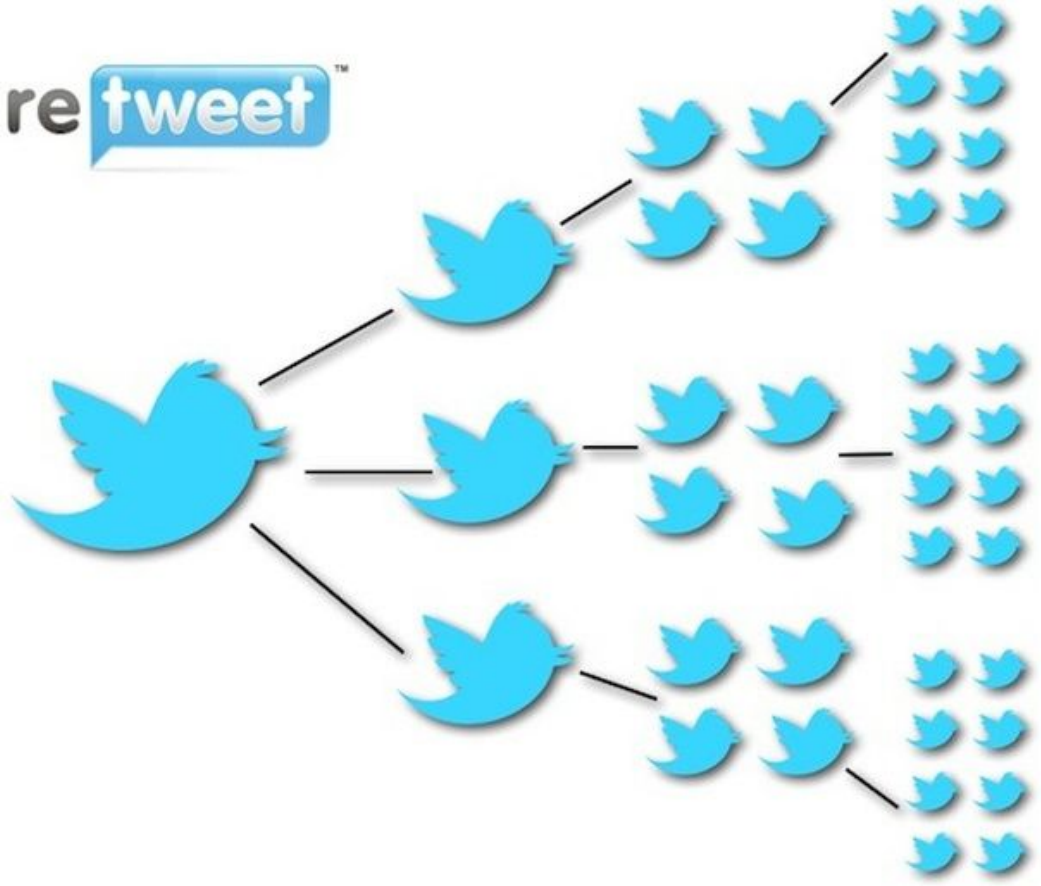
Redes Complexas
2017-2018
Project - Part 2



Grupo 3
Daniel Correia 80967
Pedro Orvalho 81151
Renato Cardoso 81530

1. Introdução

Neste projeto, decidimos explorar um grande volume de dados do Twitter recolhidos durante o Mundial de 2014[1] no âmbito de projetos orientados pelo professor Alexandre Francisco.

A exploração dos dados do Twitter teve os seguintes objetivos:

1. Estudar as propriedades gerais da rede obtida através dos retweets
2. Analisar os perfis de utilizadores durante o Mundial
3. Avaliar a propagação de hashtags e de modelos de epidemia

O principal problema que tivémos em todas as etapas do projeto (extração dos dados comprimidos, pré-processamento, análise e visualização) resume-se a uma única questão: “Como processar grandes volumes de dados em tempo útil?”.

O código desenvolvido será disponibilizado em anexo no zip entregável (onde estarão também gráficos adicionais).

2. Caracterização dos dados

Os dados do Twitter utilizados foram recolhidos entre 13 de Março e 15 de Julho de 2014 no âmbito do Mundial de 2014. Tivémos acesso aos dados através de 419 arquivos comprimidos de formato GZIP, sendo que cada arquivo tinha em média 1 milhão de tweets (400-500 MB cada), formando um total de 166 GB de dados comprimidos no estado inicial do projeto.

