Segundo exercício de programação

[PDMN2020-2] Exercício de programação (Covid-19)

Aluno: Danilo Pimentel de Carvalho Costa
Matrícula: 2016058077

Introdução

O segundo exercício de programação tem como objetivo introduzir o processamento de dados massivos utilizando o ambiente Spark. Este ambiente fornece ferramentas para construção de fluxos de trabalho para diversas tarefas de análise de dados. Para este trabalho são propostas 4 análises em uma base de dados de tweets públicos entre 29/03/2020 e 30/04/2020. A seguir, se encontram explicações, resultados e como cada análise foi construída, bem como uma breve análise do conjunto de dados.

Conjunto de dados

Os primeiros esforços foram dedicados a conhecer melhor sobre a base de dados. Na especificação do trabalho, foram relatados possíveis problemas de formatação dos dados.

A primeira dificuldade encontrada foi a agilidade no fluxo de trabalho geral para execução de tarefas no cluster. Algumas soluções foram tentadas, como a configuração de um plugin do editor de texto Visual Studio Code para editar arquivos remotamente via ssh. A solução se mostrou demorada devido aos atrasos para salvar e para digitar na linha de comando.

A segunda solução, que foi adotada pelo resto do exercício, foi a experimentação e o desenvolvimento em ambiente local, e o deploy e execução definitiva no cluster. Foi feito o download de parte do conjunto de dados para o computador pessoal. Neste, foi configurado também o spark e a ferramenta Jupyter Notebook. Tal solução possibilitou a experimentação rápida e acelerou o desenvolvimento do exercício.

Problema 1

O conteúdo do tweet começa na primeira coluna e sobrescreve as outras colunas até o fim. Em seguida, surge uma string do tipo Twitter%, mostrando a plataforma em que o tweet foi postado. Não foi encontrada uma relação entre este tipo de má formatação e a plataforma que o tweet foi publicado (Web, IPhone, Android, etc).

	status_id	user_id	created_at	screen_name	text	source	reply_to_statu
0	La opinión de Federico @Berrueto en #MILENIOAl...	Twitter Media Studio	None	None	None	FALSE	FA
1	https://t.co/ufBtMLTWil #NoticiasTVN #COVID19 ...	Twitter Media Studio	None	None	None	FALSE	FA
2	#البيان_القارئ_دائما https://t.co/EWvEvNfNt8"	Twitter Media Studio	None	None	None	FALSE	FA
3	#COVID19 #Coronavirus #MéxicoUnido #SanaDistan...	Twitter Media Studio	None	None	None	FALSE	FA
4	#VirgenMaría #Covid19 #28Mar."	Twitter for Android	None	None	None	FALSE	FA

	status_id	user_id	created_at	screen_name	text	source	reply_to_status_id	reply_to_user_i
0	qu'en ce moment	alors que le monde entier se bat contre la te...	les français	eux	ne se battent que contre une simple #grippett...	Twitter Web App	None	Non
1	>Cualquier virus del género #Coronavirus	incluyendo los 7 que infectan al ser humano (...	no solo existe el #SARSCoV2	entérate)	son agentes que se trabajan en laboratorios d...	Twitter for Android	1244710485108445185	119959656376666521
2	Americans are scared	sick&dying and chump is a narcissistic	lying	disgusting	traitor pig 😡"	Twitter for Android	None	Non
3	Also on Monday	one death reported in Hubei Province	and 44 new suspected cases	all imported ones	were reported on the mainland. #coronavirus h...	Twitter Web App	None	Non
4	Italia ha registrado en el último día mil 648 ...	la cifra más baja de los últimos 20 días	los muertos ascendieron a 11 mil 591	con 812 más respecto al domingo	según datos de hoy de Protección Civil ⬇⬇ htt...	Twitter Web App	None	Non

