Wyatt Queirolo

**Instagram Scraper V0.1 (Proof of Concept)**

**Features:**

* Extracts first ~12 posts from Instagram user
* Parses user metadata
* Inserts all data into neo4j database

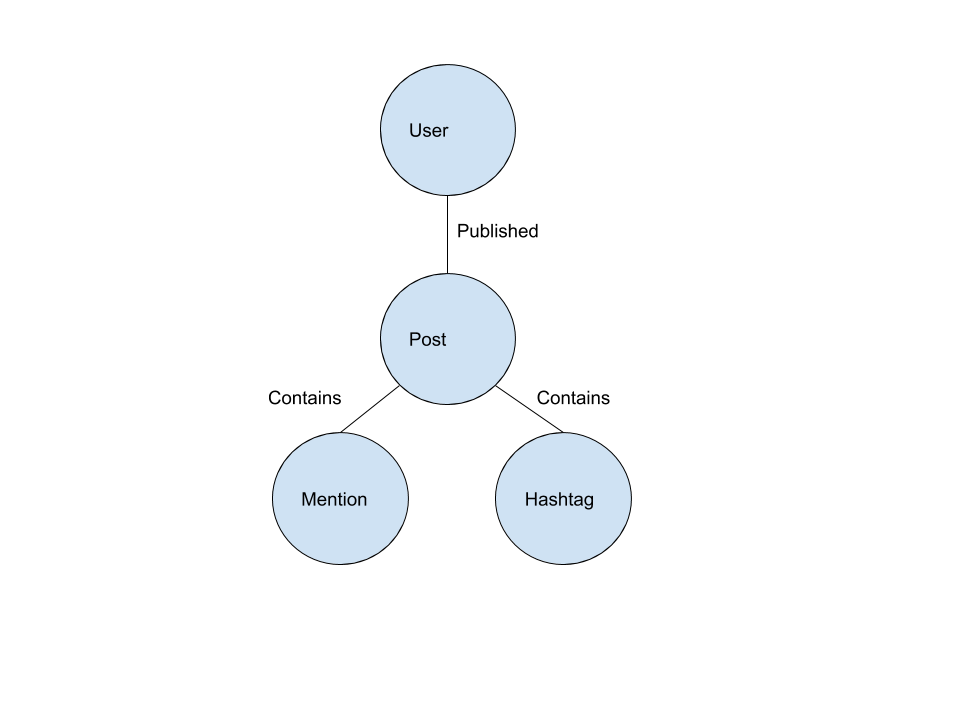
**Next Steps:**

* Iterate through rest of each users posts to extract them all
* Buff out error handling
* Build out neo4j driver functionality to allow for updating of data
* Format hashtags/other text so they are not case sensitive
* Decide on deployment architecture
* Various scraping improvements

**Instructions/requirements:**

* Currently uses python3, neo4j running locally, housed in a virtual environment
* Edit config.py file with local configuration
* run python scraper.py

**Neo4j Schema**



Note – each node is a json file containing data on the user or a post

**Example Queries:**

Number of Sponsored Posts from lifedevivi(can easily be modified to return the specific posts):

WITH ['#ad','#sponsored','#spon'] as tags MATCH (:Person{username:'lifedevivi'})--(a:Post)--(h:Hashtag) where h.tag in tags RETURN COUNT(a)

Result: 9

What influencers has @forever21 been mentioned by?

MATCH (p:Person)--(:Post)--(:Mention{mention:"@forever21"}) return p

Result: list of influencers (can order by number of posts that mention)

What hashtags are most often posted with “@forever21”

MATCH (m:Mention{mention:'@forever21'})--(p:Post)--(h:Hashtag) with count(p) as num,h return h,num ORDER BY num DESC

Result: list of hashtags, highest ranked ones were #mammothmountain, #f21xme

What influencers use a hashtag in common with “thenotoriouspug”?

WITH ['#ad','#sponsored'] as ad MATCH (:Person{username:'thenotoriouspug'})--(:Post)--(h:Hashtag)--(:Post)--(p:Person) WHERE NOT h.tag in ad with h,p,count(h) as num RETURN h,p,num ORDER BY num DESC

Result: itsdanieljae, itsbowtie both took the top spots using #danielwellington. In this way graph databases are often used as recommendation engines in that they can find similarities between different nodes.