Cada tweet disponível nos dados recolhidos apresentava uma formatação do tipo JSON definida pela API do Twitter[2] contendo informação sobre o conteúdo do tweet, o autor do tweet, relações com outros tweets ou utilizadores do sistema (e.g. retweet e reply), entre outras informações.

Neste projeto, escolhemos analisar apenas os retweets não só por questões de performance dos processos de tratamento de dados mas também por ser o mecanismo mais interessante para avaliar situações de propagação e modelos de epidemia relativamente aos hashtags.

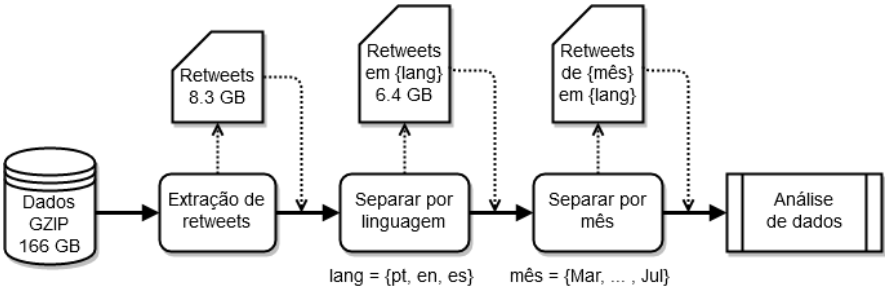
Assim, os atributos dos dados extraídos foram:

Atributo	**Significado**
id_str	id do retweet
user_id_str	id do utilizador que criou este retweet
created_at	timestamp da criação do retweet
lang	linguagem atribuída ao retweet
hashtags	hashtags associados ao retweet
retweeted_status_id_str	id do tweet original (“retweetado”)
retweeted_status_user_id_str	id do utilizador do tweet original
retweeted_status_created_at	timestamp de criação do tweet original

[1] https://en.wikipedia.org/wiki/2014_FIFA_World_Cup
[2] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.html

3. Extração e pré-processamento dos dados

Devido ao grande volume de dados, todo o processo de extração e pré-processamento teve como objetivo reduzir o volume de dados para que as ferramentas de análise funcionem em tempo útil.



O primeiro passo foi extrair apenas os retweets de todos os tweets existentes nos arquivos GZIP. Para isso utilizámos um script[3] de Python e uma biblioteca de paralelismo para distribuir a computação pelo máximo de cores possíveis. No entanto, nos nossos computadores pessoais com 8 cores e 8 GB de RAM, a extração de um ficheiro GZIP de 400 MB demorava mais de 30 minutos.

Por isso, utilizando um cluster disponibilizado pelo professor Alexandre Francisco que já tinha 64 cores e 256 GB de RAM conseguimos optimizar o processo de extração para 5 minutos por GZIP, resultando num total de 36 horas para processar os 419 arquivos GZIP.

De seguida, realizámos o pré-processamento dos retweets extraídos, o que nos permitiu obter apenas os retweets escritos em Inglês, Português, e Espanhol (en,pt,es) e separá-los por mês. Deste modo, reduzimos de 166 GB para 6.4 GB de dados separados por ficheiros em média de 500 MB.

4. Estatísticas dos dados

Relativamente aos dados extraídos, o número total de retweets foi 61 milhões, sendo que 56% foram feitos em Junho e 26% em Julho. Para além disso, de todos os retweets aqueles que foram feitos em Inglês, Espanhol, Português (en,es,pt) correspondem a 77% da amostra, o que nos permitiu ter alguma confiança na nossa decisão de pré-processamento por estas linguagens.

A linguagem mais representada nos dados foi o Inglês, contendo 52% da amostra total de dados disponível com 32 milhões de retweets, seguida do Espanhol com 20% seguida do Português com 4%.

Com base nas contagens, podemos verificar que o número de tweets originais distintos utilizados como fonte de retweet foi em média de 14% em todos os meses. Isto significa que a maioria dos retweets foram feitos a partir dos mesmo tweets originais (o que poderá ser significativo quando analisarmos situações de propagação de informação na rede ou atividade dos utilizadores).

