Trabalho 4 - Desenvolvimento de Software para Nuvem

1. Ambiente utilizado

- AWS S3 para armazenamento dos dados.
- Jupyter Notebooks com Linguagem Scala e Framework Spark para o processamento dos dados.
- Jupyter Notebooks com Linguagem Python e Framework Pandas para criação dos gráficos e tabelas.
- Os notebooks Scala foram usados em um cluster do AWS Elastic Map Reduce com a seguinte configuração:
    - 1 Master m5.xlarge
    - 2 Cores m5.xlarge
- Versão do EMR: 6.1.0
- Versão do Spark: 3.0.0
- Versão do Pandas: 1.0.5

---

2. Primeira Questão

2.1 Pré-Processamento dos Dados

O dataset da primeira questão possui 32 colunas, mas a maioria delas não seria útil para responder às perguntas propostas, por conta disso várias colunas foram dropadas. Em seguida, foram feitas operações para gerar as colunas *date* e *id*. O dataset pré-processado foi salvo no S3 no formato *parquet*, que é o formato favorito do Spark.

No começo do notebook questao-1 foi criada uma coluna chamada *hashtags*, que possui as hashtags utilizadas em cada tweet.

O dataset a ser utilizado possui as colunas: *id*, *tweets*, *date* e *hashtags*.

2.2 a) Quais foram as hashtags mais usadas pela manhã, tarde e noite?

Para responder essa pergunta foi necessário:

- Retirar os campos que não tinham hashtags.
- Pegar os registros com *date* entre 5 e 12 horas (manhã) ou 13 e 18 horas (tarde) ou 19 e 4 horas (noite).

- Agrupar e agregar pela coluna *hashtags*.

As 15 hashtags mais usadas pela manhã

hashtag	contagem
#EMABiggestFansJustinBieber	23385
#EMABiggestFans1D	23058
#VoteVampsTeenAwards	736
#trndnl	603
#QueroNoTVZ	523
#AustinMahone	492
#TwOff	435
#EMABiggestFansJustinBieber"""	423
"#EMABiggestFans1D"	416
"#EMABiggestFansJustinBieber"	408
#EMABiggestFansJustinBieber	401
"#EMABiggestFans1D"""	353
#bomdia	322
#bomdia	298
#NowIsGoodNaGlobo	274

As 15 hashtags mais usadas pela tarde

	hashtag	contagem
#EMABiggestFans1D		59270
#EMABiggestFansJustinBieber		50209
#StealMyGirl		5476
#QueroNoTVZ		4266
#bigpaynodanceoff		1353
#EMABiggestFans1D"		1284
"#EMABiggestFansJustinBieber"""		1283
#demiyourstorydoesntdefineyou		961
#HottieOfTheWeek		875
"#EMABiggestFansJustinBieber"		853
#AustinMahone		788
#LuanSantanaNaHoraDoFaro		719
#EMABiggestFansArianaGrande		714
#AMAs		693
#DomingoPreguiçosoDoSDVcomValentino		678

As 15 hashtags mais usadas pela noite

	hashtag	contagem
#EMABiggestFans1D		114729
#EMABiggestFansJustinBieber		114277
#camilasayshi		9043
#EMABiggestFansJustinBieber"""		3331
#DebateNoSBT		2773
#CartersNewVideo		2681
#bigpaynodanceoff		2476
#Vote5HEMA		2410
#TheVoiceBrasil		2379
#LuanSantanaNaHoraDoFaro		2191
#AssistamODR		2123
"#EMABiggestFans1D"		2102
"#EMABiggestFansJustinBieber"		2039
#DebateNaRecord		1815
#LinkinParkNoMultishow		1764

2.3 b) Quais as hashtags mais usadas em cada dia?

Para responder essa pergunta foi necessário:

- Retirar os campos que não tinham hashtags.
- Pegar os registros de cada dia específico (15, 16, 17, 18, 19 e 20).
- Agrupar e agregar pela coluna *hashtags*.

