# Twitter Collection 2024

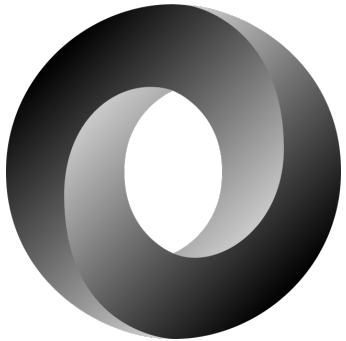
CS4264
Multimedia, Hypertext, and Information Access
Dr. Edward Fox

Chris Lam, Enk Naran, Ilya Mruz, Sushen Kolakaleti,
Kevin D'Alessandro

Virginia Tech, Blacksburg VA 24061
4/30/2024

# Outline

- Introduction / Recap
- The Sources
- Events Archive Spreadsheet
- Timeline
- DMI-TCAT
- YTK
- SFM
- System Design
- Machine Learning
- Challenges Faced
- Lessons Learned
- Future Work
- Acknowledgements, References, Q&A



# Introduction / Recap of Project Details

- The Digital Library Research Laboratory collected billions of tweets
- Collected in 3 formats (YTK, DMI-TCAT, SFM)
- Goal: convert them to one format (JSON)
- 2021 team wrote initial scripts
- 2022 team optimized the scripts, began converting, and introduced a machine learning model to classify tweets
- 2024 team delayed by a credentials mix up

# The Sources

Social Feed Manager (SFM)

- JSON file
- One Tweet per line

yourTwapperKeeper (YTK)

- MySQL database

Digital Methods Initiative Twitter Capture and Analysis Toolset (DMI-TCAT)

- MySQL database
- 1 tweet split into 7 tables

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | created_at | twitter_id | screen_nam | followers_co | friends_cour | retweet_cou | hashtags | in_reply_to_ | twitter_url | coordinates | text | url1 | url1_expand | url2 |
| 2 | 2016-06-01 0 | 7.378E+17 | TheDemocra | 508058 | 1103 | 59 | | | http://twitter.com/TheDe | | RT @AfAmD | https://t.co/ | http://theatln.tc/1O2KYGj | |
| 3 | 2016-05-31 2 | 7.3778E+17 | TheDemocra | 508058 | 1103 | 186 | | | http://twitter.com/TheDe | | Stand with D | https://t.co/ | http://bit.ly/1WvhZNP | |
| 4 | 2016-05-31 2 | 7.3774E+17 | TheDemocra | 508058 | 1103 | 143 | | | http://twitter.com/TheDe | | Democrats a | https://t.co/ | http://nyti.ms/1Uf4xHF | |

# Events Archive Spreadsheet

- Running compilation of all collections
- Key Columns:
  - Source
  - Term used to collect
  - Count of tweets in the collection
- YTK is comprehensively compiled
- SFM and DMI are not

| | ID | Source | Collection Terms | Wikipedia | Description |
|---|---|---|---|---|---|
| Collect_yTK | 1 | yTK | #egypt | https://en.wikipedia.org/wik | Originally for Egyptian revolu |
| Collect_yTK | 2 | yTK | #libya | https://en.wikipedia.org/wik | " In the second Libyan Civil v |
| Collect_yTK | 3 | yTK | #blacksburg | https://en.wikipedia.org/wik | " Blacksburg High School, wh |
| Collect_yTK | 4 | yTK | #jan25 | https://en.wikipedia.org/wik | January 25th 2011 was the d |
| Collect_yTK | 5 | yTK | #bahrain | https://en.wikipedia.org/wik | " In December 1994, a group |
| Collect_yTK | 6 | yTK | #yemen | https://en.wikipedia.org/wik | " According to the 2009 inte |
| Collect_yTK | 7 | yTK | japan earthquake | https://en.wikipedia.org/wik | "This is a list of earthquakes |
| Collect_yTK | 8 | yTK | #syria | https://en.wikipedia.org/wik | " Syria is ranked last on the G |
| Collect_yTK | 9 | yTK | OccupyWallStreet | https://en.wikipedia.org/wik | "== Origins ==The original pr |
| Collect_yTK | 10 | yTK | #nrv | | new river valley (blacksburg) |
| Collect_yTK | 11 | yTK | virginia tech | https://en.wikipedia.org/wik | A 23-year-old student, Seung |
| Collect_yTK | 12 | yTK | iran earthquake | https://en.wikipedia.org/wik | " As a result, earthquakes in |
| Collect_yTK | 13 | yTK | diabetes | https://en.wikipedia.org/wik | health category |
| Collect_yTK | 14 | yTK | heart attack | https://en.wikipedia.org/wik | health category |
| Collect_yTK | 15 | yTK | foursquare | https://en.wikipedia.org/wiki/Foursquare | |
| Collect_yTK | 16 | yTK | #Isaac | https://en.wikipedia.org/wik | Hurricane Isaac formed on A |
| Collect_yTK | 17 | yTK | turkey syria | https://en.wikipedia.org/wik | violence between Turkey and |
| Collect_yTK | 18 | yTK | emergency preparedness | https://www.ready.gov/ | "Emergency management or |
| Collect_yTK | 19 | yTK | emergency response | https://www.ready.gov/busi | "The emergency services in v |
| Collect_yTK | 20 | yTK | emergency recovery | https://www.fema.gov/med | "Emergency management or |
| Collect_yTK | 21 | yTK | emergency mitigation | https://en.wikipedia.org/wik | "Emergency management or |
| Collect_yTK | 22 | yTK | emergency management | https://en.wikipedia.org/wik | "Emergency management or |

