BIG DATA ANALYSIS

LAB01: Data gathering and storage from Twitter

For this practical work, the following hashtags have been used, and a total number of **10.300 tweets** have been captured and stored in a Mongo Database.

#agriculture
#agricultura

Capturing that number of tweets has been very slow at a speed of 2 to 3 tweets per minute, as shown in the following log:

```
…
Tweet collected at Sat Oct 19 07:00:00 +0000 2019 from user @agricultura2000
Tweet collected at Sat Oct 19 07:00:19 +0000 2019 from user @KellyHewsonF
Tweet collected at Sat Oct 19 07:00:25 +0000 2019 from user @CopCVL
Tweet collected at Sat Oct 19 07:02:06 +0000 2019 from user @sdgnigeria
Tweet collected at Sat Oct 19 07:02:35 +0000 2019 from user @GdeMolliens
Tweet collected at Sat Oct 19 07:04:26 +0000 2019 from user @ABMT_ke
Tweet collected at Sat Oct 19 07:04:34 +0000 2019 from user @DebbieSparkes
Tweet collected at Sat Oct 19 07:05:04 +0000 2019 from user @PatrickHUGUES
Tweet collected at Sat Oct 19 07:07:21 +0000 2019 from user @themoonwalking
Tweet collected at Sat Oct 19 07:07:31 +0000 2019 from user @farm_yellow
Tweet collected at Sat Oct 19 07:08:15 +0000 2019 from user @farming_
Tweet collected at Sat Oct 19 07:08:23 +0000 2019 from user @NBTWORLDNews
Tweet collected at Sat Oct 19 07:08:29 +0000 2019 from user @farming_
Tweet collected at Sat Oct 19 07:08:49 +0000 2019 from user @WattcomFR
...
```

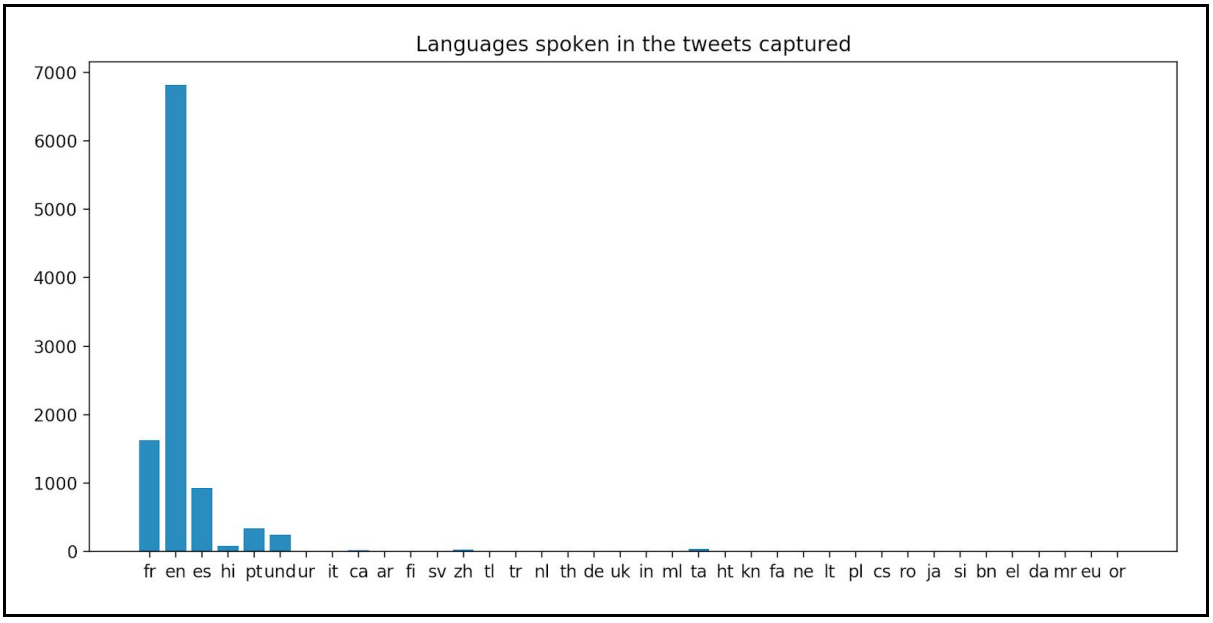
From the tweets captured, the plots shown in the following pages have been generated, as well as recycled from the given code.

This report and the Python files can be found at the following git repository:
https://github.com/ovalls/mai_bda/tree/master/Lab01

Languages spoken in the tweets captured

As expected, given the chosen hashtags, english is the most common language for the tweets, followed by many distance by french and spanish. Far from them is portuguese, although there are also some instances of hindi and chinese.