Problema 2

Dado o problema da disposição do conteúdo em múltiplas colunas, o segundo problema foi encontrar um método confiável para filtrar resultados bem formatados. A melhor forma encontrada foi explorar a estrutura bem definida do campo created_at.

```
tweets.filter("created_at like '2020-%'").limit(20).toPandas()
```

	status_id	user_id	created_at	screen_name	text	source	reply_to_status_id	reply_to_user_i
0	1244776423073542144	2722502906	2020-03-31T00:00:00Z	GradaNorteMX	Cuando mejor iban las cosas en el circuito de ...	None	None	Non
1	1244776423593775105	860252856829587457	2020-03-31T00:00:00Z	IMSS_SanLuis	Para prevenir el COVID-19 la distancia es impo...	TweetDeck	None	Non
2	1244776422037692417	24969337	2020-03-31T00:00:00Z	Milenio	"▶ ""Es deseable que en este momento	AMLO se tome las cosas en serio y seguramente...	None	Non
3	1244776422079705092	91430932	2020-03-31T00:00:00Z	NewsweekEspanol	"Tú eres tu pareja sexual más segura": la guía...	None	None	Non
4	1244776422209536002	299693451	2020-03-31T00:00:00Z	tvnnoticias	Los taxis tienen un nuevo horario de circulaci...	None	None	Non
5	1244776421953806337	89225092	2020-03-31T00:00:00Z	TUDNUSA	Sin futbol hasta que todos los jugadores se va...	None	None	Non
6	1244776422582833154	44728980	2020-03-31T00:00:00Z	ANCALERTS	Here are some tips on how you can best protect...	Twitter Media Studio	None	Non
7	1244776421257629698	44728980	2020-03-31T00:00:00Z	ANCALERTS	New York welcomes hospital ship as coronavirus...	TweetDeck	None	Non
8	1244776421274406914	132225222	2020-03-	SSalud_mx	#ConferenciaDePrensa sobre el #Coronavirus	TweetDeck	None	Non

Problema 3

O terceiro problema é determinar o local de onde o tweet foi publicado. O campo contry_code tem valores de todo o tipo. O código numérico 55 foi investigado, mas não foi um bom método, pois claramente o conteúdo dos tweets estavam em outros idiomas como inglês e russo.

A próxima tentativa foi utilizar o campo lang. Não se pode inferir que se o idioma do tweet é "pt", o tweet foi publicado no Brasil. Porém, ao filtrar pelo campo lang com o valor "pt" e agrupar por "country_code", foi possível perceber que o "country_code" com valor "BR" tem a maioria dos tweets em "pt".

O resultado acima foi o motivo da escolha de filtrar tweets por "country_code" com valor "BR" ou algum outro valor para outro país.

	country_code	count
0	None	54227
1	BR	3601
2	PT	96
3	US	79
4	ES	54

Como melhoria futura, é sugerida a combinação de diversos métodos para filtragem dos dados, como utilizar o place_full_name e place_type. Por simplicidade, foi utilizado somente o método descrito acima.

Análises propostas

Análise 1

Quais foram as 15 hashtags mais populares no mundo? E no Brasil?

A estratégia geral foi transformar o conteúdo do tweet para extrair as palavras e hashtags utilizadas. O processo aqui foi feito manualmente, partindo o conteúdo nos espaços em branco e pontuações comuns como vírgula e ponto final. Depois, foi realizada a filtragem dos tokens para manter somente hashtags. Agora, o resultado tem somente as hashtags, que se repetem com a quantidade de vezes que foram utilizadas. Para finalizar, é feito o agrupamento, a contagem de elementos do grupo, e a ordenação pela contagem. Assim, está montado o ranking das hashtags mais populares.