# Timeline


**Finalize Plan**

- Meet with client
- Plan objectives
- Get approval

**Process**

- Start conversions
- Start ML

**Analyze/Document**

- Continue conversions
- Validate data
- Analyze
- Machine Learning
- Presentation

**Wrap Up Project**

- Documentation
- Submit to VTechWorks

# DMI-TCAT Database

- Hosted on the tweets.cs.vt.edu
- Raw SQL files are accessible
- Transferred some of the whole SQL file into a database
    - db1track

```
Query OK, 5963 rows affected (0.14 sec)
Records: 5963  Duplicates: 0  Warnings: 0

Query OK, 5250 rows affected (0.10 sec)
Records: 5250  Duplicates: 0  Warnings: 0

Query OK, 5132 rows affected (0.18 sec)
Records: 5132  Duplicates: 0  Warnings: 0

Query OK, 5586 rows affected (0.15 sec)
Records: 5586  Duplicates: 0  Warnings: 0

Query OK, 5410 rows affected (0.23 sec)
Records: 5410  Duplicates: 0  Warnings: 0

Query OK, 5774 rows affected (0.15 sec)
Records: 5774  Duplicates: 0  Warnings: 0

Query OK, 5803 rows affected (0.11 sec)
Records: 5803  Duplicates: 0  Warnings: 0
```

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| 2024_dmi_new       |
| db1_track          |
| db2_onepercent     |
| dmi                |
| mysql              |
| new_dmi_database   |
| performance_schema |
| sys                |
| test               |
| twitter            |
+--------------------+
```

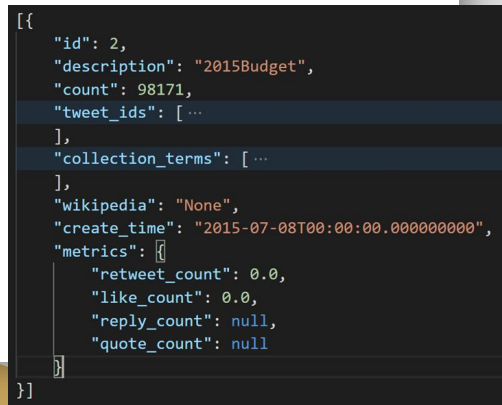
# DMI-TCAT Conversion

## Individual Level

- Establishes connection
- Retrieves data
- Transforms to JSON schema

## Collection Level

- Determines tweet IDs
- Determines number of tweets
- DMI-TCAT does not store collection data
- Attributes pulled from Events Archive spreadsheet

# YTK Conversion

```
mysql> describe z_1;
+--------------------+---------------+------+-----+---------+-------+
| Field              | Type          | Null | Key | Default | Extra |
+--------------------+---------------+------+-----+---------+-------+
| archivesource      | varchar(100)  | NO   |     | NULL    |       |
| text               | varchar(1000) | NO   | MUL | NULL    |       |
| to_user_id         | varchar(100)  | NO   |     | NULL    |       |
| from_user          | varchar(100)  | NO   | MUL | NULL    |       |
| id                 | varchar(100)  | NO   | MUL | NULL    |       |
| from_user_id       | varchar(100)  | NO   |     | NULL    |       |
| iso_language_code  | varchar(10)   | NO   | MUL | NULL    |       |
| source             | varchar(250)  | NO   |     | NULL    |       |
| profile_image_url  | varchar(250)  | NO   |     | NULL    |       |
| geo_type           | varchar(30)   | NO   | MUL | NULL    |       |
| geo_coordinates_0  | double        | NO   |     | NULL    |       |
| geo_coordinates_1  | double        | NO   |     | NULL    |       |
| created_at         | varchar(50)   | NO   |     | NULL    |       |
| time               | int           | NO   | MUL | NULL    |       |
+--------------------+---------------+------+-----+---------+-------+
14 rows in set (0.00 sec)
```

Individual Level

- Python script that converts the MySQL YTK data to JSON
- Values that aren't found set to NULL
- Most fields are NULL due to limited number of fields available from YTK

Example conversion →


