Projeto 2: Analisando dados com Spark

Projeto em duplas

Entrega: 4/12

Neste projeto vamos analisar as páginas da internet brasileira para a descoberta de padrões. Escolha dois itens da mesma categoria - por exemplo: duas marcas de carro, duas celebridades, dois tipos de fruta, o que você quiser. Vamos tentar analisar como esses itens são vistos nas páginas da internet brasileira.

Análise do vocabulário comum

Construa uma tabela de frequencia inversa de documentos (*inverse document frequency – IDF*): usando Spark construa uma tabela que mapeia cada palavra para o logaritmo da frequencia inversa do numero de documentos em que aquela palavra aparece:

$$\text{IDF}(\text{palavra}) = \log_{10}\left(\frac{N}{\text{df}(\text{palavra})}\right)$$

onde $\text{df}(\text{palavra})$ é o número de documentos em que a palavra aparece, e N é o número total de documentos da base.

Palavras com baixo IDF são palavras comuns, que pouco agregam para a análise de textos em relação a alguma entidade que se queira caracterizar. Palavras com IDF muito alto são palavras exóticas, que podem ser devidas simplesmente a erros de grafia ou são palavras específicas, como códigos numéricos ou marcações de página. Em ambos os casos, queremos ignorar estas palavras.

Análise do vocabulário específico

A frequencia normalizada de uma palavra é dada por:

$$\text{freq}(\text{palavra}) = \log_{10}\big(1 + \text{contagem}(\text{palavra})\big)$$

onde $\text{contagem}(\text{palavra})$ é o número absoluto de vezes que a palavra ocorre no conjunto completo de textos (o *corpus*)

Agora construa os seguintes dados:

- Uma tabela com as 100 palavras mais relevantes (e respectiva relevancia) em paginas onde os itens que você decidiu estudar aparecem conjuntamente. A relevância de uma palavra será dada pela frequência de ocorrência desta palavra em páginas que contem o item em estudo, multiplicada pelo IDF da palavra. Ou seja:

$$\text{relevancia}(\text{palavra}) = \text{freq}(\text{palavra}) \times \text{IDF}(\text{palavra})$$

- Duas tabelas de 100 palavras nos mesmos moldes do paragrafo anterior, mas cada uma relativa a paginas onde os itens em estudo aparecem sem a presença do outro item.

(Opcional – para rubrica A) Analise do vocabulario especifico local

Repetir o trabalho da seção anterior, mas considerar apenas palavras que ocorram a uma distância fixa (digamos, as 5 palavras mais próximas para a direita e para a esquerda) do nome do item que você está estudando.

Analise e visualização dos resultados

Você descobriu algo interessante? Construa uma explicação dos seus resultados (deixando claro em sua escrita quando você está fazendo suposições e hipóteses, e quando você está baseado em dados concretos). Para visualização você poderia construir uma nuvem de palavras (word cloud), fica bacana.

Entrega

Você deve entregar um relatório no formato PDF. Este relatório deverá conter o código Python desenvolvido para este projeto (vai ser curtinho, prometo) com a explicação de como ele funciona. Deverá conter também as análises feitas, e as tabelas de palavras mais relevantes (NÃO A TABELA DE IDF, essa é um monstrengo!)

A seguinte rubrica se aplica

I: Não entregou ou entregou nonsense

D: Entregou o relatório contendo apenas a etapa de calculo da tabela de IDF, ou mesmo somente o calculo das frequencias de documento.

C: Entregou tudo funcionando, mas meio “bruto”: não usa broadcast para a tabela de IDF, nem otimiza as pipelines (fez 3 pipelines independentes ao inves de aproveitar a parte comum da analise dos 3 textos).

B: Tudo funcionando e otimizado, relatorio bem escrito

A: Foi alem: construiu uma analise incorporando o estudo do vocabulario local, e comparou com a analise do vocabulario global. Houve algum ganho de entendimento?

Cronograma

Este é um projeto bem curto. A entrega será para o dia 4/12, mas lembrem-se que vocês tem a semana de provas no meio desse prazo! As duas aulas desta semana (16/11 e 18/11) serão estúdios.

Dicas

- Antes de qualquer coisa, construa um pequeno dataset com apenas algumas paginas, para que voce possa trabalhar no seu laptop antes de rodar um cluster.
- Instale pyspark na sua maquina. Pode trabalhar no Jupyter Notebook ou no Zeppelin, como quiser.
- Quando tudo estiver rodando, faça um script Spark e rode no cluster como “step”, não como notebook. Não esqueça de marcar o auto-termino do cluster ao final da execução.