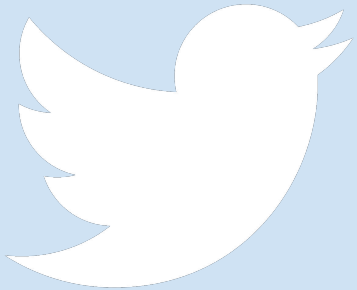
Twitter Collection

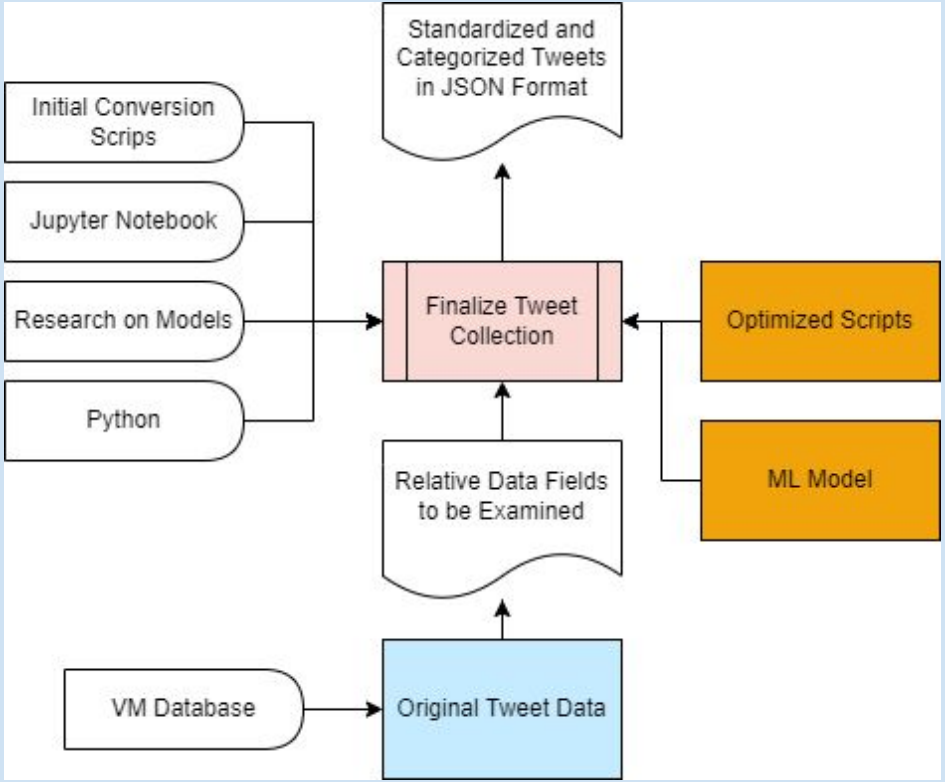


Matt Gonley, Ryan Nicholas, Griffin Knock, Nicole Fitz, Derek Bruce

Professor Edward A. Fox
CS 4624: Multimedia, Hypertext, and Information Access
Virginia Tech, Blacksburg VA 24061
5/10/2022

Outline

- Recap
- Timeline
- Deliverables
- Comparison
- Data Scale
- Work Completed
- Changes
- Optimization
- Challenges
- Acknowledgment
- References



