WEB CRAWLER AND NLP SYSTEM PROOF-OF-CONCEPT

PUBLICATION AND SYSTEM LIMITATIONS

Nikki Fitzherbert
MA5851 Master Class 1

GITHUB REPOSITORY

NikkiSarah Add files via upload — Latest commit ff2fac4 yesterday — History

1 contributor

328 lines (268 sloc) | 13.5 KB — Raw | Blame

```python
# import necessary libraries
# if the user does not already have selenium installed, then they should visit
# 'https://www.selenium.dev/documentation/en/' for installation instructions and webdriver requirement directions
from selenium import webdriver
from selenium.webdriver.firefox.options import Options
import pandas as pd
import time

# establish webdriver and set desired options. Any driver webdriver can be used but the default is Firefox
firefox_options = Options()
firefox_options.add_argument("--incognito")  # browser operates in private/incognito mode
#firefox_options.headless = True  # crawler runs without a visible browser
driver = webdriver.Firefox(options=firefox_options)

# an extra line of code is required if the webdriver is named anything other than 'driver'
# driver = ...

# list of BBC News urls for the scraper to visit
url_list = ['https://www.bbc.com/news/world',
            'https://www.bbc.com/news/world/asia',
            'https://www.bbc.com/news/world/australia',
            'https://www.bbc.com/news/uk',
```

NikkiSarah / JCU-MDS-MA5851-A3 — Private — Unwatch 1

Code | Issues | Pull requests | Actions | Projects | Security | Insights | Settings

main — 1 branch — 0 tags — Go to file — Add file — Code

NikkiSarah Added/updated files — 6393cc2 2 hours ago — 82 commits

images	Added/updated files	2 hours ago
.gitattributes	Initial commit	26 days ago
Document 1 - Overview.pdf	Align desktop and online repository	3 hours ago
Document 1.ipynb	Add files via upload	yesterday
Document 2 - Web Crawler and EDA.i...	Added/updated files	2 hours ago
Document 2 - Web Crawler and EDA....	Added/updated files	2 hours ago

About

Repository for a web crawler/NLP project involving a preliminary investigation into the sensationalism of news by broadcasters.

Readme

Releases

No releases published
Create a new release

NikkiSarah Add files via upload — Latest commit 7c81c24 yesterday — History

1 contributor

68 lines (68 sloc) | 5.68 KB — Raw | Blame

Document 1 - Overview {-}

The phrase "if it bleeds, it leads" can allegedly be first attributed to journalist Eric Pooley as the author of a 1989 New York Times article titled "Grins, Gore and Videotape - The Trouble with Local TV News". At the time, he was angry about the volume of stories with grim and menacing storylines and the fact they were being prioritised over more thoughtful and optimistic pieces (Evaluating Conversations, n.d.). Historically, journalists and news producers aimed to report events as they actually happened, with an even balance between the depressing or otherwise negative stories and those showcasing the more positive side of life. Over time, however, it has evolved into a race to find the most sensationalist, stirring and spectacular stories to maintain ratings and profits (Serani, 2008). As result, the media has become less a source of information and more of another form of entertainment (Dmitrieva, 2017), where those working-class heroes quietly go about their life with nary a mention (Williams, 2005).

THE PROBLEM


"Oh, my God—I'm having an affair!"