As 15 hashtags mais usadas no dia 15/10/2014

hashtag	contagem
#EMABiggestFans1D	29711
#EMABiggestFansJustinBieber	24981
#StealMyGirl	5823
#bigpaynodanceoff	3737
#AssistamODR	2015
#BuyLoveMeHarderOniTunes	1276
#UnlockMockingjay	1112
#Vote5HEMA	1058
#EMABiggestFansJustinBieber\"\"\"	987
#QueroMuitosSeguidoresComValentino	807
#StealMyGirl	488
#EMABiggestFansJustinBieber”	472
#EMABiggestFans5SOS	469
"#EMABiggestFansJustinBieber\"	433
#StealMyGirl	397

As 15 hashtags mais usadas no dia 16/10/2014

hashtag	contagem
#EMABiggestFans1D	63198
#EMABiggestFansJustinBieber	53203
#camilasayshi	9087
#DebateNoSBT	2689
#CartersNewVideo	2681
#AustinMahone	1420
#EMABiggestFansJustinBieber\"\"\"	1350
"#EMABiggestFans1D\"	1287
#MasterChefBR	1158
"#EMABiggestFans1D\"\"\"	1147
#HottieOfTheWeek	1098
"#EMABiggestFansJustinBieber\"	1000
#askjadelittlemix	989
#LuaBlancoNoAgoraETarde	954
#camilasayhi	883

As 15 hashtags mais usadas no dia 17/10/2014

hashtag	contagem
#EMABiggestFansJustinBieber	44672
#EMABiggestFans1D	44050
#QueroNoTVZ	2868
#TheVoiceBrasil	2370
#FlyNoMixDiario	1380
#EMABiggestFansJustinBieber\"\"\"	1334
#LançamentoDoClipeVocêSeFoiFLY	1316
#AustinMahone	1123
"#EMABiggestFans1D\"	904
#HottieOfTheWeek	769
#SextaTodosSDVcomValentino	725
"#EMABiggestFansJustinBieber\"	685
#ChristianNoEncontroComFatima	674
#AmorESexo	671
#FlyNoMixDiario	664

As 15 hashtags mais usadas no dia 18/10/2014

hashtag	contagem
#EMABiggestFans1D	25039
#EMABiggestFansJustinBieber	24936
#QueroNoTVZ	2336
#demiyourstorydoesntdefineyou	952
#WeWantZaynsSongsInFOUR	700
#MaratonaTwilight	672
#SigaTiuMarkitoNesseSabadoSdv	592
#trndnl	591
#EMABiggestFans1D\"	570
"#EMABiggestFansJustinBieber\"\"\"	526
#premiomultishow	447
#AMAs	420
#MCGuimeNoRaulGil	374
#KCAArgentina	365
"#EMABiggestFansJustinBieber\"	359

As 15 hashtags mais usadas no dia 19/10/2014

hashtag	contagem
#EMABiggestFansJustinBieber	30375
#EMABiggestFans1D	27948
#LuanSantanaNaHoraDoFaro	2910
#LinkinParkNoMultishow	1767
#demiyourstorydoesntdefineyou	1150
#VoteVampsTeenAwards	881
#ComeToBrazilMattEspinosa	845
#DomingoPreguiçosoDoSDVcomValentino	820
#ChatApimentadoDoBiel	698
#EMABiggestFansJustinBieber\"	680
"#EMABiggestFansJustinBieber\"\"\"	638
#VamosLaU	616
#trndnl	574
#AcreditamosEmVoceLL	561
#FOLLOWMELOHANTHONY	557

As 15 hashtags mais usadas no dia 20/10/2014

hashtag	contagem
#EMABiggestFansJustinBieber	9704
#EMABiggestFans1D	7111
#DebateNaRecord	1760
#debatenarecord	1556
#QueroDilmaTreze	1418
#EmTodoBrasilAecio45	638
#AecioEmTodoBrasil	607
#CongratsOn1MChris	447
#Los80	381
#PreOrderVampsUSA	363
#KCAArgentina	320
#twdstrangers	315
#austinmahonechile	281
#BuyNobodyOniTunes	280
#NoEntiendo	247

2.4 c) Qual o número de tweets por hora a cada dia?

