実験詳細

1. Twitterで有料モードで収集されたツイートからジオタグ付きのツイートを取ってくる。counting\_geo.py

有料モードで収集されたものはrawファイル。そこからファイルごとにカウントする。

count\_geo\_per\_file method で集められたのがTotal 3004877でユーザ毎に上で作られたファイルから（geotag付きのみ）ツイート数をカウントする。

sum\_count\_geo method そこからトップユーザー４万人を選出する。ジオタグ付きのツイートをよく発しているトップユーザー。top\_geo\_existing\_user\_id\_40000.json

blacklistはbotとかが入ってる。つまり2000ツイート以上を発しているアカウント。

1. check\_users.py で実際に生きているユーザーリストを作る。db.lookupsに保存された。67268 existing users (como guarde encima del experimento anterior es mas de 40000 usuarios. Tengo que hacer de vuelta???)
2. collecting\_tweet\_with\_API.py ~~でこれらのユーザのツイートを2018 年4月13日から2018年4月21日までの期間収集した~~．

保存されたCollectionは以下の通り：

db.tweets\_from\_follow\_users0: 2926 registros

db.tweets\_from\_follow\_users1: 1459 registros

db.tweets\_from\_follow\_users2: 1021 registros

db.tweets\_from\_follow\_users3: 890 registros

db.tweets\_from\_follow\_users4: 665 registros

db.tweets\_from\_follow\_users5: 2672 registros

db.tweets\_from\_follow\_users6: 1949 registros

db.tweets\_from\_follow\_users7: 839 registros

Se puede recolectar en paralelo utilizando 4 cuentas de API cambiando los parámetros de la cuenta, el nombre del archivo

El comando line se escribe de esta forma: python collecting\_tweet\_with\_API.py 0

API 1

0: 0 a 5000

1: 5000 a 10000

API 2

2: 10000 a 15000

3: 15000 a 20000

API 3

4: 20000 a 25000

5: 30000 a 35000

API 4

6: 35000 a 40000

それらのツイートを取った後collect\_follow\_relationships methodでフォローユーザーのデータもとる。そしてdb.idsに格納。これもまた複数のAPIアカウント並行で行う。

それからmake\_user\_list\_for\_experiment methodでタイムウィンドウを作ってjson fileに出力する。