```
{"contributors":null,"created_at":"2012-10-07T22:02:43.000Z","entities":{"hashtags":[{"text":"blind","indices":null},{"text":"Blacksburg","indices":null}],"media":[],"user_mentions":[],"url
{"contributors":null,"created_at":"2012-10-07T20:45:02.000Z","entities":{"hashtags":[{"text":"blacksburg","indices":null}],"media":[],"user_mentions":[],"urls":[]},"geo":{"country":null,"la
{"contributors":null,"created_at":"2012-10-07T20:44:49.000Z","entities":{"hashtags":[{"text":"vt","indices":null},{"text":"va","indices":null},{"text":"virginia","indices":null},{"text":"bl
{"contributors":null,"created_at":"2012-10-07T19:46:56.000Z","entities":{"hashtags":[{"text":"falltime","indices":null},{"text":"blacksburg","indices":null}],"media":[],"user_mentions":[],"
{"contributors":null,"created_at":"2012-10-07T19:20:27.000Z","entities":{"hashtags":[{"text":"vt","indices":null},{"text":"va","indices":null},{"text":"blacksburg","indices":null},{"text":"
{"contributors":null,"created_at":"2012-10-07T18:46:51.000Z","entities":{"hashtags":[{"text":"Blacksburg","indices":null}],"media":[],"user_mentions":[],"urls":[]},"geo":{"country":null,"la
{"contributors":null,"created_at":"2012-10-07T18:06:59.000Z","entities":{"hashtags":[{"text":"blacksburg","indices":null},{"text":"yum","indices":null}],"media":[],"user_mentions":[],"urls"
{"contributors":null,"created_at":"2012-10-07T15:49:11.000Z","entities":{"hashtags":[{"text":"Blacksburg","indices":null}],"media":[],"user_mentions":[],"urls":[]},"geo":{"country":null,"la
{"contributors":null,"created_at":"2012-10-07T15:47:11.000Z","entities":{"hashtags":[{"text":"Blacksburg","indices":null}],"media":[],"user_mentions":[],"urls":[]},"geo":{"country":null,"la
{"contributors":null,"created_at":"2012-10-07T15:24:26.000Z","entities":{"hashtags":[{"text":"Blacksburg","indices":null}],"media":[],"user_mentions":[],"urls":[]},"geo":{"country":null,"la
{"contributors":null,"created_at":"2012-10-07T15:22:39.000Z","entities":{"hashtags":[{"text":"Blacksburg","indices":null}],"media":[],"user_mentions":[],"urls":[]},"geo":{"country":null,"la
{"contributors":null,"created_at":"2012-10-07T15:21:42.000Z","entities":{"hashtags":[{"text":"Blacksburg","indices":null}],"media":[],"user_mentions":[],"urls":[]},"geo":{"country":null,"la
{"contributors":null,"created_at":"2012-10-07T15:18:24.000Z","entities":{"hashtags":[{"text":"Blacksburg","indices":null}],"media":[],"user_mentions":[],"urls":[]},"geo":{"country":null,"la
```

# YTK Conversion Continued

Collection Level

- Made some improvements to the script speed up the conversions from last team
- Doesn’t account for additional fields past tweet ID
- Places tweet IDs as a list after conversion

Example conversion




```
{"id": 279, "description": "\" The engines were air-cooled until the introduction of the Type 996 in 1998, with Porsche's \"993\" series, produced in model years 1994â€“1998, being the last of the air-cooled Porsches\"", "count": 225047, "tweet_ids": ["1000102042051010560", "1000762645656424449", "1001999007940521984", "1002013255148032000", "1002221440769318912",
"1002594478576857089", "1002596170315784193", "1002627773960282114", "1002972823860797440", "1003608658121392128", "1003804065359384576", "1004283983964692482", "1004284197433819138",
"1004955657207009280", "1005233909678759936", "1005251069545406464", "1005254798583578625", "1005655906565087232", "1005656088727810049", "1005807722028298240", "1006155053542137857",
"1006425979886153728", "1006425981349867520", "1006544086051024902", "1006702890642235393", "1006782226246316035", "1007043268675358720", "1007227405793091584", "1007474014120566784",
"1007501364673564672", "1007825618958979072", "1008521472715808769", "1009127368638828544", "1009240768010162176", "1009467973072015361", "1009774114205290496", "1009774232627183616",
"1009774502442565633", "1009775325578977281", "1009776013012856833", "1009776500055277568", "1009778296702238720", "1009780554810699776", "1009783236623568896", "1009789718228275201",
"1009792328192622592", "1009796292615208961", "1009796540129513472", "1009799550104756226", "1009801397980655617", "1009803460626321408", "1009833031593615360", "1009833467784433664",
"1009837269723832320", "1009840546402193410", "1009841040310714368", "1009842254075195393", "1009842966754553856", "1009854645636466017", "1009926300872691713", "1010242166705262594",
"1011314572639293441", "1011315606308544512", "1011317500666171392", "1011318117580935168", "1011318809179901954", "1011319374697893893", "1011323362667368451", "1011323546180706304",
"1011352459640496133", "1011640848021377025", "1011692327226703872", "1011711592856244225", "1011718942015082496", "1011719155983187968", "1011839406918062081", "1011839721113378816",
"1011839789119795200", "1011839800788377600", "1011839931604520960", "1011839931961012226", "1011839943839137792", "1011840019764428802", "1011840046704492545", "1011840138283012097",
"1011840170776256512", "1011840631797211136", "1011840740564127744", "1011840838727585792", "1011841025797705728", "1011841526517981184", "1011841540497473536", "1011841879766388736",
"1011841882249400321", "1011842115179982848", "1011842300660539393", "1011842370298507264", "1011842384215207937", "1011842508773609474", "1011842706061000704", "1011842706186924032",
"1011842884935434240", "1011843130025496576", "1011843185390350336", "1011843191031517185", "1011843598415851520", "1011843664916729861", "1011843664950071296", "1011843705689518080",
"1011844340434386944", "1011844419870322688", "1011844803015905280", "1011844994649518080", "1011845088610156544", "1011845210043748352", "1011845249008668672", "1011845250464112640",
"1011845576986464256", "1011845657664073734", "1011846056785661952", "1011846208464326657", "1011846452757266432", "1011846569082216448", "1011846656034263040", "1011846809092644864",
```

# SFM Camelot Virginia Tech Machine Info

- 24 Terabytes of Storage for SFM tweet JSON data, Raid Node
- Virtual Machine to run SFM (sfm1.cs.vt.edu)


