INF0613 – Aprendizado de Máquina Não Supervisionado

Trabalho 2 - Redução de Dimensionalidade

Daniel Noriaki Kurosawa Eric Uyemura Suda Fernando Shigeru Wakabayashi

O objetivo deste trabalho é exercitar o conhecimento de técnicas de redução de dimensionalidade. Essas técnicas serão usadas tanto para obtenção de características quanto para visualização dos conjuntos de dados. Usaremos a base de dados `speech.csv`, que está disponível na página da disciplina no Moodle. A base contém amostras da pronúncia em inglês das letras do alfabeto.

Atividade 0 – Configurando o ambiente

Antes de começar a implementação do seu trabalho configure o *workspace* e importe todos os pacotes e execute o preprocessamento da base:

```
# Adicione os demais pacotes usados neste trabalho:

library(umap)
library(Rtsne)

# Configure ambiente de trabalho na mesma pasta
# onde colocou a base de dados:
setwd("/Users/nkuros/Documents/mineiracao_dados_complexos/Aprendizado de Maquina Nao Supervisionado/Tra

# Pré-processamento da base de dados
# Lendo a base de dados
speech <- read.csv("speech.csv", header = TRUE)

# Convertendo a coluna 618 em characteres
speech$LETRA <- as.factor(speech$LETRA)
```

Atividade 1 – Análise de Componentes Principais (*3,5 pts*)

