TRABALHO LABORATORIAL

RATE BEER DATASET

Ana Duarte – Nº48963
Gonçalo Lamelas – Nº48971
Pedro Silva – Nº48965

ISEL
INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Objetivo:

- Determinar a qualidade de uma cerveja baseado no que foi escrito sobre a mesma
- Classificação Binária: pretende-se saber se o crítico considera a cerveja muito boa ou muito má
- Classificação Multi-Classe: Prever a pontuação de três aspetos das críticas (smell, taste e overall)

Construção do vocabulário:

- **Modelo Bag of Words (BoW):**

1. **Tokenization**: Este processo consiste em dividir cada documento em palavras
2. **Construção** do Vocabulário: Construir um vocabulário constituído por todas ou por um sub-conjunto das palavras presentes no corpus.
3. **Codificação**: Contar o número de vezes que cada palavra do vocabulário aparece em cada documento. Representar cada documento por um vetor de d dimensões, uma por cada palavra no vocabulário, com valores proporcionais ao número de ocorrências dessa palavra no documento.

Construção do vocabulário:

- **Limpeza do texto:** expressão regular r"\b\w\e\w+\b". Significa que serão extraídas sequências de caracteres compostas por 2 ou mais letras ou números (\w) e que estão separadas por caracteres de pontuação ou espaços (\b).
- **Stop Words:** palavras que ocorrem frequentemente em uma dada língua

Construção do vocabulário:

- **Stemming:** técnica que consiste no processo de transformar uma palavra na sua raiz, o que permite mapear palavras semelhantes numa única palavra. Por exemplo, palavras como "studies", "studying", "studied" seriam mapeadas para "studi".
- **Repesentação tf-idf:** atribui importância às palavras com base em quão frequentemente elas aparecem em poucos documentos, associando a essas palavras um valor mais elevado.

Construção do vocabulário:

- **N-gramas:** Uma das limitações desta representação é que ela descarta informações sobre a ordem das palavras. Frases como "não é bom, é mau" e "não é mau, é bom" têm a mesma representação, apesar de terem significados opostos. Incluimos sequências de duas ou mais palavras que aparecem frequentemente nos documentos na representação BoW. Conjuntos de duas palavras são chamados de bi-gramas, de três palavras são tri-gramas e, em geral, sequências de várias palavras são denominadas n-gramas. Nós vamos utilizar ngram_range=(1, 4).

Construção do vocabulário:

- Utilizando então todas as funcionalidades descritas anteriormente e um classificador Regressão Logística L2 temos as seguintes 10 palavras mais positivas e as 10 mais negativas.

- As palavras mais positivas são como é obvio adjetivos e sabores que costumamos considerar agradáveis (como chocolate, uva ou caramelo). O contrário dá-se nas palavras negativas como a cerveja ser considerada aguada, metálica ou de cartão.

```
Palavras Mais Positivas:
nice: -3.8075345315530202
citrus: -2.8310733110964916
chocolate: -2.6843294397928927
great: -2.4101678525353503
smooth: -2.387737983932004
hops: -2.2429446481906083
dark: -2.160977048729679
grapefruit: -2.0624411181567592
rich: -1.9411634323698237
caramel: -1.8934190394320012

Palavras Mais Negativas:
watery: 4.512610984791502
bad: 3.8108595101992724
water: 3.764554697498566
drain: 3.6771487845878803
corn: 3.3463819354519333
alky: 3.3252551711496596
nasty: 3.3204947861430103
metallic: 3.318816692149275
skunky: 2.8424869267035353
cardboard: 2.8400720728408224

Tempo decorrido: 32.01 segundos
```

```
Precisão: 0.29

Matriz de Confusão:
[[  96   28  142  225   22   22   13    8    1    0]
 [  31   25  146  421   58   46   21    6    2    1]
 [  21   14  250 1228  216  202  100   30    3    0]
 [  17    4  232 2782  937 1009  419   94   14    2]
 [   8    3   58 1512  967 1441  684  164   24    1]
 [   8    3   38 1096  953 2045 1315  395   32    0]
 [   6    0   18  614  595 1550 1540  663   67    1]
 [   2    1    9  234  221  845 1083  749   96    5]
 [   1    1    5   75   64  244  480  520  135   11]
 [   3    1    2   37   16   63  102  189   98   19]]

Relatório de Classificação:
              precision    recall  f1-score   support

           1       0.50      0.17      0.26       557
           2       0.31      0.03      0.06       757
           3       0.28      0.12      0.17      2064
           4       0.34      0.50      0.41      5510
           5       0.24      0.20      0.22      4862
           6       0.27      0.35      0.31      5885
           7       0.27      0.30      0.28      5054
           8       0.27      0.23      0.25      3245
           9       0.29      0.09      0.13      1536
          10       0.47      0.04      0.07       530

    accuracy                           0.29     30000
   macro avg       0.32      0.20      0.21     30000
weighted avg       0.29      0.29      0.27     30000
```