Finalmente, analisando as contagens de utilizadores distintos, verificámos que em média o número total de retweets em cada mês corresponde ao dobro do número de utilizadores, ou seja, em média cada utilizador fez dois retweets por mês.

[3] Ver em anexo ficheiro parse_tweet.py

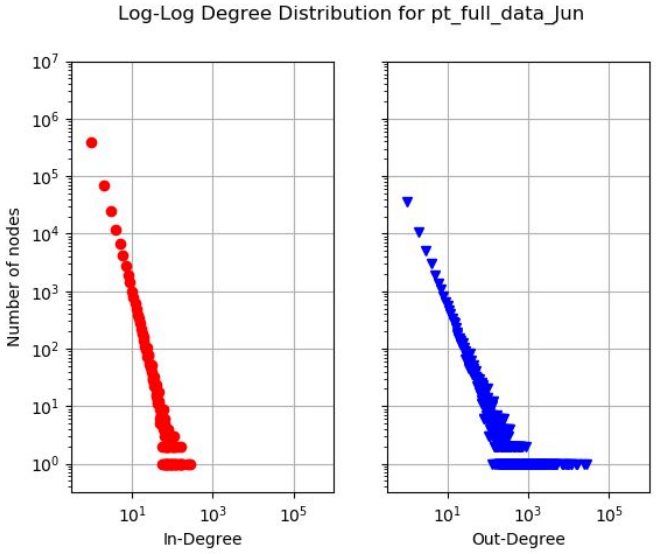
5. Análise da rede

Nesta fase do projeto, decidimos construir uma rede a partir dos retweets utilizando os utilizadores com nós e o retweet como ligação. Escolhemos usar um grafo direcionado em que as ligações têm origem no utilizador do tweet original e destino no utilizador que fez o retweet.

Em termos de tecnologias, tentámos usar a package Networkx (tema do 1º projeto) mas não foi possível porque as redes obtidas em cada mês eram muito grandes e o Networkx carrega a rede toda em memória. Por isso, recorremos a duas packages alternativas: igraph[4] e graph-tool[5].

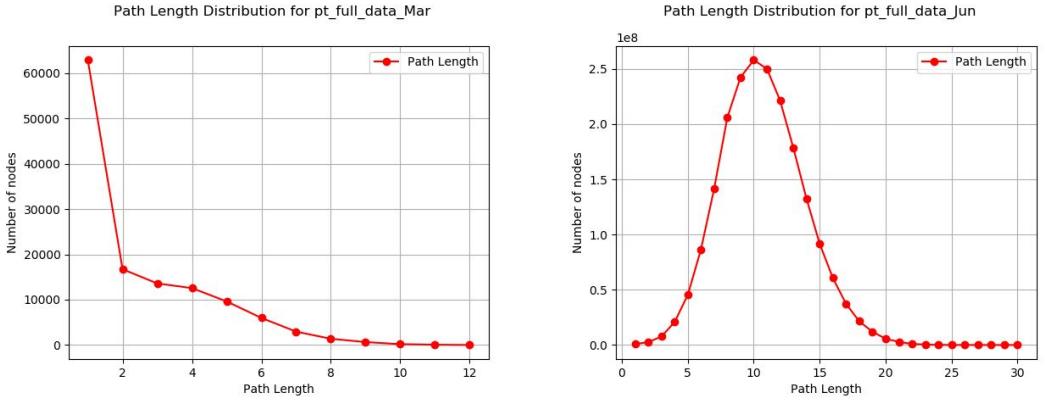
5.1. Degree

Analisando o degree da rede, verificámos a rede se aproxima de uma rede scale-free, o que seria expectável dado que estamos lidar com uma rede social onde prevalecem as propriedades de preferential attachment e porque a rede foi construída a partir dos retweets (logo, os hubs serão os utilizadores originais dos tweets). Como podemos observar no gráfico seguinte, a degree distribution é semelhante a uma power-law, como seria de esperar.



5.2. Shortest Path Length

Analisámos também a distância entre nós na rede e verificámos que para um número baixo de retweets a distância era próxima de 1 devido à existência de poucos hubs. No entanto, à medida que aumentamos a quantidade de retweets, o valor da distância satura em 10 porque a maior parte dos nós na rede são eles próprios hubs de tweets o que significa que a inserção de novos nós tem um menor impacto na distância.

