Projeto

Parte 1: Representação TFIDF

Frameworks para Big Data e Programação para Ciência dos Dados

Prof. Charles Ferreira e Paulo Nietto

1 Descrição

Grande parte da informação produzida pela humanidade ocorre digitalmente e uma parcela considerável está em formato textual. Consequentemente, existe uma necessidade de extrair informações da textos digitais de forma automática permitindo a realização de diversas análises dos dados.

Muitas técnicas de Aprendizado de Máquina são capazes de aprender informações de textos e prover as mais diferentes análises sobre o conjunto de documentos. Contudo, a grande maioria dessas técnicas não são capazes de lidar com textos em seu formato bruto, ou seja, da maneira em que nós seres humanos lemos. Desta maneira, precisamos transformar os documentos em uma estrutura numérica que seja conveniente para utilizar os algoritmos de Aprendizado de Máquina.

Neste projeto, vocês deverão fazer o pré-processamento de uma base de textos utilizando o framework Hadoop e a abordagem de Map Reduce para calcular o TFIDF de cada palavra.

TFIDF é uma peso (valor) que atribuímos para associar uma palavra com um documento de forma a representar sua importância. Logo, podemos calcular o TFIDF de uma termo (palavra) t associado a um documento d através da equação 1.

$$TFIDF_{t,d} = tf_d \times idf_t \tag{1}$$

tf_d (term_frequency) = número de vezes que o termo t aparece no documento d

$$idf_t \text{ (inverse document frequency)} = \log_{10}\left(\frac{N}{(1+df_t)}\right) \tag{2}$$

em que df_t = número de documentos em que o termo t aparece

e N é o número total de documentos.

2 Base de Textos

Vocês deverão trabalhar com uma base de textos de notícias (Disponibilizada no Blackboard). A base consiste de um único arquivo em que cada linha contém informações de um documento diferente.

Exemplo → trecho da base de textos:

```
1,3,Fears for T N pension after talks,Unions representing workers at Turner Newall say they
    are 'disappointed' after talks with stricken parent firm Federal Mogul.
2,4,The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (
    SPACE.com),"SPACE.com - TORONTO, Canada -- A second\team of rocketeers competing for the
    #36;10 million Ansari X Prize, a contest for\privately funded suborbital space flight,
    has officially announced the first\launch date for its manned rocket."
3,4,Ky. Company Wins Grant to Study Peptides (AP),"AP - A company founded by a chemistry
    researcher at the University of Louisville won a grant to develop a method of producing
    better peptides, which are short chains of amino acids, the building blocks of proteins."
4,4,Prediction Unit Helps Forecast Wildfires (AP),"AP - It's barely dawn when Mike
    Fitzpatrick starts his shift with a blur of colorful maps, figures and endless charts,
    but already he knows what the day will bring. Lightning will strike in places he expects.
    Winds will pick up, moist places will dry and flames will roar."
```

As informações de cada linha estão separadas por virgula, logo, para cada documento temos quatro informações (colunas):

1. a primeira coluna contém o id do documento;
2. a segunda coluna tem o assunto do documento representado por um valor numérico;
3. a terceira coluna possui o título do documento;
4. e a última coluna mostra o conteúdo do documento.

Ao todo, a base de textos possui 7600 documentos organizados em 4 assuntos diferentes. Não será necessário utilizar o assunto para calcular a métrica TFIDF, entretanto, essa informação pode ser relevante para fazer uma futura análise dos dados.

3 Exemplo: ToyData

Para exemplificar o cálculo do TFIDF vamos utilizar uma base de dados simples com o seguinte conteúdo:

```
1,1,,hadoop hadoop hadoop hadoop hadoop
2,2,,the book is on the table
3,1,,hadoop hadoop
4,2,,the book
```

Neste exemplo, temos 4 documentos os quais não possuem título (terceira coluna está vazia).

A Tabela 1 resume, de forma estatística, as informações dessa base de texto que são relevantes para o cálculo do TFIDF.

Tabela 1: Métricas da base de dados

termo (t)	id doc (d) em que aparece	número de vezes que aparece no doc (tf_d)	numero total de documentos com o termo t (df_t)
hadoop	1	5	2
book	2	1	2
is	2	1	1
on	2	1	1
table	2	1	1
the	2	1	2
hadoop	3	2	2
book	4	1	2
the	4	1	2

Com as informações da Tabela 1, podemos calcular o tfidf de qualquer termo t associado com qualquer documento d. Exemplo: TFIDF do termo *hadoop* no documento de id *1*.

$$TFIDF_{hadoop,1} = tf_1 \times idf_{hadoop} = 5 \times \log_{10}\left(\frac{4}{1+2}\right) = 0.624 \tag{3}$$

- **Saída do programa:**
  - Ao aplicar o Map Reduce sobre a base de textos pode-se gerar uma saída próxima a este formato:
  - `<chave;valor>` → `<idDoc_termo; tfidf>`
  - em que a chave seria idDoc_termo e o valor seria o cálculo do tfidf.
- **Aplicando na base de exemplo (saída final)**

```
1_hadoop;0.6246936830414996
2_book;0.12493873660829993
2_is;0.3010299956639812
2_on;0.3010299956639812
2_table;0.3010299956639812
2_the;0.24987747321659987
```

```
3_hadoop;0.24987747321659987
4_book;0.12493873660829993
4_the;0.12493873660829993
```

- **Dica:**
  - Não será possível calcular o tfidf utilizando somente um único programa com MapReduce.
  - Será necessário utilizar vários programas em que a saída de um programa irá alimentar a entrada de outro programa.

4 O que deve ser entregue

- Todos os códigos feitos em um arquivo.zip
- Um pdf contendo:
  - o nome e o RA de todos os integrantes do grupo;
  - uma breve descrição da estratégia utilizada para resolver o problema proposto.
    - ou seja, para cada programa com MapReduce o que foi usado como chave e como valor.
- Data de entrega: 27/10/2021