Trabalho Prático 3 de Mineração de Dados

Aluno: Raony Guimarães

Requisitos:

- Proponha uma técnica de agrupamento para este cenário e justifique a 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-se também entregar o conjunto de códigos 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.

Objetivos

O objetivo deste trabalho é agrupar um conjunto de músicas de acordo com as suas similaridades. Utilizando técnicas avançadas de Mineração de dados será proposta uma solução para o problema que defina uma similaridade entre as músicas.

Material e Métodos

Para está análise foi fornecido um conjunto de dados com músicas do site last.fm. Cada música possui um grupo de “tags” que foram adicionadas pelos usuários do site.

Padronização das tags

Para realizar o agrupamento as tags de cada música foram padronizadas utilizando normalização, remoção de letras repetidas, acentos, traços e outros caracteres estranhos que dificultassem a identificação de tags similares.

Métricas escolhidas

TF-IDF

Para calcular a importância de cada tag foi utilizada a métrica TF-IDF (term frequency–inverse document frequency)[http://en.wikipedia.org/wiki/Tf%E2%80%93idf].

Esta medida é uma estatistica que avalia a importância de cada tag em relação a música a qual ela pertence e também em relação as outras músicas do conjunto inicial dos dados. A importância de uma tag para sua música aumenta de acordo com quantidade de vezes que ela aparece mas é regulada pela sua frequência global em

relação a todas as músicas onde ela aparece. Esta técnica Sendo assim para cada tag dentro de cada música foi gerado um score utilizando a métrica escolhida.

Similaridade de Cosine

Para calcular a similaridade entre as músicas foi escolhida uma métrica chamada de “similaridade de Cosine”[http://en.wikipedia.org/wiki/Cosine_similarity]. Esta métrica calcula representa a similaridade entre dois vetores medindo o coseno do ângulo entre eles. O coseno do ângulo entre dois vetores determina quando os vetores estão apontando na mesma direção.

Portando sendo possível representar as tags de cada música por seu TD-IDF score, utilizamos a similaridade de cosine para comparar todas as músicas do conjunto de dados.

Cluster Hieráquico

A técnica escolhida para agrupamento foi a clusterização hieráquica. A clusterização hierárquica é uma técnica deterministica e exaustiva que calcula a distância entre todos os items e depois anda pela matrix de distância agrupando items que possuem um treshold de similaridade definido pelo usuário.

O resultado desta técnica é a construção de uma árvore que expressa a distância relativa entre os items do grupo, neste caso das músicas.

Randomização e Amostragem

Devido a grande quantidade de dados foi extraída uma amostra aleatória dos dados para realizar os cálculos das distâncias. Para cada conjunto X de dados foi necessário calcular X*X distâncias sendo portanto um problema que sofre da maldição da dimensionalidade. O número de operações cresce exponencialmente fazendo deste um dos maiores problemas para a análise de grandes conjuntos de dados. Sendo assim, foram selecionadas uma amostra do conjunto de dados de tamanhos: 100 para realizar o agrupamento.

