Aprendizado de Máquina

UFMG EST171 - 1ª Lista de exercícios

Caio Cesar De Oliveira Freitas & Eduardo Elias Ribeiro Junior

12 de setembro de 2016

Exercício 1

> *Seu objetivo é usar as técnicas de redução de dimensionalidade, de segmentação (clustering) e de regras de associação para entender melhor um banco de dados que contém textos com resenhas sobre aplicativos da App Store do Android. Para isso, use a função load para carregar o banco* `dadosReviewGoogle.RData`*. Este banco contém dois objetos, textos, que contém as diferentes resenhas sobre os aplicativos, e notas, que contém as respectivas notas atribuidas pelos usuários que escreveram essas resenhas. Seu objetivo não é o de predição de notas, mas apenas o de melhor entendimento dos reviews.*
>
> *Para os itens que seguem, você pode trabalhar com um subconjunto dos dados originais.*

Para resolução dos itens propostos nesse exercício optou-se por utilizar apenas 200 resenhas, para que a computação dos algoritmos não ficasse demasiadamente lenta.

> ***a)*** *Mostre 5 resenhas do banco juntamente com suas respectivas notas.*

Tabela 1: Cinco primeiras resenhas com notas distintas

Resenhas	Notas
Não rodou Não rodou no meu galaxy note...perdi $$	1.00
Bugs pra consertar Não consigo jogar bem pq o campo aparece completamente com textura desconfigurada. Gostaria que resolvessem para aproveitar bem o jogo. Da forma que está os jogos gratis compensam mais... :(	2.00
Mais ou menos No galacxy y nem abril o jogo	3.00
Muito bom, só não gostei do alvo não centraliza, e deveria ter uma versão em português	4.00
Galax Note 10.1 Melhor jogo arcada que já joguei em algum smartphone ou tablet! Gráficos e jogabilidade quase perfeitos, o único defeito dele era não poder jogar sem acesso a internet, como resolveram isso ficou mto bom!	5.00

> ***b)*** *Use o código fornecido para converter os textos em uma matriz documento-termo binária (isto é, cada entrada da matriz indica se um termo está presente ou não no respectivo texto).*

Para transformar os textos das resenhas em matriz documento-termo binário realizou-se um processo de higienização de texto. Esse processo consistiu em remover as *"palavras de parada"* do idioma (preposições, artigos, conjunções, etc.), acentuação, pontuação, espaços em branco excedentes e números, pois essas são características de textos desinteressantes para o propósito de compreender o que textos diz com relação ao aplicativo avaliado. Após higienizado, aplicou-se o processo de radicalização do texto, ou seja, reduzir as

palavras ao seu radical (e.g. `avaliar, avaliação, avaliando -> avali`). Nesse processo de utilizou-se dois algoritmos de radicalização: o algoritmo de Poter[1] (MPTER) e o algoritmo RSLP (Removedor de Sufixos da Língua Portuguesa)[2] (MRSLP), cujo resultados são exibidos na Table 2 e na Figure 1.

Tabela 2: Estatísticas do número de radicais extraídos por cada algoritmo de radicalização

	Nº radicais	Média	Mediana	1º Quartil	3º Quartil	Desvio Padrão
PTER	645.00	2.60	1.00	1.00	2.00	6.32
RSLP	609.00	2.68	1.00	1.00	2.00	6.54

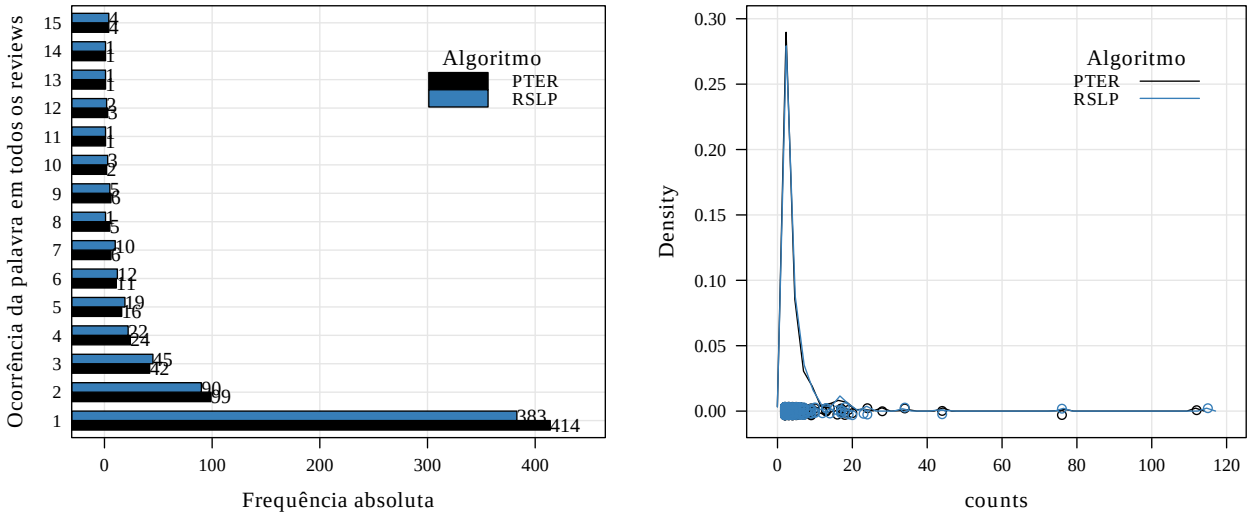

Figura 1: Número de ocorrências das palavras nas 200 resenhas. Frequências das ocorrências mais frequentes (esquerda) e Densidade empírica do número de ocorrências maiores que 2 (direita).

Note que ambos os resultados indicam que os algoritmos tiveram desempenho bastante similar. O comportamento das contagens das palavras* –no decorrer do texto denotaremos por palavras* os radicais indicados pelos algoritmos– foi bastante similar, ambos apresentando contagens baixas para a maioria das palavras*, como é o de se esperar em textos dessa natureza. Uma visualização das palavras* encontradas por cada algoritmo é realizada na Figure 2, novamente nota-se a similaridade dos algoritmos, pois as palavras* que se destacam na nuvem são as mesmas para ambos.

[1] http://snowball.tartarus.org/algorithms/portuguese/stemmer.html
[2] http://www.inf.ufrgs.br/~viviane/rslp/

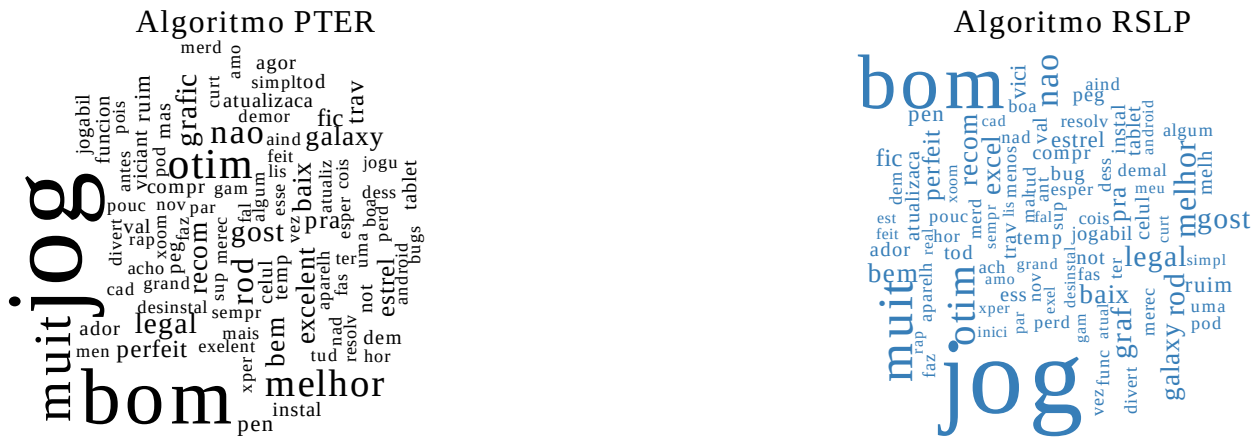

Figura 2: Nuvens de palavras* resultantes de cada algoritmo de radicalização.

Nesse trabalho preferiu-se prosseguir com a radicalização realizada pelo algoritmo de Poter, pois embora as radicalizações tenham sido similares, em alguns casos identificou-se *overstemming* no algoritmo RSLP. Para o algoritmo de POTER, em contraste com RSLP, houve casos de *understemming*, mas preferiu-se mantermos dessa forma para que não fossem agrupadas no mesmo radical palavras com sentidos distintos.

A matriz documento-termo binária referentes aos textos ficou então, com dimensão 200 documentos por 645 termos. A Figure 3 exibe a esparsidade dessa matriz.