Durante a redução de dimensionalidade, espera-se que o poder de representação do conjunto de dados seja mantido, para isso é preciso realizar uma análise da variância mantida em cada componente principal obtido. Use função `prcomp`, que foi vista em aula, para criar os autovetores e autovalores da base de dados. Não use a normalização dos atributos, isto é, defina `scale.=FALSE`. Em seguida, use o comando `summary`, analise o resultado e os itens a seguir:

```
# Executando a redução de dimensionalidade com o prcomp
speech_pca <- prcomp(speech[,-618], scale.=FALSE)
# Analisando as componentes com o comando summary
options(max.print=200)
summary(speech_pca)
```

```
## Importance of components:
##                             PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
##                              PC9   PC10   PC11   PC12   PC13   PC14   PC15   PC16
##                              PC17   PC18    PC19    PC20    PC21    PC22    PC23
##                              PC24   PC25    PC26    PC27    PC28    PC29    PC30
##                              PC31    PC32    PC33    PC34    PC35   PC36    PC37
##                              PC38    PC39    PC40    PC41    PC42    PC43    PC44
##                              PC45    PC46    PC47    PC48    PC49    PC50    PC51
##                              PC52    PC53    PC54    PC55    PC56    PC57    PC58
##                              PC59   PC60    PC61   PC62    PC63    PC64   PC65
##                              PC66    PC67    PC68    PC69    PC70    PC71    PC72
##                              PC73   PC74    PC75    PC76    PC77    PC78    PC79
##                              PC80    PC81    PC82    PC83    PC84    PC85    PC86
##                              PC87    PC88    PC89    PC90    PC91    PC92  PC93
##                              PC94    PC95    PC96    PC97    PC98    PC99   PC100
##                            PC101   PC102   PC103   PC104   PC105   PC106  PC107
##                             PC108   PC109   PC110   PC111   PC112   PC113   PC114
##                             PC115   PC116  PC117  PC118   PC119   PC120   PC121
##                             PC122   PC123   PC124   PC125   PC126   PC127   PC128
##                             PC129   PC130  PC131   PC132   PC133   PC134   PC135
##                             PC136   PC137   PC138   PC139   PC140   PC141   PC142
##                             PC143   PC144   PC145  PC146   PC147   PC148   PC149
##                             PC150   PC151   PC152   PC153   PC154   PC155   PC156
##                             PC157   PC158   PC159   PC160   PC161   PC162   PC163
##                             PC164   PC165   PC166   PC167   PC168  PC169   PC170
##                             PC171   PC172   PC173   PC174   PC175   PC176   PC177
##                             PC178   PC179   PC180   PC181   PC182   PC183   PC184
##                             PC185   PC186   PC187   PC188   PC189   PC190   PC191
##                             PC192   PC193   PC194   PC195   PC196  PC197  PC198
##                            PC199  PC200   PC201   PC202   PC203   PC204   PC205
##                             PC206   PC207   PC208   PC209   PC210   PC211   PC212
##                             PC213   PC214   PC215   PC216   PC217   PC218   PC219
##                             PC220   PC221   PC222   PC223   PC224   PC225   PC226
##                             PC227   PC228   PC229   PC230   PC231   PC232   PC233
##                             PC234   PC235   PC236   PC237   PC238   PC239   PC240
##                             PC241   PC242   PC243   PC244   PC245  PC246  PC247
##                            PC248  PC249  PC250  PC251  PC252   PC253   PC254
##                             PC255   PC256   PC257   PC258   PC259   PC260   PC261
##                             PC262   PC263   PC264   PC265   PC266   PC267   PC268
##                             PC269   PC270   PC271   PC272   PC273   PC274   PC275
##                             PC276   PC277   PC278   PC279   PC280   PC281   PC282
##                             PC283   PC284   PC285   PC286   PC287   PC288   PC289
##                             PC290   PC291   PC292   PC293   PC294   PC295   PC296
##                             PC297   PC298   PC299   PC300   PC301   PC302   PC303
##                             PC304   PC305   PC306   PC307   PC308   PC309   PC310
##                             PC311   PC312   PC313   PC314   PC315   PC316   PC317
##                             PC318   PC319   PC320   PC321   PC322   PC323   PC324
##                             PC325   PC326   PC327   PC328   PC329   PC330   PC331
##                             PC332   PC333   PC334   PC335   PC336   PC337   PC338
##                             PC339   PC340   PC341   PC342   PC343   PC344   PC345
##                             PC346   PC347   PC348   PC349   PC350   PC351   PC352
##                             PC353  PC354  PC355  PC356  PC357  PC358  PC359  PC360
##                            PC361  PC362  PC363  PC364  PC365  PC366  PC367  PC368
##                            PC369  PC370  PC371   PC372   PC373   PC374   PC375
```

```
##                          PC376   PC377   PC378   PC379   PC380   PC381   PC382
##                          PC383   PC384   PC385   PC386   PC387   PC388   PC389
##                          PC390   PC391   PC392   PC393   PC394   PC395   PC396
##                          PC397   PC398   PC399   PC400   PC401   PC402   PC403
##                          PC404   PC405   PC406   PC407   PC408   PC409   PC410
##                          PC411   PC412   PC413   PC414   PC415   PC416   PC417
##                          PC418   PC419   PC420   PC421   PC422   PC423   PC424
##                          PC425   PC426   PC427   PC428   PC429   PC430   PC431
##                          PC432   PC433   PC434   PC435   PC436   PC437   PC438
##                          PC439   PC440   PC441   PC442   PC443   PC444   PC445
##                          PC446   PC447   PC448   PC449   PC450   PC451   PC452
##                          PC453   PC454   PC455   PC456   PC457   PC458   PC459
##                          PC460   PC461   PC462   PC463   PC464   PC465   PC466
##                          PC467   PC468   PC469   PC470   PC471   PC472   PC473
##                          PC474   PC475   PC476   PC477   PC478   PC479   PC480
##                          PC481   PC482   PC483   PC484   PC485   PC486   PC487
##                          PC488   PC489   PC490   PC491   PC492   PC493   PC494
##                          PC495   PC496   PC497   PC498   PC499   PC500   PC501
##                          PC502   PC503   PC504   PC505   PC506   PC507   PC508
##                          PC509   PC510   PC511   PC512   PC513   PC514   PC515
##                          PC516   PC517   PC518   PC519   PC520   PC521   PC522
##                          PC523   PC524   PC525   PC526   PC527   PC528   PC529
##                          PC530   PC531   PC532   PC533   PC534   PC535   PC536
##                          PC537   PC538   PC539   PC540   PC541   PC542   PC543
##                          PC544   PC545   PC546   PC547   PC548   PC549   PC550
##                          PC551   PC552   PC553   PC554   PC555   PC556   PC557
##                          PC558   PC559   PC560   PC561   PC562   PC563   PC564
##                          PC565   PC566   PC567   PC568   PC569   PC570   PC571
##                          PC572   PC573   PC574   PC575   PC576   PC577   PC578
##                          PC579   PC580   PC581   PC582   PC583   PC584   PC585
##                          PC586   PC587   PC588   PC589   PC590   PC591   PC592
##                          PC593 PC594   PC595   PC596   PC597   PC598   PC599
##                         PC600  PC601  PC602 PC603  PC604  PC605  PC606  PC607
##                         PC608  PC609 PC610  PC611  PC612  PC613  PC614   PC615
##                          PC616  PC617
##  [ reached getOption("max.print") -- omitted 3 rows ]
```

Análise

a) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 80% do total?