```
[cs4624s24_tweet@camelot yTK_DMI]$ df -h --total
Filesystem              Size  Used Avail Use% Mounted on
devtmpfs                 16G     0   16G   0% /dev
tmpfs                    16G     0   16G   0% /dev/shm
tmpfs                    16G   26M   16G   1% /run
tmpfs                    16G     0   16G   0% /sys/fs/cgroup
/dev/sda1               100G   35G   66G  35% /
/dev/mapper/RAID-DISK    51T   26T   26T  50% /raid
tmpfs                   3.2G     0  3.2G   0% /run/user/1012
total                    51T   26T   26T  50% -
[cs4624s24_tweet@camelot yTK_DMI]$
```

# SFM UI Website

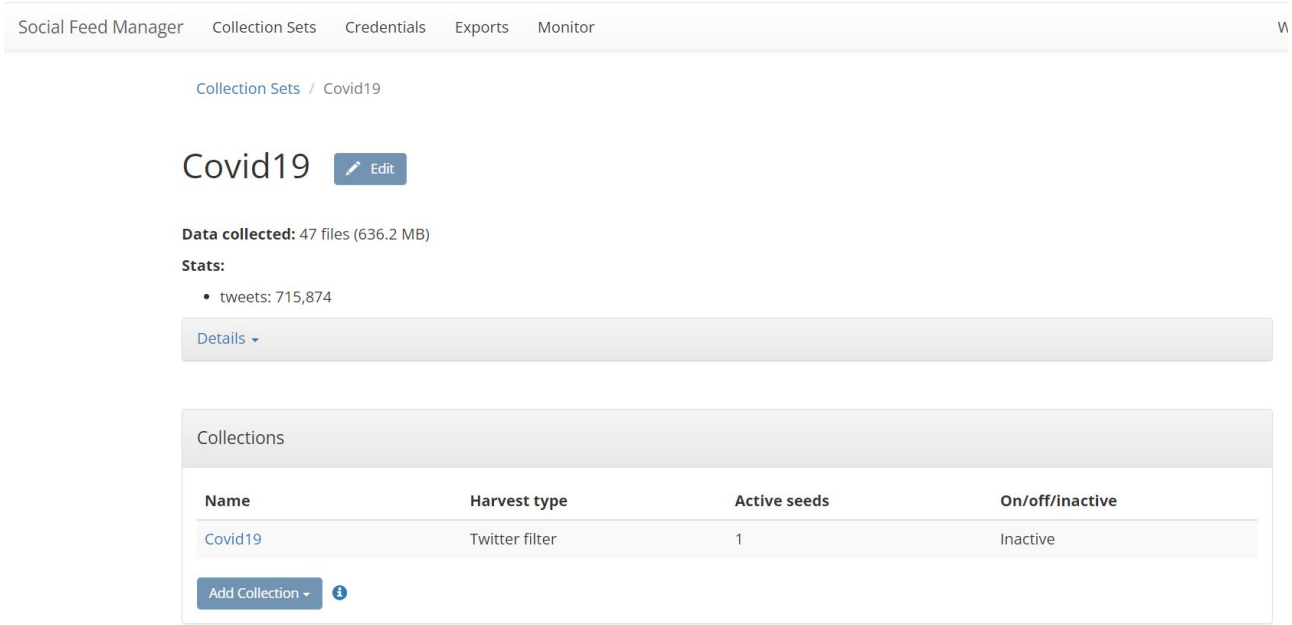
- A website to access the collections created by various research groups
- Can collect from a wide variety of websites:
  - Twitter
  - Tumblr
  - Flickr
  - Weibo
- Can collect using various methods such as: search, username, etc.
- We used the export feature to decrypt and uncompress the JSON files

# SFM UI Collection Sets

Active 0 | Shared 2 | Other Active 181 | Other Inactive 62

| Name | Collections | Date Added | Groups |
|---|---|---|---|
| Alexandre de Moraes | 1 collection | Oct. 21, 2021, 7:16:03 p.m. EDT | carolinaresearch |
| algorand sample | 1 collection | Feb. 14, 2022, 6:21:27 a.m. EST | Mary |
| #ArthurLiraGenocida 19/03/2021 17:42 | 1 collection | March 19, 2021, 4:42:51 p.m. EDT | carolinaresearch |
| Auxílio Brasil 19/10/2021 18:20 | 1 collection | Oct. 19, 2021, 5:20:07 p.m. EDT | carolinaresearch |
| Aziz 19/10/2021 10:18 | 1 collection | Oct. 19, 2021, 9:18:58 a.m. EDT | carolinaresearch |
| Bitcoin dataset | 1 collection | Feb. 16, 2022, 4:43:57 p.m. EST | Mary |
| Bitcoin sample | 1 collection | Feb. 6, 2022, 7:07:20 p.m. EST | Mary |
| #bolsaestupro 22/03/2021 22:39 | 1 collection | March 22, 2021, 9:39:57 p.m. EDT | carolinaresearch |
| #BolsonaroGenocida | 1 collection | March 16, 2021, 4:42:02 p.m. EDT | carolinaresearch |

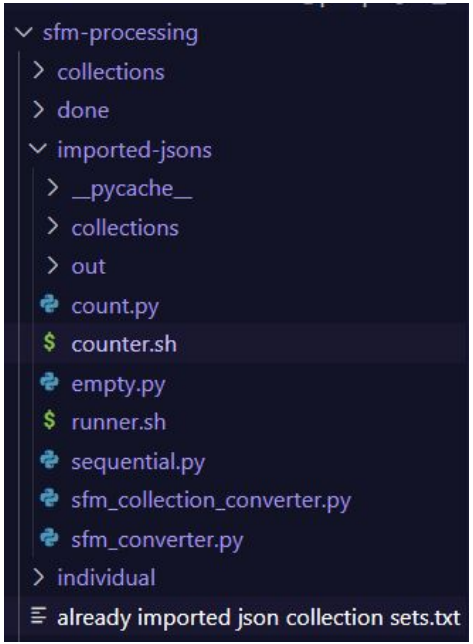
# SFM UI Collection Set Information



# SFM UI Exports

| | | |
