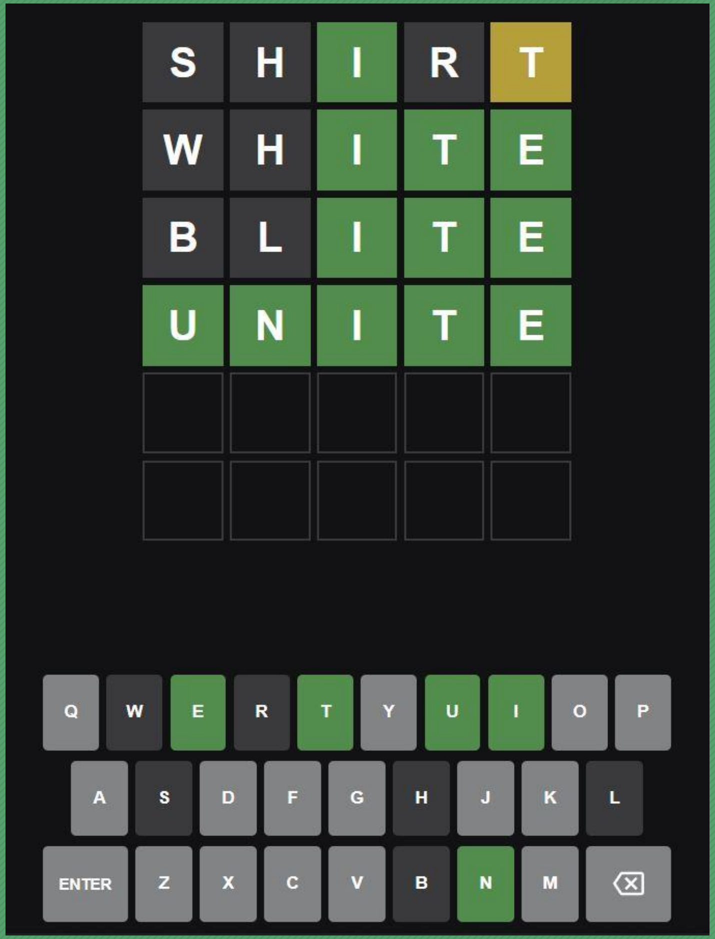
Wordle Match

Team 1 – Ayush Venkatesh, Olasunkanmi Olayinka, Sai Sumana Puppala, Jared Girouard

Wordle

- 5-letter word
- 6 guesses
- No color – letter not in word
- Yellow – letter in word but wrong position
- Green – letter in word in correct position



S H I R T

W H I T E

B L I T E

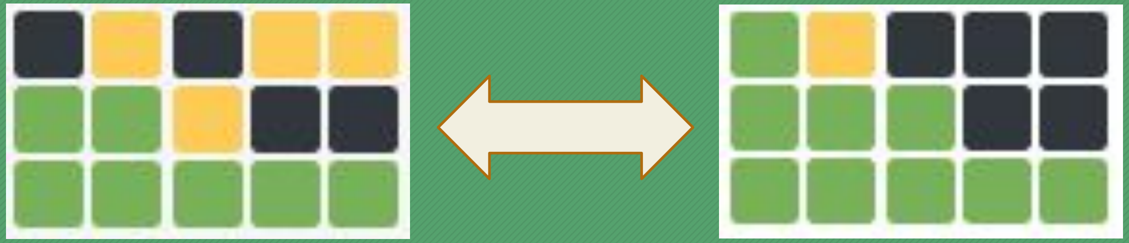
U N I T E

Use Cases

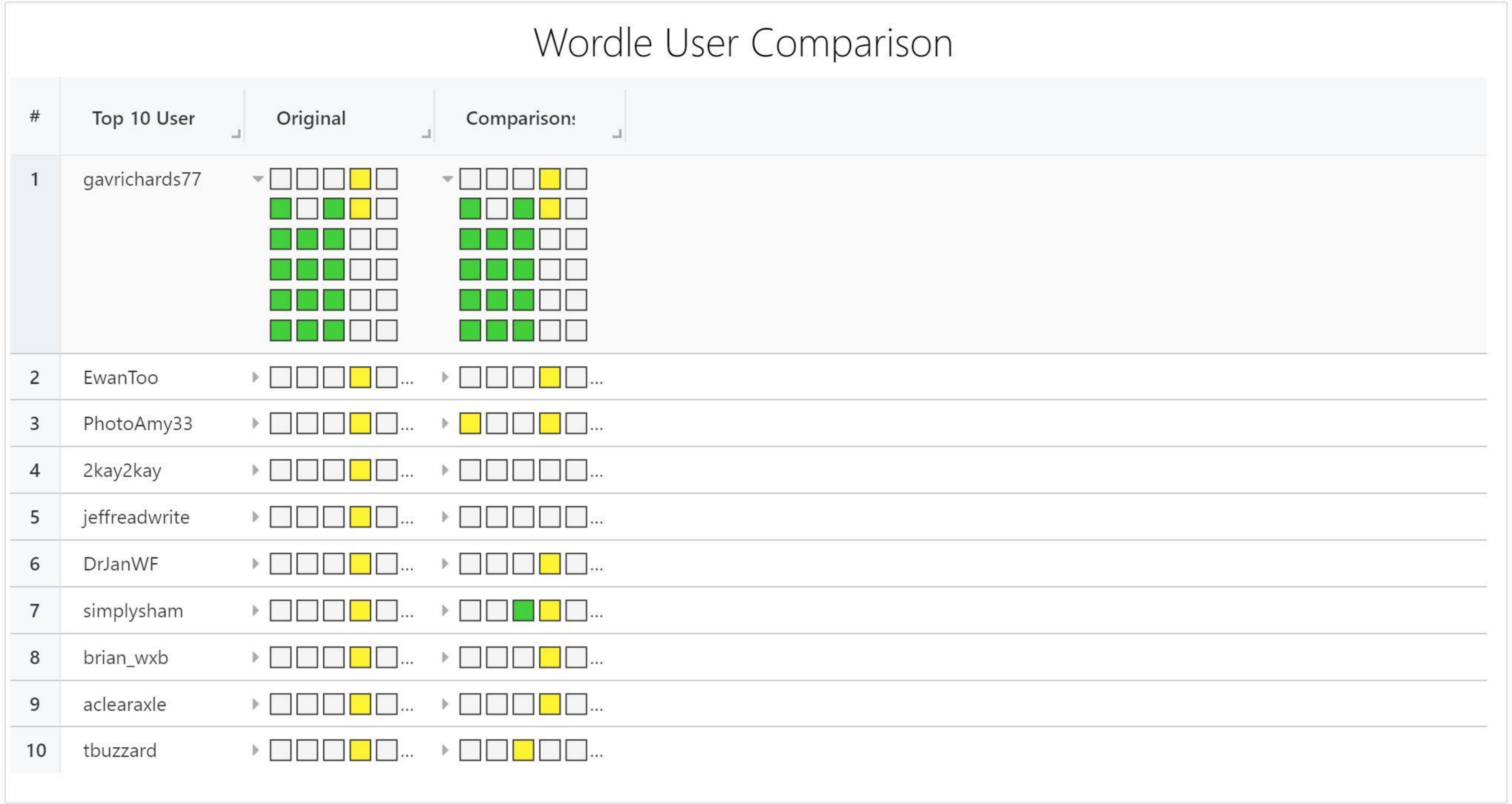
- System queries wordle tweets once each day.
- Top 10 User Search
  - Given a user, provide top 10 users with similar game pattern results for a specific day
- On a given day, compute user rankings based on scores.

