://www.visitcalifornia.com

Data Extraction - Colors

	A	B	C	D	E	F
1	Timestamp	URL	Number_Colors	Colors	Number_Background_Images	Background_Images
2	20121117	http://www.colorado.com:80/things-to-do?width=640&height=500&inline=true	1	#000000	0	
3	20130117	http://www.colorado.com:80/things-to-do?width=640&height=500&inline=true	0		0	
4	20130319	http://www.colorado.com:80/things-to-do?width=640&height=500&inline=true	2	#efe8d8, #ffffff	1	url(images/div_h_main.jpg)
5	20130519	http://www.colorado.com:80/things-to-do?width=640&height=500&inline=true	0		0	
6	20120613	http://www.colorado.com:80/things-to-do?listing=non-listing	0		0	

Note: Colors in the list can appear as plain text too i.e. white, black, blue

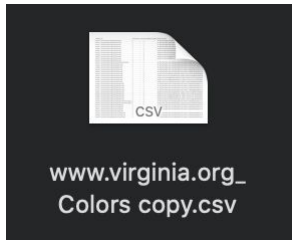
Data Extraction - Log File



```
log.txt — Edited
current parquet file being read: parquet_file_1.snappy.parquet
This is the current web name you are using: www.visitcalifornia.com
number of parquet file finished reading: 1
current parquet file being read: parquet_file_2.snappy.parquet
This is the current web name you are using: www.visitcalifornia.com
number of parquet file finished reading: 2
current parquet file being read: parquet_file_3.snappy.parquet
This is the current web name you are using: www.visitcalifornia.com
number of parquet file finished reading: 3
```

Data Visualization: Extraction for data

- Create stacked bar charts for each of the three states.
- Using Matplotlib, get colors in the Hex-Color format.
- Create a nested dictionary
- Format of dictionary, {year : {#color : freq} }



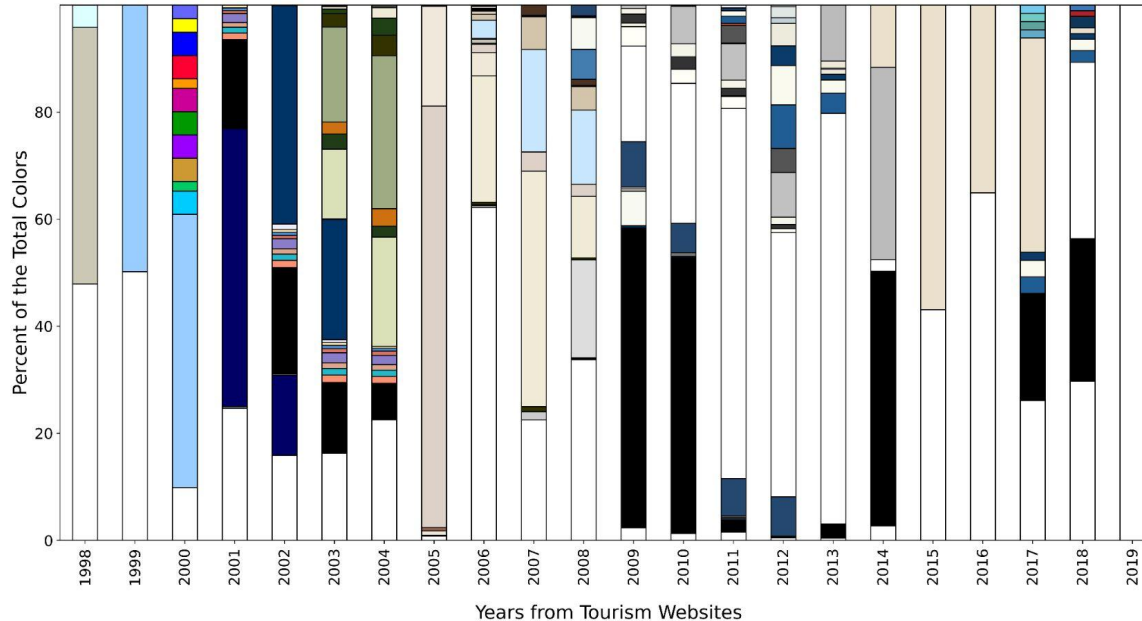
Data Visualization: JSON File

```
{
  "2001": {
    "#123123": 2
  },
  "2002": {
    "#123321": 3
  },
  "2003": {
    "#000000": 23
    "#000FFF": 1
  }
}
```