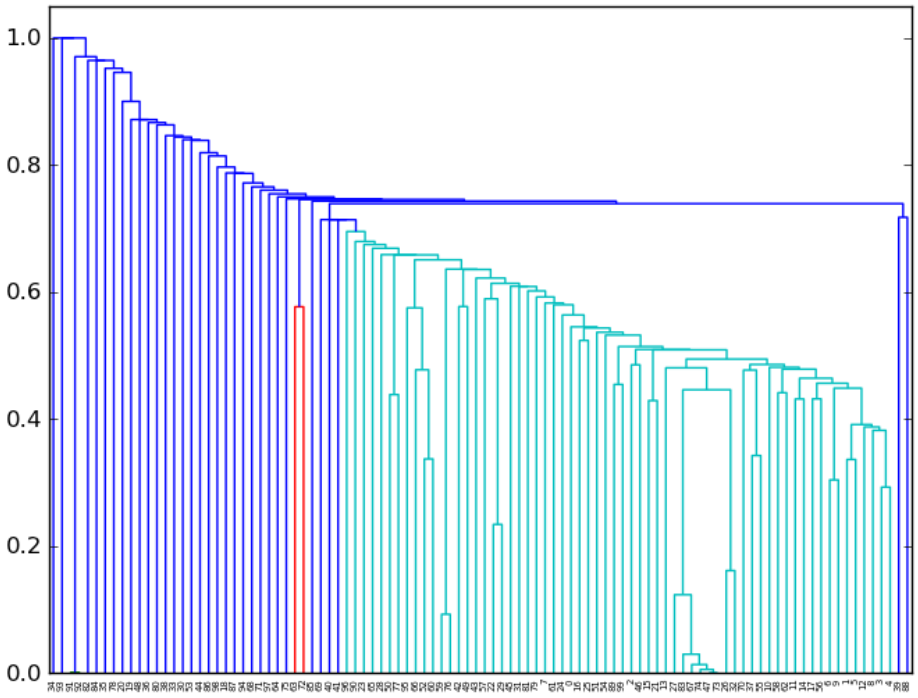
Avaliação de qualidade

Para avaliar a qualidade dos grupos foi utilizada a similaridade de cosine entre os membros de cada cluster sendo o valor 0 para músicas exatamente iguais e 1 para músicas completamente diferentes. Sendo assim os maiores resuldados são aqueles que possuem uma maior similaridade entre os membros de cada cluster.

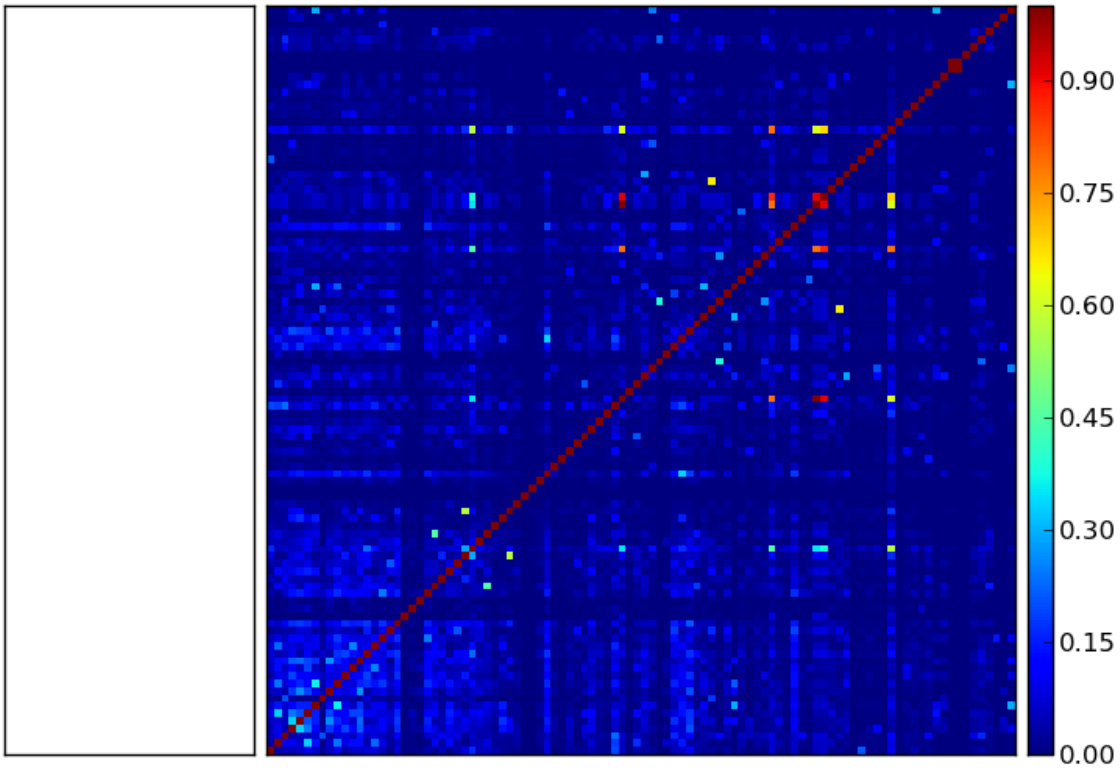
Resultados

Média de 9.43 tags por música.
1807257 tags, sendo 115730 únicas.
191538 música sendo 167857 únicas.