Methodology

- Data Collection and Analysis
  - Extract wordle game results of users from Twitter API, over a period.
  - Clean and transform data to get 5*6 pattern matrices of game results.
- User Match
  - Compute user similarity scores among user game result matrices using a Euclidean distance metric.
- Dashboards
  - Build dashboard to display top 10 similar users.



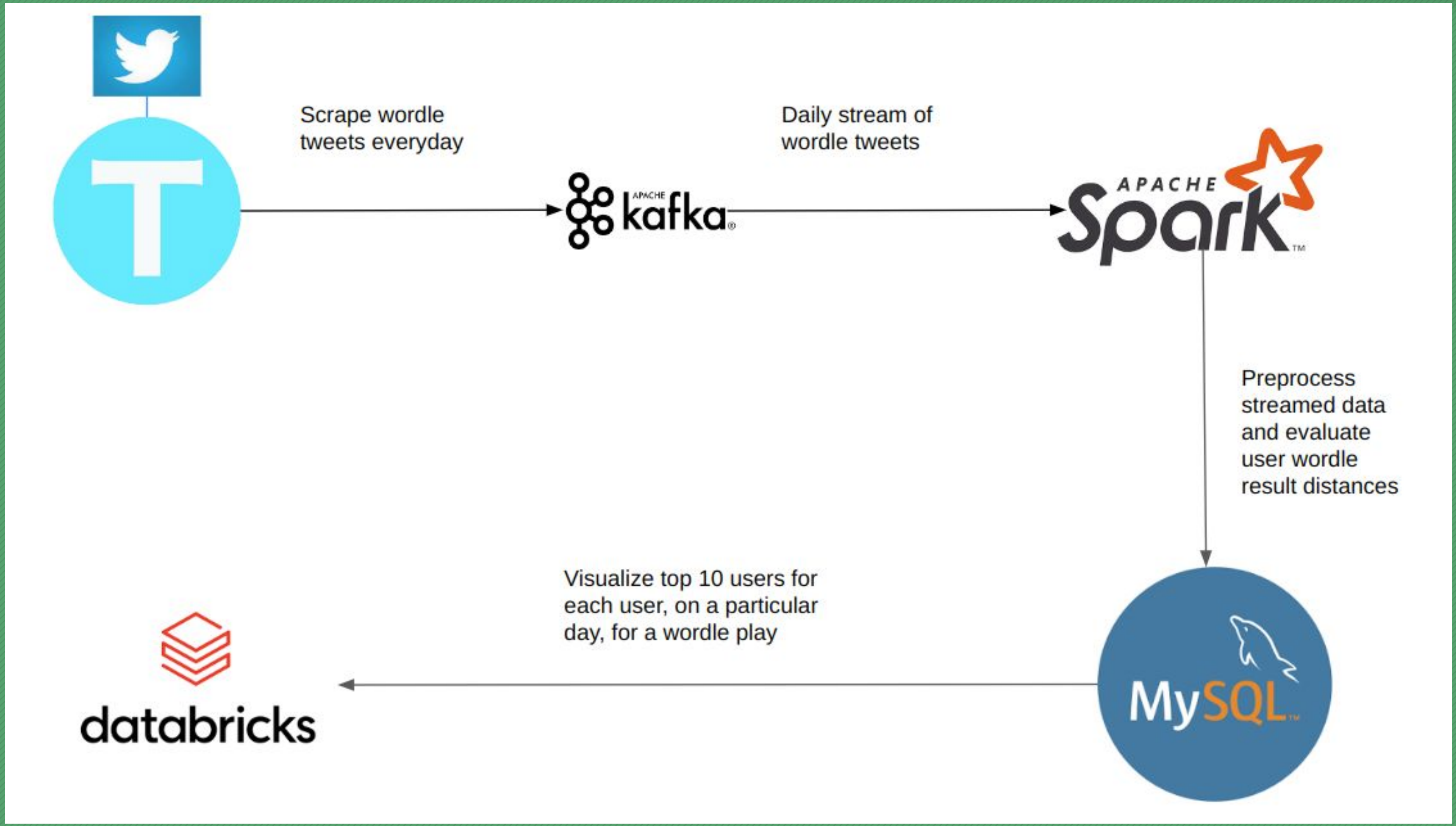
Dashboard



Wordle User Comparison

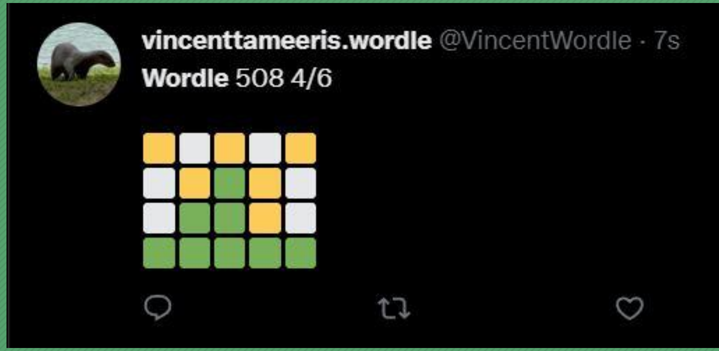
#	Top 10 User	Original	Comparisons
1	gavrichards77	⬜⬜⬜🟨⬜ 🟩⬜🟩🟨⬜ 🟩🟩🟩⬜⬜ 🟩🟩🟩⬜⬜ 🟩🟩🟩⬜⬜ 🟩🟩🟩⬜⬜	⬜⬜⬜🟨⬜ 🟩⬜🟩🟨⬜ 🟩🟩🟩⬜⬜ 🟩🟩🟩⬜⬜ 🟩🟩🟩⬜⬜ 🟩🟩🟩⬜⬜
2	EwanToo	⬜⬜⬜🟨⬜...	⬜⬜⬜🟨⬜...
3	PhotoAmy33	⬜⬜⬜🟨⬜...	🟨⬜⬜🟨⬜...
4	2kay2kay	⬜⬜⬜🟨⬜...	⬜⬜⬜⬜⬜...
5	jeffreadwrite	⬜⬜⬜🟨⬜...	⬜⬜⬜⬜⬜...
6	DrJanWF	⬜⬜⬜🟨⬜...	⬜⬜⬜🟨⬜...
7	simplysham	⬜⬜⬜🟨⬜...	⬜⬜🟩🟨⬜...
8	brian_wxb	⬜⬜⬜🟨⬜...	⬜⬜⬜🟨⬜...
9	aclearaxle	⬜⬜⬜🟨⬜...	⬜⬜⬜🟨⬜...
10	tbuzzard	⬜⬜⬜🟨⬜...	⬜⬜🟨⬜⬜...

