Aluno: Pedro Luis Saraiva Barbosa (Matrícula: 486347)
Disciplina: Desenvolvimento de Software para Nuvem
Professores: Fernando Trinta e Paulo Antonio Rego

Descrição do ambiente utilizado para execução das aplicações Hadoop:
Mac OXS Catalina version 10.15.6
Java 1.8.0_161
Eclipse IDE 2020-03

Questão 2) Dado um dataset com tweets relacionados à campanha eleitoral presidencial de 2014, responda:

a) Quais foram as hash tags mais usadas pela manhã, tarde e noite? (1,0 ponto)

Para responder essa pergunta, gerou-se dois processos, sendo que primeiramente foi necessário dividir a linha do arquivo (Map/reduce) debate-tweets.tsv por tabulação para que fosse possível ter acesso as informações dos tweets. O segundo processo foi necessário para ordenação das informações e aplicando os jobs para se obter quais foram as hashtags mais utilizadas pela manhã, tarde e noite. Nas tabelas abaixo são apresentadas as hastags e o seu total por período.

Hashtags mais usadas pela manhã	
Hashtag	**Total**
#emabiggestfans1d	14802
#emabiggestfansjustinbieber	14085
#bomdia	644
#trndnl	604
#votevampsteenawards	364
#queronotvz	317
#twoff	312

Hashtags mais usadas pela tarde	
Hashtag	**Total**
#emabiggestfans1d	58772
#emabiggestfansjustinbieber	50905
#queronotvz	4620
#stealmygirl	2380
#austinmahonechile	1894
#austinmahone	1496
#mtvhomecoming	961

Hashtags mais usadas pela noite	
Hashtag	**Total**
#emabiggestfans1d	65004
#emabiggestfansjustinbieber	55851

#stealmygirl	5233
#bigpaynodanceoff	4132
#debatenosbt	3628
#cartersnewvideo	3278
#luansantananahoradofaro	2687

b) Quais as hashtags mais usadas em cada dia? (1,0 ponto)

Da mesma forma que na questão anterior, foi utilizado dois processos. Sendo um usado para dividir o linha do arquivo por tabulação. Depois foi necessário guardar data de publicação para que se pudesse comparar com cada dia do mês. Por meio dessa comparação, foi possível gravar as informações que correspondiam ao desejado a cada laço de repetição no conjunto de dados. O segundo processo foi necessário para ordenar os dados, para se obter as hashtags mais utilizadas por dia. Na tabela abaixo são apresentadas as hashtags mais utilizadas.

Hashtags mais utilizadas por dia		
Data	**Hashtag**	**Total**
15/10/2014	#emabiggestfans1d	33509
16/10/2014	# emabiggestfans1d	65526
17/10/2014	# emabiggestfansjustinbieber	46718
18/10/2014	# emabiggestfans1d	26332
19/10/2014	# emabiggestfansjustinbieber	31521
20/20/2014	# emabiggestfansjustinbieber	9912

c) Qual o número de tweets por hora a cada dia? (1,0 ponto)

Fora utilizados dois processos, como nas questões anteriores. No primeiro foi feito a divisão da linha por tabulação para a criação de tokens, para se ter acesso aos textos e sua hora de publicação. Cada hora de publicação foi comparada com uma hora do dia, e por meio dessa iteração e comparações foi armazenado os tweets e o número de publicações por hora. A tabela abaixo apresenta os resultados.

15/10/2014			
Hora	**Qtd**	**Hora**	**Qtd**
14	34300	19	65100
15	79168	20	66820
16	77400	21	79278
17	83940	22	86001
18	77718	23	97589
16/10/2014			
Hora	**Qtd**	**Hora**	**Qtd**
00	110250	12	50163
01	163358	13	55960
02	176225	14	66157
03	122614	15	76896
04	67749	16	79292
05	42668	17	65931

06	22430	18	57105
07	10159	19	60552
08	8239	20	69170
09	24620	21	88485
10	38532	22	99567
11	42799	23	93364
20/10/2014			
Hora	**Qtd**	**Hora**	**Qtd**
00	123624	04	57010
01	134569	05	29498
02	125020	06	14510
03	92031	07	4922

d) Quais as principais sentenças relacionadas à palavra "Dilma"? (1,0 ponto)

Como nas questões anteriores, foi utilizado dois processos map/Reduce para conseguir as principais setenças relacionadas à palarva "Dilma". Primeiro foi realizado a divisão das linhas por tabulação do arquivo. Foi também transformado todas as letras em *lower case*. Foi percorrido todas as setenças em busca de textos que tivessem a palavra dilma e armazenando essas informações. No segundo processo foi realizado a ordenção dos dados que foram mais recorrentes. A tabela abaixo apresenta os dados encontrados.

Sentenças mais utilizadas com a palavra Dilma	
sentença	**quantidade**
A cara da dilma	110
Na cara da dilma	90
Que a dilma va	78
A dilma e o	71

e) Quais as principais sentenças relacionadas à palavra "Aécio"? (1,0 ponto)

Como nas questões anteriores, foi utilizado dois processos map/Reduce para conseguir as principais setenças relacionadas à palarva "Áecio". Primeiro foi realizado a divisão das linhas por tabulação do arquivo. Foi também transformado todas as letras em *lower case* e retirado e toda e qualquer pontuação e caracteres especiais. Foi percorrido todas as setenças em busca de textos que tivessem a palavra dilma e armazenando essas informações. No segundo processo foi realizado a ordenção dos dados que foram mais recorrentes. A tabela abaixo apresenta os dados encontrados.

Sentenças mais utilizadas com a palavra Dila	
sentença	**quantidade**
Vai votar no aecio	86
Entre dilma e o aecio	58
Sbt por aecio neves	54
Do sbt por aecio	54

Questão 3) Dado um dataset com tweets relacionados à visita da Torre Eiffel em Paris, responda:

a) Encontre as palavras mais utilizadas nas avaliações. (1,0 ponto)

Como nas demais questões, foi utilizado dois processos de Map/Reduce. Além disso, para a resolução desta questão, foi utilizado uma biblioteca JSON que facilitou na manipulação do arquivo eiffel-tower-reviews.json. Essa biblioteca foi necessária para pegar campos específicos do arquivo JSON. Após a utilização da bibliotecapara obter o campo *text* do arquivo foram criado tokens através da divisão dos termos por meio dos espaços. Depois foi feito a ordenação das palavras com base na frequência. A tabela abaixo apresenta as palvras mais utilizadas e sua frequência.

Palavras maus usadas	
Tower	4184
Eiffel	3246
From	2504
This	2100
There	2049
Paris	1997
That	1875
Have	1867

b) Encontre as expressões mais usadas. Considere uma expressão um conjunto de palavras na sequencia. O tamanho da sequencia pode ser determinado por você. (1,0 ponto)

c) Encontre os principais tópicos relacionados às revisões. (1,0 ponto)

Como nas demais questões, foi utilizado dois processos de Map/Reduce. Além disso, para a resolução desta questão, foi utilizado uma biblioteca JSON que facilitou na manipulação do arquivo eiffel-tower-reviews.json. Essa biblioteca foi necessária para pegar campos específicos do arquivo JSON. Após a utilização da biblioteca para obter o campo *title* do arquivo foram criado tokens através da divisão dos termos por meio dos espaços. Depois foi feito a ordenação das palavras com base na frequência. A tabela abaixo apresenta as palavras mais utilizadas e sua frequência.

Palavras maus usadas	
Eiffel tower	99
Amazing	94
Beautiful	56
Must see	45
Iconic	41

d) Mapeie a distribuição temporal das revisões. (1,0 ponto)