Para as hashtags mais populares no mundo, a intenção foi aproveitar ao máximo o conteúdo do conjunto de dados. É sabido que há a má formatação dos dados, e que nem sempre o conteúdo do tweet está na coluna "text". Por isso, antes de executar a estratégia acima, todas as colunas de um registro são concatenadas em uma só. Isso maximiza a quantidade de conteúdo analisado, que seria perdido caso fossem utilizados somente registros bem formatados. Abaixo estão os resultados encontrados:

Hashtag	Qtd. de ocorrências
#covid19	7709295

#coronavirus	5796944
#covid_19	2111701
#covid—19	346591
#stayhome	341498
#coronaviruspandemic	340881
#stayhomestaysafe	305545
#lockdown	291970
#corona	247319
#coronavirusoutbreak	229956
#covid	217314
#covid2019	203255
#stayathome	199863
#quedateencasa	173532
#pandemic	146797

O mesmo não pode ser feito para as hashtags no Brasil. O método utilizado para determinar a localização que o tweet foi publicado foi o descrito na seção anterior. Ao filtrar por "country_code" com valor "BR", tweets que foram publicados no Brasil mas estão mal formatados não são considerados. O processo após a filtragem por "country_code" segue a mesma estratégia descrita acima. Abaixo estão os resultados encontrados:

Hashtag	**Qtd. de ocorrências**
#covid19	1897
#coronavirus	1832
#covid_19	553
#coronavírus	404
#fiqueemcasa	275
#quarentena	257
#ficaemcasa	182
#forabolsonaro	179
#pandemia	143
#bbb20	128
#coronavirusbrasil	121
#covid—19	119
#brasil	117
#foraprior	94
#foramanu	90

Análise 2

Como foi a evolução de comentários sobre a pandemia ao longo das semanas nos Estados Unidos, Brasil e México (os três países mais afetados no continente americano até dezembro/2019). Para responder essa pergunta, faça um único gráfico contabilizando a média móvel do número de tweets ao longo dos dias, para cada país.

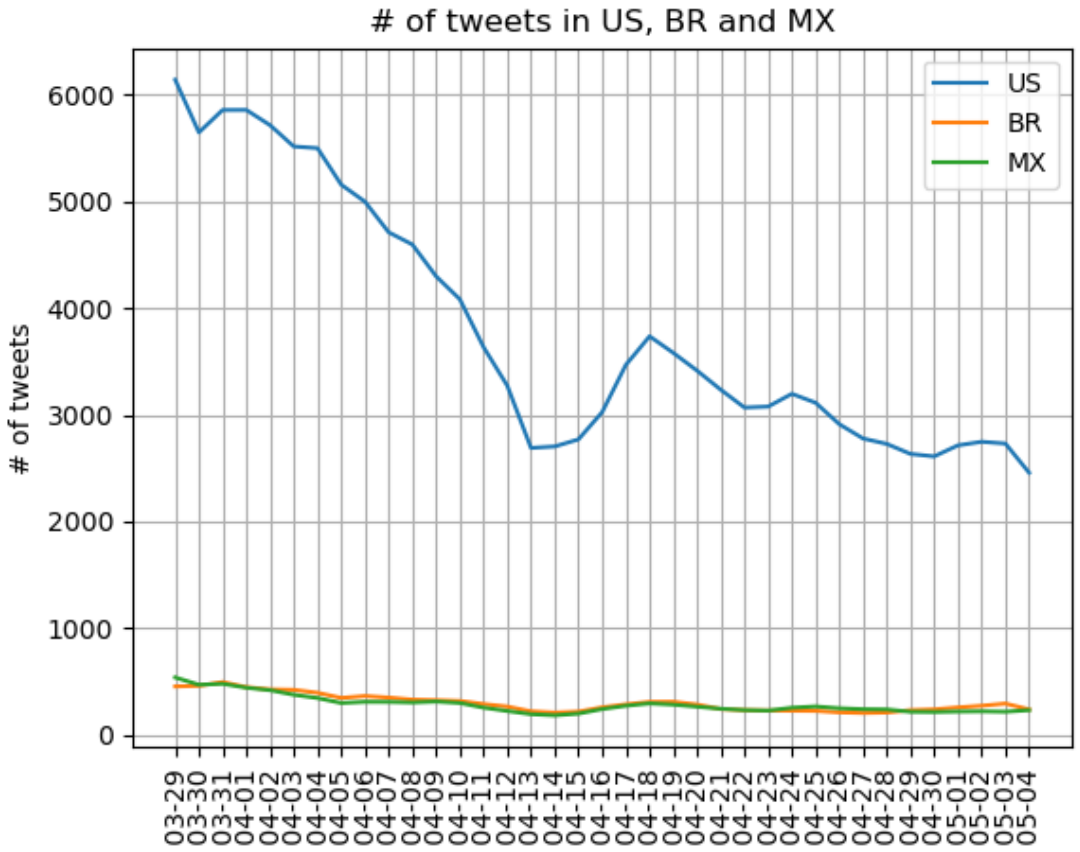
A filtragem inicial foi feita utilizando o campo "country_code" com valores "BR", "US", e "MX". Para construir o gráfico com a média móvel por dia, precisamos da data de publicação do tweet. Assim, a filtragem também exclui registros que o valor para "created_at" não seja do tipo "2020-%".