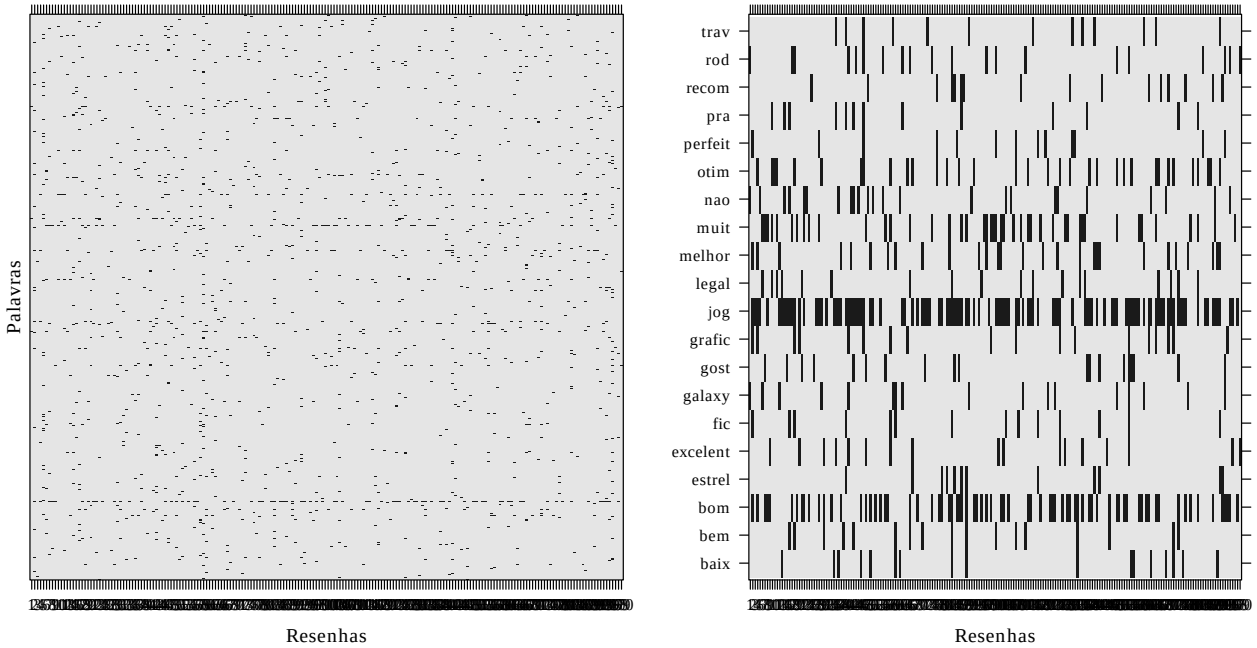

Figura 3: Visualização da matriz documento-termo binária. Regiões de preenchimento em preto representam a ocorrência do termo no documento. Matriz completa (direita) apenas os termos 3% mais frequentes.

---

c) *Use duas técnicas de clustering para criar agrupamentos dos diferentes textos (não use as notas para isso). Para a técnica que foi fornecida, faça também duas variações. Interprete os grupos obtidos por cada um dos métodos. Eles concordam entre si?.*

As técnicas de agrupamento utilizadas foram o método de *k-means*, onde utilizou-se 4 diferentes algoritmos para encontrar os k grupos cada qual com diferentes números de grupos e algoritmos hierárquicos definidos a partir de distâncias euclidianas com quatro diferentes definições de distância entre grupos.

K-means

No métodos *k-means* os quatro algoritmos utilizados para encontrar os grupos foram Hartigan-Wong[3], Lloyd[4], Forgy[5], MacQueen[6]. Para cada algoritmo de 2 a 15 grupos foram considerados. Os resultados são exibidos na Figure 4.

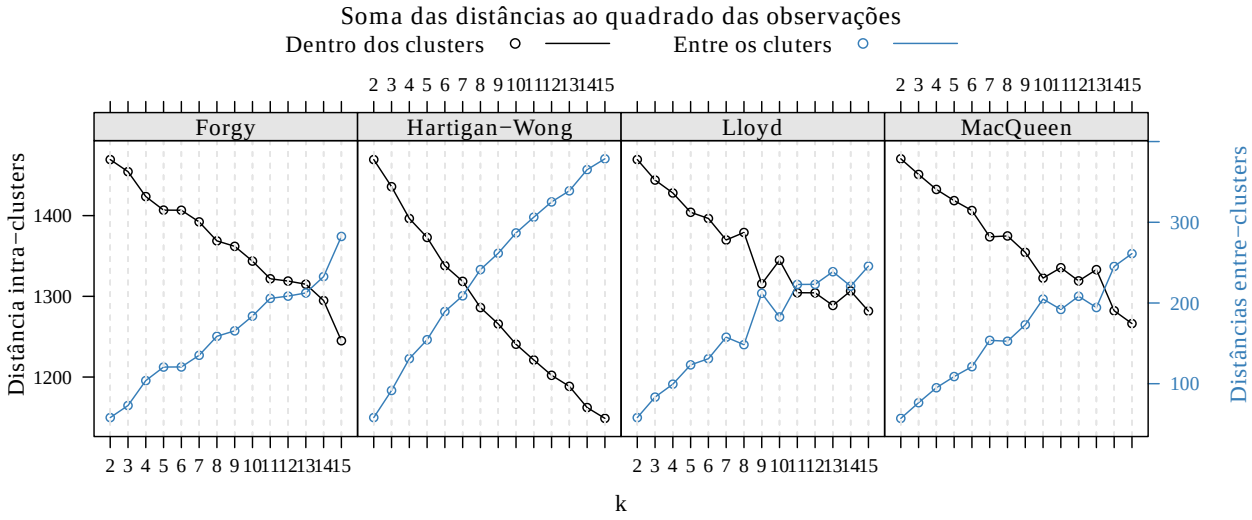

Figura 4: Distâncias intra e entre clusters para cada algoritmo definido com k clusters

Na Figure 4 são exibidas as somas das distâncias entre as observações e o centro do cluster a qual pertence, distâncias intra-clusters (em preto) e as distâncias entre os centros dos grupos, distâncias entre-clusters (em azul). Em geral, um bom agrupamento garante uma pequena distância intra-cluster e alta entre-clusters, pela figura nota-se que os algoritmos levam a diferentes escolhas do número de grupos.

A escolha do número de grupos sob o método de k-means é largamente trabalhado na literatura, para que a escolha não se torne puramente subjetiva adotou-se a utilização do índice de GAP, proposto por Robert Tibshirani em 2001[7]. No software R esse índice pode ser computado pela função `index.Gap` do pacote `clusterSim`, essa função retorna um objeto `diffu`, onde o número de grupos indicado por esse critério será o menor cujo `diffu` é maior que 0. A Figure 5 exibe os valores de `diffu` para cada um dos algoritmos.

[3]Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100–108.

[4]Lloyd, S. P. (1957, 1982) Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory 28, 128–137.

[5]Forgy, E. W. (1965) Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21, 768–769.

[6]MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. Berkeley, CA: University of California Press.

[7]Tibshirani, R., Walther, G., Hastie, T. (2001), Estimating the number of clusters in a data set via the gap statistic, "Journal of the Royal Statistical Society", ser. B, vol. 63, part 2, 411-423.

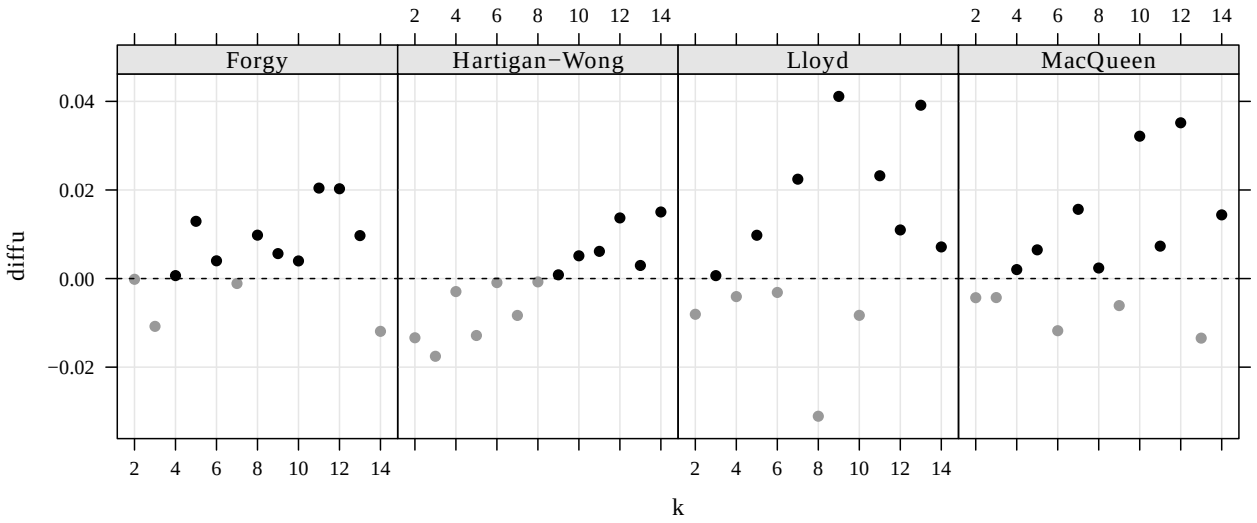

Figura 5: Diferenças de índices GAP para os diferentes algoritmos.

Novamente pelos índices de GAP há diferentes indicações do número ideal de clusters, foram indicados 4, 9, 3, 4 quando utilizados os algoritmos Forgy, Hartigan-Wong, Lloyd, MacQueen respectivamente. A Table 3 exibe os valores das distâncias intra e entre classes e nota-se que independente do número de clusters escolhido o algoritmo de Hartigan-Wong obtém os melhores resultados.