If It Bleeds, It Leads
PRESS
TV
RADIO
NEWS CHANNEL
MEDIA
TV NETWORK
PRESS
MEDIA
TV
NEWS

WEB CRAWLER LIMITATIONS

- Slow
- Not entirely stable
- Unable to handle advanced webpage elements

HARVESTED DATA LIMITATIONS

- Data sparseness
- …

		headline2	summary2	category	category2	timestamp_aus	webpage	contributor_name	contributor_team	section	subsection	
0	but these c...	'They used my picture and I should've got p...	Sean Heavey recognised his photo the mom...	nan	nan	2020-09-18 09:12:36+10:00	In Pictures	Nell Mackenzie	Business reporter, BBC News	nan	nan	https
1	medical ce...	Covid: Floral mosaic installed at Birmingha...	A giant floral mosaic at a health centre has ...	nan	nan	2021-03-02 17:44:52+11:00	Entertainment & Arts	nan	nan	BBC News	Birmingham & Black Country	https
2	notos from a...	Africa's week in pictures: 25 September-1 O...	A selection of the week's best photos from a...	Photography	nan	2020-10-02 10:13:49+10:00	In Pictures	nan	nan	nan	nan	https
3	ve a huge b...	'My pet has helped me so much during the ...	"I don\'t know what I would have done with...	Animals	Animals	2021-03-28 12:07:24+11:00	Science & Environment	Rachel Stonehouse	Newsbeat Reporter	nan	nan	https
4	n a hit as in...	Bitcoin: Elon Musk loses world's richest title ...	Tesla boss Elon Musk has lost his title as the ...	Elon Musk	Elon Musk	2021-02-24 05:55:34+11:00	Technology	nan	nan	nan	nan	https
5	assports" to ...	Scotland's papers: Covid 'passport' plans an...	nan	nan	nan	2021-03-26 18:52:54+11:00	UK	nan	nan	nan	nan	https
6	0th series b...	I'm A Celebrity: Ant and Dec 'happy' to retur...	I\'m A Celebrity stars Ant McPartlin and Dec...	Ant and Dec	Ant and Dec	2021-02-05 03:53:09+11:00	Entertainment & Arts	nan	nan	nan	nan	https
7	n was refus...	Night rider: 21 years sleeping on a London b...	For more than two decades after his asylum ...	Homelessness in the UK	Homelessness in the UK	2020-01-12 11:09:40+11:00	Long Reads	Venetia Menzies	Journalist and photographer	nan	nan	https
8	ll has beco...	The Rest and Be Thankful: Scotland's infamo...	In recent years, the Rest and Be Thankful in ...	nan	nan	2020-08-28 09:21:28+10:00	In Pictures	nan	nan	BBC News	Scotland	https
9	llegal protes...	Coronavirus: Melbourne police arrest 74 anti...	More than 70 people have been arrested in ...	Coronavirus pandemic	Coronavirus pandemic	2020-09-14 03:07:57+10:00	Australia	nan	nan	nan	nan	https
10	dlands NHS...	Coronavirus caught in Staffordshire hospital...	More than 100 patients at hospitals in Staffo...	Coronavirus pandemic	Coronavirus pandemic	2021-03-11 04:07:08+11:00	Health	nan	nan	West Midlands Express & Star	nan	https
11	p the extra...	Huawei takes HSBC to court as it tries to sto...	The Chinese telecoms giant Huawei is takin...	China	China	2021-02-12 11:13:12+11:00	Technology	Gordon Corera	Security correspondent	nan	nan	https
12	nentary on d...	Listen: Rain stalls India reply to Australia's 36...	Rain wipes out evening session\nLyon takes...	nan	nan	2021-01-16 10:30:00+11:00	Australia	nan	nan	nan	nan	https
13	rs for the 20...	In pictures: Sony World Photography Awards	The winners of the professional category of ...	Sony	Sony	2020-06-09 09:19:41+10:00	In Pictures	nan	nan	nan	nan	https
14	up on mobil...	Fun but doomed: LG's most memorable sm...	There had been chatter for years that LG wo...	Mobile phones	Mobile phones	2021-04-05 23:24:52+10:00	Business	Leo Kelion	Technology desk editor	nan	nan	https
15	ars of angui...	LGBT+ History Month: Farmer feared he 'wo...	A farmer from Cornwall who hid his sexualit...	Penzance	Penzance	2021-02-24 19:23:51+11:00	Science & Environment	Hannah Stacey	BBC Radio Cornwall	Falmouth Packet	Pirate FM	https
16	files to dat...	Line of Duty: All you need to know to get up...	Line of Duty, BBC One's hit drama about pol...	BBC	BBC	2021-03-20 03:42:17+11:00	Entertainment & Arts	nan	nan	nan	nan	https
17	nics giant h...	Toshiba confirms $20bn takeover bid from B...	Japanese conglomerate Toshiba has receive...	Japan	Japan	2021-04-07 12:47:20+10:00	World	nan	nan	nan	nan	https
18		Fran Lebowitz: 'Being offended is part of lea...	The pandemic may have put a stop to travel ...	nan	nan	2021-02-19 19:51:36+11:00	Entertainment & Arts	nan	nan	BBC News	Entertainment & Arts	https

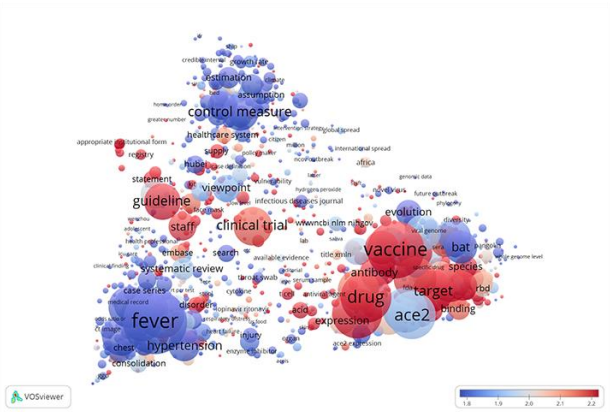
NLP TASK LIMITATIONS

SENTIMENT ANALYSIS

- Availability of a news-specific lexicon
- NRC Emotion lexicon
- Sentiment rating thresholds
- Comparison to a machine-learning approach

TOPIC MODELLING

- Hyperparameter specification
- NMF vs other models
- Performance assessment



Questions?