Data Visualization: Stacked Bar Chart

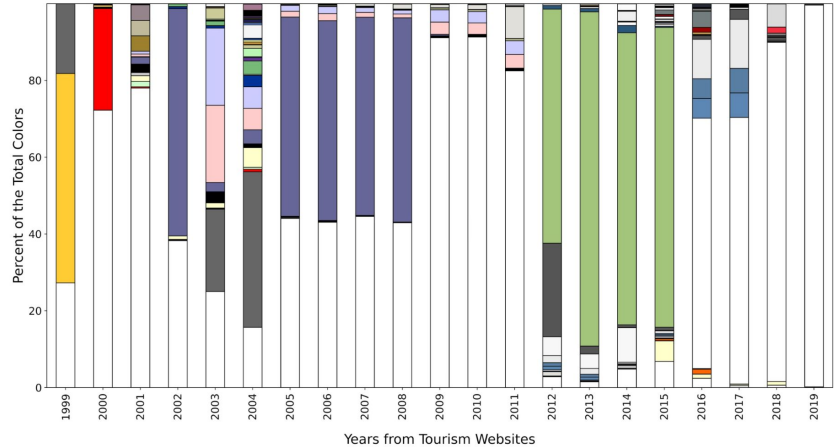
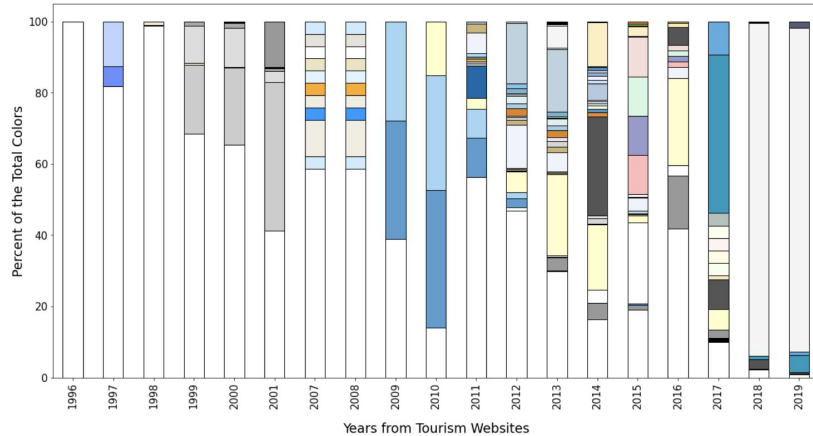
Colorado's Yearly Color Frequency



#FFFFFF	#684A88
#CBC886	#525D76
#DDFFFF	#A59787
#95C0FF	#6D5D4C
#00C0FF	#E3D4BF
#00CC66	#D4C7B6
#CC9933	#C9E6FF
#9900FF	#D3C5AB
#009900	#E2D4BE
#CC0099	#99392A
#FF9900	#5376A0
#FF0033	#688765
#0000FF	#875A6E
#FFFF00	#B96512
#6666FF	#EAE6A1
#000066	#902638
#CCCCCC	#E3E2A2
#330066	#774409
#003300	#8D0C0C
#330000	#00C6BC
#000000	#003399
#D9D272	#173777
#2389C9	#9E0309
#DEA58A	#181818
#8A7ECA	#4C3321
#DE7462	#437DB0
#3F97E3	#F8F9F1
#FFE9AB	#97928B
#EEEEFF	#E5E5CC
#003366	#DCDCDC
#669999	#254870
#663366	transparent
#DDDDDD	transparent
#EEEEEE	transparent
#DAE8B7	#333333
#203F15	#4F4E8
#213C18	#C1C1C2
#CE7110	#EBEBEB
#9FAB83	#6F67EF
#333300	#555555
#204116	#DA521E
#EFEBD6	#205D95
#C0BF97	#FBFAEE
#F0F3E2	#0B3C69
#996633	#F1F1F1
#336699	#EDEB0C
#E5DED4	#C3D1D8
#DED7C1	#E1E7E4
#EFE8D8	#E1E7E5
#936248	#F2F2F2
#DED2C8	#CEC9D8
#FEF8D8	#888888
#D2C2A6	#EBE0CC
#C1C0C1	#61ABC9
#F0F0F0	#5DA59D
#C7C1B2	#73CCC3
#B31B34	#73CCF0
#999999	#0F375A
#FAE491	#AD232B
#FFFFCC	#3E78BA

Data Visualization for Virginia and California

California's Yearly Color Frequency



Lessons Learned

- Full Online Collaboration
- Python
- Jupyter Notebook for Data Analysis
- Understanding Parquet Files
- BeautifulSoup Parsing



BeautifulSoup

Future Work



- Refactor code for performance
- Refactor code to be more abstract and easier to navigate and understand
 - Cleaning up comments / Redundant coding
- Create more error checking
- Process frequency count file extensions from Image CSV
- Process top keywords from Raw Text CSV

Acknowledgements

- Florian Zach, PhD, <florian@vt.edu>, Assistant Professor, Howard Feiertag Department of Hospitality and Tourism Management, Pamplin College of Business, Virginia Tech, Wallace Hall 362, Blacksburg VA 24061 USA
- Edward Fox, PhD, <fox@vt.edu>, CS4624 Professor, Department of Computer Science, College of Engineering, 2160G Torgersen Hall, Blacksburg VA 24061 USA
- US State Tourism Spring 2020 Team, <http://hdl.handle.net/10919/98257>
- US State Tourism Spring 2019 Team, <http://hdl.handle.net/10919/92622>
- NSF IIS-1619028, Global Event and Trend Archive Research (GETAR)
- NSF CMMI-1638207, Coordinated, Behaviorally-Aware Recovery for Transportation and Power Disruptions (CBAR-tpd)

References

- Doan, Viet, et al. “Tourism Destination Websites.” VTechWorks, Virginia Tech, 8 May 2019, <https://vtechworks.lib.vt.edu/handle/10919/92622>
- Shere, Danya, et al. “US State Tourism Websites.” VTechWorks, Virginia Tech, 11 May 2020, <https://vtechworks.lib.vt.edu/handle/10919/98257>



Questions?