Tabela 3: Distâncias intra e entre clusters para os k's indicados

Algoritmo	k	Dist. intra	Dist. entre
Hartigan-Wong	3.00	1435.85	91.61
Lloyd	3.00	1444.08	83.38
Forgy	3.00	1454.47	72.99
MacQueen	3.00	1451.01	76.45
Hartigan-Wong	4.00	1396.43	131.03
Lloyd	4.00	1428.05	99.41
Forgy	4.00	1423.60	103.86
MacQueen	4.00	1432.44	95.02
Hartigan-Wong	9.00	1265.82	261.64
Lloyd	9.00	1315.43	212.03
Forgy	9.00	1361.96	165.50
MacQueen	9.00	1354.47	172.99

Clusterização Hierárquica

Considerando o método de clusterização hierárquica utilizou-se diferentes abordagens na definição de distâncias entre os grupos formados a cada passo do algoritmo. Foram utilizadas o método de `ward.D2`, que quantifica a diferença entre a soma dos erros quadráticos de cada padrão e a média da partição em que está a observação está contida; `single` distância é dada pela menor distância entre dois elementos de grupos distintos; `complete` distância é dada pela maior distância entre dois elementos de grupos distintos; e `average` distância é dada pela média das distâncias entre todos os elementos de grupos distintos.

Os dendrogramas referentes aos agrupamentos utilizando os diferentes tipos de ligação (definição de distâncias entre grupos) são exibidos na Figure 6. Para agrupamentos hierárquicos a escolha de grupos é realizada de forma subjetiva à interpretação gráfica. Nesse caso nota-se que, assim como no *k-means*, diferentes configurações do método levam a diferentes agrupamentos, no caso da ligação `ward.D2` parece que pode-se dividir os textos em dois grupos, já para os demais essa divisão não é clara indicando que não há um agrupamento claro.

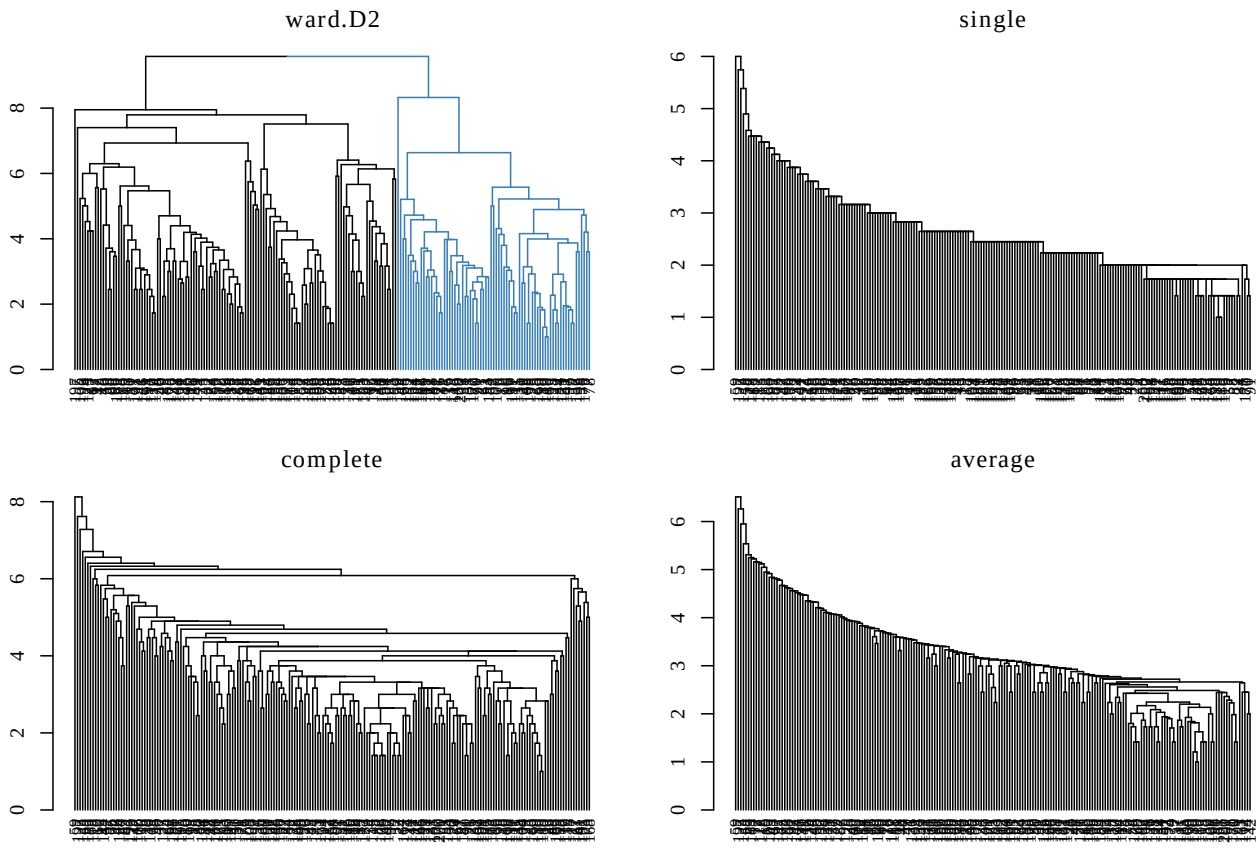

Figura 6: Dendogramas referentes aos agrupamentos hierárquicos com diferentes tipos de ligação.

Para comparação dos métodos escolheu-se construir dois grupos utilizando o agrupamento hierárquico com ligação de grupos dada pelo método de `ward.D2` e utilizando o agrupamento por *k-means* pelo algoritmo de Hartigan-Wong. Foram agrupadas 88 observações no 1º cluster e 112 no 2º pelo *k-means* e considerando o agrupamento hierárquico foram 75 e 125 observações nos 1º e 2º clusters respectivamente.

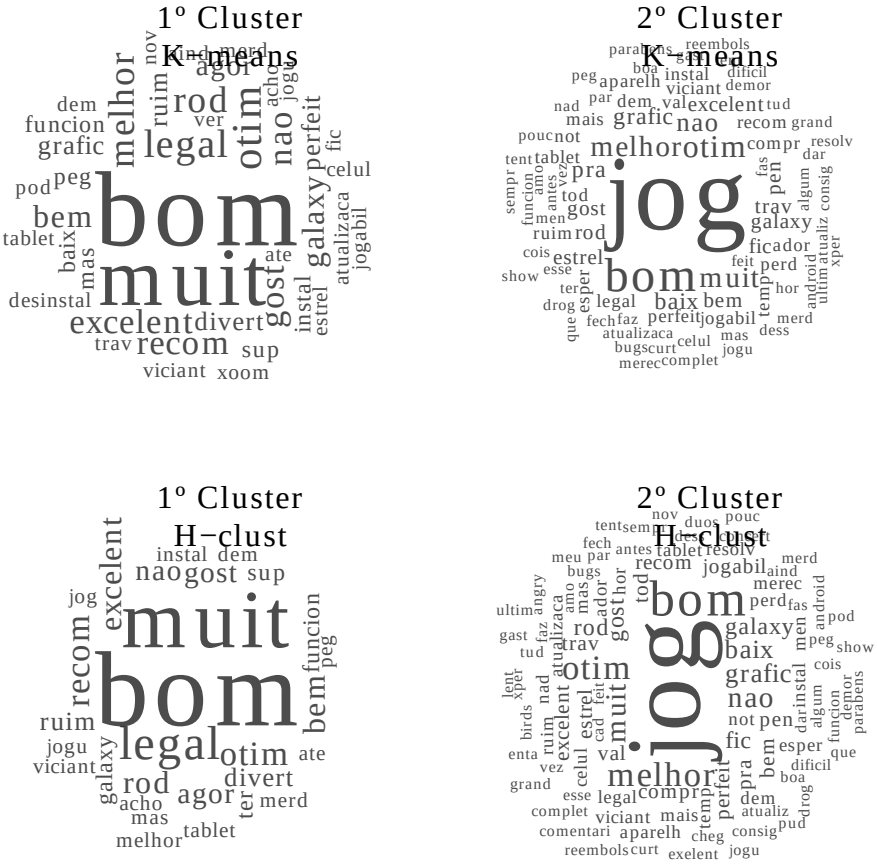

Figura 7: Nuvens de palavras* mais frequentes em cada cluster definido por cada método.

Na Figure 7 são exibidas as palavras* masi comuns em cada um dos clusters formados por cada algoritmo. Nota-se que os grupos parecem ter a mesma característica, seja por um agrupamento similar ou pela predominância dessas palavras* na amostra selecionada. Pode-se inferir sobre os grupos formados que o 1º se refere a resenhas essencialmente positivas sobre os aplicativos enquanto o 2º cluster tem a predominância de palavras relacionadas à jogabilidade do aplicativo, sendo resenhas positivas ou negativas.