Construção do vocabulário:

O máximo de precisão que conseguimos foi 29%, quer nós aumentássemos o número de amostras, o número do n-grams ou mudássemos o discriminante logístico por isso conseguimos otimizar a duração do processo a menos de 1 minuto.

Acabámos por escolher o Ridge pois este era o que nos dava melhores valores de precisão nas reviews muito boas (score>=9) e muito más (score<=2), ambos os overalls 1 e 10, o min e o max, com 50% o que é o nosso melhor valor.

```
*** Regressão Logística ***
Precisão: 0.92

Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.93      0.86      0.89      1996
           1       0.91      0.96      0.93      3043

    accuracy                           0.92      5039
   macro avg       0.92      0.91      0.91      5039
weighted avg       0.92      0.92      0.92      5039


Matriz de Confusão:
[[1712  284]
 [ 132 2911]]

*** Naive Bayes ***
Precisão: 0.90

Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.95      0.78      0.86      1996
           1       0.87      0.97      0.92      3043

    accuracy                           0.90      5039
   macro avg       0.91      0.88      0.89      5039
weighted avg       0.90      0.90      0.89      5039


Matriz de Confusão:
[[1557  439]
 [  82 2961]]
```

Classificação Binária - Treino:

Vamos começar por dividir o conjunto de treino nas reviews muito boas e muito más e descartar as restantes. Vamos usar os mesmos parâmetros que utilizámos na fase anterior mas desta vez utilizamos também o classificador Naive Bayes.

O seu princípio básico envolve a aplicação do Teorema de Bayes, assumindo independência condicional entre as características, daí o termo "Naive".

Como podemos ver pelos resultados em cima ambos os modelos dão nos bastante confiança tendo resultados de 90%. Não foi preciso modicar nada na pipeline apenas mudou o conjunto de teste. Com estas observações podemos usar ambos estes classificadores com todas as certezas.

```
*** Regressão Logística ***
Precisão: 0.91

Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.95      0.91      0.93      2429
           1       0.82      0.90      0.86      1111

    accuracy                           0.91      3540
   macro avg       0.89      0.91      0.90      3540
weighted avg       0.91      0.91      0.91      3540


Matriz de Confusão:
[[2209  220]
 [ 106 1005]]

*** Naive Bayes ***
Precisão: 0.87

Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.97      0.84      0.90      2429
           1       0.73      0.94      0.82      1111

    accuracy                           0.87      3540
   macro avg       0.85      0.89      0.86      3540
weighted avg       0.89      0.87      0.88      3540


Matriz de Confusão:
[[2044  385]
 [  68 1043]]
```

Classificação Binária - Teste:

Como é lógico os resultados são inferiores aos anteriores mas isto não é por muito sendo a maior queda de 3% no classificador Naive Bayes.

Com estes testes gerais conseguimos chegar à conclusão que ambos os classificadores estão a funcionar bastante bem mas vamos aprofundar um pouco a questão.