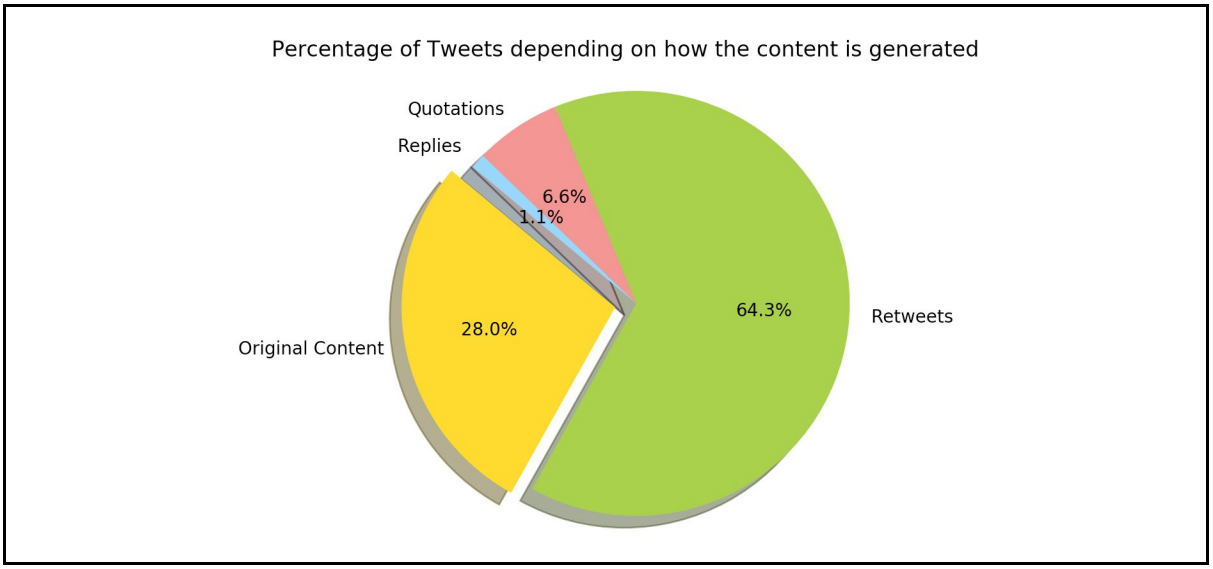
Some tweets don't have a determined language; those are marked as "und", as well as some have "ta" language which couldn't be found in twitter developer help: https://developer.twitter.com/en/docs/developer-utilities/supported-languages/api-reference/get-help-languages



Percentage of tweets depending on how the content is generated

Using the given code, it's shown that only 28% of the captured tweets are original content, while more than double of it (64.3%) are retweets; quotations and replies are 7.7% of the generated content.

This shows that most of the content on Twitter is generated via retweeting other's original tweets.

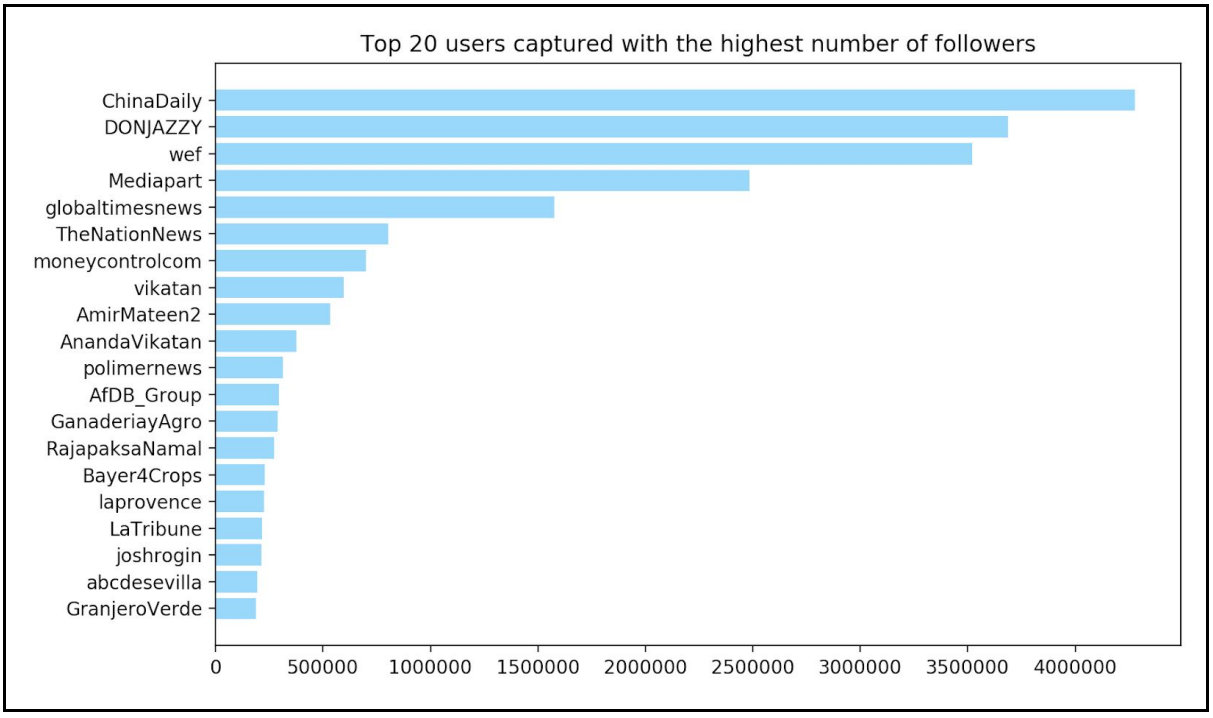


Top 20 users captured with the highest number of followers

From the captured tweets, the most influencer user, China Daily, has over 4 Million followers, followed by DONJAZZY and wef with more than 3 Million followers.

