

Cuidados na aleatoriedade

Transcrição

Nosso avaliador já funciona e como resultado anterior tivemos um valor próximo de 0.29. Mas ao tentarmos executar novamente, percebemos uma variedade de resultados de 0.29 e passamos a receber `NaN` (*Not a Number*), depois 3, 1.5, 1.25 e 1.20. O que está acontecendo?

O que ocorre é que na implementação do Mahout do Java, a cada iteração do algoritmo os 90% e 10% da amostra são escolhidos aleatoriamente. Perceba que isso é oriundo da amostra total.

Por este motivo é possível perceber que não é uma boa ideia executar o algoritmo várias e várias vezes a fim de obter o valor que a gente queira diretamente, principalmente em testes. Em testes de unidade por exemplo, é uma prática comum fixar a aleatoriedade.

Para resolver esse problema, o Mahout tem uma classe chamada `RandomUtils` onde podemos utilizar o método `useTestSeed` para fixar o valor randômico da análise.

```
public class Avaliador{
    public static void main(String[] args) throws IOException, TasteException{

        RandomUtils.useTestSeed();

        File file = new File("dados.csv");
        FileDataModel model = new FileDataModel(file);

        RecommenderEvaluator evaluator = new AverageAbsoluteDifferenceRecommenderEvaluator();
        RecommenderBuilder builder = new RecomendadorDeProdutosBuilder();
        double erro = evaluator.evaluate(builder, null, model, 0.9, 1.0);
        System.out.println(erro);
    }
}
```

Ao fazer essa alteração não adianta quantas vezes o algoritmo será executado, o resultado será sempre o mesmo, que neste caso foi algo como 1.339.

O que precisamos considerar fortemente é que ao executar as análises sempre de forma aleatória encontraremos padrões que nos atendam ou não da forma que é desejada. Em alguns casos o algoritmo pode ser perfeito para um determinado conjunto de dados, porém péssimo para outros. Um exemplo? Se você perguntar os números da loteria e eu te disser sempre os últimos números que saíram, eu sempre estarei certo, mas muito provavelmente você jamais acertará as apostas.

Com isso já vimos os cuidados necessários e também colocamos em prática os algoritmos de recomendações e de avaliação do Mahout. Vamos utilizar dados reais agora?