Classificação Binária - Teste:

- Vamos fazer um pequeno teste com reviews fabricadas e testamos uma má uma intermédia e uma boa:

1. Má : "This beer is bad! The flavor sucks but the aroma is nice."
2. Média: "This beer is alright. The flavor is not good but the aroma smells like caramel."
3. Boa: "This beer is great! The flavor sucks but the aroma is nice."

```
*** Review - Mau ***

** Regressão Logística **
The predicted sentiment for the review is: Muito Mau

** Naive Bayes **
The predicted sentiment for the review is: Muito Mau

*** Review - Médio ***

** Regressão Logística **
The predicted sentiment for the review is: Muito Mau

** Naive Bayes **
The predicted sentiment for the review is: Muito Bom

*** Review - Bom ***

** Regressão Logística **
The predicted sentiment for the review is: Muito Bom

** Naive Bayes **
The predicted sentiment for the review is: Muito Bom
```

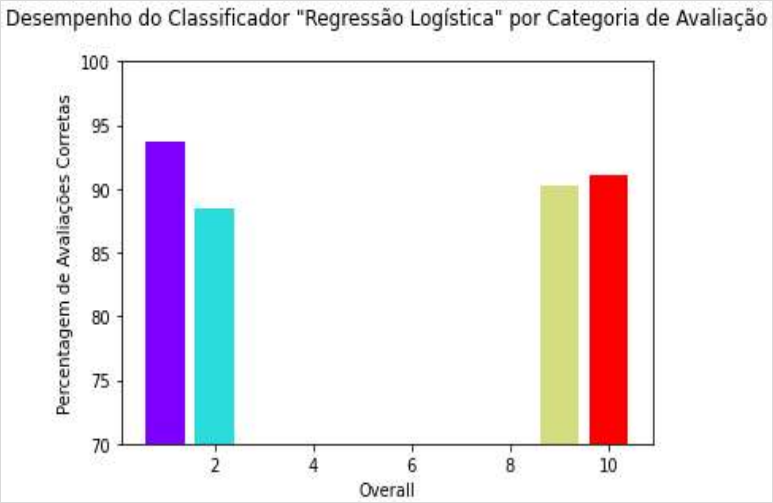
Classificação Binária - Teste:

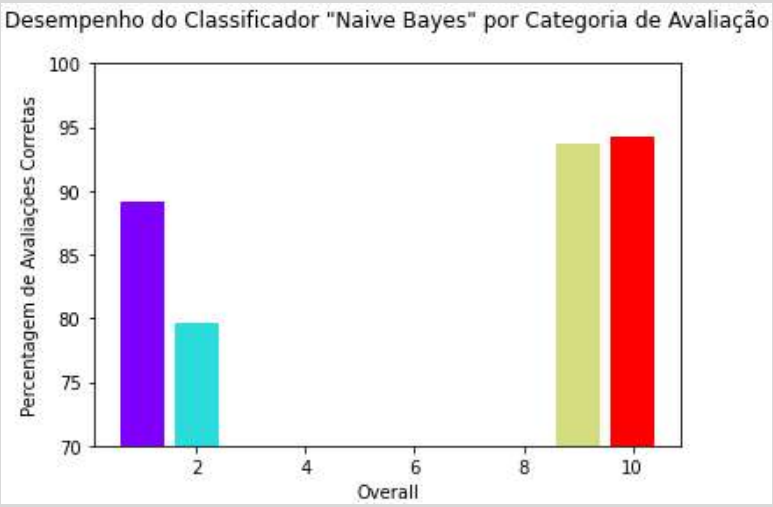
- Os resultados dos 2 extremos são os esperados mas existe alguma indecisão no que toca a reviews que não declaram com confiança o que sentem. A diferença nos resultados para a análise do meio entre os classificadores de Regressão Logística e Naive Bayes pode ser atribuída à forma como cada classificador faz previsões com base nas características fornecidas.
- No caso do classificador Naive Bayes, ele faz previsões usando o teorema de Bayes, assumindo que as características (palavras, neste caso) são condicionalmente independentes dado o rótulo da classe. O Naive Bayes pode ser melhor em capturar a probabilidade de certas combinações de palavras numa análise e não leva em consideração as interações entre as características.
- Por outro lado, a Regressão Logística considera a combinação linear das características e aplica uma função logística para fazer previsões. Ela pode capturar relações mais complexas entre características e pode ter um desempenho diferente com base na natureza dos dados.
- Se formos analisar os valores dos coeficientes entre os 2 classificadores conseguimos encontrar uma maior diferença entre os mesmos no classificador de Regressão Logística comparando com o Naive Bayes sendo esta a razão provável de neste ser considerado muito mau e no outro o contrário pois naturalmente "not good" vai ter um maior valor comparanando com um cheiro a caramelo.
- Concluímos que numa review que explique a sua opinião com adjetivos mais fortes conseguimos decifrar com toda a certeza a sua opinião enquanto que com uma mais no meio seja mais duvidoso.

```
*** Review - Mau ***

** Regressão Logística **
The predicted sentiment for the review is: Muito Mau

** Naive Bayes **
The predicted sentiment for the review is: Muito Mau

*** Review - Médio ***

** Regressão Logística **
The predicted sentiment for the review is: Muito Mau

** Naive Bayes **
The predicted sentiment for the review is: Muito Bom

*** Review - Bom ***

** Regressão Logística **
The predicted sentiment for the review is: Muito Bom

** Naive Bayes **
The predicted sentiment for the review is: Muito Bom
```