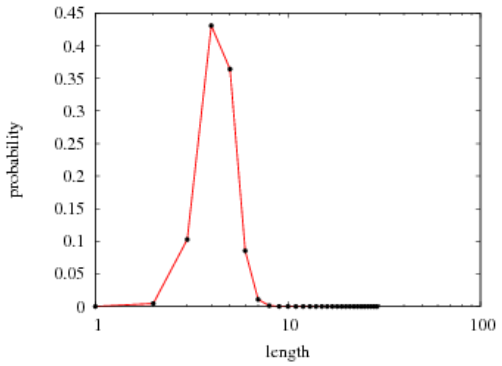


[4] http://igraph.org/python/
[5] https://graph-tool.skewed.de/

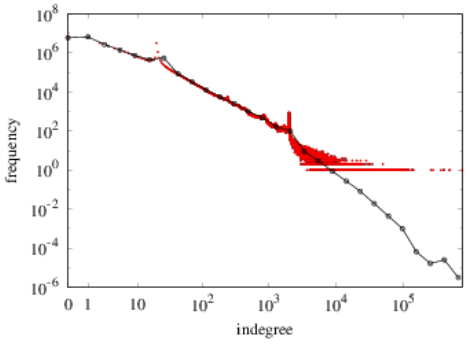
5.3. Comparação com dados do WebGraph

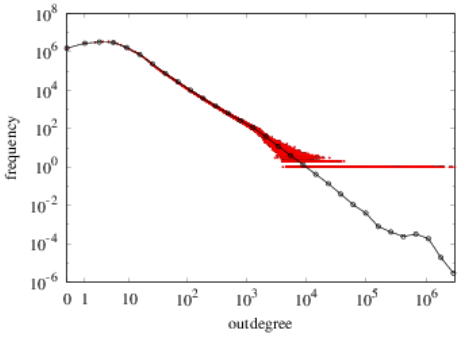
Comparando os nossos resultados com os disponíveis no WebGraph[6] para um dataset do Twiiter, podemos verificar que o valor da distância mais curta entre nós satura mais cedo do que na nossa rede em valores 4-5.

Isto deve-se ao facto de a nossa rede ter sido construída a partir de retweets e a rede do WebGraph ser uma rede dos followers. Logo, cada hub da rede WebGraph tem muito mais ligações que os hubs da nossa rede, o que resulta numa menor distância entre nós.



A nível do degree, os resultados foram bastantes semelhantes obtendo-se degree distributions com curvas de power-law no mesmo grau de magnitude da nossa rede. Existe apenas um aumento no indegree no caso do WebGraph, pelas mesmas razões da distância.

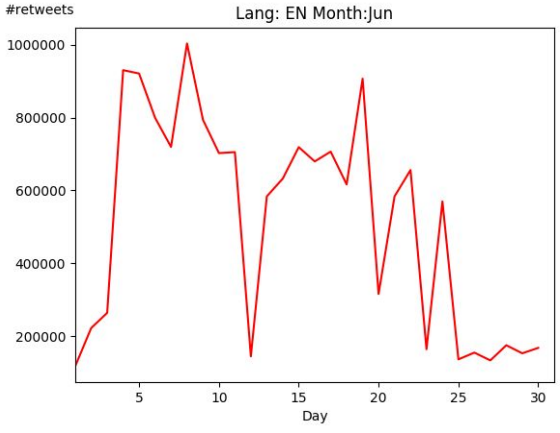




6. Análise de utilizadores

6.1. Processamento dos Dados

Após termos separado os dados por mês e posteriormente por língua, procedemos à contagem de retweets por cada utilizador em cada dia de cada mês, gerando gráficos dos respectivos meses para cada língua, os dados obtidos são maioritariamente da língua inglesa e a maioria dos retweets feitos em inglês ocorreram em Junho, cuja distribuição se encontra no gráfico em baixo.