Tabela 4: Amostras de resenhas retiradas de cada cluster

Método	Cluster	Amostra de texto 1	Amostra de texto 2
K-means	1	É virus Eu baixei tals e meu anti virus foi ger se tinha algum virus e detectou que ele era eu desinstalei e meu celular ficou bem melhor! Meu celulalr é o \samsung galaxy ace duos\	Nota 10 Jogo exelente bom grafico com boa jogabilidade recomendo
K-means	2	Excelente!!! Muito bom...Recomendo!	Galaxy s Um dos melhores jogos de puzzle
H-clust	1	Bom e ruim Gráficos muito bons, mas a jogabilidade no touchscreen é ruim. Joguei um pouco, matei a curiosidade e desinstalei.	Nao gostei Nao gostei pq nao pego no meu tablet, pq esse jogo é pra pc
H-clust	2	Excelente Rodou liso no s3...	Muito bom amei viciante nao paro d jogar

Na Table 4 são exibidos duas amostras de resenhas retiradas aleatoriamente dos cluster formados por cada

um dos métodos. Note que nessas amostras não há um claro padrão nos textos. assim como já observado na Figure 7 ambos os grupos são similares, diferendo somente na predominância das palavras que o compõe.

---

d) *Mostre as 5 regras de associação encontradas (não use as notas para isso) usando o algoritmo a priori com maior suporte, as 5 com maior confiança e as 5 com maior lift. Interprete o valor do suporte, lift e confiança de um regra de sua escolha. Mostre ao menos 3 maneiras distintas essas regras visualmente.*

Nesse exercício foram encontradas as regras de associação cujo contém um suporte, proporção do conjunto de palavras* ocorrer simultaneamente em um documento, superior a 0,04, e confiança, probabilidade do conjunto de itens consequente (RHS) dado a ocorrência do conjunto de itens antecedente (LHS), superior a 0,60. Com essas configurações foram encontrados um conjunto de 21 regras. Na Table 8 são exibidas as cinco regras com melhor desempenho considerando seu suporte, confiança e lift (razão da confiança pela probabilidade do consequente).

Tabela 5: Algumas regras de associação com melhores suporte, confiança e lift

Regras	Suporte	Confiança	Lift
`Melhores Suporte`			
{muit} => {bom}	0.15	0.68	1.79
{otim} => {jog}	0.12	0.68	1.21
{melhor} => {jog}	0.10	0.68	1.21
{nao} => {jog}	0.08	0.67	1.19
{jog,muit} => {bom}	0.08	0.84	2.22
`Melhores Confiança`			
{jogabil} => {grafic}	0.04	1.00	11.11
{pra} => {jog}	0.06	0.92	1.65
{pen} => {jog}	0.04	0.90	1.61
{val} => {jog}	0.04	0.89	1.59
{tod} => {jog}	0.04	0.89	1.59
`Melhores Lift`			
{jogabil} => {grafic}	0.04	1.00	11.11
{grafic,jog} => {otim}	0.04	0.62	3.62
{jog,muit} => {bom}	0.08	0.84	2.22
{muit} => {bom}	0.15	0.68	1.79
{pra} => {jog}	0.06	0.92	1.65

Para exemplificar a interpretação dessas medidas calculadas para cada regra tome a regra de maior `{grafic,jog} => {otim}`, essa regra diz que resenhas que contém os radicais `grafic` e `jog` tendem a também conter o radical `otim`. Essa regra tem um suporte de 0.04, ou seja, em 4% dos textos essas palavras ocorrem simultaneamente. A confiança dessa regra é de 0.6153846, isso diz que dentre as resenhas que contém `grafic` e `jog`, 61.5384615% contém também `otim`. E finalmente dado que a resenha contém os radicais `grafic` e `jog`, há um aumento de 3.6199095 na probabilidade de se observar `otim` na resenha.

Acima exibimos textualmente as regras e suas medidas de qualidade, porém são várias as visualizações propostas para regras de associação[8]. Na Figure 8 são apresentadas três formas de visualização. Ambas são exprimem os mesmos resultados, porém com ênfases distintas. No gráfico superior direito temos a dispersão dos dados com relação ao suporte e confiança nos eixos x e y e o lift é apresentado em escola de cores dos

[8]Hahsler, M., Grün, B., Hornik, K. (2005). arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. Journal of Statistical Software, 14(15), 1 - 25.

pontos. Note que a maioria das regras tem suporte baixo e confiança superior a 0.7, ainda pode-se observar que as regras de maior lift estão associadas a menores valores de suporte. No gráfico superior esquerdo temos a representação em forma de grafo e podemos observar claramente que há a predominância de regras que levam a consequência `jog`. No gráfico inferior também nota-se a predominância de regras com RHS `jog` e com relação ao antecedente temos essencialmente regras formadas por apenas um radical.

---

e) *Implemente componentes principais para esses dados (não use as notas para isso). Mostre quais são as 5 variáveis que recebem os maiores coeficientes (cargas) no primeiro componente. Mostre também as 5 variáveis que recebem os menores coeficientes (cargas) no primeiro componente. É possível interpretar essas palavras? Faça o mesmo com o segundo componente. Faça um diagrama de dispersão dos dois primeiros componentes principais. Use uma cor para cada ponto de acordo com a nota atribuída. Há um relação entre os componentes encontrados e as notas atribuídas? Você consegue encontrar outliers com base nesses gráficos? Mostre ao menos três textos outliers. Repita o procedimento usando os três primeiros componentes, isto é, usando um gráfico em 3d.*

Na Table 6 e Table 7 são exibidos os cinco maiores carregamentos para os 2 e 3 primeiros componentes principais respectivamente. Note que as palavras com maiores cargas nesses componentes se repetem, isso se dá pela predominância de ocorrências dessas palavras nos textos. Esse é o principal fator que dificulta a interpretação dos componentes. Todavia pode-se notar na 2º componente que todas as cargas são positivas e relacionas a radicais que exprimem uma noa avaliação de um aplicativo com relação a jogabilidade e gráficos. Já a 1ª componente apresenta o radical `jog` com carga negativa o que significa que textos que contém esse radical terão valores menores dessa componente assim como para os radicais `nao` e `fic`.

Tabela 6: Menores carregamentos

1º Comp.		2º Comp.		3º Comp.	
word	loading	word	loading	word	loading
magnif	-2.1E-04	simpl	-3.4E-06	orrivel	5.9E-05
kkkk	-2.7E-04	estand	1.6E-05	shopping	5.9E-05
affz	-2.7E-04	conect	1.6E-05	sum	5.9E-05
che	-3.2E-04	pouquinh	1.6E-05	etinh	5.9E-05
ficarej	-3.2E-04	period	1.6E-05	kkkk	1.1E-04

Tabela 7: Maiores carregamentos

1º Comp.		2º Comp.		3º Comp.	
word	loading	word	loading	word	loading
bom	0.586	jog	0.686	otim	0.494
jog	-0.478	bom	0.499	grafic	0.383
muit	0.447	otim	0.226	nao	-0.324
nao	-0.165	muit	0.150	jog	-0.236
fic	-0.136	grafic	0.129	excelent	0.228

Na Figure 9 exibi-se a dispersão das componentes principais calculadas para cada resenha. À esquerda são as duas primeiras componentes principais a à esquerdas as três primeiras. Não há uma visível relação das componentes com as notas atribuídas pelos autores dos textos, há um número muito maior de notas extremas (considerando notas 5 e 1 têm-se aproximadamente 75% das observações) e essas notas estão dispersas de forma não padronizadas nos gráficos. Nos gráficos também foram destacados os 3 pontos *outliers*, a definição desses cinco pontos se deu pelas maiores distâncias com relação ao ponto mediano das componentes principais. Embora com esse definição possa se ordenar as observações quanto a dissimilaridade com relação a característica comum e identificar quantas observações menos semelhantes forem necessárias, apenas duas observações de destacam nos gráficos tanto considerando 2 como 3 componentes.