Classificação Binária - Teste:

- Podemos observar pelos 2 gráficos acima que um dos valores é claramente mais dificil de fazer a previsão e este é o 2. Isto deve-se ao facto de os reviewers usarem um palavreado mais rico e neutro quando comparando com os outros valores. Vamos ter como exemplo esta review:
- "Light cocoa powder, grassy faint citrus hops and perhaps a light sweet spiciness in the nose. Clear amber brown coloured body sports a quickly receding off white head that leaves a ring around the edge of the glass. Nutty and toasty malts with herbal hops, light spices and perhaps a touch of fruit character. Dry body with somewhat low carbonation. Somehow this isnt the celebration I was expecting from Christmas in the Big Easy. Bottle sampled with Beershine and Oakes."
- Como podemos ver este utilizador não utilizou nenhuma das palavras características de uma má review logo fica mais dificil para os classificadores fazerem essa distinção.
- Os restantes resultados não oferecem muitas introspeções apenas o facto de o classificador Naive Bayes preferir reviews positivas enquanto que a Regressão Logística prefere negativas.





Classificação Multi-Classe - Treino:

- Vai ter um procedimento exatamente igual à fase anterior só que não vamos ter de filtrar as reviews e vamos usar também o smell e taste. Como não temos muito tempo para realizar este trabalho vamos utilizar o mesmo classificador que viemos a utilizar até agora: a Regressão Logística.

Cheiro

Precisao: 0.46

Relatório de Classificação:

	precision	recall	f1-score	support
1	0.58	0.44	0.50	804
2	0.38	0.27	0.32	1398
3	0.45	0.62	0.52	2651
4	0.47	0.47	0.47	2041
5	0.50	0.21	0.29	606
accuracy			0.46	7500
macro avg	0.48	0.40	0.42	7500
weighted avg	0.46	0.46	0.45	7500

\Matriz de Confusão:

```
[[ 355  201  204   39    5]
 [ 137  384  743  128    6]
 [  81  311 1631  601   27]
 [  24   87  890  952   88]
 [  11   16  142  311  126]]
```

Sabor

Precisao: 0.46

Relatório de Classificação:

	precision	recall	f1-score	support
1	0.58	0.42	0.49	1678
2	0.38	0.27	0.32	2710
3	0.46	0.60	0.52	5276
4	0.46	0.49	0.48	4126
5	0.49	0.21	0.29	1210
accuracy			0.46	15000
macro avg	0.47	0.40	0.42	15000
weighted avg	0.46	0.46	0.45	15000

Matriz de Confusão:

```
[[ 706  426  461   78    7]
 [ 304  734 1390  272   10]
 [ 148  600 3191 1286   51]
 [  44  141 1731 2021  189]
 [  17   23  220  700  250]]
```

Overall

Precisao: 0.29

Relatório de Classificação:

	precision	recall	f1-score	support
1	0.47	0.19	0.27	286
2	0.24	0.02	0.04	384
3	0.29	0.14	0.19	1021
4	0.35	0.51	0.41	2794
5	0.23	0.20	0.21	2402
6	0.28	0.34	0.31	2967
7	0.27	0.31	0.29	2482
8	0.28	0.24	0.26	1633
9	0.25	0.10	0.14	759
10	0.48	0.05	0.09	272
accuracy			0.29	15000
macro avg	0.31	0.21	0.22	15000
weighted avg	0.29	0.29	0.28	15000