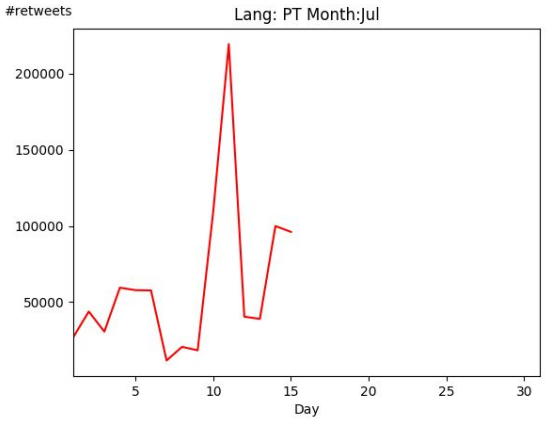


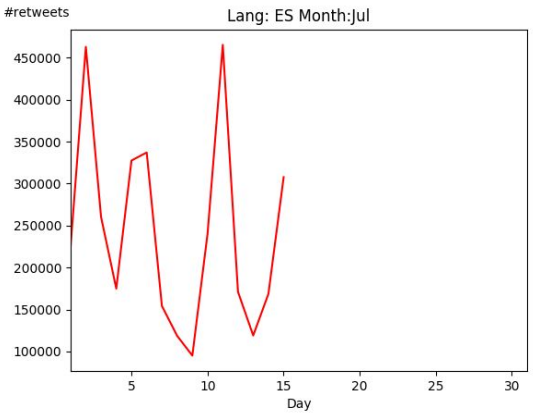
[6] http://law.di.unimi.it/webdata/twitter-2010/

Posteriormente procedemos à recolha de quais os utilizadores que fizeram mais de 50, 75 ou 150 retweets por mês. Para isto utilizámos a biblioteca pandas do python, para fazer a escolha dos dados.
A ideia era comparar a distribuição de retweets no mês com a os retweets feitos pelos utilizadores mais activos, escolhidos aleatoriamente.

6.2. Análise Mensal dos Retweets por Língua

A nossa análise mensal de retweets em cada língua consistiu em, sem retirar quaisquer utilizadores, verificar quantos retweets foram feitos em cada dia de cada mês e tentar entender essas ocorrências.
Por exemplo como podemos ver pelo gráfico anterior, o maior pico de retweets que houve foi dia 8 de Junho que correspondeu à Fan Fest no Brazil[7]. Podemos também observar que outro grande pico foi dia 19 de Junho que correspondeu ao jogo entre Inglaterra e Uruguai[8].
Nas outras línguas as distribuições são semelhantes mudando somente os picos, pois estes normalmente ocorrem quando são jogos de equipas das línguas respectivas ou no dia anterior.
Por exemplo o caso da língua portuguesa o maior número de retweets foi no dia 11 de Julho dia anterior à final entre o 3º e 2º lugar, Brasil e Holanda[9], como podemos ver no gráfico seguinte à esquerda. No caso da língua espanhola, a maior propagação de retweets também se deu em Julho, com dois dias passando os 450 K retweets, como se pode ver no gráfico seguinte à direita.





Nestes meses, Junho e Julho, houve uma grande propagação de tweets comparando com os meses anteriores. Isto muito provavelmente deve-se a estes meses, Junho e Julho, corresponderem à fase final do campeonato enquanto os outros meses corresponderem à fase de qualificação.
A diferença entre estas duas fases é bastante significativa, (como se pode ver nos gráficos em anexo):

- Na língua inglesa houve um aumento no máximo de de retweets passando de 200-225K nos primeiros meses no campeonato para 900K-1M na fase final do campeonato.
- Na língua espanhola passou-se de 100-160K de retweets na fase de qualificação para 300-450K nos meses da fase final.
- Na língua portuguesa houve um aumento de 12-20K de retweets na fase de qualificação para 100-210K na fase final do campeonato.