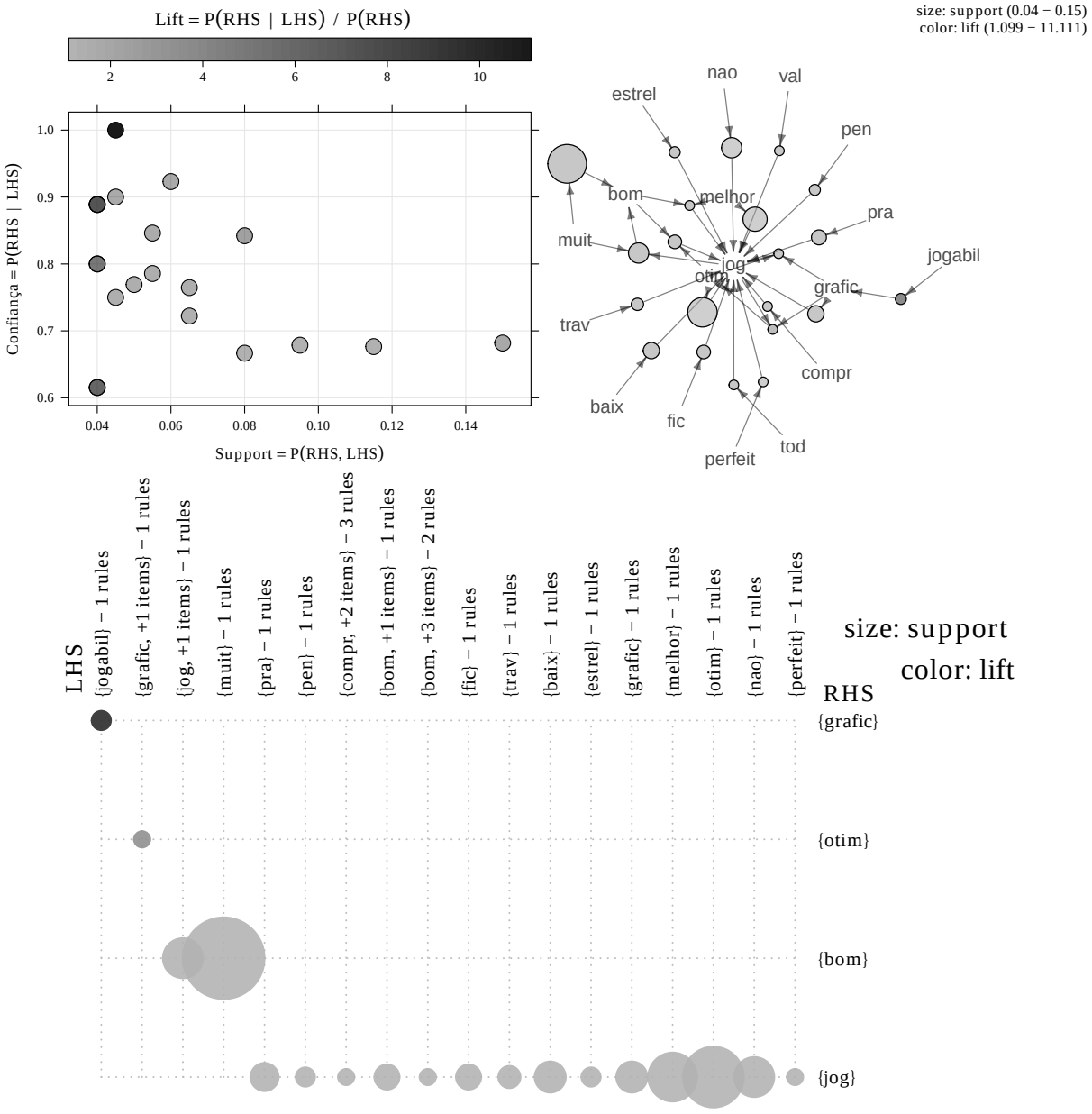

Figura 8: Visualização das regras de associação dos radicais presentes em cada resenha. Gráfico de dispersão com escala de cores para lift (superior à esquerda), visualização baseada e grafos com radicais e regras como vértices e Visualização baseada em matriz agrupada de regras (inferior).

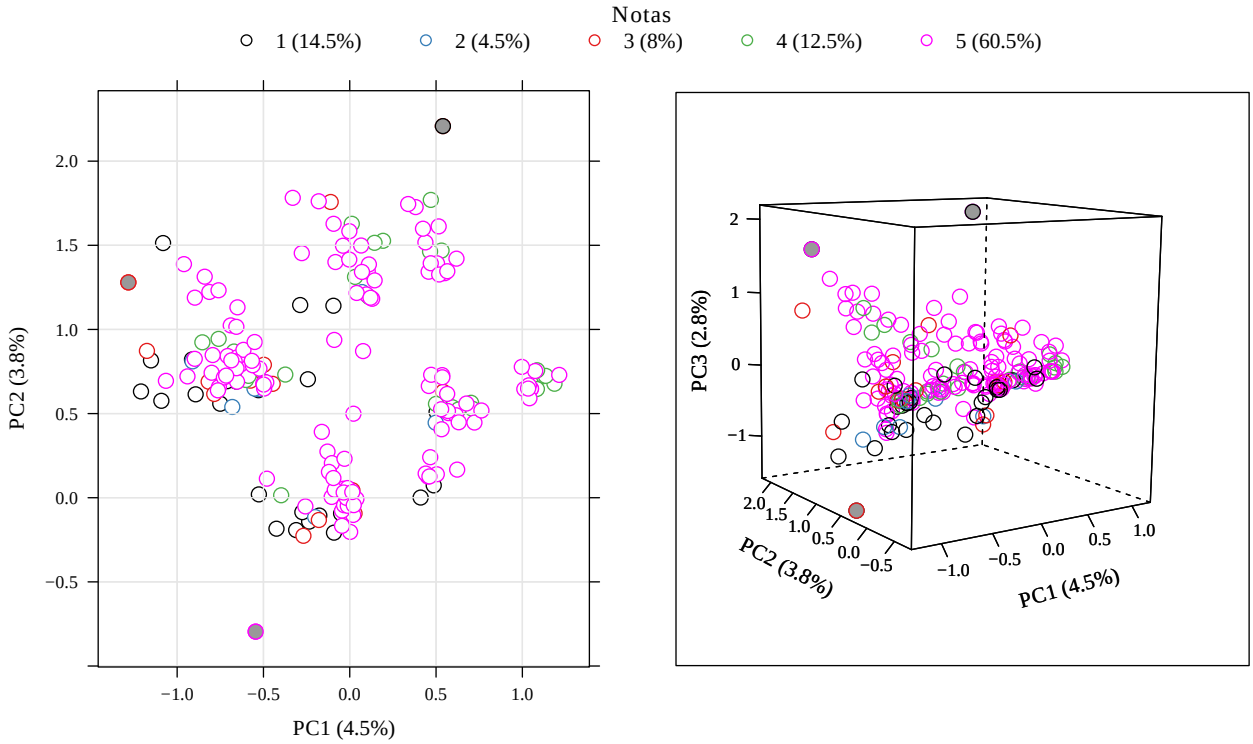

Figura 9: Dispersão das componentes com a indicação das notas atrucuídas a cada texto.

Tabela 8: Algumas regras de associação com melhores suporte, confiança e lift

Id	Distância	Texto
`Considerando 2 componentes`		
183	1.63	Otimo Muito bom esse jogo so que, tudo o que vc faz gasta a energia , e se vc naum quer comprar tem q esperar um eternidade pra encher . O que eu acho mais legal eh dar festas pra conseguir novos membros kkk
59	1.57	Xoom 2 e Sansug Galaxy Tablet 2 7.0 meus dois tablets acontece de travar no inicio da parte 11 do Smligglers´ DEM justamente quanto é apresentado o pássaro amarelo. uma pena, no iPhone funciona que é uma maravilha. ///****** RESOLVIDO ****//// vá em Configurações => uso de dados (ative a função) => definir limite de dados móveis (ative esta função). Após, reinicie dica do \esdrascps\ abraços
192	1.37	Estranho Tem uns bugs estranhos, as vezes na hora de pular fica travando e morre se ñ fosse por isso daria 5 estrelas porque o jogo é otimo.... pena espero melhorias.....
`Considerando 3 componentes`		
58	2.27	Motorola Atrix 4G Vi muitos comentários negativos, o que me deixou com medo de atualizar. Depois dessa atualização o que era ótimo ficou excelente. Alterações nos gráficos e cenários excelente, exige mais precisão. Se pudesse dava nota 6.
127	2.18	Nem abre no LG 4X HD O jogo parece ter ficado com excelentes gráficos e o preço é ótimo. Mas se ele nem ao menos abrir no seu aparelho é complicado. Espero que os donos de aparelhos com Tegra 3 não percam tempo comprando esse jogo.

5	1.86	Decepciona! Quando chega na fase de desmontar as bombas do lagarto o jogo simplesmente fecha! Nessa parte do jogo se eu der um tapinha em qlqr bandido dda cidade o jogo simplesmente fecha! E o engraçado é que tem varias reclamaçoes de usuarios sobre esse problema e o desenvolvedor simplesmente caga ou nao dá nenhuma satisfação! Decepcionante

---

f) *Implemente kernel PCA para esses dados, e trabalhe com ao menos duas variações dela. Plote novamente o gráfico de dispersão para essas novas técnicas. Eles são muito diferentes entre si? E com relação a componentes principais? Repita o procedimento usando os três primeiros componentes, isto é, usando um gráfico em 3d.*

Para esse exercício utilizou-se os kernels Gaussiano $K(x_i, x_k) = \exp(-\sigma \|x_i, x_k\|^2)$, com o hiperparâmetro σ fixado em 0.3 e o kernel Polinomial $K(x_i, x_k) = (\alpha + \gamma \langle x_i, x_k \rangle)^\delta$ com os hiperparâmetros $alpha = 1$, $\gamma = 1$ e $\delta = 1.2$. Os resultados das componentes principais não lineares utilizando as expansões por kernels definidas acima são exibidos na **??** e **??** utilizando duas e três componentes.