Overview

Optimization

SQL scripts from previous semesters capstone group

Classify Data

Classify Tweets to Corresponding Collections

Convert Data

Convert Data to Finalized JSON Schema

Timeline

February

Finalize Schemas

YTK, DMI-TCAT, SFM

March

Optimize Scripts

Improve upon last semesters conversion scripts

April

Classify & Convert Data

Classify tweets to corresponding collections and run scripts

May

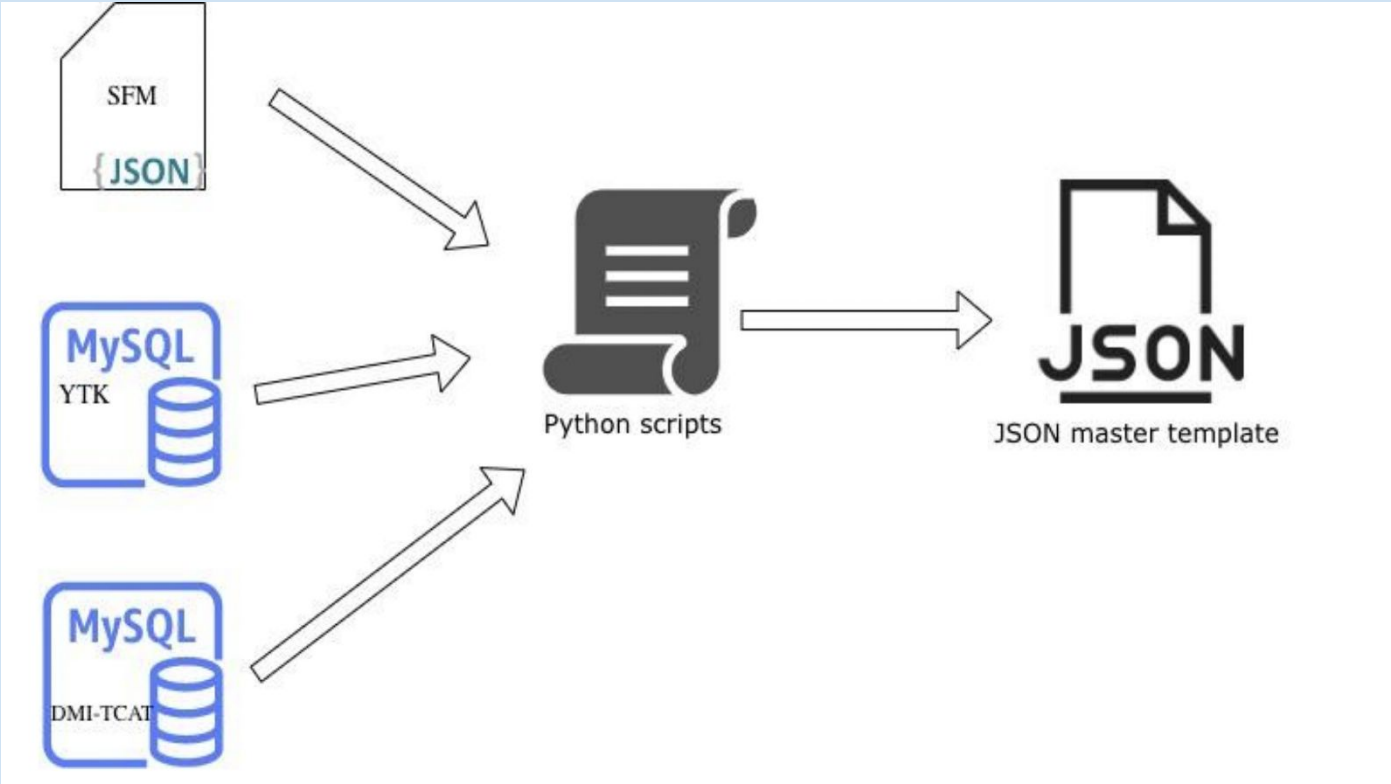
Write Documentation

Create final report and thoroughly comment code

Deliverables

- Optimized Scripts
  - Code cleaned up & documented
  - Fixed missing components
- Converted JSON data
- ML Model
- Report & Presentation

Comparison: DMI vs. YTK vs. SFM



Data Scale: DMI

Table	Num Tweets	Num Hashtags	Num URLs	Num Mentions	Num Places	Num Media	Time
paris_shooting	0	0	0	0	0	0	0:00:00
yemen_cyclone_megh	0	0	0	0	0	0	0:00:00
wild_fire	31	39	12	26	26	15	0:00:00.021124
Race_Together	423	931	219	497	497	67	0:00:00.410534
Umpqua_Community_College	10317	5567	6539	6888	6888	2514	0:00:06.343043
Japan_earthquake	19289	106992	2887	2891	2891	9673	0:00:08.658198
ashleymadison_hack	72651	124796	37750	58001	58001	21901	0:00:34.384757
Budget2015	98171	154542	38430	104784	104784	61171	0:00:49.651242
Mecca	132963	402126	36504	113195	113195	117814	0:01:07.067494
pothole	196665	78720	94062	156131	156131	57923	0:01:25.503444
Joaquin	234905	143345	75552	156938	156938	58899	0:01:37.426643
wdbj7_shooting	236260	176615	131883	209296	209296	71774	0:01:47.749376
confederate_flag	463688	174497	372647	483605	483605	225075	0:03:21.426796
Oregon	871656	531812	574755	594040	594040	310549	0:06:17.818379
El_Chapo_escape	1317720	284900	668319	980211	980211	357202	0:09:29.053893
car_crash	2686609	550537	2058403	2091385	2091385	1461125	0:20:11.323309
Climate_Change	5057264	2214727	3354582	4621623	4621623	994597	0:39:32.065757
nuclear	10248267	6246232	6812467	8224626	8224626	2307057	1:20:14.584924
MEAN:	1202604	622021	792500	989118	989118	336519	
SUM:	21646879	11196378	14265011	17804137	17804137	6057356	2:46:43.540603