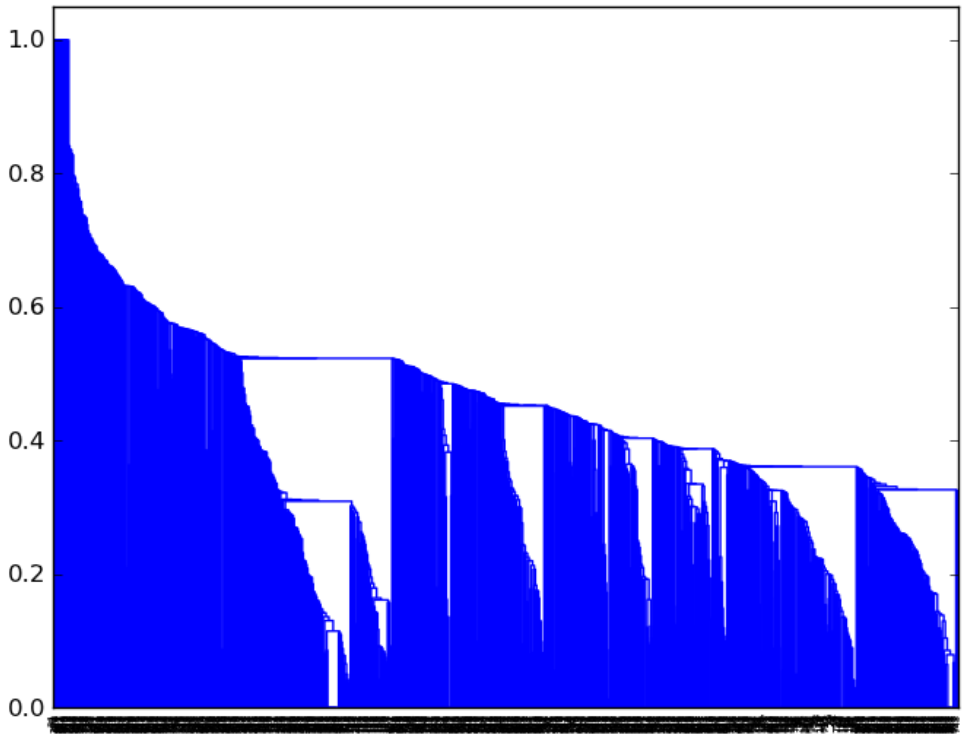
Dendograma gerado a partir da clusterização hierárquica de 100 músicas contidas no arquivo tags2.txt



