Text, logo

Description automatically generated

Exercise: Grouping and ranking Twitter posts

As background, on the chance you’ve never used Twitter, people will often “retweet” or re-post messages from another user’s timeline to their own timeline. This then allows that user’s followers to see a message they might not otherwise have seen. A few years ago, Twitter implemented an official mechanism to support this practice, but some people persist in the “old school” method of retweeting by prepending the characters “RT” and re-posting the message manually.

For instance, if @BarackObama posts:

Time’s running out: Sign up by March 31st to have health coverage in 2014. ofa.bo/rLw #LastCall2014

Someone doing an old-style retweet might post:

RT @BarackObama: Time’s running out: Sign up by March 31st to have health coverage in 2014. ofa.bo/rLw

#LastCall2014

to his or her own stream of Twitter posts.

Of course, as might already be apparent, this creates a lot of Twitter posts that are, at a database or technical level, distinct, but that are, conceptually, the same Twitter post appearing many times. This is where you come in. Attached to this email is a JSON file with a few hundred posts from Twitter with a subset of the metadata returned by Twitter. Many of these posts are old-style manual retweets.

We’d like you to tackle two questions:

1. Finding the “true” number of posts, by grouping the Twitter posts together if they are retweeting or commenting on the same message. Your method should be robust to small changes, so that if someone introduces a small typo or uses a retweet prefix or postfix that’s not in the set of data provided, your technique for grouping would still work.
2. Ranking these groups of Twitter posts by popularity.

Neither question has an obviously correct answer. In that vein, it would be useful for us if you explained how and why you chose the rules you did for excluding or including posts in a given group, and how and why you defined the popularity of a group of posts they way you did.