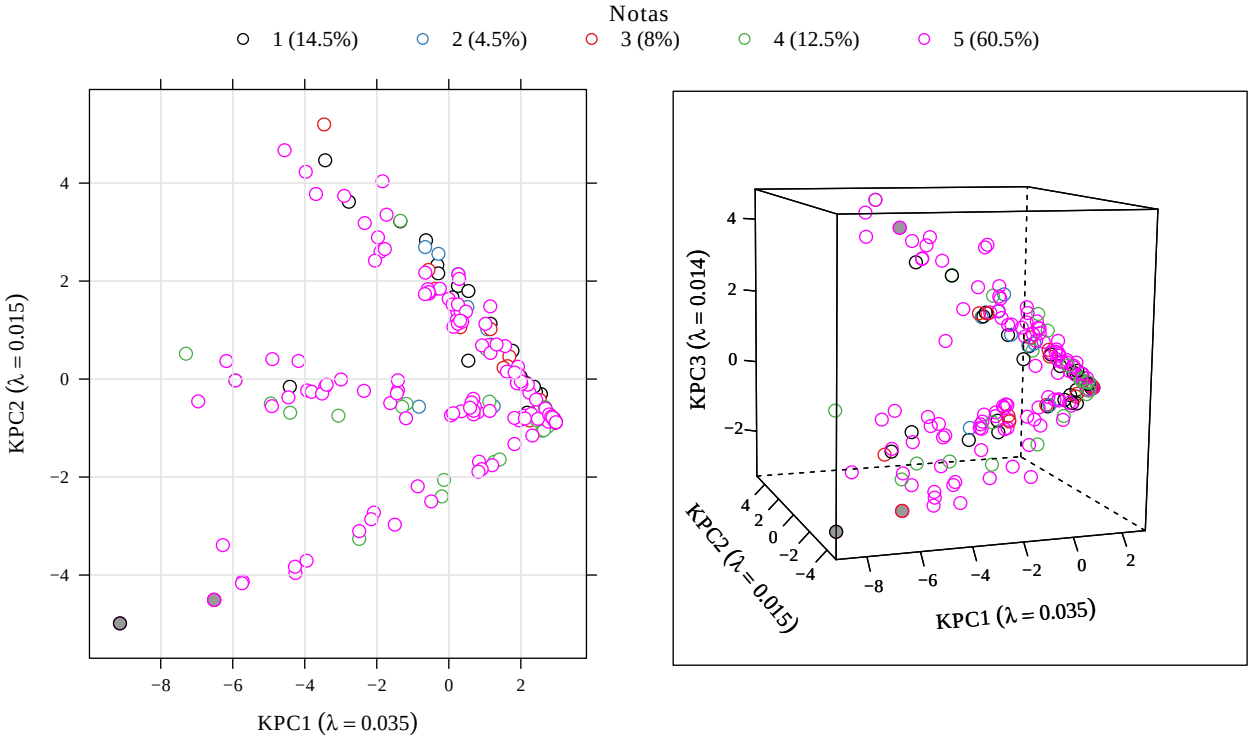

Figura 10: Dispersão das kernel componentes gaussianas com a indicação das notas atribuídas a cada texto.

Considerando o kernel Gaussiano, Figure 10, observa-se um comportamento bastante incomum, em forma de cone. Mesmo com três componentes esse comportamento permanece. Essa abordagem se mostrou bastante distinta dos componentes principais lineares, porém ainda não identifica-se claramente uma relação das componentes com as notas atribuídas embora que as notas mais baixas estejam essencialmente com valores positivos da segunda componente.

Ao considerar o kernel Polinomial, Figure 11, também temos um comportamento distinto das demais abordagens. Ainda com essa transformação não há relação das componentes com as notas, porém com

essa transformação pode-se identificar mais claramente um outlier, o texto 59. Assim como nas demais abordagens a consideração da terceira componente contribuiu pouco para a discriminação visual.

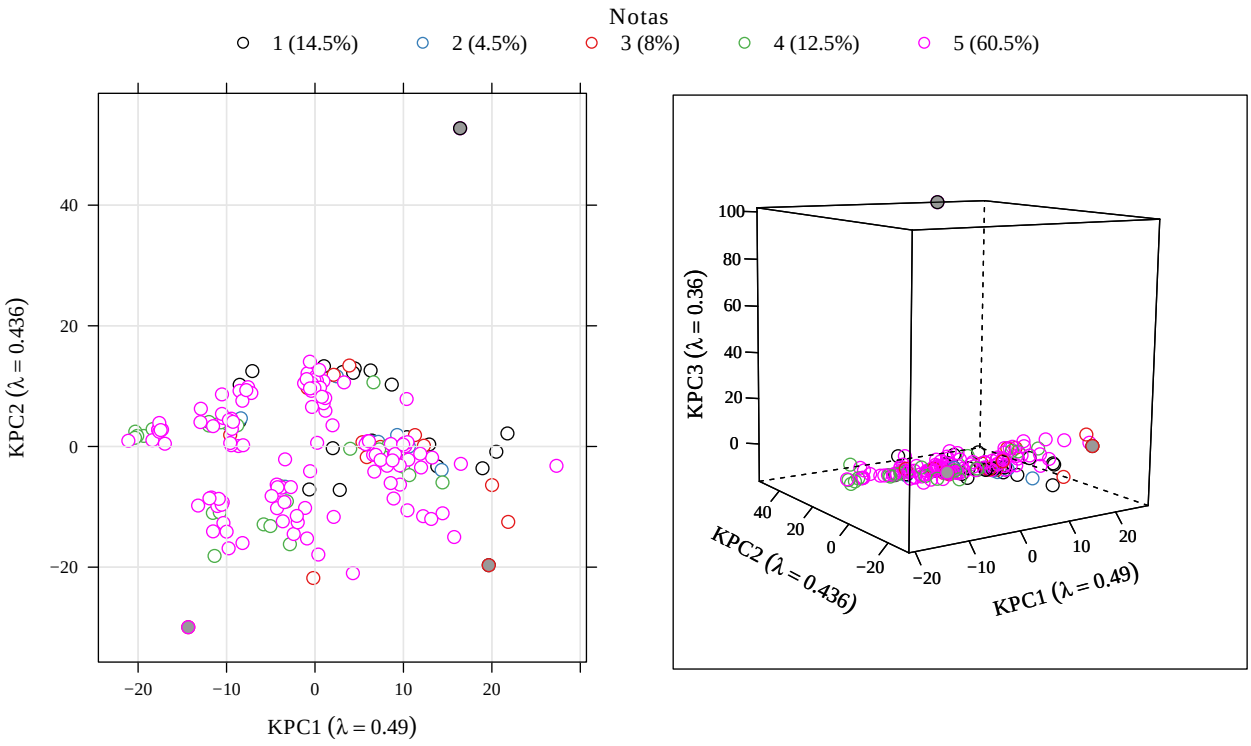

Figura 11: Dispersão das kernel componentes polinomiais com a indicação das notas atribuídas a cada texto

Exercício 2

Baixe o arquivo lista4.R. Ele mostra um código para baixar o banco de dados IncomeESL, que será utilizado neste exercício. Este banco mede diversas covariáveis em indivíduos americanos, como salário, origem e nível de educação. O código fornecido coverte este banco para o formato transactions, que será usado para implementar as regras de associação vistas em aula. Em particular, o código discretiza as variáveis numéricas. Usando o algoritmo a priori:

O problema consiste em realizar uma análise relacionada a determinados atributos de indivíduos americanos, como o salário, origem, nível de educação, ocupação, idade, entre outros. O objetivo é encontrar regras de associação que seja relevante, a fim de compreender mais a fundo o perfil dos indivíduos americanos.

As informações estão contidas no conjunto `IncomeESL`, disponível no pacote `arules`, do *software* R. No conjunto, dispomos de 8993 observações e de 14 variáveis, cujo exibimos abaixo:

- income: renda
- sex: sexo
- martial status: estado civil
- age: idade
- education: nível educacional
- ocupation: ocupação
- years in bay area: anos na baía
- dual incomes: dupla renda
- number in household: qtd. de moradores
- number of children: qtd. de crianças
- householder status: status do proprietário
- type of home: tipo de casa
- ethnic classification: classificação étnica
- language in home: idioma na casa.

Vale ressaltar que todas as variáveis ou são qualitativas ou foram categorizadas.

---

a) *Mostre as 10 regras (juntamente com suporte, confiança e lift) com maior lift entre regras com suporte de ao menos 0.001, confiança ao menos 0.8, e tamanho máximo 3.*

Tabela 9: Dez melhores regras de associação com relação ao lift, sujeitas a suporte > 0.001, confiança > 0.08 e com no máximo três itens

Regras	Suporte	Confiança	Lift
{marital status=divorced,language in home=spanish} => {ethnic classification=hispanic}	0.00	0.97	7.61
{occupation=laborer,language in home=spanish} => {ethnic classification=hispanic}	0.01	0.94	7.37
{occupation=retired,language in home=spanish} => {ethnic classification=hispanic}	0.00	0.92	7.22
{occupation=unemployed,language in home=spanish} => {ethnic classification=hispanic}	0.00	0.91	7.19
{number of children=1+,language in home=spanish} => {ethnic classification=hispanic}	0.03	0.90	7.13
{income=$0-$40,000,language in home=spanish} => {ethnic classification=hispanic}	0.04	0.90	7.08
{number in household=2+,language in home=spanish} => {ethnic classification=hispanic}	0.03	0.90	7.06
{type of home=house,language in home=spanish} => {ethnic classification=hispanic}	0.03	0.89	7.03
{occupation=student,language in home=spanish} => {ethnic classification=hispanic}	0.01	0.89	7.01
{marital status=widowed,language in home=spanish} => {ethnic classification=hispanic}	0.00	0.89	7.00

Com base em Table 9, temos que a regra com maior *lift* fornece a informação que se o indivíduo é divorciado e o idioma em sua casa é o espanhol, então a probabilidade dele ser hispânico aumenta em 7.61 vezes. No geral, as regras com os 10 maiores *lifts* implicam no consequente em que a classificação étnica do indivíduo é

hispânico. Além disso, tais regras possuem confianças altas. No mais, conclui-se que todos os antecedentes contém a informação de que o idioma falado na residência é espanhol, e isto aumenta a chance de que a etnia do indivíduo seja hispânico, o que é de se esperar.