Resposta: 38 componentes

b) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 90% do total?

Resposta: 91 componentes

c) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 95% do total?

Resposta: 170 componentes

d) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos 99% do total?

Resposta: 382 componentes

e) Faça um breve resumo dos resultados dos itens *a)-d)* destacando o impacto da redução de dimensionalidade.

Resposta: Ao avaliar a variância acumulada, percebemos que usando um número tão baixo quanto 38 variáveis, teríamos uma variância acumulada de 80% do total, o que pode ser considerado suficiente para alguns domínios. Mesmo caso necessitemos de uma análise mais criteriosa, com 382 componentes teríamos 99% do total, contra as 617 componentes originais, uma redução significativa quando pensamos no poder computacional necessário para processar o dataset.

Atividade 2 – Análise de Componentes Principais e Normalização (*3,5 pts*)

A normalização de dados em alguns casos, pode trazer benefícios. Nesta questão, iremos analisar o impacto dessa prática na redução da dimensionalidade da base de dados `speech.csv`. Use função `prcomp` para criar os autovetores e autovalores da base de dados usando a normalização dos atributos, isto é, defina `scale.=TRUE`. Em seguida, use o comando `summary`, analise o resultado e os itens a seguir:

```
# Executando a redução de dimensionalidade com o prcomp
 # com normalização dos dados
speech_pca <- prcomp(speech[,-618], scale.=TRUE)
# Analisando as componentes com o comando summary
options(max.print=20)
summary(speech_pca)
```

```
## Importance of components:
##                            PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
##                            PC9   PC10   PC11   PC12   PC13   PC14   PC15   PC16
##                           PC17   PC18    PC19    PC20    PC21   PC22    PC23
##                            PC24    PC25   PC26    PC27    PC28    PC29    PC30
##                            PC31    PC32    PC33    PC34    PC35    PC36    PC37
##                            PC38    PC39    PC40    PC41    PC42    PC43    PC44
##                            PC45    PC46   PC47    PC48   PC49    PC50    PC51
##                           PC52   PC53    PC54    PC55    PC56    PC57    PC58
##                            PC59    PC60    PC61    PC62    PC63    PC64    PC65
##                            PC66    PC67    PC68    PC69    PC70    PC71   PC72
##                           PC73    PC74    PC75    PC76    PC77    PC78    PC79
##                           PC80    PC81    PC82    PC83    PC84    PC85    PC86
##                            PC87    PC88    PC89   PC90    PC91    PC92    PC93
##                            PC94    PC95    PC96    PC97    PC98    PC99   PC100
##                           PC101   PC102   PC103   PC104   PC105   PC106   PC107
##                           PC108   PC109   PC110  PC111   PC112   PC113   PC114
##                           PC115   PC116   PC117   PC118   PC119  PC120   PC121
##                           PC122   PC123   PC124   PC125   PC126   PC127   PC128
##                           PC129   PC130  PC131   PC132   PC133   PC134   PC135
##                           PC136   PC137   PC138   PC139   PC140   PC141   PC142
##                           PC143   PC144  PC145  PC146   PC147   PC148   PC149
##                           PC150   PC151   PC152   PC153   PC154   PC155   PC156
##                           PC157   PC158   PC159   PC160   PC161   PC162   PC163
##                           PC164  PC165   PC166   PC167   PC168   PC169   PC170
```

```
##                          PC171   PC172   PC173   PC174   PC175   PC176   PC177
##                          PC178   PC179   PC180   PC181   PC182   PC183   PC184
##                          PC185   PC186   PC187  PC188  PC189  PC190   PC191
##                          PC192   PC193   PC194   PC195   PC196   PC197   PC198
##                          PC199   PC200   PC201   PC202   PC203   PC204   PC205
##                          PC206   PC207   PC208   PC209   PC210   PC211   PC212
##                          PC213   PC214   PC215   PC216   PC217   PC218   PC219
##                          PC220   PC221   PC222   PC223  PC224  PC225  PC226
##                         PC227  PC228  PC229   PC230   PC231   PC232   PC233
##                          PC234   PC235   PC236   PC237   PC238   PC239   PC240
##                          PC241   PC242   PC243   PC244   