Para responder essa pergunta foi necessário:

- Pegar os registros de cada dia específico (15, 16, 17, 18, 19 e 20).
- Criar uma coluna chamada *hora* a partir da coluna *date*.
- Agrupar pela coluna *hora* e agregar pela coluna *tweets*.

Número de tweets por hora no dia 15/10/2014

hora	número de tweets
14	34378
15	79157
16	78353
17	83950
18	77713
19	65095
20	66813
21	79270
22	86030
23	97574

Número de tweets por hora no dia 16/10/2014

hora	número de tweets
0	110232
1	163338
2	176211
3	124599
4	77743
5	42661
6	22228
7	10157
8	8327
9	23616
10	37528
11	42716
12	49156
13	55958
14	66154
15	76891
16	79276
17	65824
18	57098
19	58547
20	68148
21	88473
22	99443
23	93350

Número de tweets por hora no dia 17/10/2014

hora	número de tweets
0	107984
1	142770
2	153981
3	122205
4	76951
5	42222
6	20985
7	10695
8	8016
9	20559
10	35147
11	40918
12	48140
13	59851
14	66972
15	74748
16	74909
17	65007
18	58841
19	57831
20	65602
21	53146
22	53497
23	54053

Número de tweets por hora no dia 18/10/2014

hora	número de tweets
0	55404
1	61050
2	63501
3	56562
4	42778
5	37148
6	27473
7	17724
8	12108
9	11539
10	17456
11	27491
12	41485
13	58383
14	71402
15	75206
16	72334
17	69001
18	63700
19	61727
20	59165
21	60665
22	65102
23	68663

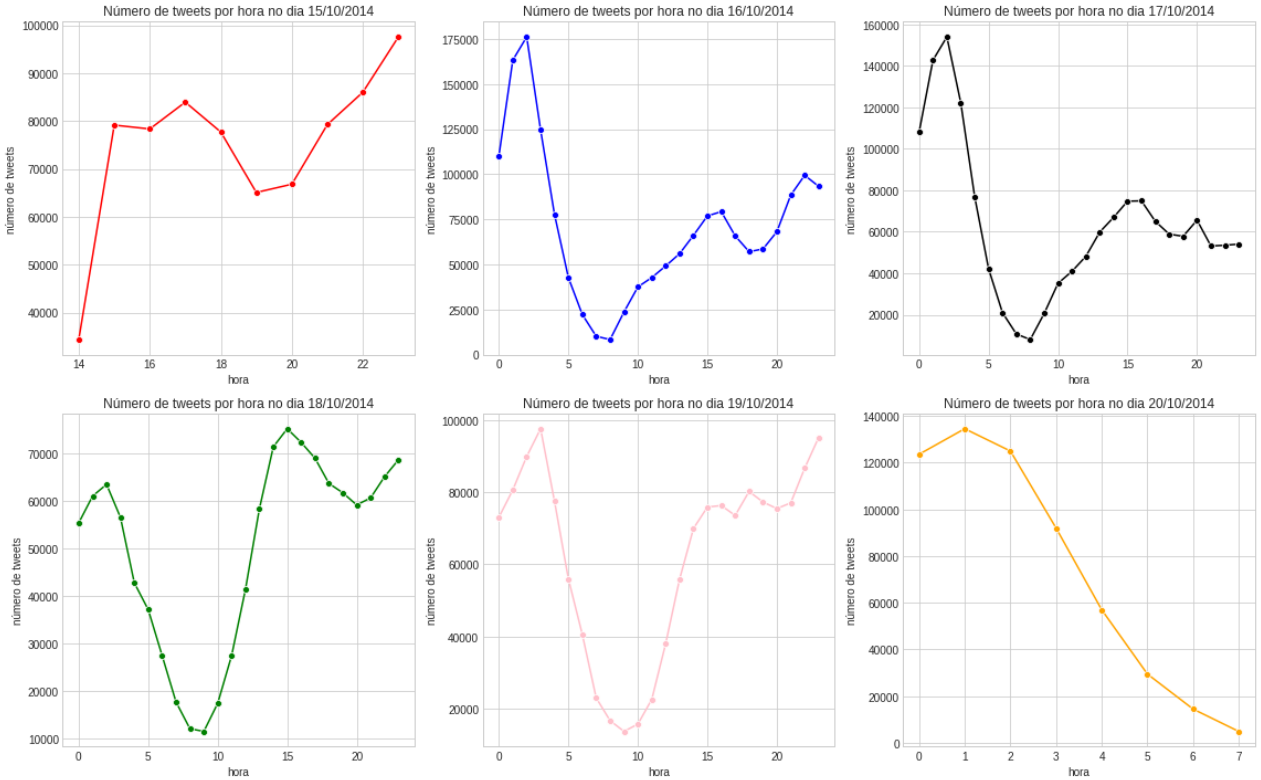
Número de tweets por hora no dia 19/10/2014

