Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Trabalho Prático 3

Sua habilidade em análises envolvendo Mineração de Dados possui agora uma repercussão internacional. Sabendo disso, uma relevante comunidade de música *online*, o *Last.fm*, resolveu contratar seus serviços para uma importante tarefa. O *Last.fm* possui em seu sistema milhões de músicas distintas e um grande problema que eles precisam tratar é o de agrupar semanticamente tais músicas. Uma vez que os usuários podem fazer *upload* de músicas próprias no sistema, a diversidade de músicas existentes é gigantesca. Dessa forma, não há como especialistas definirem para cada música gêneros tradicionais tais como sertanejo, pop, rock, dentre outros. Por outro lado, definir grupos de itens semelhantes é de extrema importância para o sistema de recomendação do *Last.fm*, bem como para os usuários, uma vez que eles podem navegar por itens pertencentes a um mesmo gênero. De forma a tentar resolver este problema, o *Last.fm* forneceu a você uma amostra da base deles contendo quase 200.000 músicas distintas. Cada música deste conjunto possui um conjunto de tags que representam as tags mais frequentemente atribuídas a tal música pelos usuários do sistema. A idéia seria identificar grupos distintos de músicas a partir das tags assinaladas a elas. Dessa forma, pede-se:

- Proponha uma técnica de agrupamento para este cenário e justifique sua escolha;
- Proponha uma estratégia de avaliação de qualidade dos grupos encontrados;
- Implemente sua proposta de agrupamento e avaliação de qualidade;
- Gere como saída os grupos identificados com maior qualidade, bem como os valores de qualidade de tais grupos.

Deve ser entregue uma documentação com todas as análises e decisões tomadas. Os critérios de correlação do presente trabalho serão todos embasadas na qualidade das discussões realizadas, bem como nas justificativas para as decisões tomadas. Vocês estão livres para utilizar qualquer tipo de

algoritmo de agrupamento, bem como utilizar implementações disponíveis na Web, ou bibliotecas de ferramentas tais como *Weka*, *R*, *RapidMiner*, *Pentaho* dentre outras. A documentação deve conter no máximo 10 páginas, com espaçamento simples entre linhas, tamanho de fonte 12 e fonte *Times New Roman*.

Deve-se também entregar o conjunto de código ou scripts utilizados para achar os grupos e a medida de qualidade proposta, os grupos identificados e o valor de qualidade associado a tais grupos para o conjunto de entrada disponibilizado.

Data de entrega: 10 de Outubro