|---|---|---|
| Brexit > Brexit hashtags | requested | April 8, 2024, 1:25:02 p.m. EDT |
| Brexit > Scottish National Party | requested | April 8, 2024, 1:24:55 p.m. EDT |
| Brexit > Liberal Democrats Party | requested | April 8, 2024, 1:24:46 p.m. EDT |
| Brexit > UK Independence Party | completed success | April 8, 2024, 1:24:37 p.m. EDT |
| Brexit > Conservative Party | completed success | April 8, 2024, 1:24:28 p.m. EDT |
| Brexit > test | running | April 8, 2024, 1:24:19 p.m. EDT |
| Brexit > Labour Party | completed success | April 8, 2024, 1:24:02 p.m. EDT |
| brandbuilding > Donald Trump | completed success | April 8, 2024, 1:23:42 p.m. EDT |
| Bolsonaro > Bolsonaro | completed success | April 8, 2024, 1:23:24 p.m. EDT |
| Belgian elected representatives > Seedlist 1 | completed success | April 8, 2024, 1:22:51 p.m. EDT |
| Viva 64 > Viva 64 | completed success | April 4, 2024, 10:15:31 a.m. EDT |
| Viagra 11/04/2022 17h55 > Viagra 11/04/2022 17h55 | completed success | April 4, 2024, 10:15:17 a.m. EDT |

# SFM New Scripts / Automation




- How It Works:
  - Import the raw JSON files from the exports in SFM UI
  - Run the runner.sh script (runs the collection and individual conversion scripts sequentially)
  - Collection level goes to collections folder
  - Individual level goes to out folder
  - Move the finished JSON files to collections and done folder once scripts are complete

- Scripts:
  - runner.sh, counter.py, sequential.py, sfm_converter and sfm_collection_convert.py

# SFM New Scripts / Automation Continued

# SFM Tweet Conversion Data

Completed 9744468 SFM tweet conversions as of 4/28.

Below is part of an output for the individual conversion script

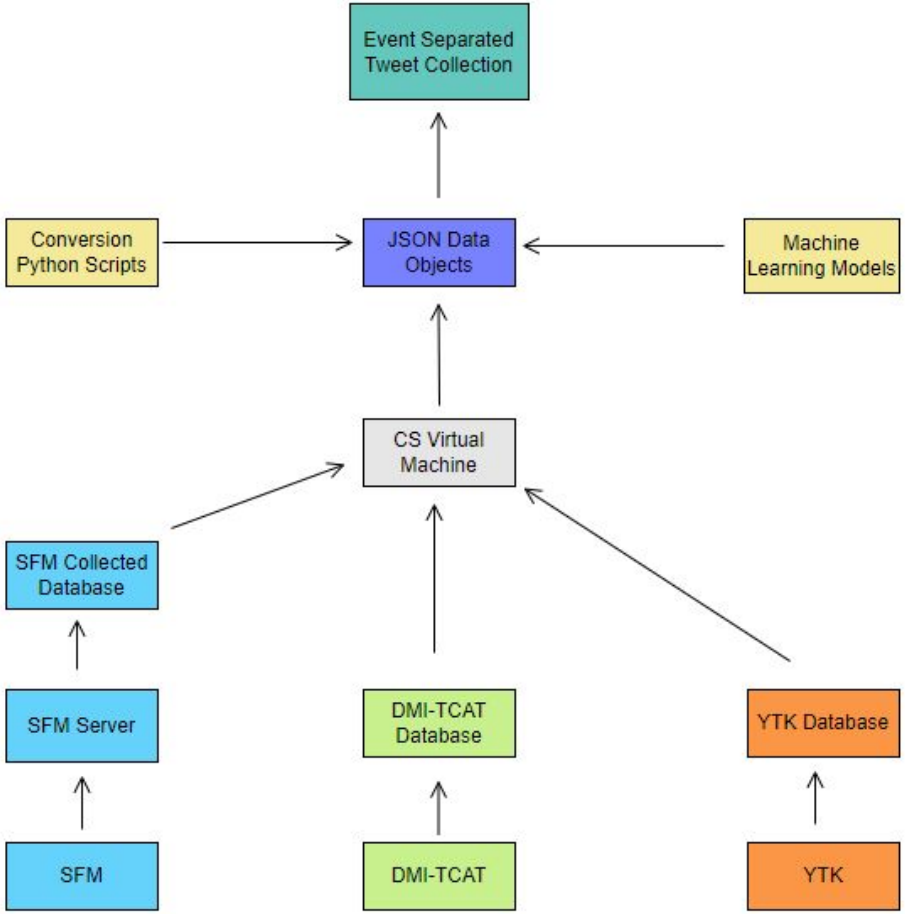
```
[cs4624s24_tweet@camelot sfm-processing]$ bash counter.sh
Started.
Total Count: 9744468
[cs4624s24_tweet@camelot sfm-processing]$
```

sfm-processing > done > {} 3c78c42e34f54259bdfcad886dd14261_001.json > ...