PC245   PC246   PC247
##                          PC248   PC249   PC250   PC251   PC252   PC253   PC254
##                          PC255   PC256   PC257   PC258   PC259   PC260   PC261
##                          PC262   PC263   PC264   PC265   PC266   PC267   PC268
##                          PC269   PC270   PC271   PC272   PC273   PC274   PC275
##                          PC276   PC277   PC278   PC279   PC280   PC281   PC282
##                          PC283  PC284  PC285  PC286  PC287  PC288  PC289  PC290
##                         PC291  PC292   PC293   PC294   PC295   PC296   PC297
##                          PC298   PC299   PC300   PC301   PC302   PC303   PC304
##                          PC305   PC306   PC307   PC308   PC309   PC310   PC311
##                          PC312   PC313   PC314   PC315   PC316   PC317   PC318
##                          PC319   PC320   PC321   PC322   PC323   PC324   PC325
##                          PC326   PC327   PC328   PC329   PC330   PC331   PC332
##                          PC333   PC334   PC335   PC336   PC337   PC338   PC339
##                          PC340   PC341   PC342   PC343   PC344   PC345   PC346
##                          PC347   PC348   PC349   PC350   PC351   PC352   PC353
##                          PC354   PC355   PC356   PC357   PC358   PC359   PC360
##                          PC361   PC362   PC363   PC364   PC365   PC366   PC367
##                          PC368   PC369   PC370   PC371   PC372   PC373   PC374
##                          PC375   PC376   PC377   PC378   PC379   PC380   PC381
##                          PC382   PC383   PC384   PC385   PC386   PC387   PC388
##                          PC389   PC390   PC391   PC392  PC393  PC394  PC395
##                         PC396  PC397  PC398  PC399  PC400  PC401  PC402  PC403
##                         PC404  PC405  PC406   PC407   PC408   PC409   PC410
##                          PC411   PC412   PC413   PC414   PC415   PC416   PC417
##                          PC418   PC419   PC420   PC421   PC422   PC423   PC424
##                          PC425   PC426   PC427   PC428   PC429   PC430   PC431
##                          PC432   PC433   PC434   PC435   PC436   PC437   PC438
##                          PC439   PC440   PC441   PC442   PC443   PC444   PC445
##                          PC446   PC447   PC448   PC449   PC450   PC451   PC452
##                          PC453   PC454   PC455   PC456   PC457   PC458   PC459
##                          PC460   PC461   PC462   PC463   PC464   PC465   PC466
##                          PC467   PC468   PC469   PC470   PC471   PC472   PC473
##                          PC474   PC475   PC476   PC477   PC478   PC479   PC480
##                          PC481   PC482   PC483   PC484   PC485   PC486   PC487
##                          PC488   PC489   PC490   PC491   PC492   PC493   PC494
##                          PC495   PC496   PC497   PC498   PC499   PC500   PC501
##                          PC502   PC503   PC504   PC505   PC506   PC507   PC508
##                          PC509   PC510   PC511   PC512   PC513   PC514   PC515
##                          PC516   PC517   PC518   PC519   PC520   PC521   PC522
##                          PC523   PC524   PC525   PC526   PC527   PC528   PC529
##                          PC530   PC531   PC532   PC533   PC534   PC535   PC536
##                          PC537   PC538   PC539   PC540   PC541   PC542   PC543
##                          PC544   PC545   PC546   PC547   PC548   PC549   PC550
```

```
##                            PC551   PC552   PC553   PC554   PC555   PC556   PC557
##                            PC558   PC559   PC560   PC561   PC562   PC563   PC564
##                            PC565   PC566   PC567   PC568   PC569   PC570   PC571
##                            PC572   PC573   PC574   PC575   PC576   PC577   PC578
##                            PC579   PC580   PC581   PC582   PC583   PC584   PC585
##                            PC586   PC587   PC588   PC589   PC590   PC591   PC592
##                            PC593   PC594   PC595   PC596   PC597   PC598   PC599
##                          PC600   PC601   PC602   PC603   PC604  PC605  PC606
##                           PC607  PC608  PC609  PC610  PC611  PC612 PC613  PC614
##                           PC615  PC616 PC617
##  [ reached getOption("max.print") -- omitted 3 rows ]
```

Análise

a) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `80%` do total?