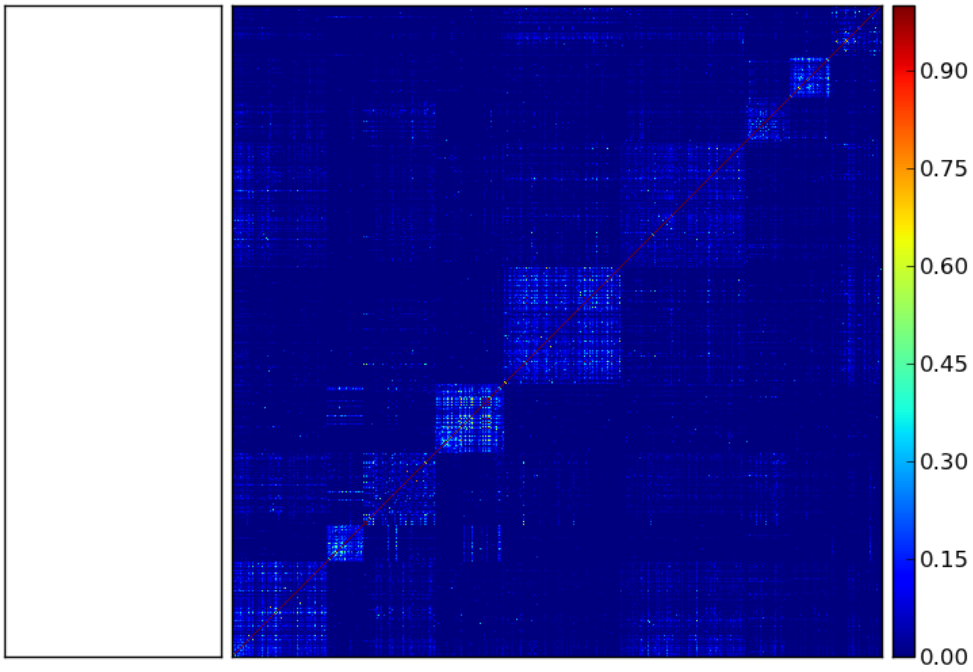
Matrix de distância para 100 músicas do banco de dados



Dendograma gerado para 1000 músicas do banco de dados



Matrix de distância para 1000 músicas do banco de dados



Os pontos azuis representam score 1 de distância, ou seja 0% de similaridade e os pontos vermelhos representam score 1 de distância ou seja 100% de similaridade. Os pontos amarelos representam os objetos similares.

Para uma melhor compreensão do problema, foram geradas algumas estatísticas do conjunto inicial como por exemplo a frequência das músicas e tags.

20 Tags Mais Frequentes
tag/frequência
rock	38549
alternative	22990
electronic	21510
indie	20916
pop	20484
hiphop	14592
femalevocalists	13968
favorites	11794
electronica	11411
alternativerock	11402
metal	11358
chilout	10925
indierock	10316
punk	10192
singersongwriter	9758

ambient	9626
dance	9611
love	8932
80s	8219
clasicrock	7902

20 Músicas Mais Frequentes
música/frequência
rock|466|
hardcore|400|
hiphop|368|
trance|367|
punk|311|
indie|301|
blackmetal|233|
electronic|230|
country|197|
heardonpandora|190|
pop|190|
deathmetal|170|
dance|169|
jaz|162|
ambient|158|
regae|152|
ska|149|
emo|138|
alternative|137|
|folk|117|

Clusters Gerados
Para o conjunto de dados com 100 músicas:
Usando treshold: 0.95
Número de clusters: 7
Melhores Clusters:
cluster sizes: [93, 2, 1, 1, 1, 1, 1]
Cluster 2
Tamanho 93
Cluster Distance: 1.37500079684
Items [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 83, 85, 86, 87, 88, 89, 90, 94, 95, 96, 97, 98, 99]
Cluster 3
Tamanho 2
Cluster Distance: 0.499188576284
Items [91, 92]

Para o conjunto de dados com 1000 músicas:

Usando treshold: 0.95
numero de clusters: 17
cluster sizes: [984, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Cluster 2
Tamanho 984
Cluster Distance: 6.09648257472

Discussão

Para o cluster hierárquico é necessário definir um valor de treshold para a geração dos clusters. Embora o dendograma do cluster tenha sido gerado não foi encontrado um valor de treshold que fosse suficiente para separar os cluster em conjuntos relevantes.

Conclusão

Através do uso de técnicas de Mineração de Dados, tais como clusterização hierárquica e utilizando métricas de similaridade como Cosine foi calcular a similaridade entre grupos de músicas embora a extensão dos cálculos para o conjunto total não seja possível por ser um problema np-hard ou seja, seus cálculos crescem exponencialmente.