Projeto de Implementação 3

Publicado em **16/01/2023 - Entrega em 30/01/2023**

Executar em Duplas

Para este PI serão utilizados dois ambientes de execução diferentes: Na Parte 1, como os demais PIs, será usado o Google Collaboratory. Na Parte 2, vamos utilizar um subconjunto dos PCs do Laboratório 1 do IComp como se fosse um cluster Spark.

Apresentação

A tarefa a ser executada neste PI é um experimento de agrupamento (clustering) de textos usando o algoritmo K-Means estudado em sala de aula, implementado em PySpark.

Serão fornecidos com entrada um conjunto de *tweets* reais constituídos de texto e *hashtags*. Neste conjunto de dados, cada tweet está rotulado como contendo ou não discurso de ódio, ou seja, expressar racismo, sexismo ou outro tipo de preconceito.

O algoritmo deverá realizar um agrupamento usando K-means com valor de K=2 e, em seguida, deve-se avaliar a qualidade do resultado obtido comparando os dois clusters gerados pelo K-means. As métricas de precisão e revocação, vistas em aula, devem ser usadas na avaliação.

Dataset Experimental

O dataset experimental que vamos utilizar nesse PI deve ser obtido no site "Twitter Sentiment Analysis - Detecting Hatred Tweets" hospedado no Kaggle. Devem ser usados somente os dados disponíveis no arquivo "train.csv". Um pequeno trecho do arquivo é mostrado na figura abaixo.

id	label	tweet
1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans. #disapointed #getthanked
14	1	@user #cnn calls #michigan middle school 'build the wall' chant '' #tcot
15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins

Neste trecho, a coluna label indica se o tweet for considerado como trazendo dicussor de ódio (label=1) ou não (label=0).

Variações do Algoritmo

Deverão ser implementas 3 versões diferentes do algoritmo:

Versão 1: Considera somente o texto dos tweets e descarta as hashtags
Versão 2: Considera omente as hashtags dos tweets e descarta todo o texto

Versão 3: Considera tanto o texto quanto as hashtags dos tweets

Execução dos Experimentos

- Execução 1: Utilizando o Google Colab, onde o experimento será excutado uma única vez, como normalmente tem sido feitos nos PIs anteriores.

- Execução 2: Utilizando o Cluster do Laboratório. Neste caso, o experimento deve ser executado usado 3 configurações diferentes: com 2, 3 e 4 workers, e o tempo de execução deve ser medido.

O Que Entregar

- Jupyter Notebook com código completo da implementação das 3 versões do Algorirmo.
    - **PI3-BGD-2022-02-NOME1-NOME2**
        - NOME1 : nome do(a) primeiro(a) aluno(a) da dupla
        - NOME2 : nome do(a) segundo(a) aluno(a) da dupla
    - O notebook deverá estar pronto para execução sem erros.
    - O resultado das métricas de precisão e revocação, que deve ser gerada na ultima página do notebook para cada uma das versões, sendo os resultados das três versões comparadas

- Conteúdo das páginas

  http://XXXXX-m:18080/history/application_NNNN_YYYY/executors/

  que são geradas a partir da execução do job no cluster Spark, para cada uma das três configurações do cluster. Esse link é gerado como a última linha resultante da execução do comando de submissão do script no terminal.

  Esse conteúdo deverá ser copiado em uma célula "Markdown" do Notebook

4. Como Entregar

- **P**ara a entrega, os alunos devem compartilhar o Notebook no Google Colab compartilhado com *alti@icomp.ufam.edu.br*. O professor não solicitará compartilhamento. O notebook compartilhado deve ser "pinado" (Save and Pin Revision) e não deve ser modificado depois disso. O professor verificará a "Revision History" do notebook e não corrigirá o projeto se houverem versões posteriores ao prazo de entrega.

- **Correção:** A correção será feita da seguinte forma: eu vou acessar o notebook compartilhado, executá-lo e verificar se a saída está de acordo. Também vou examinar e avaliar o código do notebook. Os alunos devem preparar o notebook de forma que ele inclua a criação do cluster Spark virtual.