Architecture


Scrape wordle tweets everyday
kafka
Daily stream of wordle tweets
APACHE Spark
Preprocess streamed data and evaluate user wordle result distances
MySQL
Visualize top 10 users for each user, on a particular day, for a wordle play
databricks

Data Sources

- Data was extracted via twint, a twitter tool, each day.
  - User who posted the tweet
  - Created date
  - Wordle score
  - Date/Time
- Live stream of data preprocessed before storing it in mysql database.
- Extracted data for the last 8 days, accounting to 117k user distances.

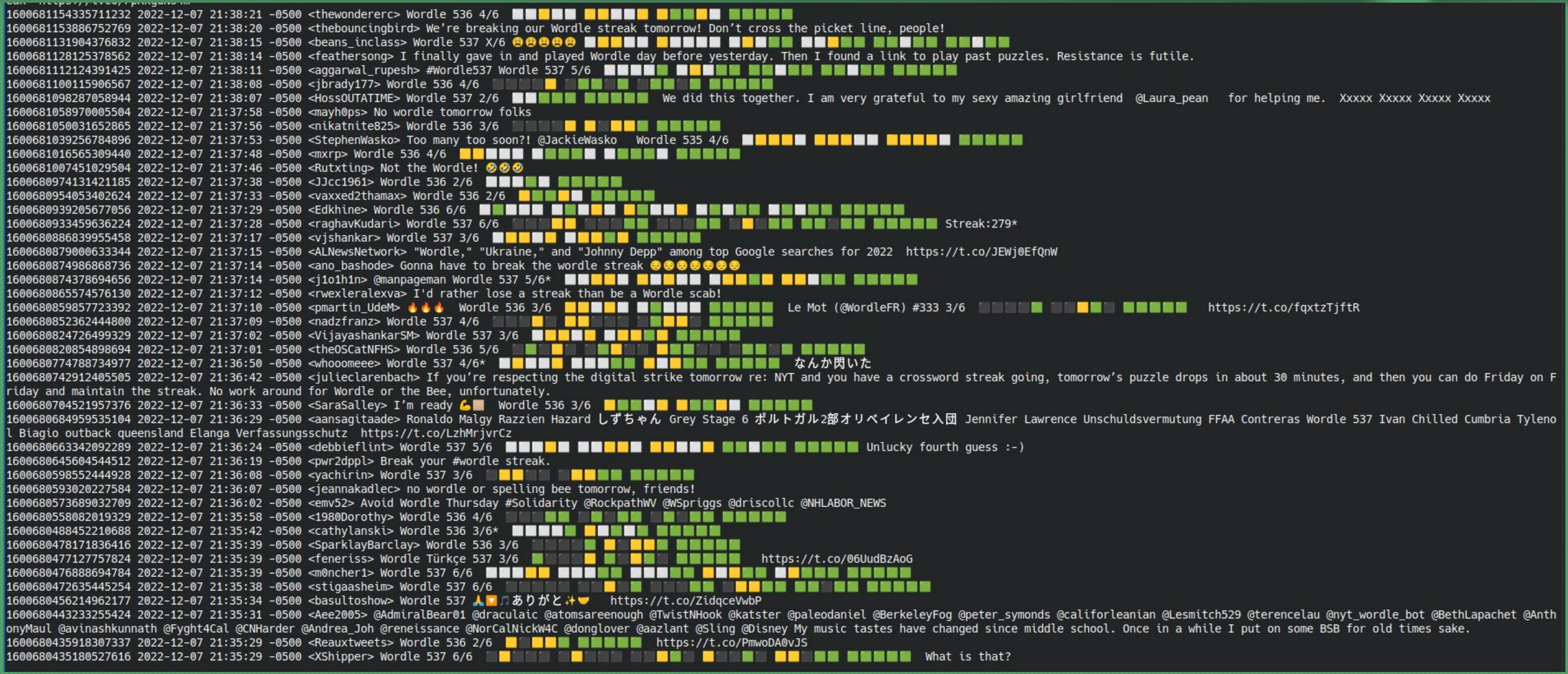


vincenttameeris.wordle @VincentWordle · 7s

Wordle 508 4/6

🟨⬜🟨⬜🟨
⬜🟨🟩🟨⬜
⬜🟩🟩🟨⬜
🟩🟩🟩🟩🟩

Data Sources: Raw Stream

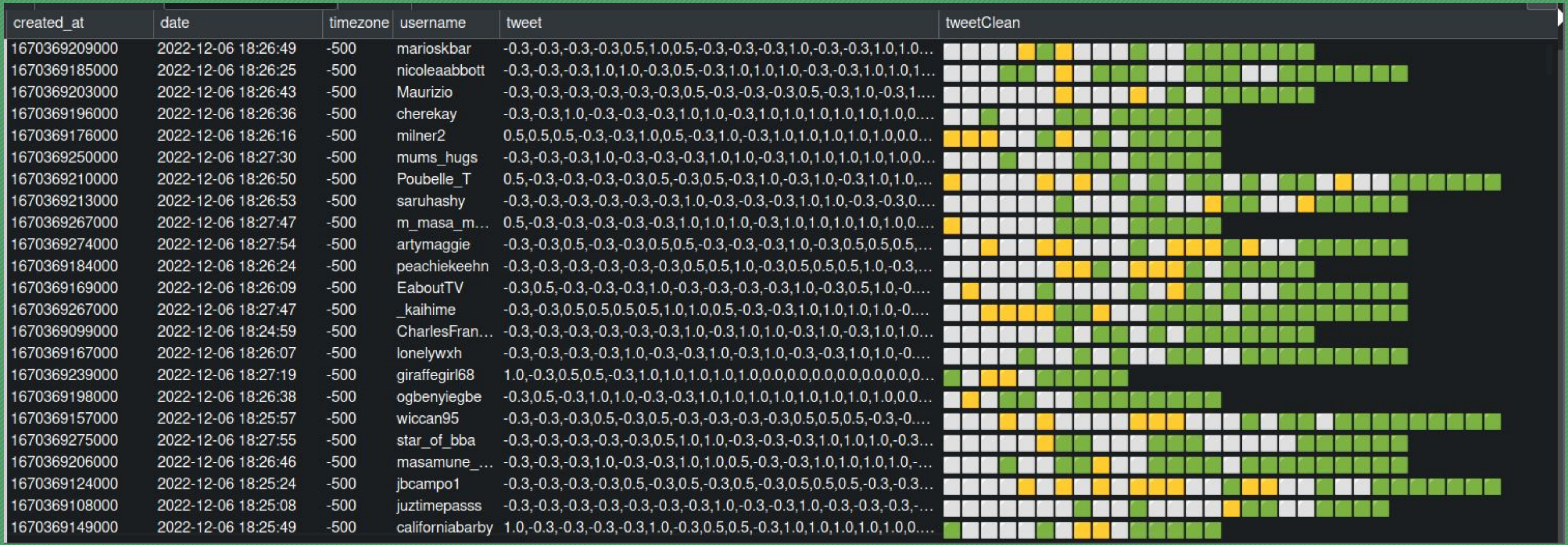


```
1600681154335711232 2022-12-07 21:38:21 -0500 <thewondererc> Wordle 536 4/6  ⬜⬜🟨⬜⬜ 🟨🟨⬜⬜🟨 🟨🟩🟩🟨⬜ 🟩🟩🟩🟩🟩
1600681153886752769 2022-12-07 21:38:20 -0500 <thebouncingbird> We’re breaking our Wordle streak tomorrow! Don’t cross the picket line, people!
1600681131904376832 2022-12-07 21:38:15 -0500 <beans_inclass> Wordle 537 X/6 😩😩😩😩😩 ⬜🟨🟨⬜⬜ 🟨⬜⬜⬜⬜ ⬜🟨⬜🟩🟩 ⬜⬜🟨🟩🟩 🟩🟩⬜🟩🟩 🟩🟩⬜🟩🟩
1600681128125378562 2022-12-07 21:38:14 -0500 <feathersong> I finally gave in and played Wordle day before yesterday. Then I found a link to play past puzzles. Resistance is futile.
1600681112124391425 2022-12-07 21:38:11 -0500 <aggarwal_rupesh> #Wordle537 Wordle 537 5/6  ⬜⬜⬜⬜🟩 ⬜🟨⬜🟩🟩 🟩🟩⬜🟩🟩 🟩🟩⬜🟩🟩 🟩🟩🟩🟩🟩
1600681100115906567 2022-12-07 21:38:08 -0500 <jbrady177> Wordle 536 4/6  ⬛⬛⬛⬛🟨 ⬛🟩🟩⬛🟩 ⬛🟩🟩⬛🟩 🟩🟩🟩🟩🟩
1600681098287058944 2022-12-07 21:38:07 -0500 <HossOUTATIME> Wordle 537 2/6  ⬜⬜🟩🟩🟩 🟩🟩🟩🟩🟩  We did this together. I am very grateful to my sexy amazing girlfriend  @Laura_pean   for helping me.  Xxxxx Xxxxx Xxxxx Xxxxx
1600681058970005504 2022-12-07 21:37:58 -0500 <mayh0ps> No wordle tomorrow folks
1600681050031652865 2022-12-07 21:37:56 -0500 <nikatnite825> Wordle 536 3/6  ⬛⬛⬛⬛🟨 🟨⬛🟨🟨🟩 🟩🟩🟩🟩🟩
1600681039256784896 2022-12-07 21:37:53 -0500 <StephenWasko> Too many too soon?! @JackieWasko   Wordle 535 4/6  ⬜🟨🟨🟨⬜ 🟨🟨🟨⬜⬜ 🟨🟨🟨🟨⬜ 🟩🟩🟩🟩🟩
1600681016565309440 2022-12-07 21:37:48 -0500 <mxrp> Wordle 536 4/6  🟨🟨⬜⬜⬜ ⬜🟩🟩🟩⬜ ⬜🟩🟩🟩⬜ 🟩🟩🟩🟩🟩
