

“Music Twitch” Crawling Project – Text Technology

University of Stuttgart, Germany

Seminar: Text Technology

WiSe 2018/19

Date: 28/01/2019

Meghdut Sengupta, Eduardo Galuppi,

Nana Agyei-kena, Lisa Schütz

Table Of Content

1. Basic Idea of “Music Twitch”
2. Mechanism
3. Code Snippets
4. Interface
5. Problems Encountered
6. Conclusion And Future Scopes

Music Twitch – Basic Idea

Idea: Create a „music **ensemble**“

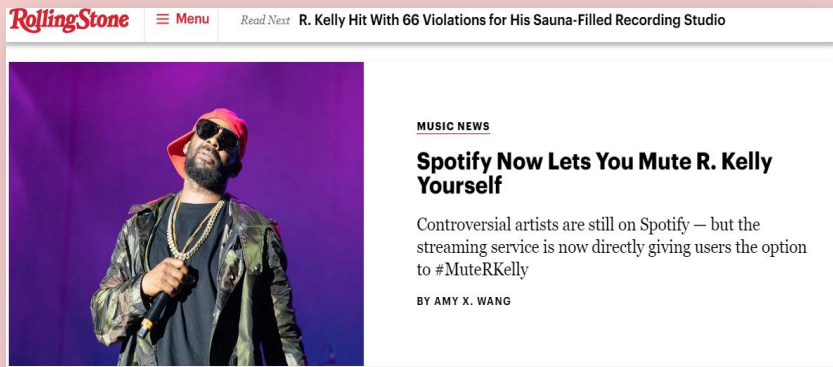
- The user can read about the latest news of their favourite artists
- Access to Spotify:
 - albums released
 - biography
 - related artists
- Information on the artist via his Wikipedia page

Music Twitch – Mechanism

RollingStone

1. Crawling www.rollingstone.com

- Latest news about music (titles, images, etc.)



In This Article: R. Kelly, Spotify

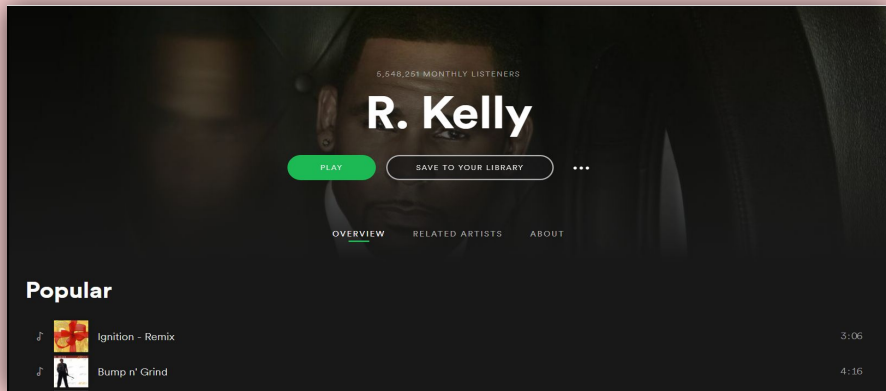
- Getting the related data based on the previous info

Music Twitch – Mechanism



2. From this data we are making additional requests to Spotify API
→ detailed information about artist

Official Spotify page of the artist



Music Twitch – Mechanism



3. The relevant Wikipedia URL of the artist


A screenshot of a web browser displaying the Wikipedia page for R. Kelly. The browser's address bar shows the URL "https://en.wikipedia.org/wiki/R._Kelly". The page features the Wikipedia logo on the left, a navigation bar with "Article" and "Talk" tabs, and a search bar. The main content area shows the title "R. Kelly" and a notice about the article's reliability, stating it is about a person involved in a current event and may be unreliable. The notice includes a link to "improve this article" and a link to the "talk page".

← → ↻ 🔒 https://en.wikipedia.org/wiki/R._Kelly

Not logged in Talk Contributions Create account

Article Talk

Read View source View history Search Wikipedia


 WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

R. Kelly

From Wikipedia, the free encyclopedia

For the album, see R. Kelly (album).



This article is about a person involved in a [current event](#). Information may change rapidly as the event progresses, and initial news reports may be **unreliable**. The [last updates](#) to this article [may not reflect](#) the most current information. Please feel free to [improve this article](#) or discuss changes on the [talk page](#). (January 2019) ([Learn how and when to remove this template message](#))

Code Snippet Of Our Spider



```
def parse(self, response):

    urls = response.xpath('//div[@class="l-blog_item l-blog_item--spacer-s"]/ul/li/article/a/@href').extract()

    for url in urls:
        request = scrapy.Request(url, callback=self.parse_inner)
        yield request

    # ----- Redirections to the next page -----
    next_page = response.xpath('//*[id="site_wrap"]/div[3]/div[1]/main/div[4]/div/a[2]/@href').extract()

    # ----- Double check if next_page is not empty -----
    if next_page == []:
        next_page = response.xpath('//*[id="site_wrap"]/div[3]/div[1]/main/div[4]/div/a[1]/@href').extract()

    # ----- Assign the URL to the next page to start the crawl again -----

    if next_page:
        next_href = next_page[0]
        next_page_url = next_href
        request = scrapy.Request(url=next_page_url)
        yield request
```

