FYI Code Files & Process:

Files:

1. process.py (Preprocessing file)
2. kewsearch.py (Create Dictionary)
3. search\_engine.py (Fuzzy String Matching) --Only this file and JSON needed to run code

Process:

1. Start with preprocessing Instagram and twitter data.
2. Clean text (lower case, stop words, slangs, punctuations, digits), remove duplicates, drop empty and unknown rows.
3. Inner-Join Twitter & Instagram data.
4. Send preprocess data to ‘kewsearch.py’ module.
5. Create a dictionary of Influencer\_ID as key and text as values.
6. Create dictionary of products in ‘kewsearch.py’ by combining data from 3 major sources:
7. Amazon product database
8. Best Buy ecommerce database
9. Dataturk product database
10. CountVectorize (Similar to Jaswanth’s algorithm) values to find keywords from product and user dictionaries.
11. Save new dictionaries thus created.
12. Send these dictionaries to search\_engine.py
13. Gather input from user.
14. Find the input in product dictionary using Fuzzy Token Set Ratio.
15. Use the values so obtained in creating a new set of product keyword values.
16. Find the Influencer\_IDs by using Fuzzy Partial Ratio over the set of product keywords.
17. Count the number of occurrences of user keywords.
18. Sort them by highest occurrences.
19. Return highest ‘n’ occurred values.