hora	número de tweets
0	73049
1	80626
2	89885
3	97575
4	77489
5	55958
6	40591
7	22988
8	16698
9	13707
10	15743
11	22588
12	38047
13	55888
14	69949
15	75866
16	76299
17	73534
18	80192
19	77265
20	75453
21	77033
22	86629
23	95083

Número de tweets por hora no dia 20/10/2014

hora	número de tweets
0	123610
1	134555
2	124997
3	92015
4	57005
5	29497
6	14504
7	4922

Séries temporais do número de tweets por dia



2.5 d) Quais as principais sentenças relacionadas à palavra “Dilma”?

Para responder essa pergunta foi necessário:

- Definir uma sentença como um conjunto(sub-string) de 40 caracteres dentro dos tweets.
- Dividir os tweets em conjuntos de 40 caracteres.
- Como a divisão por 40 caracteres pode gerar uma última sub-string com menos de 40 caracteres, foi necessário pegar somente as sentenças com 40 caracteres.
- Pegar as sentenças que continham *"dilma"* e que não fossem hashtags.
- Agrupar e agregar pelas sentenças.

As 15 principais sentenças relacionadas à palavra “Dilma”

sentença dilma	contagem
No debate de hj, Dilma foi orientada pel	12
Relatórios usados por Dilma para critica	10
Dilma que fala tanto em um projeto de se	9
em sobre Dilma. A previsão é que ele ven	8
Dilma perde o rumo no meio da entrevista	7
@os_deminas @BlogDilmaBR2 @jossimarfaria	7
No debate Dilma Rousseff está pensando q	7
1. #QueroDilmaTreze 2. #EmTodoBrasilAeci	7
Se eu fosse a Dilma sabendo que iria fic	7
Só pra lembrar: Aécio 45 Dilma 13 CPM 22	6
NANCIOU CAMPANHA DE DILMA. É O MAR DE LA	6
— Dilma ou Aécio? — Dois hambúrgueres, a	6
Dilma se atrapalha e passa mal em entrev	5
@JairoRochaFilho @JornalOGlobo @dilmabr	5
Aécio passa Dilma Rousseff e coordenador	5

2.6 e) Quais as principais sentenças relacionadas à palavra “Aécio”?

Para responder essa pergunta foi necessário:

- Definir uma sentença como um conjunto(sub-string) de 40 caracteres dentro dos tweets.
- Dividir os tweets em conjuntos de 40 caracteres.
- Trocar as ocorrências das palavras "aécio" por "aecio" e "Aécio" por "Aecio".
- Como a divisão por 40 caracteres pode gerar uma última sub-string com menos de 40 caracteres, foi necessário pegar somente as sentenças com 40 caracteres.
- Pegar as sentenças que continham *"aecio"* e que não fossem hashtags.
- Agrupar e agregar pelas sentenças.