```
{"created_at": "Sat Apr 10 01:11:26 +0000 2021", "id": 1380689846885351426, "id_str": "1380689846885351426", "full_text": "@snaptravel
{"created_at": "Thu Apr 08 14:32:47 +0000 2021", "id": 1380166740521279494, "id_str": "1380166740521279494", "full_text": "@tedcruz @te
{"created_at": "Sat Mar 27 22:07:43 +0000 2021", "id": 1375932571805216774, "id_str": "1375932571805216774", "full_text": "RT @fortesal
{"created_at": "Fri Mar 26 19:53:01 +0000 2021", "id": 1375536286493253638, "id_str": "1375536286493253638", "full_text": "@HughThunkIt
{"created_at": "Tue Mar 23 17:30:10 +0000 2021", "id": 1374413171675435014, "id_str": "1374413171675435014", "full_text": "As a local h
{"created_at": "Tue Mar 23 13:51:51 +0000 2021", "id": 1374358233276645380, "id_str": "1374358233276645380", "full_text": "RT @ExpressN
{"created_at": "Mon Mar 22 18:00:10 +0000 2021", "id": 1374058334999998471, "id_str": "1374058334999998471", "full_text": "The Texas At
{"created_at": "Mon Mar 22 14:47:34 +0000 2021", "id": 1374009867577475072, "id_str": "1374009867577475072", "full_text": "RT @AlsoBill
{"created_at": "Mon Mar 22 14:32:02 +0000 2021", "id": 1374005954740695052, "id_str": "1374005954740695052", "full_text": "@aggierican
{"created_at": "Mon Mar 22 11:59:33 +0000 2021", "id": 1373967583641210881, "id_str": "1373967583641210881", "full_text": "Texas attorr
{"created_at": "Mon Mar 22 05:00:04 +0000 2021", "id": 1373862016847388676, "id_str": "1373862016847388676", "full_text": "The Texas At
{"created_at": "Sun Mar 21 04:31:46 +0000 2021", "id": 1373492508232085508, "id_str": "1373492508232085508", "full_text": "Family Alleg
{"created_at": "Sat Mar 20 15:44:06 +0000 2021", "id": 1373299315960123396, "id_str": "1373299315960123396", "full_text": "Texas attorr
```

# System Design

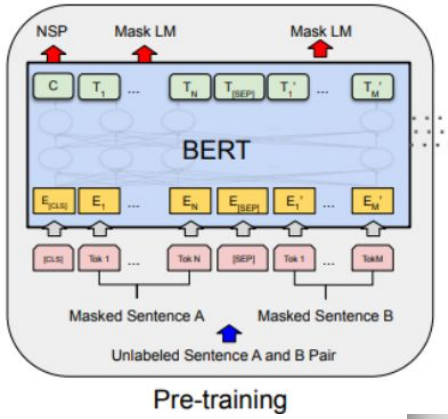
- Gather databases and store on CS Machines (Camelot, Tweets, etc.)
- Convert to JSON objects using conversion automation scripts
- Convert to JSON collections using collection automation scripts