---

b) *Mostre as 10 regras (juntamente com suporte, confiança e lift) com maior confiança entre regras com suporte de ao menos 0.001, confiança ao menos 0.8, e tamanho máximo 3.*

Tabela 10: Dez melhores regras de associação com relação a confiança, sujeitas a suporte > 0.001, confiança > 0.08 e com no máximo três itens

Regras	Suporte	Confiança	Lift
{marital status=widowed} => {dual incomes=not married}	0.03	1.00	1.67
{marital status=divorced} => {dual incomes=not married}	0.10	1.00	1.67
{marital status=single} => {dual incomes=not married}	0.41	1.00	1.67
{age=14-34,ethnic classification=east indian} => {income=$0-$40,000}	0.00	1.00	1.61
{income=$0-$40,000,ethnic classification=east indian} => {age=14-34}	0.00	1.00	1.71
{marital status=cohabitation,ethnic classification=pacific islander} => {age=14-34}	0.00	1.00	1.71
{marital status=cohabitation,ethnic classification=pacific islander} => {language in home=english}	0.00	1.00	1.10
{occupation=laborer,ethnic classification=pacific islander} => {education=no college graduate}	0.00	1.00	1.42
{occupation=clerical/service,ethnic classification=pacific islander} => {language in home=english}	0.00	1.00	1.10
{householder status=live with parents/family,ethnic classification=pacific islander} => {education=no college graduate}	0.00	1.00	1.42

Sujeito as condições impostas no enunciado, têm-se as 10 regras com maiores confianças, exibidas na Table 10, com valores iguais a 1. As primeiras regras afirmam que todos os indivíduos que eram solteiros, divorciados ou viúvos, então não possuíam dupla renda. Observe ainda que o suporte de que o indivíduo seja solteiro é de 40,91%. Outras regras informam que todos os indivíduos que tinham entre 14 e 34 anos e eram indianos, então possuíam uma renda entre $ 0 e $ 40,000. Ainda todos os indivíduos que eram de etnia pacífico islandês e com estado civil em coabitação ou cuja ocupação consistia em prestar serviços religiosos, tinham o idioma inglês falado nas respectivas residências. E aqueles cuja ocupação era operário assalariado ou se viviam com os pais/família, não possuíam nenhuma graduação no ensino superior.

---

c) *Plote as 10 regras com maior lift entre regras com suporte de ao menos 0.001, confiança ao menos 0.8, e tamanho máximo 3.*

---

d) *Mostre todas as regras (juntamente com suporte, confiança e lift) com maior confiança entre regras com suporte de ao menos 0.001, confiança ao menos 0.7, tamanho máximo 3, tamanho mínimo 2 e que tenha "ethnic classification=hispanic" do lado esquerdo da regra.*

Tabela 11: Regras de associação com relação a confiança, sujeitas a suporte > 0.001, confiança > 0.08 e com no máximo três itens, no mínimo 2 e que tenham ethnic hispanic como antecedente

Regras	Suporte	Confiança	Lift
{ethnic classification=hispanic} => {age=14-34}	0.09	0.71	1.22
{ethnic classification=hispanic} => {income=$0-$40,000}	0.10	0.78	1.26
{ethnic classification=hispanic} => {education=no college graduate}	0.11	0.86	1.22

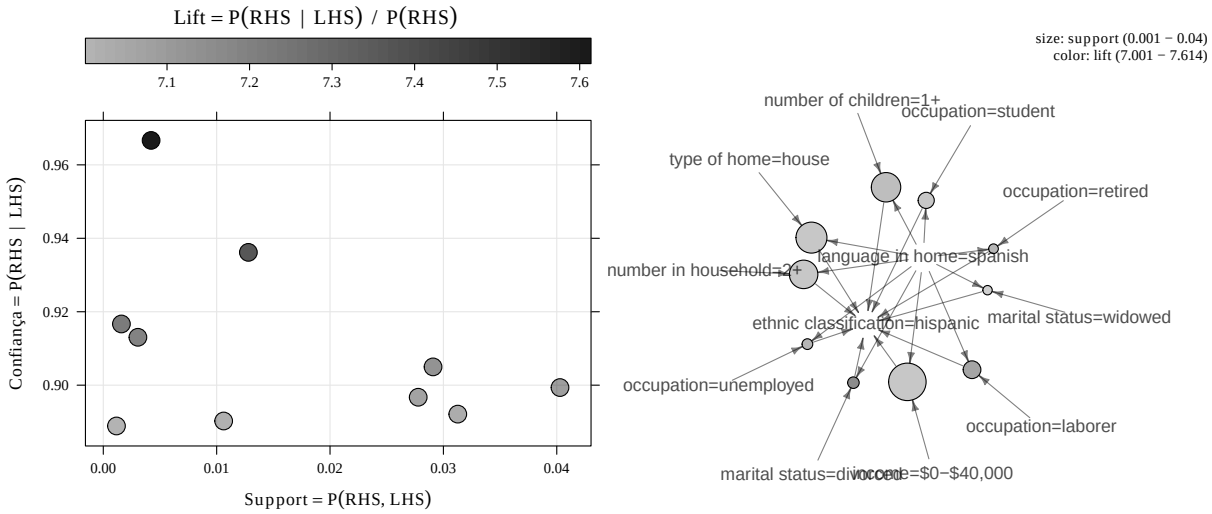

Figura 12: Visualização das dez regras de associação com maiores lifts construídas no conjunto de dados Income. Gráfico de dispersão com escala de cores para lift (esquerda), visualização baseada e grafos (direita).

Diante das restrições, as três regras que relacionam etnia hispânica como antecedente, são apresentadas na Table 11. Com a informação de que o indivíduo é hispânico, as probabilidades de que ele tenha entre 14 e 34 anos, com renda entre $0 e $40,000 e sem graduação em faculdade aumentam de maneira semelhante, 1.21, 1.25 e 1.22, respectivamente. Em geral, as confianças não foram tão significantes, todas abaixo de 90%.

---

e) *Explore os dados você mesmo. Mostre ao menos duas regras de associação que você achou interessante além das já apresentadas na lista. Justifique porque as achou interessantes.*

Exercício 3

Neste exercício você irá explorar alguns sistemas de recomendação para o MovieLens Dataset. Para tanto, instale a biblioteca `recommenderlab`, e carregue os dados usando data(MovieLense).

---

a) *Usando 75% dos dados para treinamento e assumindo que são dadas 12 avaliações por usuário, compare a performance dos seguintes métodos com relação a quão boas a predições das notas são:.*

- *Filtro colaborativo com base nos produtos com k = 2*
- *Filtro colaborativo com base nos produtos com k = 5*
- *Filtro colaborativo com base nos produtos com k = 8*
- *Filtro colaborativo com base nos usu´arios com k = 2*
- *Filtro colaborativo com base nos usu´arios com k = 5*
- *Filtro colaborativo com base nos usu´arios com k = 8*

Neste problema, o objetivo é explorar alguns sistemas de recomendação para um conjunto relacionado a avaliações de filmes. Os dados foram coletados através do site da MovieLens (https://movielens.org/), durante um período de sete meses entre 1997 e 1998. O conjunto contém 100,000 avaliações, de 1 a 5, de 943 usuários a respeito de 1664 filmes.

Para este problema, 75% dos dados foram utilizados para treinamento. Assumiu-se que 12 avaliações são fornecidas por cada usuário. Assim, a fim de obter boas predições das notas, foram empregados os filtros colaborativos com base no produto e com base no usuário e, para cada filtro, utilizando k igual a 2, 5 e 8. Por fim, as medidas MSE (EQM), RMSE (REQM) e MAE são calculadas, tais medidas avaliam o quão boas as predições das notas estão.

Na Table 12 são apresentados as estimativas para as medidas de risco MSE, RMSE e MAE. De modo geral, pode-se observar que as mediadas são relativamente baixas, o que fornece indícios de que as notas estão bem preditas. No mais, as medidas IB crescem de acordo com o aumento do k, enquanto as medidas do UB decrescem de acordo com o aumento do k, isso é mais visível na Figure 13, onde os mesmos valores são exibidos em forma de barras para comparação.

Tabela 12: Valores estimados das medidas de risco MSE, RMSE e MAE.