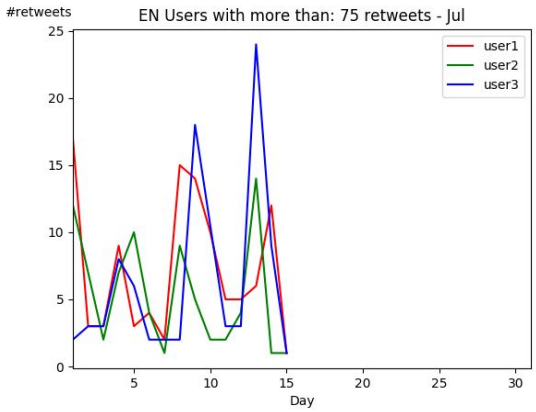
[7] http://www.fifa.com/worldcup/news/y=2014/m=6/news=ronaldo-kicks-off-fifa-fan-fest-2360141.html
[8] https://www.fifa.com/worldcup/matches/round=255931/match=300186486/report.html
[9] https://www.fifa.com/worldcup/matches/round=255957/match=300186502/report.html

6.3. Análise Mensal de Retweets por Utilizador

Analisámos os números de retweets feitos pelos utilizadores mais activos de cada língua em cada mês, isto é, os utilizadores que fizeram mais de 50, 75 ou 150 retweets num mês. Desses dados escolhemos sempre 3 utilizadores escolhidos aleatoriamente[10] para verificarmos se a distribuição dos seus retweets consistia com a distribuição mensal de todos os utilizadores.

Tentámos também descobrir utilizadores que sofressem do actualmente denominado como "*Trump Effect*" um utilizador que quando está activo faz muitos tweets/retweets e depois ausenta-se durante umas horas/dias, ou simplesmente tem uma actividade bastante mais reduzida.

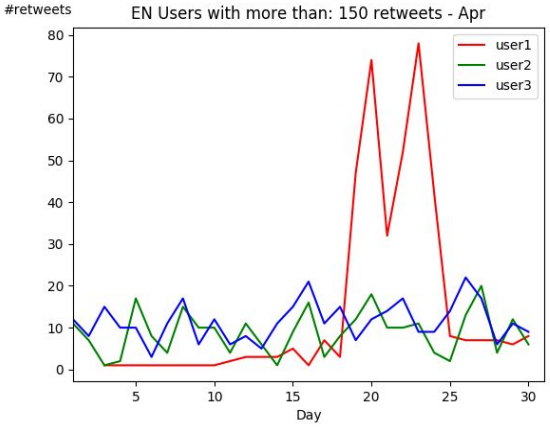
6.3.1. Retweets dos Utilizadores mais activos VS Mensais



Ao extrairmos os dados aleatoriamente dos utilizadores mais activos fomos comparar se as distribuições dos seus retweets eram similares às distribuições dos retweets de todos os utilizadores, o que nem sempre se verifica.

Porém pudemos verificar que entre estes utilizadores as distribuições dos retweets algumas vezes eram similares como podemos ver pelo gráfico ao lado que representa os retweets ingleses de 3 utilizadores que fizeram pelo menos 75 retweets no mês de Julho, e podemos verificar que os picos de retweets acontecem nos mesmos dias ou em dias consecutivos o que pode corresponder aos eventos finais do campeonato.

6.3.2. Trump Effect vs Bots



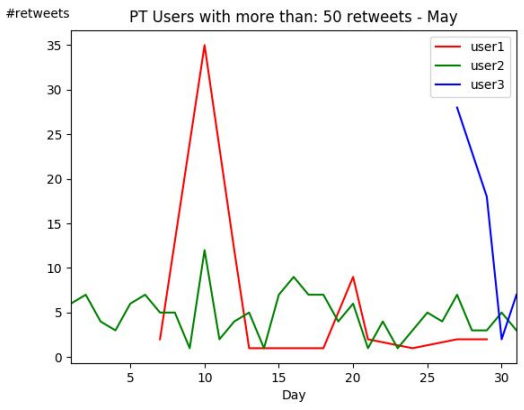
Em relação ao *Trump Effect* já mencionado analisámos este efeito só nos utilizadores mais activos pois os outros não tinham grande interesse pois a grande maioria dos utilizadores nem chega a 50 retweets por mês.

Ao analisar os gráficos obtidos encontramos alguns utilizadores que em alguns dias passavam de 0-5 retweets para 70-80 como é o caso do user 1 no gráfico ao lado de 3 utilizadores que fizeram mais de 150 retweets em inglês no mês de Abril.