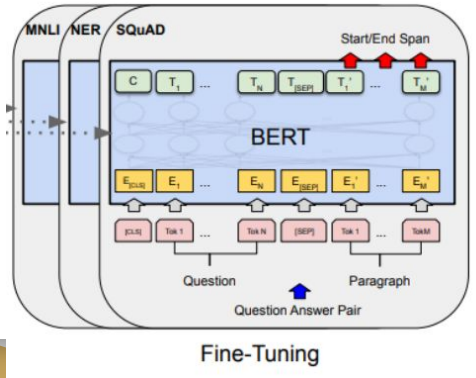
# Machine Learning

Hate Speech Detection

- Detect hate speech on tweets
- Tested a GloVe model with a Naive Bayes classifier
- And a BERT model with a Naive Bayes classifier




Pre-training

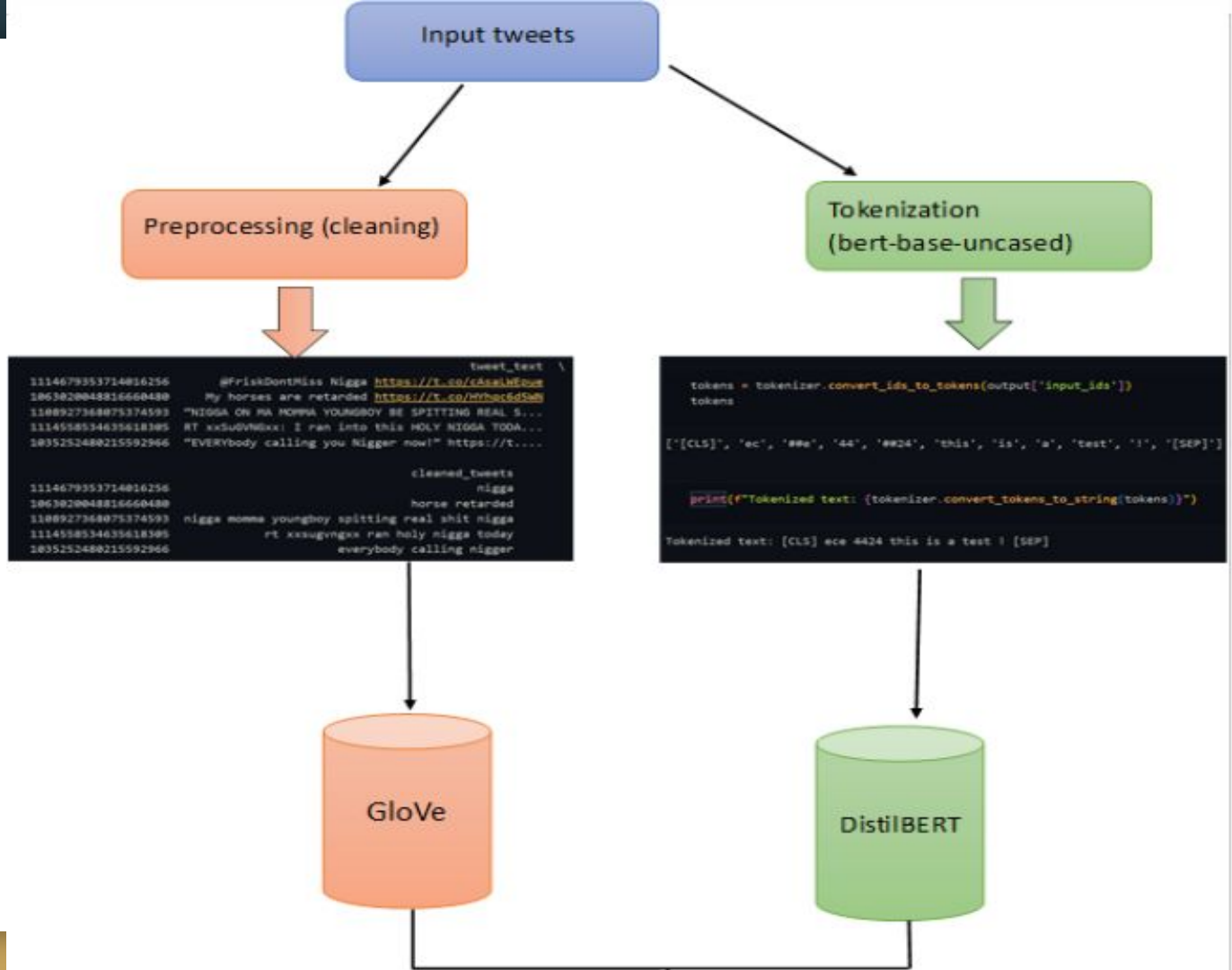



Fine-Tuning

# ML Results

| **DistilBERT** | **GloVe** |
| --- | --- |
| Running Time: 4m 02s | Running Time: 58s |
| Accuracy: 83.12% | Accuracy: 70.01% |
| Cost Ratio: 284.10 | Cost Ratio: 82.85 |

$$R = \frac{RT_{sec}}{Acc} \cdot 100$$


Input tweets
Preprocessing (cleaning)
Tokenization
(bert-base-uncased)
tweet_text \
1114679353714016256 @FriskDontMiss Nigga https://t.co/cAaaLWEpue
1063020048816660480 My horses are retarded https://t.co/HYhqoc6d5W
1108927368075374593 "NIGGA ON MA MOMMA YOUNGBOY BE SPITTING REAL S...
1114558534635618305 RT xxSuGVNBxx: I ran into this HOLY NIGGA TODA...
1035252480215592966 "EVERYbody calling you Nigger now!" https://t....
cleaned_tweets
1114679353714016256 nigga
1063020048816660480 horse retarded
1108927368075374593 nigga momma youngboy spitting real shit nigga
1114558534635618305 rt xxsugvngxx ran holy nigga today
1035252480215592966 everybody calling nigger
tokens = tokenizer.convert_ids_to_tokens(output['input_ids'])
tokens
['[CLS]', 'ec', '##e', '##4', '##24', 'this', 'is', 'a', 'test', '!', '[SEP]']
print(f"Tokenized text: {tokenizer.convert_tokens_to_string(tokens)}")
Tokenized text: [CLS] ece 4424 this is a test ! [SEP]
GloVe
DistilBERT

# Challenges Faced

- Documentation out of date / not fully covering the whole project
- Getting access to all the different VT servers and services for the project
- Communication within the team members and client
- Equally dividing the work between the team members
- Working through team member schedules throughout the semester

# Lessons Learned

- Ask questions about parts of previous team’s work earlier rather than later
- Spend more time reviewing old code and documentation earlier
- Don’t be afraid to ask questions where you’re confused