Matriz de Confusão:

```
[[  54   15   74  111   12   11    6    3    0    0]
 [  19    8   84  209   31   19   14    0    0    0]
 [  16    7  145  598  106   86   52   11    0    0]
 [  11    1  117 1416  486  519  181   56    5    2]
 [   7    1   38  769  471  701  313   85   16    1]
 [   3    2   12  555  492 1017  658  204   23    1]
 [   1    0   10  291  265  755  777  331   50    2]
 [   2    0    7  104  116  372  564  395   69    4]
 [   0    0    6   40   32  112  238  250   76    5]
 [   2    0    3   11    5   32   57   81   67   14]]
```

Classificação Multi-Classe - Teste:

- Os resultados não são bons mas, mais uma vez, era esperado já que nem os docentes conseguiram bons modelos.
- No entanto, isto não nos impede de tirar conclusões sobre os mesmos como iremos ver a seguir

```
Precisao (Smell): 0.45

Matriz de Confusão (Smell):

[[2309 1336 1098  144   11]
 [ 728 1783 2937  463   27]
 [ 342 1139 4983 1592   67]
 [ 111  244 2211 2086  177]
 [  43   32  318  653  166]]

Relatório de Classificação (Smell):

              precision    recall  f1-score   support

           1       0.65      0.47      0.55      4898
           2       0.39      0.30      0.34      5938
           3       0.43      0.61      0.51      8123
           4       0.42      0.43      0.43      4829
           5       0.37      0.14      0.20      1212

    accuracy                           0.45     25000
   macro avg       0.45      0.39      0.40     25000
weighted avg       0.46      0.45      0.45     25000
```

```
Precisao (Taste): 0.46

Matriz de Confusão (Taste):

[[2222 1291 1163  131   13]
 [ 668 1534 2729  445   20]
 [ 278  972 5087 1785   58]
 [  92  218 2244 2538  179]
 [  37   39  327  715  215]]

Relatório de Classificação (Taste):

              precision    recall  f1-score   support

           1       0.67      0.46      0.55      4820
           2       0.38      0.28      0.32      5396
           3       0.44      0.62      0.52      8180
           4       0.45      0.48      0.47      5271
           5       0.44      0.16      0.24      1333

    accuracy                           0.46     25000
   macro avg       0.48      0.40      0.42     25000
weighted avg       0.47      0.46      0.46     25000
```

```
Precisao (Overall): 0.29

Matriz de Confusão (Overall):

[[ 238   50  323  447   32   40   15    8    1    0]
 [  39   42  300  727   78   52   23   12    1    1]
 [  38   34  336 1746  241  192   82   25    3    2]
 [  29   20  276 3040  946  961  350   66   12    2]
 [  10    3   58 1389  770 1137  493  102   14    1]
 [   2    4   38  828  605 1425  813  258   35    2]
 [   1    2   26  439  356 1022  878  401   70    4]
 [   1    1   16  173  130  486  597  385   73   11]
 [   3    1    6   82   46  167  244  233   73   11]
 [   2    0    3   19   10   30   59   64   46   12]]

Relatório de Classificação (Overall):

              precision    recall  f1-score   support

           1       0.66      0.21      0.31      1154
           2       0.27      0.03      0.06      1275
           3       0.24      0.12      0.16      2699
           4       0.34      0.53      0.42      5702
           5       0.24      0.19      0.21      3977
           6       0.26      0.36      0.30      4010
           7       0.25      0.27      0.26      3199
           8       0.25      0.21      0.22      1873
           9       0.22      0.08      0.12       866
          10       0.26      0.05      0.08       245

    accuracy                           0.29     25000
   macro avg       0.30      0.21      0.22     25000
weighted avg       0.29      0.29      0.27     25000
```

Classificação Multi-Classe - Teste:

- Vamos fazer um pequeno teste com reviews fabricadas :