1600681007451029504 2022-12-07 21:37:46 -0500 <Rutxting> Not the Wordle! 🤣🤣🤣
1600680974131421185 2022-12-07 21:37:38 -0500 <JJcc1961> Wordle 536 2/6  ⬜⬜⬜🟩⬜ 🟩🟩🟩🟩🟩
1600680954053402624 2022-12-07 21:37:33 -0500 <vaxxed2thamax> Wordle 536 2/6  🟨🟩🟩🟨⬜ 🟩🟩🟩🟩🟩
1600680939205677056 2022-12-07 21:37:29 -0500 <Edkhine> Wordle 536 6/6  ⬜🟩⬜⬜⬜ ⬜🟩⬜🟨⬜ 🟨🟩⬜⬜🟨 ⬜🟩⬜🟩🟩 ⬜🟩⬜🟩🟩 🟩🟩🟩🟩🟩
1600680933459636224 2022-12-07 21:37:28 -0500 <raghavKudari> Wordle 537 6/6  ⬛⬛⬛🟨🟨 ⬛⬛⬛🟩🟩 ⬛⬛⬛🟩🟩 ⬛🟨⬛🟩🟩 🟩🟩⬛🟩🟩 🟩🟩🟩🟩🟩 Streak:279*
1600680886839955458 2022-12-07 21:37:17 -0500 <vjshankar> Wordle 537 3/6  ⬜🟨🟨⬜🟨 ⬜🟨🟨🟩🟨 🟩🟩🟩🟩🟩
1600680879000633344 2022-12-07 21:37:15 -0500 <ALNewsNetwork> "Wordle," "Ukraine," and "Johnny Depp" among top Google searches for 2022  https://t.co/JEWj0EfQnW
1600680874986868736 2022-12-07 21:37:14 -0500 <ano_bashode> Gonna have to break the wordle streak 😔😔😔😔😔😔😔
1600680874378694656 2022-12-07 21:37:14 -0500 <j1o1h1n> @manpageman Wordle 537 5/6*  ⬜⬜🟨🟨⬜ 🟨⬜🟨⬜⬜ ⬜🟨🟨🟩🟨 🟨🟨⬜🟩🟩 🟩🟩🟩🟩🟩
1600680865574576130 2022-12-07 21:37:12 -0500 <rwexleralexva> I'd rather lose a streak than be a Wordle scab!
1600680859857723392 2022-12-07 21:37:10 -0500 <pmartin_UdeM> 🔥🔥🔥  Wordle 536 3/6  🟨🟨⬜🟨⬜ ⬜🟩⬜⬜⬜ 🟩🟩🟩🟩🟩  Le Mot (@WordleFR) #333 3/6  ⬛⬛⬛⬛🟩 ⬛⬛🟨🟩⬛ 🟩🟩🟩🟩🟩   https://t.co/fqxtzTjftR
1600680852362444800 2022-12-07 21:37:09 -0500 <nadzfranz> Wordle 537 4/6  ⬛⬛⬛🟨⬛ 🟨🟨⬛⬛⬛ ⬛🟩🟨🟨⬛ 🟩🟩🟩🟩🟩
1600680824726499329 2022-12-07 21:37:02 -0500 <VijayashankarSM> Wordle 537 3/6  ⬜🟨🟨⬜🟨 ⬜🟨🟨🟩🟨 🟩🟩🟩🟩🟩
1600680820854898694 2022-12-07 21:37:01 -0500 <theOSCatNFHS> Wordle 536 5/6  ⬛🟩⬛🟨⬛ ⬛🟩🟨⬛⬛ 🟨🟩🟩⬛⬛ ⬛🟩🟩⬛🟩 🟩🟩🟩🟩🟩
1600680774788734977 2022-12-07 21:36:50 -0500 <whooomeee> Wordle 537 4/6*  ⬜🟨⬜⬜🟨 ⬜⬜⬜🟩🟩 🟨⬜🟨🟩🟩 🟩🟩🟩🟩🟩  なんか閃いた
1600680742912405505 2022-12-07 21:36:42 -0500 <julieclarenbach> If you’re respecting the digital strike tomorrow re: NYT and you have a crossword streak going, tomorrow’s puzzle drops in about 30 minutes, and then you can do Friday on F
riday and maintain the streak. No work around for Wordle or the Bee, unfortunately.
1600680704521957376 2022-12-07 21:36:33 -0500 <SaraSalley> I’m ready 💪🏼  Wordle 536 3/6  🟨🟩🟩⬜🟨 🟨🟩🟩🟨⬜ 🟩🟩🟩🟩🟩
1600680684959535104 2022-12-07 21:36:29 -0500 <aansagitaade> Ronaldo Malgy Razzien Hazard しずちゃん Grey Stage 6 ポルトガル2部オリベイレンセ入団 Jennifer Lawrence Unschuldsvermutung FFAA Contreras Wordle 537 Ivan Chilled Cumbria Tyleno