Para computar a quantidade de tweets por dia, é preciso transformar os valores da coluna "created_at" para extrair o dia que o tweet foi publicado. A escolha aqui foi retirar tudo o que há depois de "T". O campo created_at está no formato ISO, e tudo depois de T se refere ao momento do dia. Uma vez extraída a data do tweet, podemos ordenar os tweets por data, e agrupar por data para determinar a quantidade de tweets em cada dia.

O cálculo da média móvel para um determinado dia é feito tirando a média da quantidade de tweets dos últimos 5 dias (valor escolhido arbitrariamente). Para fazer tal cálculo, foi feito um flatMap no resultado parcial com datas e quantidades. Para cada data, foram gerados outros 6 registros com a própria data e datas dos próximos 5 dias e a quantidade de tweets daquele dia. Por exemplo, para o registro (2020-03-20), seriam gerados 2020-03-20, 2020-03-21, 2020-03-22, 2020-03-23, 2020-03-24 e 2020-03-25.

Assim, para uma determinada data, temos atreladas à sua chave todos os valores para a média móvel. Podemos agora realizar um reduceByKey, e de fato calcular a média para o dia. Um detalhe importante é que os primeiros 4 dias não terão os 5 dados anteriores. A média é calculada com a quantidade de registros disponíveis. Para o primeiro dia, o valor é a própria quantidade do dia. Para o segundo dia, o valor é a média da própria quantidade e da quantidade do dia anterior.

Calculada a média móvel, temos os dados para a construção do gráfico. Aqui, a decisão tomada foi de não gerar o gráfico de dentro do ambiente Spark. O resultado parcial é armazenado no HDFS, extraído para a pasta home do usuário, e depois utilizado em um programa auxiliar para gerar o gráfico. O resultado se encontra a seguir:



Análise 3

Houve usuários que se beneficiaram da pandemia para aumentar o número de seguidores? Mapeie os 100 usuários que mais cresceram em número de seguidores nesse intervalo da amostra. Quais desses usuários também estão na lista dos 100 usuários ativos?

Para esta análise, é preciso determinar com confiança a quantidade de seguidores de um perfil no momento do tweet. Este é um problema nesta base de dados, pois nem todos os registros têm esta informação. Para os registros que têm valor em "followers_count", nem todos os valores são válidos. Após uma breve exploração, foi encontrada uma forma de filtrar os registros que têm valores válidos para "followers_count". Observa-se que registros com valor bem formatado para "account_created_at" têm valor válido para "followers_count". Este foi o método utilizado para remover dados inválidos.

