Universidade Federal do ABC
Programa de Pós-graduação em Ciência da Computação
Disciplina de Inteligência na Web e Big Data

PROVA

Patrícia Dias dos Santos
RA 23201810211
patricia.santos@ufabc.edu.br

São Paulo, 04 de Maio de 2018

Questão 2: Apresentação do projeto.

O meu projeto se trata da implementação da versão serial e paralela do algoritmo *DCDistance: A Supervised Text Document Feature extraction based on class labels* [1], um algoritmo de extração e redução de atributos supervisionados, que cria recursos baseados na distância entre um documento e um representante de cada etiqueta de classe.

Eu utilizei a base *DOHMH New York City Restaurant Inspection Results*[1], com dados em formato .csv e com um tamanho de pouco mais de 133 megabytes. Essa base contém o resultado da inspeção sanitária em restaurantes da cidade de Nova Iorque entre os anos de 2014 e 2018.Para aplicar o DC Distance eu selecionei as colunas "VIOLATION DESCRIPTION"e "CRITICAL FLAG". Minha ideia era treinar a base para ler a descrição da violação e decidir se ela era crítica ou não.

Abaixo uma breve descrição do que eu fiz até agora:

1. Carreguei a base no contexto do Spark utilizando o SQLContext
2. Apliquei alguns filtros na base utilizando comandos do SQL para só usar as duas colunas da tabela que me interessavam
3. Apliquei funções para tokenizar e remover stop-words na coluna "VIOLATION DESCRIPTION"
4. Dividi a base entre treino (30 por cento) e teste (70 por cento)
5. Calculei o TF-IDF nas linhas da coluna "VIOLATION DESCRIPTION"
6. Verifiquei através de regressão logística que meu modelo estava conseguindo prever as categorias
7. Somei os vetores de cada linha para criar os vetores de cada classe sendo classe C1 a categoria "critical"e classe C2 a categoria "not critical", representadas pelos labels 0.0 e 1.0, respectivamente.
8. Calculei a distância euclidiana entre estes dois vetores C1 e C2 e a coluna com os valores vetorizados TF-IDF anteriores e assim consegui gerar os atributos.

O que falta:

(a) Paralelizar a versão serial do algoritmo.
(b) Calcular quanto tempo cada versão do algoritmo demora para rodar.
(c) Comparar o tempo das duas versões e dizer qual é melhor e se valeu a pena paralelizar.

[1]Disponível em: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/data

O código está disponível em: https://github.com/patyDSantos/BIGDATA2018/blob/master/Projeto/versao-serial-04-05.ipynb.

Referências

[1] Charles Henrique Porto Ferreira, Debora Maria Rossi de Medeiros, and Fabricio Olivetti de França. Dcdistance: A supervised text document feature extraction based on class labels. *arXiv preprint arXiv:1801.04554*, 2018.