l Biagio outback queensland Elanga Verfassungsschutz  https://t.co/LzhMrjvrCz
1600680663342092289 2022-12-07 21:36:24 -0500 <debbieflint> Wordle 537 5/6  ⬜⬜⬜🟨⬜ ⬜⬜🟨🟨⬜ 🟨🟨⬜⬜🟨 🟩🟩⬜🟩🟩 🟩🟩🟩🟩🟩 Unlucky fourth guess :-)
1600680645604544512 2022-12-07 21:36:19 -0500 <pwr2dppl> Break your #wordle streak.
1600680598552444928 2022-12-07 21:36:08 -0500 <yachirin> Wordle 537 3/6  ⬛🟨🟨⬛⬛ ⬛🟨🟨🟩🟩 🟩🟩🟩🟩🟩
1600680593020227584 2022-12-07 21:36:07 -0500 <jeannakadlec> no wordle or spelling bee tomorrow, friends!
1600680573689032709 2022-12-07 21:36:02 -0500 <emv52> Avoid Wordle Thursday #Solidarity @RockpathWV @WSpriggs @driscollc @NHLABOR_NEWS
1600680558082019329 2022-12-07 21:35:58 -0500 <1980Dorothy> Wordle 536 4/6  ⬛⬛⬛🟩🟩 ⬛🟩⬛🟩🟩 ⬛🟩⬛🟩🟩 🟩🟩🟩🟩🟩
1600680488452210688 2022-12-07 21:35:42 -0500 <cathylanski> Wordle 536 3/6*  ⬜⬜⬜⬜🟩 🟨⬜🟩⬜🟩 🟩🟩🟩🟩🟩
1600680478171836416 2022-12-07 21:35:39 -0500 <SparklayBarclay> Wordle 536 3/6  ⬛⬛⬛⬛🟩 🟨⬛🟨🟨🟩 🟩🟩🟩🟩🟩
1600680477127757824 2022-12-07 21:35:39 -0500 <feneriss> Wordle Türkçe 537 3/6  🟩⬛⬛⬛🟨 🟩⬛🟨🟩⬛ 🟩🟩🟩🟩🟩   https://t.co/06UudBzAoG
1600680476888694784 2022-12-07 21:35:39 -0500 <m0ncher1> Wordle 537 6/6  ⬜⬜⬜🟨🟨 ⬜⬜⬜🟩🟩 ⬜⬜⬜🟩🟩 🟨⬜🟨🟩🟩 ⬜🟨🟩🟩🟩 🟩🟩🟩🟩🟩
1600680472635445254 2022-12-07 21:35:38 -0500 <stigaasheim> Wordle 537 6/6  ⬛⬛⬛⬛⬛ ⬛⬛🟨⬛🟩 ⬛⬛⬛🟩🟩 ⬛🟨🟨🟩🟩 🟩🟩⬛🟩🟩 🟩🟩🟩🟩🟩
1600680456214962177 2022-12-07 21:35:34 -0500 <basultoshow> Wordle 537 🙏🔽🎵ありがと✨🤝   https://t.co/ZidqceVwbP
1600680443233255424 2022-12-07 21:35:31 -0500 <Aee2005> @AdmiralBear01 @draculaic @atomsareenough @TwistNHook @katster @paleodaniel @BerkeleyFog @peter_symonds @califorleanian @Lesmitch529 @terencelau @nyt_wordle_bot @BethLapachet @Anth
onyMaul @avinashkunnath @Fyght4Cal @CNHarder @Andrea_Joh @reneissance @NorCalNickW4C @donglover @aazlant @Sling @Disney My music tastes have changed since middle school. Once in a while I put on some BSB for old times sake.
1600680435918307337 2022-12-07 21:35:29 -0500 <Reauxtweets> Wordle 536 2/6  🟨⬛🟨🟨🟩 🟩🟩🟩🟩🟩   https://t.co/PmwoDA0vJS
1600680435180527616 2022-12-07 21:35:29 -0500 <XShipper> Wordle 537 6/6  ⬛🟨⬛⬛⬛ ⬛🟨⬛⬛⬛ ⬛⬛🟨🟩⬛ 🟨⬛⬛🟩⬛ 🟨🟨⬛🟩🟩 🟩🟩🟩🟩🟩  What is that?
```