1. Mau Cheiro/Bom Sabor : “This beer has a bad smelly aroma with hints of citrus but a nice caramel taste.”
2. Bom Cheiro/Mau Sabor : "This beer has a nice chocolate aroma with hints of citrus but a bad watery flavour."
3. Má: "This beer has a bad chocolate aroma with hints of cardboard and a bad flavour.“
4. Boa: "This beer has a nice chocolate aroma with hints of citrus and great flavour.“

```
Predictions for Review 1:
{'smell': 3, 'taste': 2, 'overall': 4}

Predictions for Review 2:
{'smell': 2, 'taste': 2, 'overall': 4}

Predictions for Review 3:
{'smell': 1, 'taste': 2, 'overall': 4}

Predictions for Review 4:
{'smell': 3, 'taste': 3, 'overall': 6}
```

Classificação Multi-Classe - Teste:

- Como podemos ver os resultados são péssimos bem podemos estar a elogiar ambas as características, como na review 4, que os resultados não passam de medíocres.
- Também conseguimos observar que, nas 2 primeiras reviews alternámos entre qual a caracteristica elogiavamos, e não mudava absolutamente nada até fez o contrário pois quando dissémos mal do aroma tivémos o resultado melhor do que quando dissemos que era agradável.
- Isto é claramente culpa da base de dados pois por muitos testes que fizéssemos quer mudássemos o classificador ou as suas características estes foram os melhores resultados.

```
Predictions for Review 1:
{'smell': 3, 'taste': 2, 'overall': 4}

Predictions for Review 2:
{'smell': 2, 'taste': 2, 'overall': 4}

Predictions for Review 3:
{'smell': 1, 'taste': 2, 'overall': 4}

Predictions for Review 4:
{'smell': 3, 'taste': 3, 'overall': 6}
```

PCA:

- Infelizmente não tivemos tempo de acabar este tópico mas fica aqui uma breve introdução. A aplicação de PCA (Principal Component Analysis) pode ter impactos diferentes nas tarefas de classificação binária, especialmente quando se lida com dados textuais.

- Sem PCA: A precisão pode ser razoavelmente boa, especialmente se os dados já estiverem num formato que permita um bom desempenho do classificador.

- Com PCA: A redução de dimensionalidade pode ajudar a simplificar o modelo, removendo redundâncias nos dados. Em dados textuais, onde a dimensionalidade pode ser alta, PCA pode ajudar a focar nas principais características, resultando num modelo mais eficiente.

```
Dimensão original do conjunto de dados: 6718
Precisão sem PCA: 0.93
Melhor Precisão com PCA (71 componentes): 0.91
```

PCA:

- Os resultados indicam que o PCA reduziu para 81 componentes o que fez com que houvesse uma ligeira redução na precisão do modelo. Podemos tirar algumas conclusões:
- Sem PCA (Precisão: 0.93): O modelo sem PCA atingiu uma precisão de 93%, o que é bastante bom. Isto sugere que os recursos originais eram informativos e suficientes para obter um bom desempenho.
- Com PCA (Melhor Precisão com 81 componentes - 0.91): A aplicação de PCA com uma redução para 81 componentes resultou numa precisão ligeiramente inferior (91%). Isso pode indicar que a informação contida nas primeiras 81 componentes não foi tão discriminativa quanto as características originais.
- A aplicação do PCA é um trade-off entre a simplificação do modelo e a preservação da informação. Em suma, apesar da redução na precisão, uma diminuição de 6718 dados para 71 é uma excelente redução e ajuda imenso na simplificação do modelo oferecendo uma maior rapidez. Isto não é necessário nesta classificação binária devida ao já baixo número de características mas pode ser benéfico noutros casos ou até mesmo na nossa classificação multi-classe.

```
Dimensão original do conjunto de dados: 6718
Precisão sem PCA: 0.93
Melhor Precisão com PCA (71 componentes): 0.91
```

Conclusões:

- Este trabalho prático deu-nos a possibilidade de consolidar todos os campos lecionadas ao longo da disciplina de Aprendizagem Automática como por exemplo: os diferentes tipos de classificação e os seus diversos classificadores, o modelo Bag of Words ou PCA.
- Apesar de não termos conseguido os melhores resultados acreditamos que conseguimos atingir todos os requisitos demonstrando o domínio que temos sobre a matéria lecionada.
- Contudo também acreditamos que se tivéssemos mais tempo ou a base de dados fosse melhor escolhida, conseguíamos facilmente atingir resultados mais satisfatórios.