Mini-Projeto 3 - Recomendação

1. Introdução e Descrição dos Dados

Ler é uma atividade que traz uma série de benefícios, tais como estimular a criatividade e o senso crítico. Naturalmente, existem muitas formas de literatura sendo cada forma preferida por um grupo de pessoas. A empresa BookCrossing pensou: será que pessoas gostariam de outros livros que tenham sido bem avaliados por outras pessoas com características similares? Para responder esta pergunta, resolverem convidar você!

Para tanto, você irá dispor de dados referentes aos usuários, e.g., a localidade e a idade, dados referentes aos livros, e.g., ISBN e ano de publicação e as avaliações, em que cada usuário deu uma nota para os livros lidos entre 0 e 10. Esses dados foram anteriormente particionados em treino e teste pela empresa que já garantiu uma coisa: vocês não vão enfrentar o problema do novo usuário (usuário que não aparece no treino, só no teste) durante a predição das notas no teste, mas nada foi dito a respeito do problema do novo item.

2. Fases

O mini-projeto é dividido em duas fases. A primeira será mais longa (uma semana) e terá como objetivo principal o aprendizado das estruturas do Mahout desde a leitura dos dados até a escolha dos algoritmos, avaliação e predição de notas. A segunda parte, mais curta, terá apenas modificação nos algoritmos requisitando um estudo mais aprofundado dos algoritmos do Mahout e da seleção de atributos para o mesmo.

Em ambas as fases serão utilizados os mesmos dados de entrada (já disponibilizados desde a liberação desse documento) e serão produzidos os mesmos artefatos, sendo o da segunda fase uma evolução do anterior.

Ao término do mini-projeto os modelos propostos pelos alunos serão comparados em sala de aula com relação ao seu erro na predição.

Fase 1: Treinando, validando e predizendo notas para pares (usuário, livro)

Objetivos

Implementar e comparar alguns modelos para predição de notas de livros para usuários.

Descrição

Você deverá definir pelo menos dois modelos, um baseado no algoritmo **aleatório** e outro baseado em um algoritmo de **filtragem colaborativa pura**, ou seja, baseada na similaridade das preferências dos usuários pelos itens (baseada no item) ou dos itens pelos usuários (baseada no usuário).

Os modelos desenvolvidos deverão ser validados utilizando a base de treino disponibilizada (*mp3_book_crossing_treino.csv*) em **20 execuções** treino/teste calculando o RMSE, os resultados serão armazenados em um arquivo CSV (ver artefato 1 abaixo).

O melhor modelo será escolhido para estimar as notas para os livros dos usuários no conjunto de teste (*mp3_book_crossing_teste.csv*). As notas estimadas devem também ser armazenadas em um arquivo CSV (ver artefato 2 abaixo).

Deverá ser redigido também um documento relatando o processo utilizado para selecionar o melhor modelo (ver artefato 3 abaixo).

Artefatos

Todos os artefatos abaixo devem ser empacotados em um arquivo .zip com o nome no seguinte formato: `MP3_<NomeDoAlunoNesseFormato>_Fase1.zip`.

1. Arquivo CSV com Validação dos Modelos

Nome do arquivo

`<NomeDoAlunoNesseFormato>_Validacao_Modelos.csv`

Cabeçalho

```
"nome_modelo","execucao","rmse"
```

Requisitos

- O arquivo deverá ter 20 linhas para cada modelo proposto + 1 com o cabeçalho;
- O RMSE deve ser calculado pelo Mahout.

2. Arquivo CSV com Predições das Notas

Nome do arquivo

`<NomeDoAlunoNesseFormato>_<Nome-Do-Modelo-Nesse-Formato>_Predicoes.csv`

Cabeçalho

```
"nota"
```

Requisitos

- Lembre-se de manter a ordem e a quantidade de linhas do arquivo de teste (*mp3_book_crossing_teste.csv*). Exemplo, se a 3ª linha do arquivo de teste tem o usuário A com o livro X, a 3ª linha do arquivo de predição deve ter a predição correspondente;
- O arquivo deve ter 135.488 linhas com notas (tamanho do arquivo de teste) + 1 com o cabeçalho;
- As notas devem estar no intervalo [0, 10].

3. Documento

Nome do arquivo

`MP3_Analise.pdf`

Requisitos

O documento deve responder as seguintes perguntas:

- Qual foi o processo de seleção do melhor modelo de predição de notas?
    - Essa resposta deve incluir os algoritmos e hiper-parâmetros (tamanho da vizinhança, etc.) variados.
- Qual o melhor modelo em termos de *RMSE*? E em termos de *tempo* para predição? Por quê?
- Quais as maiores dificuldades no aprendizado e na utilização do Mahout até agora?
- O ad2_home ajudou na comparação dos modelos?

4. Código Fonte

Código fonte utilizado em uma parta `src`.

Entrega

16/12 (segunda-feira) até 12h59 (antes da aula)

Fase 2: Comparando novos algoritmos

Objetivos

Comparar dois novos algoritmos para a predição de notas.

Descrição

devem ser implementados ao menos dois novos modelos baseados nos seguintes algoritmos: **filtragem colaborativa** baseada na similaridade dos **atributos** dos itens e na **fatoração de matrizes** (também conhecido como baseado no SVD).

Artefatos

Os mesmos da primeira fase, com os mesmos nomes e requisitos. Devendo adicionar a validação para os novos modelos, as predições de notas (caso encontre outro melhor modelo), o documento de análise atualizado como também o código fonte.

Entrega

19/12 (quinta-feira) até 23h59 (dia anterior a última aula do ano)

3. Aplicações Web de apoio

A **ad2_home** o ajudará a selecionar o melhor modelo através de Intervalos de Confiança da média do RMSE por modelo. A aplicação receberá como entrada um arquivo no mesmo formato do que será submetido com os resultados da validação por modelo. A aplicação será liberada até a sexta (13/12).

A GUI **ad2_class** terá o mesmo gráfico da ad2_home, permitindo também carregar o arquivo de análise. Além disso, terá um gráfico que comparará os resultados das predições de todos os alunos entre si .

4. Nota

- 6.0 pontos : Fase 1
- 3.0 pontos : Fase 2

- 0.5 ponto : Presença em sala
- 0.5 ponto : Participação na lista do piazza

Atrasos

Os artefatos da Fase 1 poderão ser entregues com uma penalização pelo atraso até o momento final da entrega da Fase 2. No entanto, a cada dia de atraso será descontado 1 ponto (na terça valerá 5.0, quarta valerá 4.0 e na quinta valerá 3.0). A Fase 2 tem *hard deadeline* (não poderá ser atrasada).

5. Aulas

Aula 17/12: Terça-Feira no Reenge-08

Apresentação do resultado da Fase 1. Atenção! Você será argüido(a) neste momento pelos professores, o que pesará na sua avaliação! Esteja preparado(a)!

Aula 20/12: Sexta-Feira no Reenge-08

Apresentação do resultado final da Fase 2. Atenção! Você será argüido(a) neste momento pelos professores, o que pesará na sua avaliação! Esteja preparado(a)!

6. Lembretes

- Esse mini-projeto é **individual**.
- Todas as aulas terão lista de presença.

Bom aprendizado, Boas festas e Boas férias!