Data Transformation

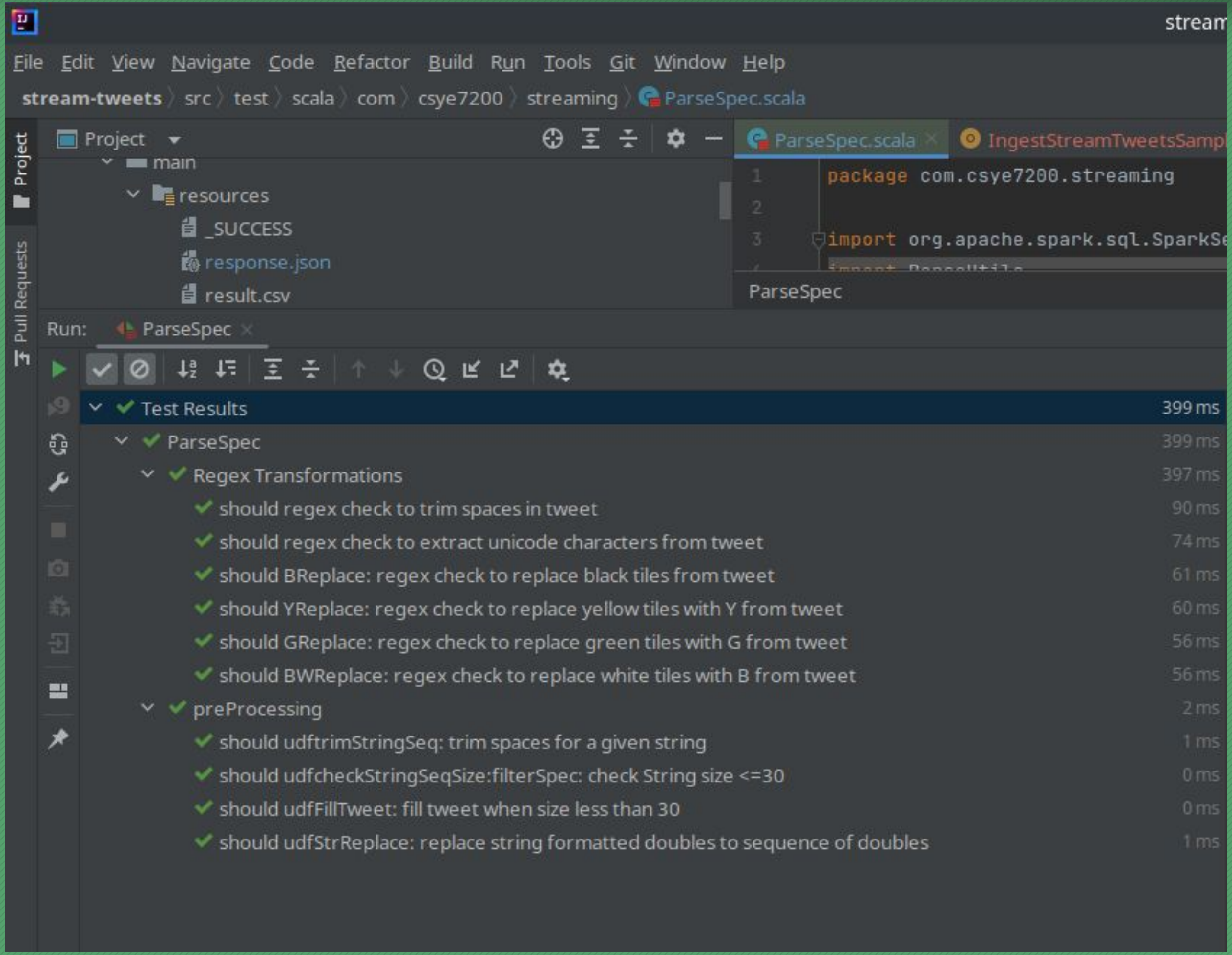
- Data extracted via twint is extremely raw; unicode formatted data. A series of transformations done to compute wordle sequence vectors.
  - Regex transformations: series of regex extract and replacements to decode unicode characters.
  - Formatting the extracted string to the right format to computer vectors; fill a partial wordle result to a full flattened matrix (5*6): vector size 30
  - Computing a vector of doubles for distance calculation from a string sequence representing wordle result.
- Compute all pair user distances for users each day before writing to database