Resposta: 48 componentes

b) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `90%` do total?

Resposta: 112 componentes

c) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `95%` do total?

Resposta: 200 componentes

d) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `99%` do total?

Resposta: 400 componentes

e) Quais as principais diferenças entre a aplicação do PCA nesse conjunto dados com e sem normalização?

Resposta: Aplicando a técnica de PCA com a nromalização dos dados garante que todas as varíaveis tenham igual importância ao normalizar as suas escalas, evitando desvios padrões muito distoantes. Ao fazer isso porém, percebemos que a técnica necessitou um maior número de componentes para uma dada variância acumulada. f) Qual opção parece ser mais adequada para esse conjunto de dados? Justifique sua resposta.

Resposta: Sem normalização, pois dada uma mesma variância acumulada, necessita menos componentes

Atividade 3 – Visualização a partir da Redução (*3,0 pts*)

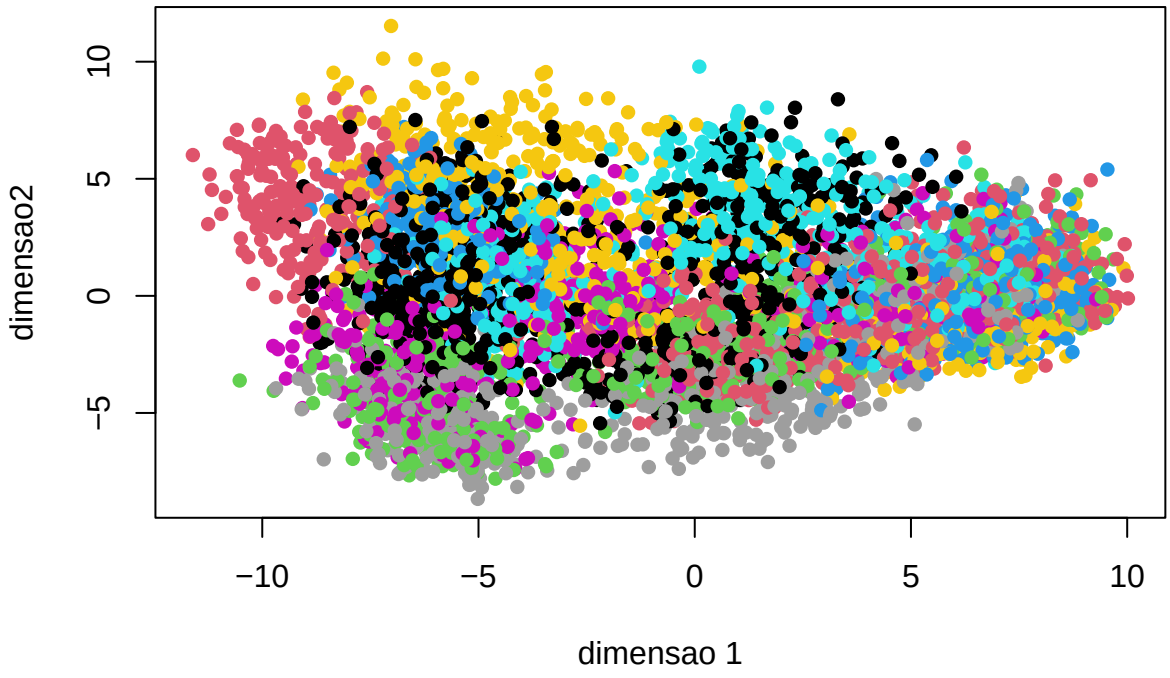
Nesta atividade, vamos aplicar diferentes métodos de redução de dimensionalidade e comparar as visualizações dos dados obtidos considerando apenas duas dimensões. Lembre de fixar uma semente antes de executar o T-SNE.