Event Data Scale: YTK

	Event Name	Number of Tweets
770	Myocardial infarction	20812918.0
771	Hassan Rouhani	21562860.0
772	Tropical cyclone	22262127.0
773	Bahrain	23012550.0
774	2012HurricaneSandy	23179590.0
775	2016HurricaneMatthew	24001324.0
776	Bomb	24391505.0
777	Negar Mottahedeh	26241037.0
778	Sinkhole	26739475.0
779	Israel	26945844.0
780	Tornado	27363113.0
781	Blacksburg, Virginia	28844286.0
782	Obesity	31477896.0
783	Diabetes mellitus	33365886.0
784	Syria	34464914.0
785	Pothole	35336155.0
786	Flood	42291044.0
787	Tsunami	43950091.0
788	Foursquare	44569577.0
789	Earthquake	62302902.0

YTK Organization

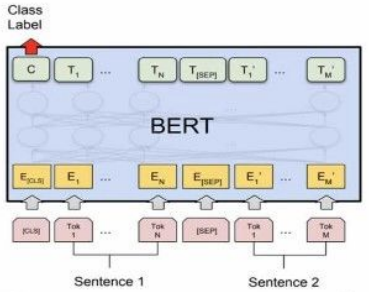
Database	ID	Source	Collection Terms	Wikipedia	Description
Collect_yTK	1	yTK	#egypt	https://en.wikipedia.org/wiki/Egyptian_revo	Originally for Egyptian r
Collect_yTK	2	yTK	#libya	https://en.wikipedia.org/wiki/Libya	" In the second Libyan C
Collect_yTK	3	yTK	#blacksburg	https://en.wikipedia.org/wiki/Blacksburg,_V	" Blacksburg High Schoo
Collect_yTK	4	yTK	#jan25	https://en.wikipedia.org/wiki/Egyptian_revo	January 25th 2011 was
Collect_yTK	5	yTK	#bahrain	https://en.wikipedia.org/wiki/Bahrain	" In December 1994, a g
Collect_yTK	6	yTK	#yemen	https://en.wikipedia.org/wiki/Yemen	" According to the 2009
Collect_yTK	7	yTK	japan earthquake	https://en.wikipedia.org/wiki/2011_T%C5%[illegible]	"This is a list of earthqu
Collect_yTK	8	yTK	#syria	https://en.wikipedia.org/wiki/Syria	" Syria is ranked last on
Collect_yTK	9	yTK	OccupyWallStreet	https://en.wikipedia.org/wiki/Occupy_Wall_	"== Origins ==The origin
Collect_yTK	10	yTK	#nrv		new river valley (blacksb
Collect_yTK	11	yTK	virginia tech	https://en.wikipedia.org/wiki/Virginia_Tech	A 23-year-old student, S

```
| z_319              |
| z_32               |
| z_320              |
| z_321              |
| z_322              |
| z_323              |
| z_324              |
| z_325              |
| z_326              |
| z_327              |
| z_327_ne2          |
| z_328              |
| z_329              |
| z_329_ne2          |
| z_33               |
| z_330              |
| z_331              |
| z_6                |
| z_7                |
| z_9                |
+--------------------+
```

Machine Learning

- **Goal**: Classify Tweets to Events
- Naive Bayes Model
- BERT + Transformer Neural Network
  - Learns contextual relations between words (or sub-words) in a text bidirectionally

```
Paleontology                                59
ClimateChange                               50
2016HurricaneMatthew                        34
Transit                                     25
Flood                                       23
                                            ..
Kenya                                        1
Afghanistan men's national volleyball team   1
ThanksKilling                                1
Shut 'Em Down (album)                        1
PowerOutage                                  1
Name: Event Name, Length: 809, dtype: int64
```



PyTorch

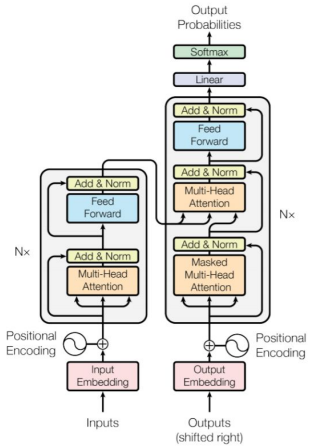


Figure 1: The Transformer - model architecture.