Os valores em "followers_count" eram strings. Para utilizá-lo, foi preciso realizar o casting para o tipo inteiro. Assim, o objetivo agora é entender qual foi o crescimento de seguidores por usuário. Os registros filtrados são agrupados por user_id, e é calculada a quantidade de tweets por usuário e o valor máximo e mínimo de "followers_count". A diferença nos fornece o crescimento de seguidores e a quantidade de tweets por usuário no período. Podemos assim chegar aos seguintes resultados:

2 - Top 100 usuários com maior crescimento

user_id	min	max	count	growth
3171712086	12930400	13649972	5	719572
759251	46692373	47402413	8	710040
14499829	6950550	7637195	48	686645
1154180084078501889	315912	896488	57	580576
122658025	4024406	4496280	3	471874

Mostrando os 5 primeiros resultados. Todos os resultados em 2 - Top 100 usuários com maior crescimento

2 - Top 100 usuários mais ativos

user_id	min	max	count	growth
4913320595	29334	29729	8643	395
717039627916484608	972	999	3886	27
231726084	5376	5421	3856	45
1198228761910611969	182	526	3460	344
301831339	479373	485889	2489	6516

Mostrando os 5 primeiros resultados. Todos os resultados em 2 - Top 100 usuários mais ativos

Com estes 2 rankings, podem ser determinados os usuários com maior crescimento que também estão entre os usuários mais ativos:

2 - Usuários com maior crescimento mais ativos

user_id
15872418
355989081
1214315619031478272
1943418931
37034483
7587032

Análise 4

Quais foram os 3 principais assuntos discutidos nos Estados Unidos? Extraia 5 palavras utilizadas em cada assunto (desconsiderando as hashtags).

A solução desta análise foi baseada nas seguintes referências:

- https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3741049972324885/3783546674231782/4413065072037724/latest.html
- https://databricks.com/blog/2015/09/22/large-scale-topic-modeling-improvements-to-lda-on-apache-spark.html

A solução para determinar os 3 principais assuntos discutidos nos Estados Unidos foi criar um modelo utilizando o LDA (Latent Dirichlet allocation). O Spark fornece a implementação do LDA em sua biblioteca de aprendizado de máquina.

Os dados foram filtrados pelo campo "country_code" com valor "US" para manter somente os tweets publicados nos Estados Unidos. Foram utilizados somente os valores para o campo "text" dos registros.

Para identificar os tópicos dos tweets, é preciso fazer o pré-processamento do texto para extrair as palavras relevantes. Hashtags, menções, números, links, stop words, entre outros precisam ser filtrados. Foi utilizado o RegexTokenizer para fazer a extração dos tokens, e o StopWordsRemover para remoção de stop words.

O algoritmo LDA recebe como entrada uma lista, com um item para cada documento. Cada item contém o id do documento e um vetor, com a quantidade de vezes que cada palavra do vocabulário foi citada. Para colocar os dados dos tweets produzidos até agora neste formato, foi utilizado o CountVectorizer.

Com os dados formatados, o modelo do algoritmo LDA pode ser criado. Houve dificuldade para encontrar os melhores hiper-parâmetros. A quantidade de iterações foi variada entre 3, 10, 30, e 100. Os outros hiperparâmetros foram deixados com os valores padrão. A determinação do melhor resultado foi arbitrária, sendo o modelo com 100 iterações. Os tópicos e os respectivos termos são encontrados a seguir:

topic_id	term
0	people
0	like
0	time
0	today
0	home
1	para
1	package
1	urge

1	debt
1	student
2	florida
2	california
2	miami
2	beach
2	puerto

Melhorias futuras seriam otimizar as configurações de hiperparâmetros, tentar outros tipos de otimizadores, e até experimentar outros algoritmos para modelagem de tópicos.

Conclusão

Pude aprender sobre o ambiente Spark e suas diversas ferramentas, e uma introdução suave a ambientes de computação distribuída. O tema escolhido para o primeiro exercício de programação foi complexo o suficiente para a aprendizagem dos conceitos, e simples o suficiente para implementação no prazo descrito.