As 15 principais sentenças relacionadas à palavra “Aecio”

sentença aecio	contagem
r Aecio somem do site do TCE http://t.co	10
no Debate do SBT por Aecio Neves http:/	8
@Prjuliocpc @MariaLuciadeJe3 @AecioNeves	8
Segundo pesquisa Aecio lidera com vantag	8
como sendo da esposa de Aecio. Quanta p	8
Pressionada por Aecio no debate do SBT,	7
vo no Debate do SBT por Aecio Neves http	7
vivo no Debate do SBT por Aecio Neves ht	6
Você sabia que o 'Aecio' manteve-se dura	6
— Dilma ou Aecio? — Dois hambúrgueres, a	6
@neyleprevost @pricilamarikoch @AecioNev	6
Só pra lembrar: Aecio 45 Dilma 13 CPM 22	6
Aos meus amigos que votarão em Aecio htt	6
O governo @AecioNeves vai levar tecnolog	6
@AecioNeves Mentiroso estava sentadinho	6

---

3. Segunda Questão

3.1 Pré-Processamento dos Dados

O dataset da segunda questão era um json aninhado com 16 campos no total, mas a maioria dos campos não seria útil para responder às perguntas propostas, por conta disso vários campos foram dropados. Em seguida, foram feitas operações para gerar a coluna Date. O dataset pré-processado foi salvo no S3 no formato *parquet*, que é o formato favorito do Spark.

O dataset a ser utilizado possui as colunas: *id*, *text*, *title* e *date*.

3.2 a) Encontre as palavras mais utilizadas nas avaliações

Para responder essa pergunta foi necessário:

- Criar uma coluna chamada *word* que contém todas as palavras de todos os reviews(coluna *text*) e deixar todas em Lower Case.
- Pegar somente as palavras com mais de 4 caracteres.
- Agrupar e agregar pela coluna *word*.

As 15 palavras mais utilizadas nas avaliações

palavra	contagem
tower	4184
eiffel	3246
there	2049
paris	1997
visit	1322
views	1239
tickets	1222
worth	1164
night	1107
tower.	1039
great	1038
first	945
around	840
second	835
paris.	831

3.3 b) Encontre as expressões mais usadas

Para responder essa pergunta foi necessário:

- Definir uma expressão como um conjunto(sub-string) de 45 caracteres dentro dos tweets.
- Preencher os campos nulos em *text* com " "
- Dividir os reviews em conjuntos de 45 caracteres.
- Criar uma coluna chamada *Expr* que contém todos os conjuntos de 45 caracteres de todos os reviews.
- Como a divisão por 45 caracteres pode gerar uma última sub-string com menos de 45 caracteres, foi necessário pegar somente as sentenças com 45 caracteres.
- Agrupar e agregar pela coluna *Expr*.

As 15 expressões mais utilizadas nas avaliações

expressão	contagem
no trip to paris would be complete without a	3
you can't go to paris and not visit the eiffe	3
you can't go to paris without visiting the ei	3
no visit to paris would be complete without a	2
we got there first thing in the morning which	2
a trip to paris is not complete without a vis	2
you cannot go to paris without visiting the e	2
the eiffel tower is to paris what the statue	2
of liberty is to new york and what big ben is	2
well what can i say that hasn't already been	2
this was my second visit to the eiffel tower.	2
the eiffel tower can be seen from all over pa	2
so worth climbing up rather than taking the e	2
we loved our stay in paris & visiting the eif	2
we visited the eiffel tower twice during our	2

3.4 c) Encontre os principais tópicos relacionados às revisões

Para responder essa pergunta foi necessário:

- Definir um tópico como uma palavra com mais de 3 caracteres e menos de 16.
- Criar uma coluna chamada *word* que contém todas as palavras da coluna *title* e deixar todas em Lower Case.
- Pegar só as palavras com com mais de 3 caracteres e menos de 16.
- Agrupar e agregar pela coluna *word*.