UJSON

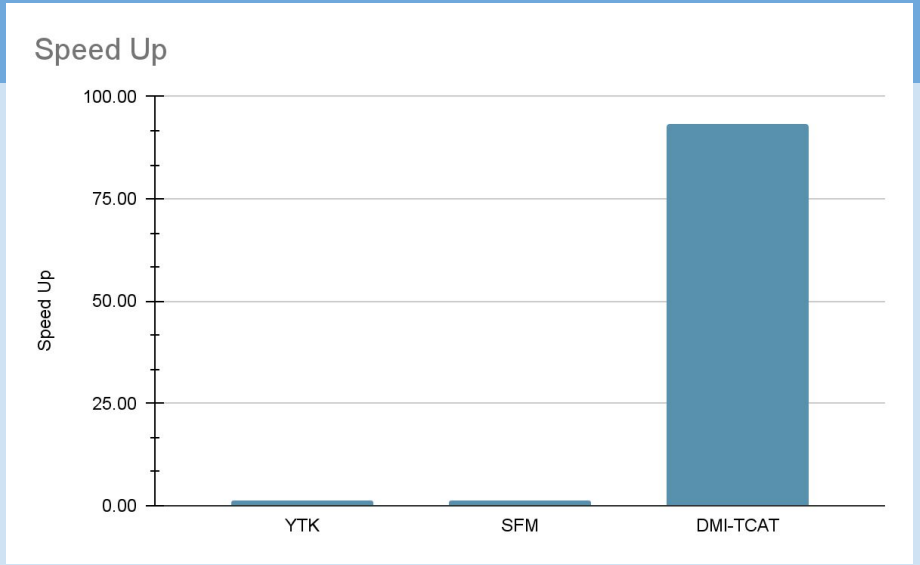
- Updated JSON API
- Tested Conversion + Writing on 1 Million Tweets
  - Finished in 14 seconds
  - Standard JSON took 112 seconds

	ujson	nujson	orjson	simplejson	json
Array with 256 doubles					
encode	22,082	4,282	76,975	5,328	5,436
decode	24,127	34,349	29,059	14,174	13,822
Array with 256 UTF-8 strings					
encode	3,557	2,528	24,300	3,061	2,068
decode	2,030	2,490	931	406	358
Array with 256 strings					
encode	39,041	31,769	76,403	16,615	16,910
decode	25,185	24,287	34,437	32,388	27,999
Medium complex object					
encode	10,382	11,427	32,995	3,959	5,275
decode	9,785	9,796	11,515	5,898	7,200

Changes

- Commented Code
- Modified for Scale
  - UJSON vs. JSON Python package
- Fixed issues in schema
- Added Hashtags for YTK
- Corrected YTK Collection

Optimization



Original Scripts:

Optimized Scripts:

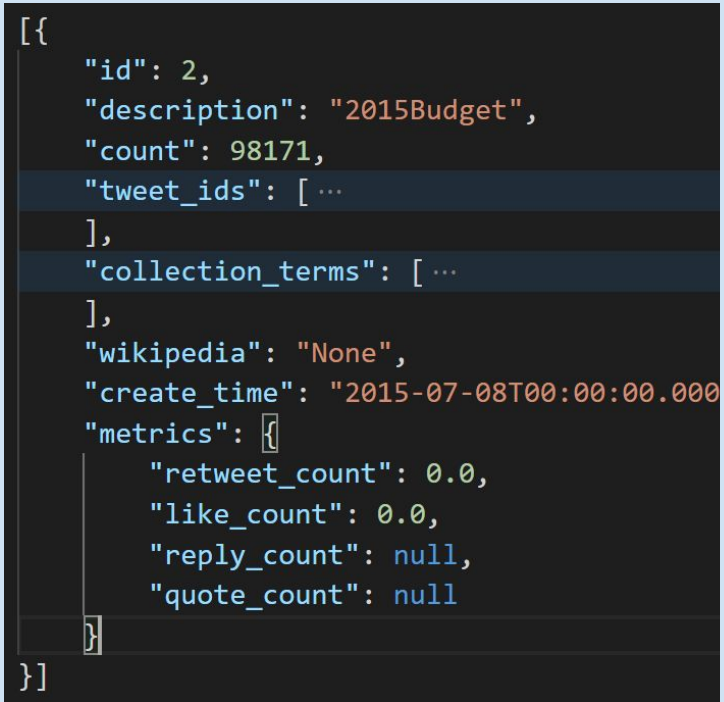
Format	Number of Tweets (Sample Data)	Time Taken (seconds)	Tweets per Second	Time Taken (seconds)	Tweets per Second
DMI	98,171	4,398.700	22.30	47.05	2086.52
YTK	750,052	179.060	4021.28	123.01	6097.48
SFM	200,171	27.999	7149.20	19.24	10403.90