# Future Work for the Project Tweet Conversion

- Add any and all missing collections to the Events Archive Spreadsheet once more conversions are completed.
- SFM: We have exported around 45 collections from the SFM machine. For future work the next team needs to continue where we left off from exporting JSON files from the SFM machine to then process on the Camelot server. The converted JSONs are hosted on the Raid Node on the Camelot Machine for data storage. Any future team can view them there for keeping track of already completed Collection Sets.
- DMI_TCAT / YTK: Need to transfer rest of SQL file to a new database to convert the remaining tweets in the DMI_TCAT.sql file and the ytk1-big_all_mysql.sql files. Currently the transfer for DMI_TCAT is underway in 2030 Torgerson Hall, Virginia Tech and the YTK transfer should be started once the DMI_TCAT is completed.
- Integrate Machine Learning into Data Pipeline of JSON Objects. Any future work should start running the machine learning models on the growing tweet data to get more accurate results with a larger training set.

# Acknowledgements

Dr. Andrea Kavanaugh, Associate Director Center for Human/Computer Interaction, kavan@vt.edu,

Dr. Mohamed Magdy Farag, Research Associate, VTTI-Sustainable Mobility, mmagdy@vt.edu,

Dr. Edward A. Fox, Professor of CS 4624 Capstone, fox@vt.edu,

Satvik Chekuri, Ph.D. student, GTA, satvikchekuri@vt.edu,

Past CS4624 Tweet Collection 2022 Team.

# References

[1] M. Gonely, R. Nicholas, N. Fitz, G. Knock, and D. Bruce, “Twitter Collections,” *vtechworks*, May 10, 2022. http://hdl.handle.net/10919/109988 (accessed Jan. 23, 2024).

[2] P. Dhakal, Y. Bhargava, A. Herms, K. Powell, and D. Burdisso, “Library Tweets Conversion,” *vtechworks*, Dec. 16, 2021. http://hdl.handle.net/10919/107086 (accessed Jan. 23, 2024).

[3] XDevelopers, “Data dictionary: Standard v1.1,” *Developer Platform*, 2024. https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet (accessed Feb. 04, 2024).

# Q&A