Configuração	RMSE	MSE	MAE
IB (k=2)	1.1493	1.3208	0.8013
UB (k=2)	1.1184	1.2509	0.8988
IB (k=5)	1.1948	1.4276	0.8482
UB (k=5)	1.1005	1.2111	0.8848
IB (k=8)	1.1992	1.4382	0.8585
UB (k=8)	1.0965	1.2024	0.8813

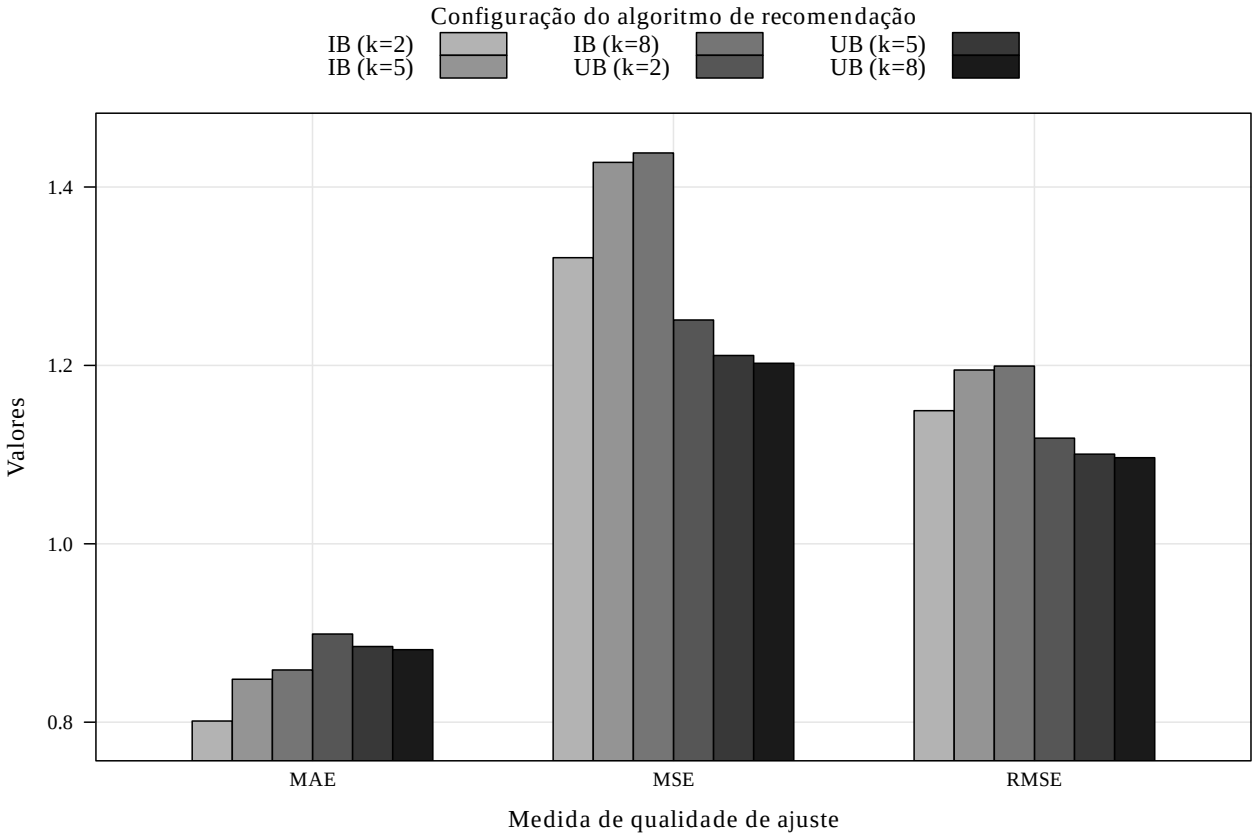

Figura 13: Valores estimados das medidas de risco MSE, RMSE e MAE.

---

b) *Compare os mesmo métodos que o descrito no item anterior, mas desta vez usando os métodos de avaliação com base nas N melhores recomendações. Você deve considerar uma avaliação como sendo boa quando sua nota é maior ou igual a 4. Você deve estimar a sensibilidade, 1-especificidade, precisão e lembrança (recall) para N = 1, 5, 10, 20, 50 e 100 recomendações.*

Com base nos filtros empregados anteriormente e assumindo que será recomendado N filmes para um determinado usuário, considerando uma avaliação quando sua nota excede ou é igual a nota 4, os filtros foram empregados para diferentes valores para N (1, 5, 10, 20, 50 e 100), cujo resultados são exibidos na Table 13 e Figure 14.

Tabela 13: Medidas de qualidade de predição

Precisão	Lembrança	Sensibilidade	Especificidade
`IB (k=2)`			
0.3263	0.0156	0.0156	0.9996
0.2737	0.0592	0.0592	0.9978
0.2318	0.0829	0.0829	0.9956
0.2067	0.1008	0.1008	0.9932
0.2016	0.1052	0.1052	0.9924
0.2016	0.1052	0.1052	0.9924
`UB (k=2)`			

0.3559	0.0145	0.0145	0.9996
0.2669	0.0464	0.0464	0.9977
0.2339	0.0823	0.0823	0.9952
0.2059	0.1348	0.1348	0.9901
0.1641	0.2156	0.2156	0.9739
0.1238	0.2762	0.2762	0.9453

IB (k=5)

0.2839	0.0086	0.0086	0.9996
0.2669	0.0505	0.0505	0.9977
0.2511	0.0873	0.0873	0.9953
0.2156	0.1372	0.1372	0.9905
0.1774	0.1874	0.1874	0.9822
0.1729	0.1920	0.1920	0.9810

UB (k=5)

0.4746	0.0190	0.0190	0.9997
0.3212	0.0645	0.0645	0.9979
0.2831	0.1026	0.1026	0.9955
0.2407	0.1495	0.1495	0.9905
0.1892	0.2599	0.2599	0.9747
0.1486	0.3488	0.3488	0.9469

IB (k=8)

0.2754	0.0089	0.0089	0.9995
0.2754	0.0429	0.0429	0.9977
0.2513	0.0850	0.0850	0.9953
0.2204	0.1347	0.1347	0.9903
0.1759	0.2266	0.2266	0.9766
0.1545	0.2555	0.2555	0.9687

UB (k=8)

0.4831	0.0210	0.0210	0.9997
0.3814	0.0760	0.0760	0.9981
0.3153	0.1093	0.1093	0.9957
0.2587	0.1568	0.1568	0.9908
0.1984	0.2617	0.2617	0.9750
0.1552	0.3698	0.3698	0.9473

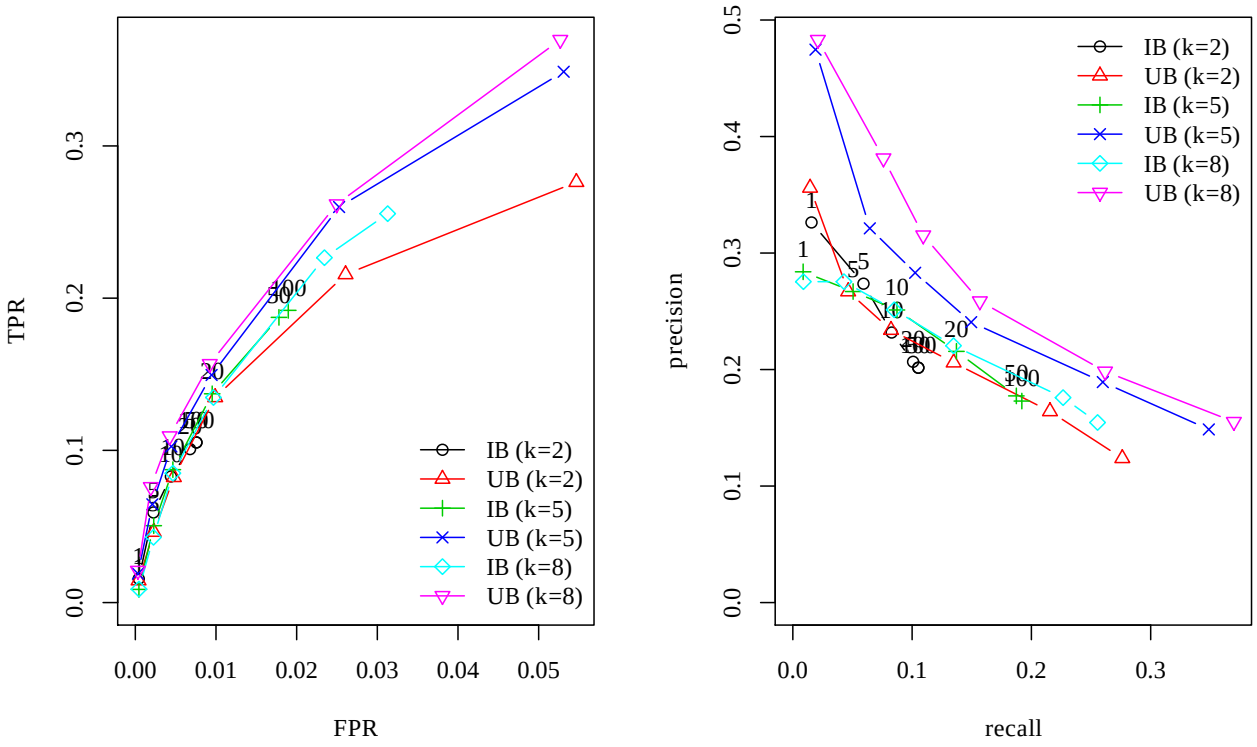

Figura 14: Medidas de qualidade de precisão.

---

c) *Algum dos métodos foi uniformemente melhor que os outros? Justifique.*

Com base nos resultados apresentados principalmente pela Figure 13 e **??** nota-se que nesse conjunto de dados os filtros colaborativos baseados em usuários tiveram um melhor desempenho com relação aos baseados em itens. Considerando o número de usuários utilizados para predição os filtro que utilizam um maior número de usuários para predição ($k = 8$ e $k = 5$) obtiveram resultados ligeiramente superiores aos demais.