The top 5 have more than 1.5 Million followers each, then the numbers descend quite progressively until the twentieth position where GranjeroVerde has almost 200.000 followers.

It's interesting to see how only analyzing 20 positions in the chart, given the previous hashtags, such a difference in number of followers is detected.



The exact number of followers for the top 20 is the following:

```
ChinaDaily: 4277145
DONJAZZY: 3687795
wef: 3520281
Mediapart: 2483297
globaltimesnews: 1574796
TheNationNews: 802448
moneycontrolcom: 700101
vikatan: 594690
AmirMateen2: 534154
AnandaVikatan: 375321
polimernews: 312557
AfDB_Group: 295968
GanaderiayAgro: 288864
RajapaksaNamal: 271678
Bayer4Crops: 229077
laprovence: 225925
LaTribune: 217396
joshrogin: 214124
abcdesevilla: 193729
GranjeroVerde: 186539
```

Top 20 hashtags for climatic, ecologic and organic issues

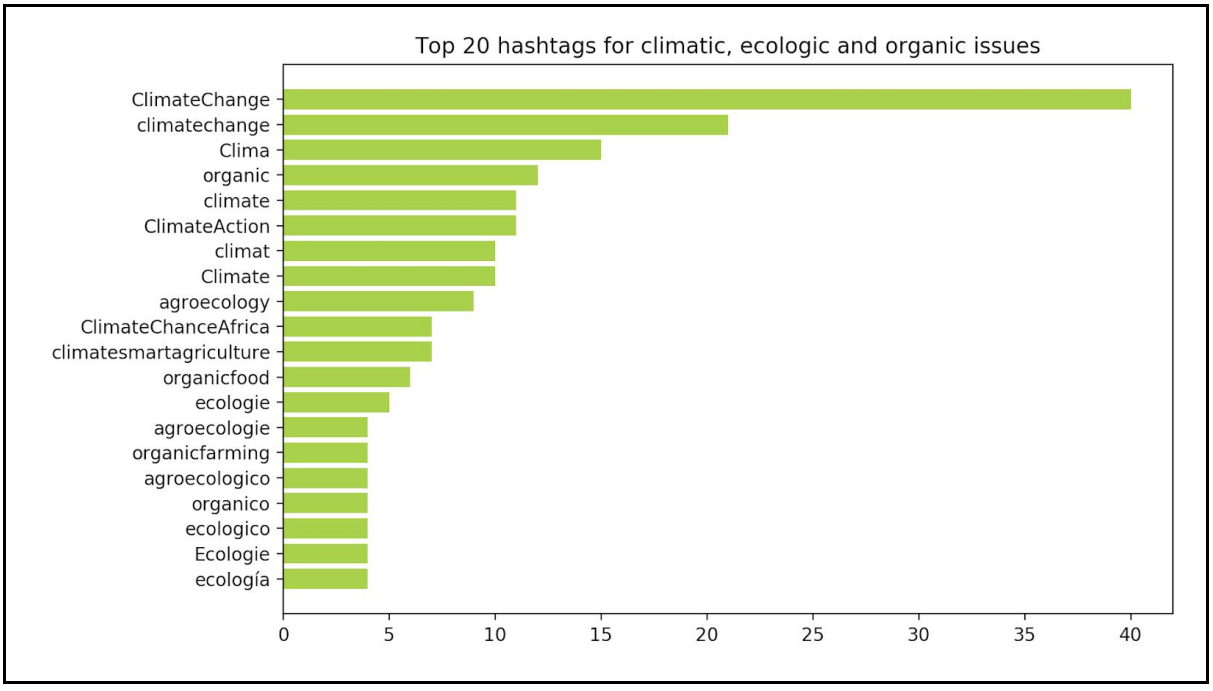
The following strings have been chosen to search for other hashtags in tweets captured:

- 'Clima': to include hashtags such as clima, climate...
- 'Ecol': to include hashtags such as ecologie, ecologic, ecológico...
- 'organic'

From the total 10.300 tweets, the top 20 of hashtags dealing with those subjects sum up 192, which is 1.86% of the total number of tweets.

ClimateChange seems to be the “trending topic” regarding those issues with a total number of 61 tweets.

The detail of the 192 tweets regarding each of the subjects would be: clima 132 (68.75%), ecology 34 (17.71%) and organic 26 (13.54%).



The exact number for the different top 20 hashtags on those subjects is the following:

```
climatechange: 21
ClimateChange: 40
Clima: 15
organic: 12
climate: 11
ClimateAction: 11
climat: 10
Climate: 10
agroecology: 9
ClimateChanceAfrica: 7
climatesmartagriculture: 7
organicfood: 6
ecologie: 5
agroecologie: 4
organicfarming: 4
agroecologico: 4
organico: 4
ecologico: 4
Ecologie: 4
ecología: 4
```