Os 15 principais tópicos relacionados às revisões

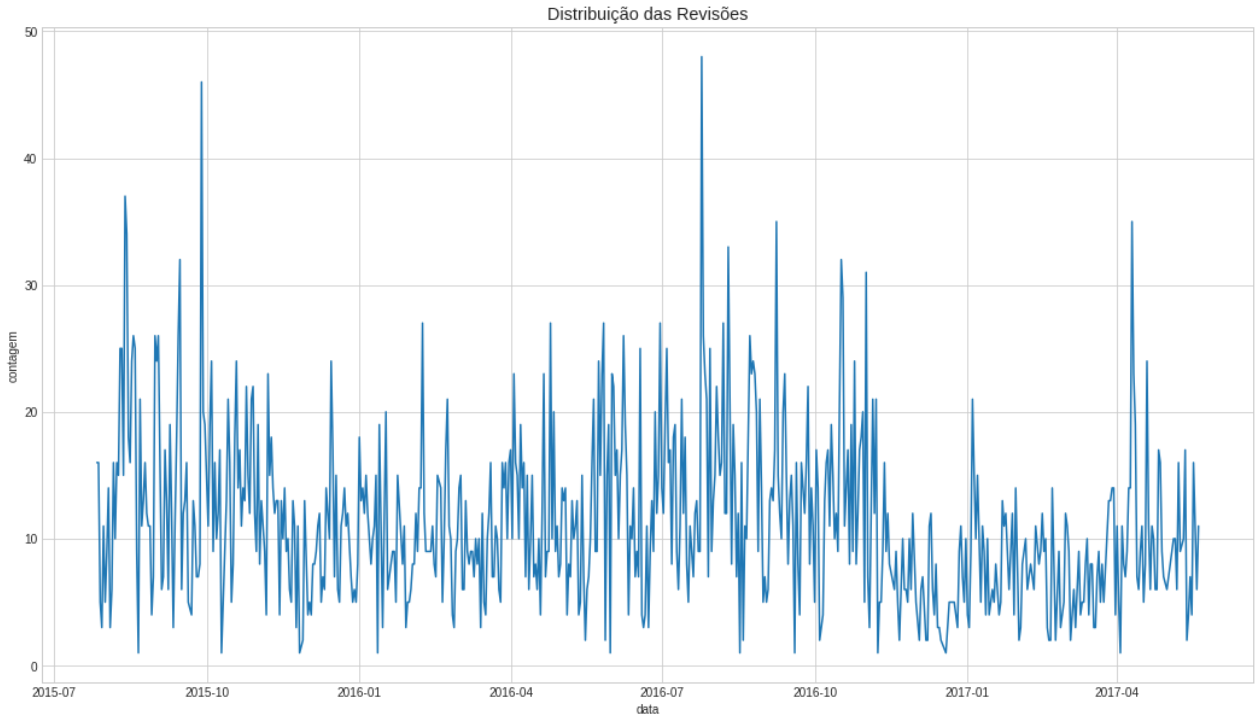
expressão	contagem
must	687
paris	634
eiffel	579
tower	556
amazing	426
view	340
beautiful	322
great	317
night	288
visit	287
views	239
worth	219
iconic	181
from	148
experience	136