Results

Format	Number of Tweets	Time Taken (seconds)	Tweets per Second	File Size (GB)
DMI	21,646,879	9,780	2,213.3	43.13
YTK	791,601,941	157,920	5,012.7	992.98
SFM	200,171	19.24	10,403.9	0.42

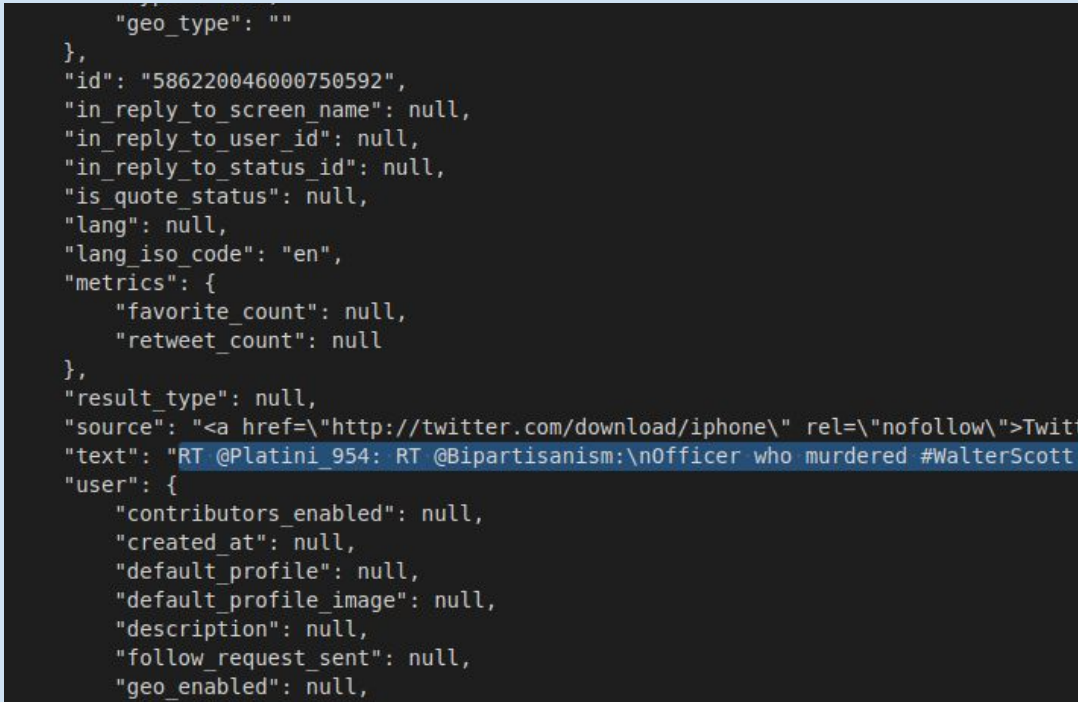
Converted JSON Data

Collection Information



```
[{
    "id": 2,
    "description": "2015Budget",
    "count": 98171,
    "tweet_ids": [ ···
    ],
    "collection_terms": [ ···
    ],
    "wikipedia": "None",
    "create_time": "2015-07-08T00:00:00.000
    "metrics": {
        "retweet_count": 0.0,
        "like_count": 0.0,
        "reply_count": null,
        "quote_count": null
    }
}]
```

Individual Conversion



```
        "geo_type": ""
    },
    "id": "586220046000750592",
    "in_reply_to_screen_name": null,
    "in_reply_to_user_id": null,
    "in_reply_to_status_id": null,
    "is_quote_status": null,
    "lang": null,
    "lang_iso_code": "en",
    "metrics": {
        "favorite_count": null,
        "retweet_count": null
    },
    "result_type": null,
    "source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twit
    "text": "RT @Platini_954: RT @Bipartisanism:\nOfficer who murdered #WalterScott
    "user": {
        "contributors_enabled": null,
        "created_at": null,
        "default_profile": null,
        "default_profile_image": null,
        "description": null,
        "follow_request_sent": null,
        "geo_enabled": null,
```

Challenges

Scalability:

- Data Size

Access:

- VPN
- Script Runtimes

Acknowledgements

Our Client, Xinyue Wang , our Mentor Pranav Chimote, TA Ryan Wood, and Professor Fox.

References

- Last Semester's Work on the Tweet Conversion
  - Library Tweet Conversion, Edward Fox & Xinyue Wang
- Events Archiving
  - http://eventsarchive.org/
- PyTorch
  - https://pytorch.org/docs/stable/index.html
- Transformers - BERT Documentation
  - https://huggingface.co/docs/transformers/index
- UJSON
  - https://mpython.readthedocs.io/en/master/library/pythonStd/ujson.html