a) Aplique a redução de dimensionalidade com a técnica PCA e gere um gráfico de dispersão dos dados. Use a coluna `618` para classificar as amostras e definir uma coloração.

```r
# Aplicando redução de dimensionalidade com a técnica PCA
# Compara com a Analise de Componentes Principais:
pca <- princomp(speech[,-618])$scores[,1:2]

pca
```

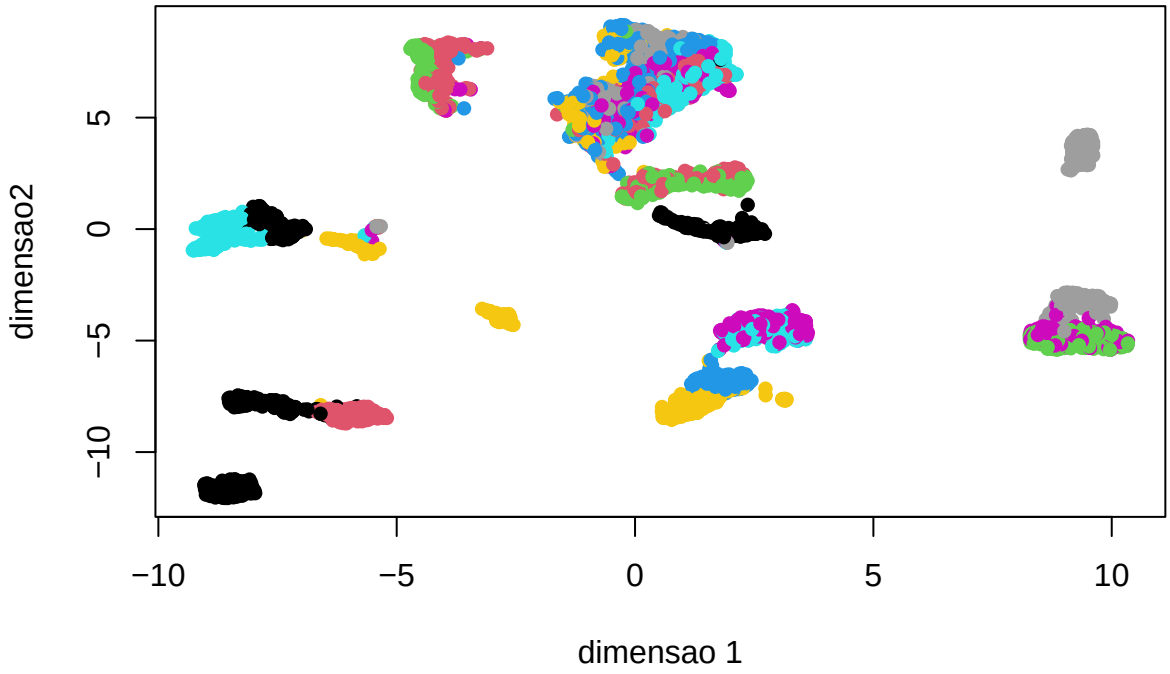
```
##            Comp.1    Comp.2
##     [1,]  3.90e+00  1.224202
##     [2,]  1.64e+00 -0.048759
##     [3,]  7.17e+00  2.241636
##     [4,]  6.53e+00  2.513841
##     [5,]  8.16e+00  2.587754
##     [6,]  8.40e+00  3.264489
##     [7,]  8.72e+00  3.350916
##     [8,]  8.33e+00  3.420890
##     [9,]  6.79e+00  4.030121
##    [10,]  6.05e+00  3.350991
##   [ reached getOption("max.print") -- omitted 6228 rows ]
```

```r
# Gerando o gráfico de dispersão

# Visualizacao de componentes principais:
plot(pca, col=speech[,618] , xlab="dimensao 1",
     ylab="dimensao2", pch =16)
```