3.5 d) Mapeie a distribuição temporal das revisões

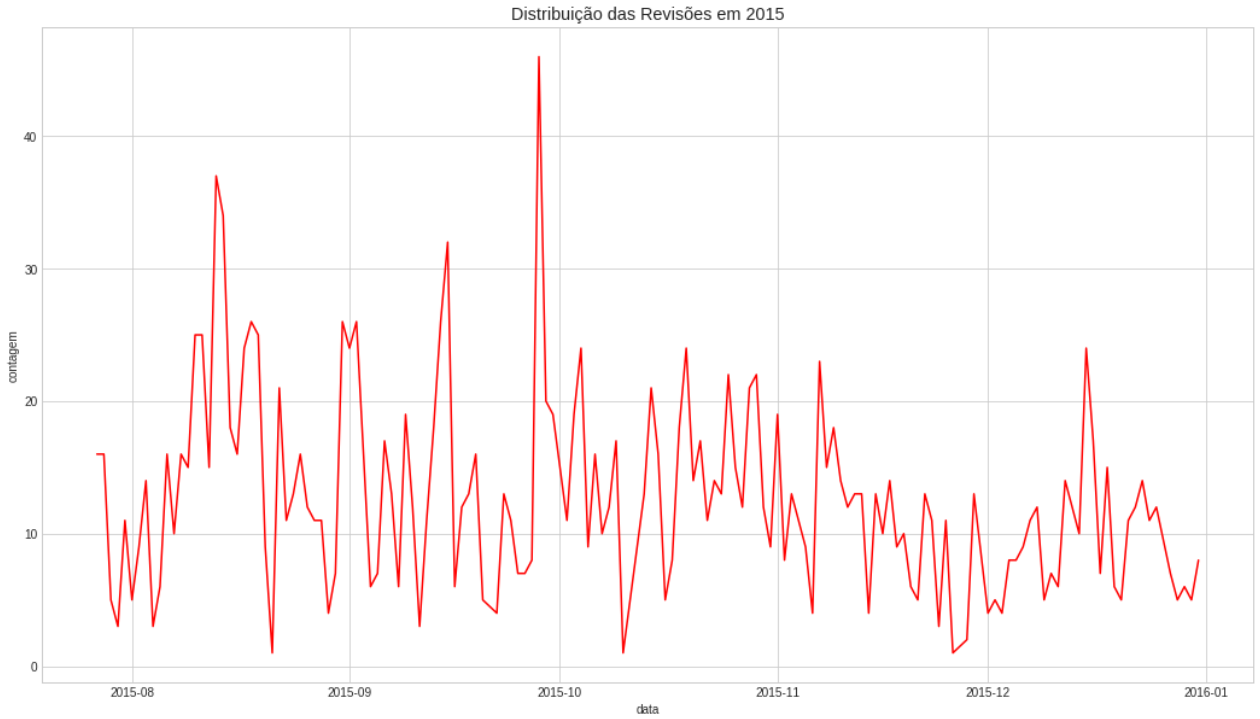
Para responder essa pergunta foi necessário:

- Agrupar e agregar pela coluna *Date*

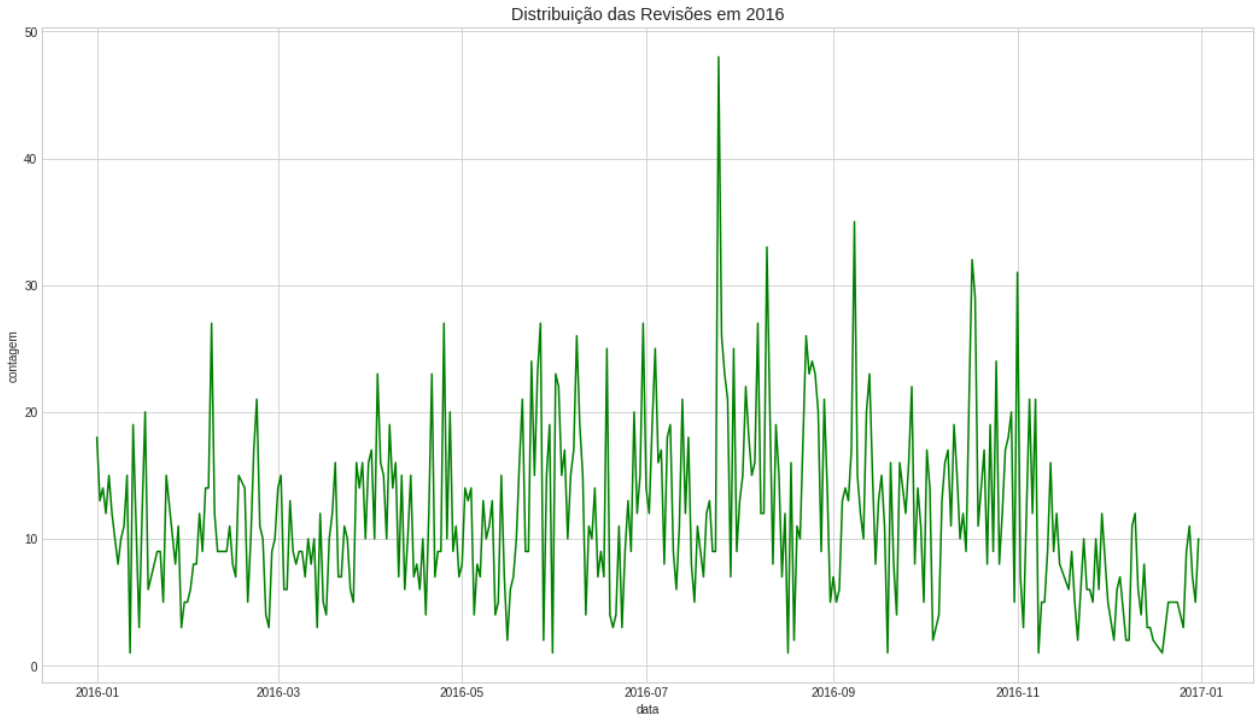
Distribuição temporal das revisões em 2015, 2016 e 2017



