



INF1771 – T3

INTELIGÊNCIA ARTIFICIAL

MARCELO PAULON, RODRIGO SILVA, GABRIEL MEDEIROS, RENAN DA FONTE

MODELO 1 - WEKA

- Pré-processamento (Weka):
 - Unificação da part1 e part2 do dataset
 - TextDirectoryLoader para importar .txt's
 - StringToWordVector
 - Lista de ~650 stop-words em inglês
 - TF-IDF
 - Snowball Stemmers Library
 - AttributeSelection – 60 atributos
 - Remoção manual de atributos – 57 palavras
 - 91,72% do dataset



Weka



MODELO 1

- Classificação:
- IBk (k-nearest neighbor) – cross validation: 10 folds
- Correctly Classified Instances 38462 76.924 %
- Incorrectly Classified Instances 11538 23.076 %
- Kappa statistic 0.5385
- Mean absolute error 0.3014
- Root mean squared error 0.4122
- Relative absolute error 60.2731 %
- Root relative squared error 82.439 %
- Total Number of Instances 50000

MODELO 1

- Classificação:
- J48 (C45) – cross validation: 10 folds
- Correctly Classified Instances 38500 77 %
- Incorrectly Classified Instances 11500 23 %
- Kappa statistic 0.54
- Mean absolute error 0.3187
- Root mean squared error 0.4059
- Relative absolute error 63.7386 %
- Root relative squared error 81.1795 %
- Total Number of Instances 50000

MODELO 1

- Classificação:
- LibSVM - cross validation: 10 folds
- Correctly Classified Instances 39458 78.916 %
- Incorrectly Classified Instances 10542 21.084 %
- Kappa statistic 0.5783
- Mean absolute error 0.2108
- Root mean squared error 0.4592
- Relative absolute error 42.168 %
- Root relative squared error 91.8346 %
- Total Number of Instances 50000

MODELO 1

- Classificação:
- SMO - cross validation: 10 folds
- Correctly Classified Instances 39480 78.96 %
- Incorrectly Classified Instances 10520 21.04 %
- Kappa statistic 0.5792
- Mean absolute error 0.2104
- Root mean squared error 0.4587
- Relative absolute error 42.08 %
- Root relative squared error 91.7388 %
- Total Number of Instances 50000

MODELO 1

- Classificação:
- MultiLayerPerceptron (Rede Neural) - cross validation: 10 folds
- Correctly Classified Instances 38804 77.608 %
- Incorrectly Classified Instances 11196 22.392 %
- Kappa statistic 0.5522
- Mean absolute error 0.2732
- Root mean squared error 0.3996
- Relative absolute error 54.6316 %
- Root relative squared error 79.9156 %
- Total Number of Instances 50000

MODELO 2 – WEKA + JAVA

- Pré-processamento (Java):
 - Unificação da part1 e part2 do dataset
 - Importação dos dados – Apache Commons IO
 - Lista de ~650 stop-words em inglês
 - StrikeAMatch para corrigir erros de grafia
 - Palavras positivas e negativas mais frequentes, com um limite percentual de presença no conjunto oposto
 - 350 palavras – parâmetro: frequência em cada review
 - 99,47% do dataset





MODELO 2

- Classificação:
- IBk (k-nearest neighbor) – cross validation: 10 folds
- Correctly Classified Instances 34460 69.2817 %
- Incorrectly Classified Instances 15279 30.7183 %
- Kappa statistic 0.3856
- Mean absolute error 0.3087
- Root mean squared error 0.5443
- Relative absolute error 61.7483 %
- Root relative squared error 108.8676 %
- Total Number of Instances 49739

MODELO 2

- Classificação:
- J48 (C45) – cross validation: 10 folds
- Correctly Classified Instances 38474 77.3518 %
- Incorrectly Classified Instances 11265 22.6482 %
- Kappa statistic 0.547
- Mean absolute error 0.2845
- Root mean squared error 0.4219
- Relative absolute error 56.9075 %
- Root relative squared error 84.3813 %
- Total Number of Instances 49739

MODELO 2

- Classificação:
- LibSVM - cross validation: 10 folds
- Correctly Classified Instances 41315 83.0636 %
- Incorrectly Classified Instances 8424 16.9364 %
- Kappa statistic 0.6613
- Mean absolute error 0.1694
- Root mean squared error 0.4115
- Relative absolute error 33.8728 %
- Root relative squared error 82.3077 %
- Total Number of Instances 49739

MODELO 2

- Classificação:
- SMO - cross validation: 10 folds
- Correctly Classified Instances 41569 83.5743 %
- Incorrectly Classified Instances 8170 16.4257 %
- Kappa statistic 0.6715
- Mean absolute error 0.1643
- Root mean squared error 0.4053
- Relative absolute error 32.8515 %
- Root relative squared error 81.0574 %
- Total Number of Instances 49739

MODELO 2

- Classificação:
- MultiLayerPerceptron (Rede Neural) - cross validation: 10 folds
- Tempo de execução > 3 dias, não executado

MODELO 3 – WEKA + JAVA

- Pré-processamento (Java):
 - Unificação da part1 e part2 do dataset
 - Importação dos dados – Apache Commons IO
 - Lista de ~650 stop-words em inglês
 - StrikeAMatch para corrigir erros de grafia
 - Palavras positivas e negativas mais frequentes, com um limite percentual de presença no conjunto oposto
 - 250 palavras – parâmetro: TF-IDF
 - 99,28% do dataset



MODELO 3

- Classificação:
- IBk (k-nearest neighbor) – cross validation: 10 folds
- Correctly Classified Instances 31800 64.0574 %
- Incorrectly Classified Instances 17843 35.9426 %
- Kappa statistic 0.2811
- Mean absolute error 0.3596
- Root mean squared error 0.5992
- Relative absolute error 71.9234 %
- Root relative squared error 119.8417 %
- Total Number of Instances 49643

MODELO 3

- Classificação:
- J48 (C45) – cross validation: 10 folds
- Correctly Classified Instances 38227 77.0038 %
- Incorrectly Classified Instances 11416 22.9962 %
- Kappa statistic 0.5401
- Mean absolute error 0.2838
- Root mean squared error 0.4272
- Relative absolute error 56.7586 %
- Root relative squared error 85.4321 %
- Total Number of Instances 49643

MODELO 3

- Classificação:
- LibSVM - cross validation: 10 folds
- Correctly Classified Instances 40616 81.8162 %
- Incorrectly Classified Instances 9027 18.1838 %
- Kappa statistic 0.6363
- Mean absolute error 0.1818
- Root mean squared error 0.4264
- Relative absolute error 36.3677 %
- Root relative squared error 85.285 %
- Total Number of Instances 49643

MODELO 3

- Classificação:
- SMO - cross validation: 10 folds
- Correctly Classified Instances 41126 82.8435 %
- Incorrectly Classified Instances 8517 17.1565 %
- Kappa statistic 0.6569
- Mean absolute error 0.1716
- Root mean squared error 0.4142
- Relative absolute error 34.313 %
- Root relative squared error 82.8408 %
- Total Number of Instances 49643

MODELO 3

- Classificação:
- MultiLayerPerceptron (Rede Neural) - cross validation: 10 folds

OBRIGADO