Outros dois casos são o user 1 e o user 2 do gráfico ao lado em baixo, utilizadores que fizeram pelo menos 50 retweets em Português no mês de Maio. O user1 até ao dia 6 /7 não tinha feito retweets e do nada passou de 2-3 no dia 6 para 35 no dia 7. O user2 até ao dia 28 não tinha feito qualquer retweet e no dia 28 fez 27.

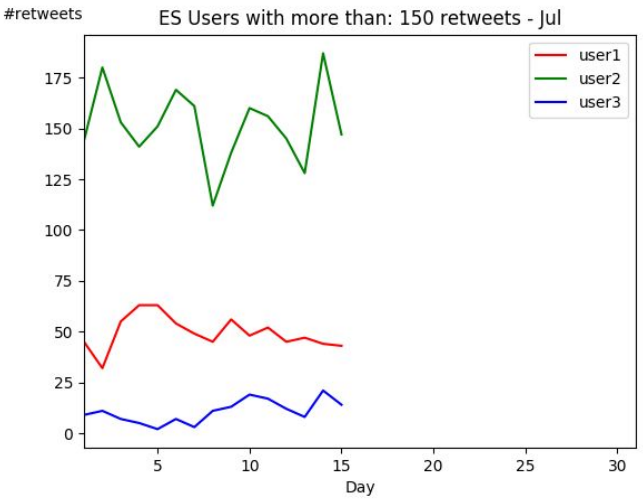
Acreditamos que estas anomalias se devem a algum evento concreto que causou grande emoção ao utilizador, por exemplo, uma vitória ou um golo.

Porém acreditamos também que nem todas estas anomalias se devem a utilizadores humanos como é o caso do próximo gráfico de utilizadores que fizeram pelos menos 150 retweets em espanhol no mês de Julho.



[10] https://docs.python.org/2/library/random.html

Acontece que se realmente se tratar de um utilizador humano, tirando o tempo para dormir (7 horas) e fazendo uma média de 170 retweets por dia, como é o caso do user2, isto significa que o user2 faz 10 retweets à hora todos os dias durante 17 horas.
Deste modo achamos que alguns dos utilizadores mais activos do twitter a que tivemos acesso se tratam de bots e não de utilizadores humanos, não contando assim para o estudado *Trump Effect.*



7. Análise de hashtags

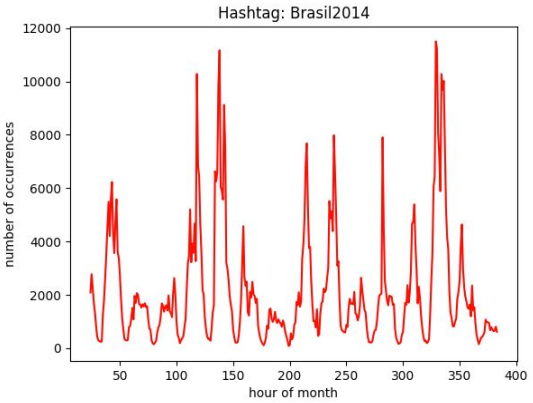
7.1. Resultados estatísticos

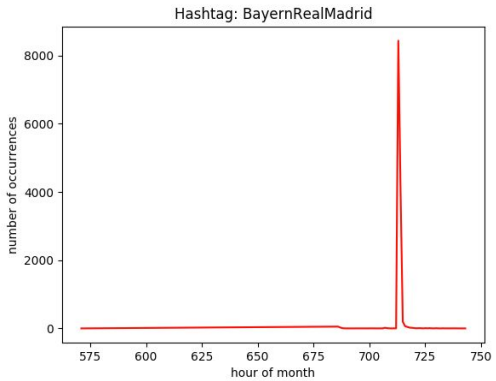
Inicialmente, realizámos uma avaliação estatística dos dados de forma a decidir qual a melhor abordagem a tomar. Começamos por contar qual o número de vezes que um determinado hashtag aparece num dataset. Reparamos que existiam hashtags que predominavam em relação a outros, desta forma decidimos fazer um ranking com o top 10 hashtags em cada mês por cada língua.

7.2. Propagação de um hashtag

Depois de decidirmos qual o top de hashtags decidimos ver a utilização de um hashtag por hora, sendo que ficamos com vários tipos de hashtags, uns que são usados durante todo o mês, outros são apenas usados num dado momento e ainda há alguns que aparecem em pares.