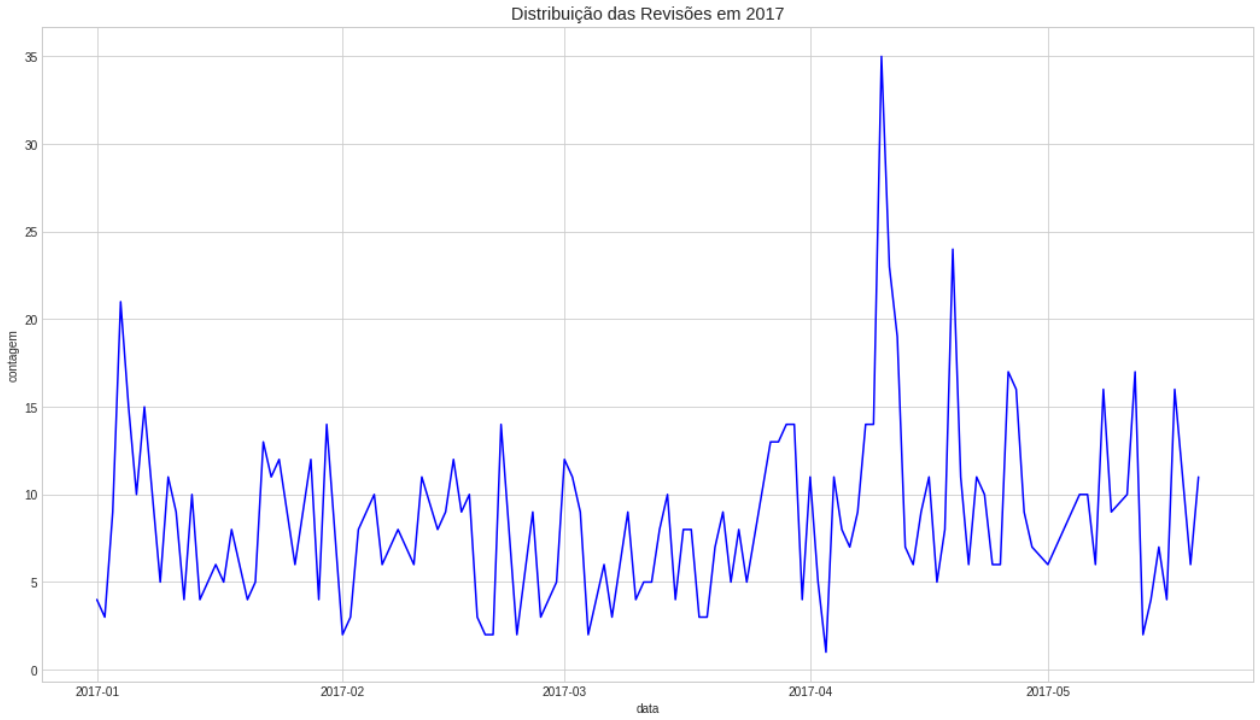
Distribuição temporal das revisões em 2015



Distribuição temporal das revisões em 2016



Distribuição temporal das revisões em 2017



4. Considerações finais

Todo o código desenvolvido e todas as imagens geradas estão disponíveis nesse repositório.