Snapshot of transformed data



created_at	date	timezone	username	tweet	tweetClean
1670369209000	2022-12-06 18:26:49	-500	marioskbar	-0.3,-0.3,-0.3,-0.3,0.5,1.0,0.5,-0.3,-0.3,-0.3,1.0,-0.3,-0.3,1.0,1.0...	⬜⬜⬜⬜🟨🟩🟨⬜⬜⬜🟩⬜⬜🟩🟩🟩🟩🟩🟩🟩
1670369185000	2022-12-06 18:26:25	-500	nicoleaabbott	-0.3,-0.3,-0.3,1.0,1.0,-0.3,0.5,-0.3,1.0,1.0,1.0,-0.3,-0.3,1.0,1.0,1...	⬜⬜⬜🟩🟩⬜🟨⬜🟩🟩🟩⬜⬜🟩🟩🟩⬜⬜🟩🟩🟩🟩🟩🟩🟩
1670369203000	2022-12-06 18:26:43	-500	Maurizio	-0.3,-0.3,-0.3,-0.3,-0.3,-0.3,0.5,-0.3,-0.3,-0.3,0.5,-0.3,1.0,-0.3,1....	⬜⬜⬜⬜⬜⬜🟨⬜⬜⬜🟨⬜🟩⬜🟩🟩🟩🟩🟩🟩
1670369196000	2022-12-06 18:26:36	-500	cherekay	-0.3,-0.3,1.0,-0.3,-0.3,-0.3,1.0,1.0,-0.3,1.0,1.0,1.0,1.0,1.0,1.0,0....	⬜⬜🟩⬜⬜⬜🟩🟩⬜🟩🟩🟩🟩🟩🟩
1670369176000	2022-12-06 18:26:16	-500	milner2	0.5,0.5,0.5,-0.3,-0.3,1.0,0.5,-0.3,1.0,-0.3,1.0,1.0,1.0,1.0,1.0,0.0...	🟨🟨🟨⬜⬜🟩🟨⬜🟩⬜🟩🟩🟩🟩🟩
1670369250000	2022-12-06 18:27:30	-500	mums_hugs	-0.3,-0.3,-0.3,1.0,-0.3,-0.3,-0.3,1.0,1.0,-0.3,1.0,1.0,1.0,1.0,1.0,0...	⬜⬜⬜🟩⬜⬜⬜🟩🟩⬜🟩🟩🟩🟩🟩
1670369210000	2022-12-06 18:26:50	-500	Poubelle_T	0.5,-0.3,-0.3,-0.3,-0.3,0.5,-0.3,0.5,-0.3,1.0,-0.3,1.0,-0.3,1.0,1.0,...	🟨⬜⬜⬜⬜🟨⬜🟨⬜🟩⬜🟩⬜🟩🟩⬜🟩⬜🟩🟩⬜🟨⬜⬜🟩🟩🟩🟩🟩🟩
1670369213000	2022-12-06 18:26:53	-500	saruhashy	-0.3,-0.3,-0.3,-0.3,-0.3,-0.3,1.0,-0.3,-0.3,-0.3,1.0,1.0,-0.3,-0.3,0....	⬜⬜⬜⬜⬜⬜🟩⬜⬜⬜🟩🟩⬜⬜🟨🟩🟩⬜⬜🟨🟩🟩🟩🟩🟩
1670369267000	2022-12-06 18:27:47	-500	m_masa_m...	0.5,-0.3,-0.3,-0.3,-0.3,-0.3,1.0,1.0,1.0,-0.3,1.0,1.0,1.0,1.0,1.0,0....	🟨⬜⬜⬜⬜⬜🟩🟩🟩⬜🟩🟩🟩🟩🟩
1670369274000	2022-12-06 18:27:54	-500	artymaggie	-0.3,-0.3,0.5,-0.3,-0.3,0.5,0.5,-0.3,-0.3,-0.3,1.0,-0.3,0.5,0.5,0.5,...	⬜⬜🟨⬜⬜🟨🟨⬜⬜⬜🟩⬜🟨🟨🟨🟩🟨⬜⬜🟩🟩🟩🟩🟩🟩
1670369184000	2022-12-06 18:26:24	-500	peachiekeehn	-0.3,-0.3,-0.3,-0.3,-0.3,-0.3,0.5,0.5,1.0,-0.3,0.5,0.5,0.5,1.0,-0.3,...	⬜⬜⬜⬜⬜⬜🟨🟨🟩⬜🟨🟨🟨🟩⬜🟩🟩🟩🟩🟩
1670369169000	2022-12-06 18:26:09	-500	EaboutTV	-0.3,0.5,-0.3,-0.3,-0.3,1.0,-0.3,-0.3,-0.3,-0.3,1.0,-0.3,0.5,1.0,-0....	⬜🟨⬜⬜⬜🟩⬜⬜⬜⬜🟩⬜🟨🟩⬜🟩⬜⬜🟩🟩🟩🟩🟩🟩🟩
1670369267000	2022-12-06 18:27:47	-500	_kaihime	-0.3,-0.3,0.5,0.5,0.5,0.5,1.0,1.0,0.5,-0.3,-0.3,1.0,1.0,1.0,1.0,-0....	⬜⬜🟨🟨🟨🟨🟩🟩🟨⬜⬜🟩🟩🟩🟩⬜🟩🟩🟩🟩🟩🟩🟩🟩🟩
1670369099000	2022-12-06 18:24:59	-500	CharlesFran...	-0.3,-0.3,-0.3,-0.3,-0.3,-0.3,1.0,-0.3,1.0,1.0,-0.3,1.0,-0.3,1.0,1.0...	⬜⬜⬜⬜⬜⬜🟩⬜🟩🟩⬜🟩⬜🟩🟩🟩🟩🟩🟩🟩
1670369167000	2022-12-06 18:26:07	-500	lonelywxh	-0.3,-0.3,-0.3,-0.3,1.0,-0.3,-0.3,1.0,-0.3,1.0,-0.3,-0.3,1.0,1.0,-0....	⬜⬜⬜⬜🟩⬜⬜🟩⬜🟩⬜⬜🟩🟩⬜⬜🟩🟩🟩🟩🟩🟩🟩🟩🟩
1670369239000	2022-12-06 18:27:19	-500	giraffegirl68	1.0,-0.3,0.5,0.5,-0.3,1.0,1.0,1.0,1.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0...	🟩⬜🟨🟨⬜🟩🟩🟩🟩🟩
1670369198000	2022-12-06 18:26:38	-500	ogbenyiegbe	-0.3,0.5,-0.3,1.0,1.0,-0.3,-0.3,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,0.0...	⬜🟨⬜🟩🟩⬜⬜🟩🟩🟩🟩🟩🟩🟩🟩
1670369157000	2022-12-06 18:25:57	-500	wiccan95	-0.3,-0.3,-0.3,0.5,-0.3,0.5,-0.3,-0.3,-0.3,-0.3,0.5,0.5,0.5,-0.3,-0....	⬜⬜⬜🟨⬜🟨⬜⬜⬜⬜🟨🟨🟨⬜⬜⬜🟩⬜🟩🟩⬜🟩🟩🟩🟩🟩🟩🟩🟩🟩
1670369275000	2022-12-06 18:27:55	-500	star_of_bba	-0.3,-0.3,-0.3,-0.3,-0.3,0.5,1.0,1.0,-0.3,-0.3,-0.3,1.0,1.0,1.0,-0.3...	⬜⬜⬜⬜⬜🟨🟩🟩⬜⬜⬜🟩🟩🟩⬜⬜⬜⬜⬜🟩🟩🟩🟩🟩🟩
1670369206000	2022-12-06 18:26:46	-500	masamune_...	-0.3,-0.3,-0.3,1.0,-0.3,-0.3,1.0,1.0,0.5,-0.3,-0.3,1.0,1.0,1.0,1.0,-...	⬜⬜⬜🟩⬜⬜🟩🟩🟨⬜⬜🟩🟩🟩🟩⬜🟩🟩🟩🟩🟩🟩🟩🟩🟩
1670369124000	2022-12-06 18:25:24	-500	jbcampo1	-0.3,-0.3,-0.3,-0.3,0.5,-0.3,0.5,-0.3,0.5,-0.3,0.5,0.5,0.5,-0.3,-0.3...	⬜⬜⬜⬜🟨⬜🟨⬜🟨⬜🟨🟨🟨⬜⬜🟩🟨🟨⬜⬜🟩⬜⬜🟩🟩🟩🟩🟩🟩🟩
1670369108000	2022-12-06 18:25:08	-500	juztimepasss	-0.3,-0.3,-0.3,-0.3,-0.3,-0.3,-0.3,1.0,-0.3,-0.3,1.0,-0.3,-0.3,-0.3,-...	⬜⬜⬜⬜⬜⬜⬜🟩⬜⬜🟩⬜⬜⬜⬜🟨🟩🟩⬜⬜🟩🟩🟩🟩
1670369149000	2022-12-06 18:25:49	-500	californiabarby	1.0,-0.3,-0.3,-0.3,-0.3,1.0,-0.3,0.5,0.5,-0.3,1.0,1.0,1.0,1.0,1.0,0....	🟩⬜⬜⬜⬜🟩⬜🟨🟨⬜🟩🟩🟩🟩🟩