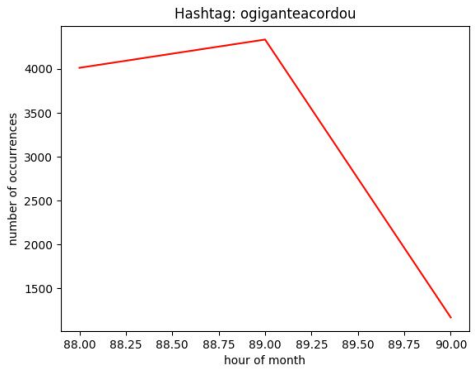
Os hashtags que sao usados durante todo o mês normalmente estão relacionados com um acontecimento que dura o mês inteiro e são mais comuns no mês de Junho e Julho, isto pode acontecer pois como o contexto dos tweets é o mundial é normal que nos meses do mundial os top hashtags sejam usados regularmente. Este tipo de hashtag apresenta um gráfico com vários picos a acontecerem regularmente.

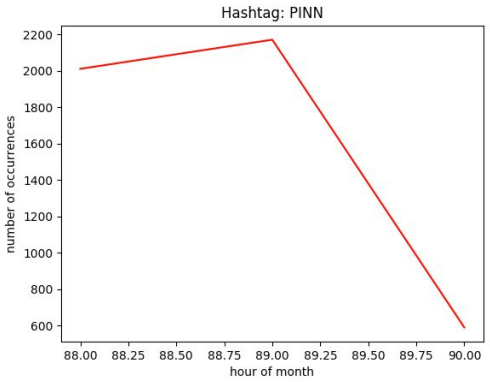




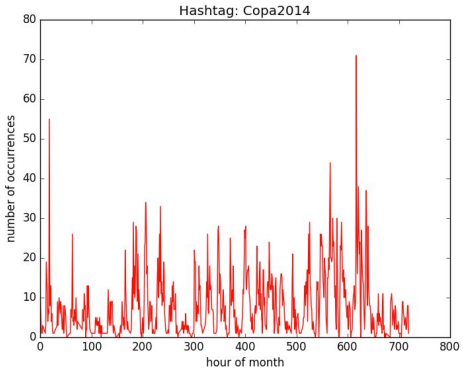
Por outro lado, os hashtags que apenas são usados num dado momento apresentam um ou poucos picos, que normalmente estão relacionados com um acontecimento específico. Este tipo de hashtag aparece mais em meses afastados do mundial, onde um acontecimento consegue chegar ao top com apenas um pico.

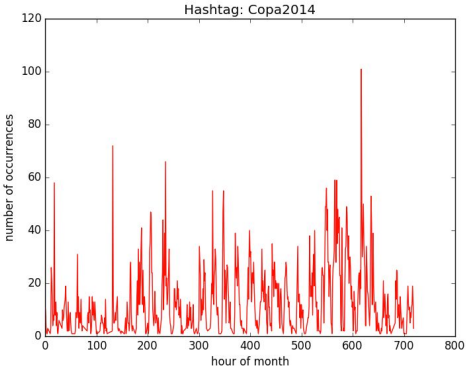
O último tipo de hashtag são hashtags correlacionados, em que os gráficos são bastante parecidos e como o anterior estão normalmente relacionados com um acontecimento específico pois aparecem ao mesmo tempo.





Para além do número de hashtags por hora vimos também o número de utilizadores únicos (utilizadores que estão a usar este hashtag pela primeira vez) por hora, em cada mês, para determinar a aderência a um hashtag.
Decidimos comparar os gráficos obtidos inicialmente com os gráficos obtidos para os utilizadores únicos.
Em regra geral, a forma do gráfico mantinha-se constantes, havendo diminuição de valores, principalmente nos picos, comparado com os gráficos obtidos anteriormente, o que quer dizer que normalmente uma pessoa faz um tweet com um hashtag, ou seja o valor de tweets é aproximadamente o número de pessoas que aderiram a um hashtag .





Porém, algo peculiar que aconteceu foi a existência de utilizadores a fazerem tweets em menos de um segundo, desta forma eles acabavam por nunca ser utilizadores únicos nos intervalos considerados, havendo uma grande discrepância entre o número de tweets com esse hashtag e o número de utilizadores únicos.