Code Snippet For Parsing The Inner Pages

```
# ----- Title -----  
  
title = response.xpath('//*[@id="site_wrap"]/div[2]/div/main/article/header/h1/text()').extract()[0].replace('\n', '').  
  
# ----- author of the news -----  
  
author = response.xpath('//div[@class="c-byline__author"]/a/text()').extract()[0].replace('\n', '').replace('\t', '')  
  
# ----- Artist Name -----  
  
artist_name = response.xpath('//*[@id="site_wrap"]/div[2]/div/main/article/div/footer/div[1]/p/a/text()').extract()  
  
# ----- Img -----  
  
img = response.xpath('//*[@id="site_wrap"]/div[2]/div/main/article/figure/div[1]/div/img/@data-src').extract()
```



Spotify Authentication



```
def __init__(self):  
    # ----- Authenticating the Spotify API with the required credentials -----  
    self.client_credentials_manager = SpotifyClientCredentials(client_id='b7cf2f5f0e074d1ca984a546822797cc',  
                                                                client_secret='c435b02f49304b7db74ec08ddb7df6e7')  
  
    self.spotify = spotipy.Spotify(client_credentials_manager=self.client_credentials_manager)
```

Request To The Spotify API



```
# ----- Collecting spotify data -----  
  
if artist_name:  
    name = artist_name  
    results = self.spotify.search(q='artist:' + name, type='artist')  
    items = results['artists']['items']  
  
    if len(items) > 0:  
        # ----- Link to the index page of the Artist in Spotify -----  
        spotify_link = items[0]['external_urls']['spotify']  
  
        # ----- Link to the about page of the Artist -----  
        spotify_about = spotify_link + '/about'
```

Link To Wikipedia



Generating the Wikipedia URL

```
inbuilt_wiki_uri = 'https://en.wikipedia.org/wiki/'  
  
wiki_link = inbuilt_wiki_uri + "_".join(artist_name.split(' '))
```

The Interface



Releases

**Weezer Follow
'Africa' With
Surprise Covers
LP, 'The Teal
Album'**

Weezer



**Red Hot Chili
Peppers, Miley**

Problems Encountered

- Inconsistent data

Output pane

Data Output Explain Messages History

	title character varying	artist_name character varying	spotify_link character varying	spotify_about character varying
1	See Conor Oberst, Phoebe Bridgers Perform 'Better Oblivion Community Center' Songs on 'CBS This Morning'	Conor Oberst	https://open.spotify.com/artist/227gv3uEhick1aBzTUCE6R	https://open.spotify.com/artist/227gv3uEhick1aBzTUCE6R
2	Michael Jackson Estate Slams 'Leaving Neverland': 'Tabloid Character Assassination'	Michael Jackson	https://open.spotify.com/artist/3fMbdgg4jU18AjLCKBhR5m	https://open.spotify.com/artist/3fMbdgg4jU18AjLCKBhR5m
3	Michel Legrand, Oscar-Winning Film Composer, Dead at 86	obit	https://open.spotify.com/artist/03w7vokQTSRwe0SLqdpWfK	https://open.spotify.com/artist/03w7vokQTSRwe0SLqdpWfK
4	Eddie Van Halen's 20 Greatest Solos			
5	Linda Perry: My Life in 15 Songs			
6	Song You Need to Know: Yebba, 'Evergreen'	Song You Need to Know		
7	Kanye West Sues Universal, EMI Over Record Contracts, Song Publishing	Kanye West	https://open.spotify.com/artist/SK4w6rqBFMDnAW6FQUK56x	https://open.spotify.com/artist/SK4w6rqBFMDnAW6FQUK56x
8	'Baby Shark' is Shooting Up the Charts, But No One Owns the Rights	Charts	https://open.spotify.com/artist/3JGrgRrVxkljndTVbryvDx	https://open.spotify.com/artist/3JGrgRrVxkljndTVbryvDx
9	'Te Boté' Was a Massive Hit - Now It's Spawned Imitators	Latin	https://open.spotify.com/artist/5WQwio3gAl1xuxQAPL1y3T	https://open.spotify.com/artist/5WQwio3gAl1xuxQAPL1y3T
10	See Conor Oberst, Phoebe Bridgers Perform 'Better Oblivion Community Center' Songs on 'CBS This Morning'	Conor Oberst	https://open.spotify.com/artist/227gv3uEhick1aBzTUCE6R	https://open.spotify.com/artist/227gv3uEhick1aBzTUCE6R
11	Michel Legrand, Oscar-Winning Film Composer, Dead at 86	obit	https://open.spotify.com/artist/03w7vokQTSRwe0SLqdpWfK	https://open.spotify.com/artist/03w7vokQTSRwe0SLqdpWfK
12	Michael Jackson Estate Slams 'Leaving Neverland': 'Tabloid Character Assassination'	Michael Jackson	https://open.spotify.com/artist/3fMbdgg4jU18AjLCKBhR5m	https://open.spotify.com/artist/3fMbdgg4jU18AjLCKBhR5m
13	Kanye West Sues Universal, EMI Over Record Contracts, Song Publishing	Kanye West	https://open.spotify.com/artist/SK4w6rqBFMDnAW6FQUK56x	https://open.spotify.com/artist/SK4w6rqBFMDnAW6FQUK56x
14	Song You Need to Know: Yebba, 'Evergreen'	Song You Need to Know		
15	Eddie Van Halen's 20 Greatest Solos			

OK

Unix Ln1, Col128, Ch128 272 rows. 134 msec

- Image retrieval



Conclusion And Future Scopes

- An ensemble of data based on the recent trends
- Add ons: Twitter, Reddit, ...
- Expansion to various domains
 - World Politics
 - Trends in Economic and Business sectors
 - Latest research in Science and Technology
 - A platform for further work on machine learning



Thank you for your attention!