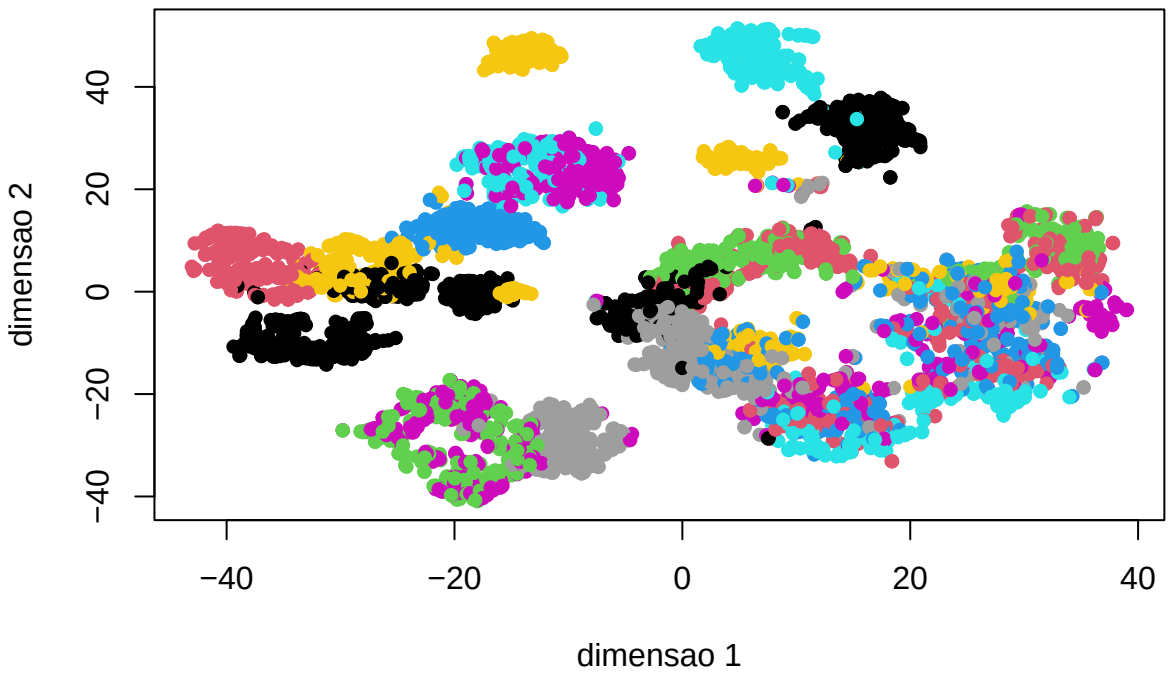
b) Aplique a redução de dimensionalidade com a técnica UMAP e gere um gráfico de dispersão dos dados. Use a coluna `618` para classificar as amostras e definir uma coloração.

```
# Aplicando redução de dimensionalidade com a técnica UMAP
# Executa UMAP:
set.seed(42) # semente fixa para reprodutibilidade
speech.umap <- umap(as.matrix(speech[,-618]))

# Gerando o gráfico de dispersão
plot(speech.umap$layout , col=speech[,618] , xlab="dimensao 1",
     ylab="dimensao2", pch =16)
```



c) Aplique a redução de dimensionalidade com a técnica T-SNE e gere um gráfico de dispersão dos dados. Use a coluna `618` para classificar as amostras e definir uma coloração.

```
# Aplicando redução de dimensionalidade com a técnica T-SNE
speech_unique_values <- unique(speech)
set.seed(42) # semente fixa para reprodutibilidade
tsne <- Rtsne(as.matrix(speech_unique_values[,-618]), perplexity = 30, dims=3)

# Gerando o gráfico de dispersão
plot(tsne$Y, col=speech_unique_values[,618], xlab="dimensao 1", ylab="dimensao 2", pch=16)
```



Análise

d) Qual técnica você acredita que apresentou a melhor projeção? Justifique.

Resposta: Analisando os plots de distribuição podemos eliminar imediatamente o método de PCA pois apresenta uma nuvem de pontos com as cores todas misturadas e difusas. Comparando o UMAP com T-SNE acreditamos que o método de UMAP gera a melhor projeção pois segmenta bem algumas cores (ex. azul claro, amarelo e preto) e as separa bem espacialmente, já o T-SNE apesar de segmentar as cores em sí, os pontos estão um pouco difusos e muitas núvens de pontos aglutinam-se (ex. a circumferência na região dim1 > 0 e dim2 < 0 com as cores cinza, rosa, azul claro, preto e verde)