Tests: Parse Tweets



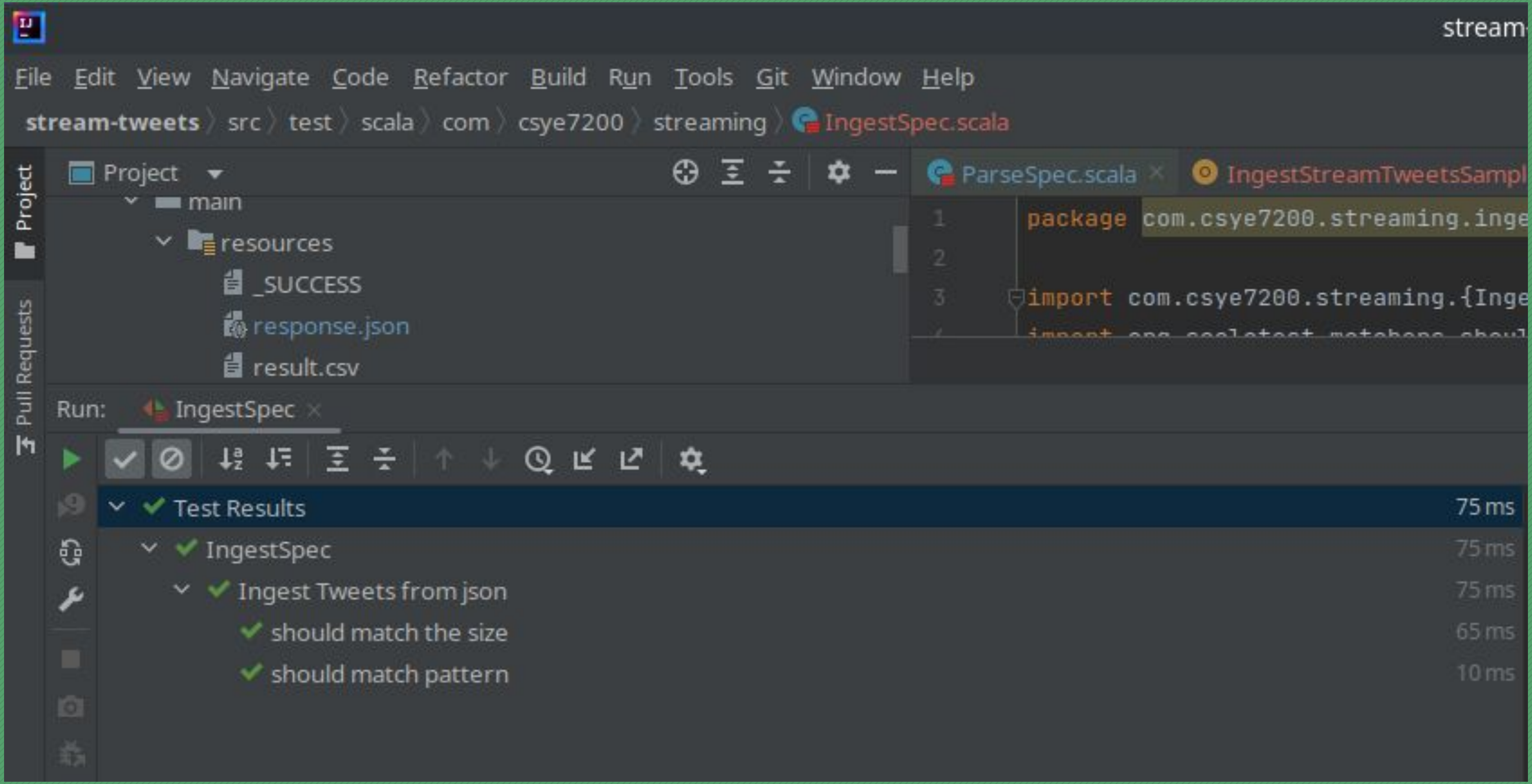
stream
File Edit View Navigate Code Refactor Build Run Tools Git Window Help
stream-tweets 〉 src 〉 test 〉 scala 〉 com 〉 csye7200 〉 streaming 〉 ParseSpec.scala
Project
Pull Requests
main
resources
_SUCCESS
response.json
result.csv
ParseSpec.scala
IngestStreamTweetsSampl
ParseSpec
Run: ParseSpec

```
package com.csye7200.streaming

import org.apache.spark.sql.SparkSe
```

```
Test Results                                                                        399 ms
  ParseSpec                                                                         399 ms
    Regex Transformations                                                           397 ms
      should regex check to trim spaces in tweet                                     90 ms
      should regex check to extract unicode characters from tweet                    74 ms
      should BReplace: regex check to replace black tiles from tweet                 61 ms
      should YReplace: regex check to replace yellow tiles with Y from tweet         60 ms
      should GReplace: regex check to replace green tiles with G from tweet          56 ms
      should BWReplace: regex check to replace white tiles with B from tweet         56 ms
    preProcessing                                                                     2 ms
      should udftrimStringSeq: trim spaces for a given string                         1 ms
      should udfcheckStringSeqSize:filterSpec: check String size <=30                 0 ms
      should udfFillTweet: fill tweet when size less than 30                          0 ms
      should udfStrReplace: replace string formatted doubles to sequence of doubles   1 ms
```

Tests: Ingest Tweets


stream
File Edit View Navigate Code Refactor Build Run Tools Git Window Help
stream-tweets 〉 src 〉 test 〉 scala 〉 com 〉 csye7200 〉 streaming 〉 IngestSpec.scala
Project
Pull Requests
ParseSpec.scala
IngestStreamTweetsSampl
main
resources
_SUCCESS
response.json
result.csv
Run:
IngestSpec

```
package com.csye7200.streaming.inge

import com.csye7200.streaming.{Inge
```

```
Test Results                         75 ms
  IngestSpec                         75 ms
    Ingest Tweets from json          75 ms
      should match the size          65 ms
      should match pattern           10 ms
```

Milestones

- Sprint 1 (11/6–11/12): Define requirements and create repository
- Sprint 2 (11/13 – 11/19): Data collection
- Sprint 3 (11/20 – 11/26): Data collection and data cleaning
- Sprint 4 (11/27 – 12/3): Data visualization
- Sprint 5 (12/4 – 12/8): Finalize Data visualization

Use of Scala

- Spark consumer gets data from Kafka
- Scala processes and transforms before writing it to the database
- Spark parallelly computes user similarity scores using Euclidean distance (best metric evaluated).
- Databricks visualizes data with Scala.

Clustering of users - Additional

- Given the users, and the distance between each pair of users, use the k-means clustering algorithm to find clusters.
- Use the in-built mllib.clustering library in Databricks.
- Idea is to find the optimal number of centroids by iterating through a range, computing the error (SSE).
- Find the number of centroids which produce the least error.
- Using this, visualize the clusters graphically.

Goals

- Demonstrate uses of Scala and Spark by creating an end-to-end project that extracts, transforms and visualizes data from an API.
- Create a similarity matrix among all users that shows how similar a player's wordle attempts were to other players.
- Visualize the top 10 performing players for a given user.
- (If time permits) Create graphs of commonly occurring words. (Did not achieve)

Acceptance Criteria

- 1 million rows are extracted. (Actual: 2500 per day, more than 117k rows after distance calculations)
- User search response in 5 seconds. (Actual: 4.73 seconds on Databricks community)
- Distance metric gives a similarity score of at least 75% in the range of a base metric. (Actual: N/A since we chose to use Euclidean distance)
- Data is extracted twice a day. (Actual: Data is extracted once a day, user distances calculated eod)