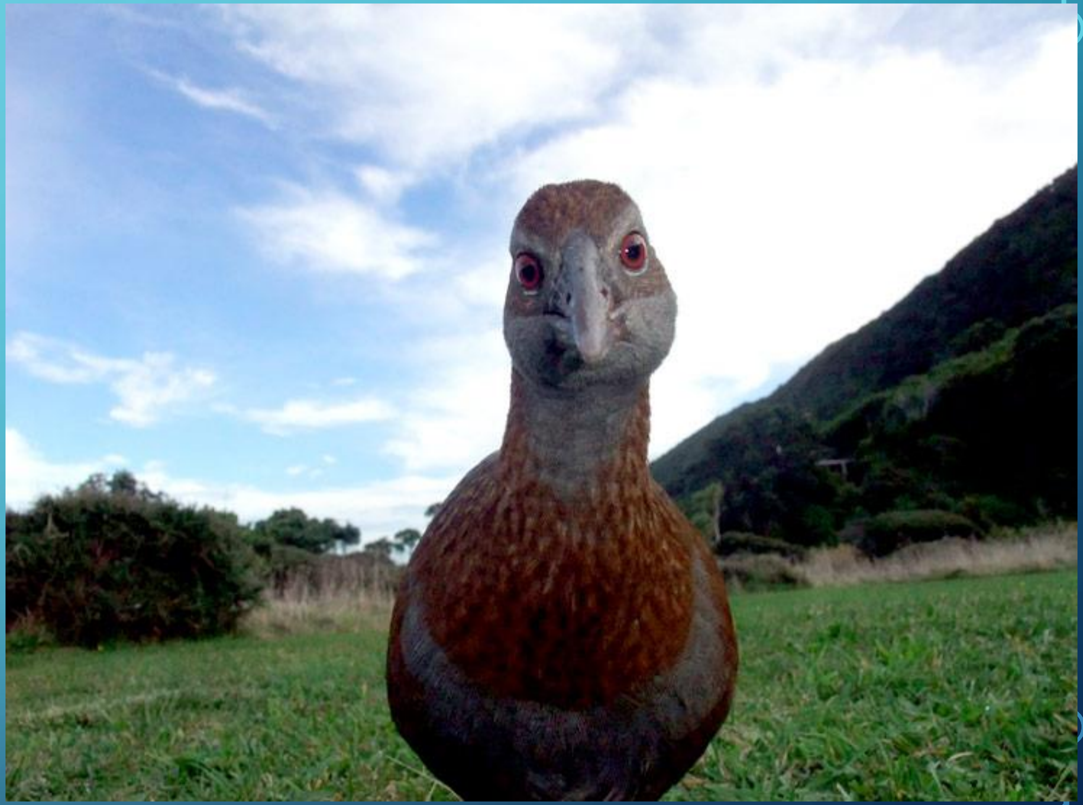
INF1771 – T3
INTELIGÊNCIA ARTIFICIAL

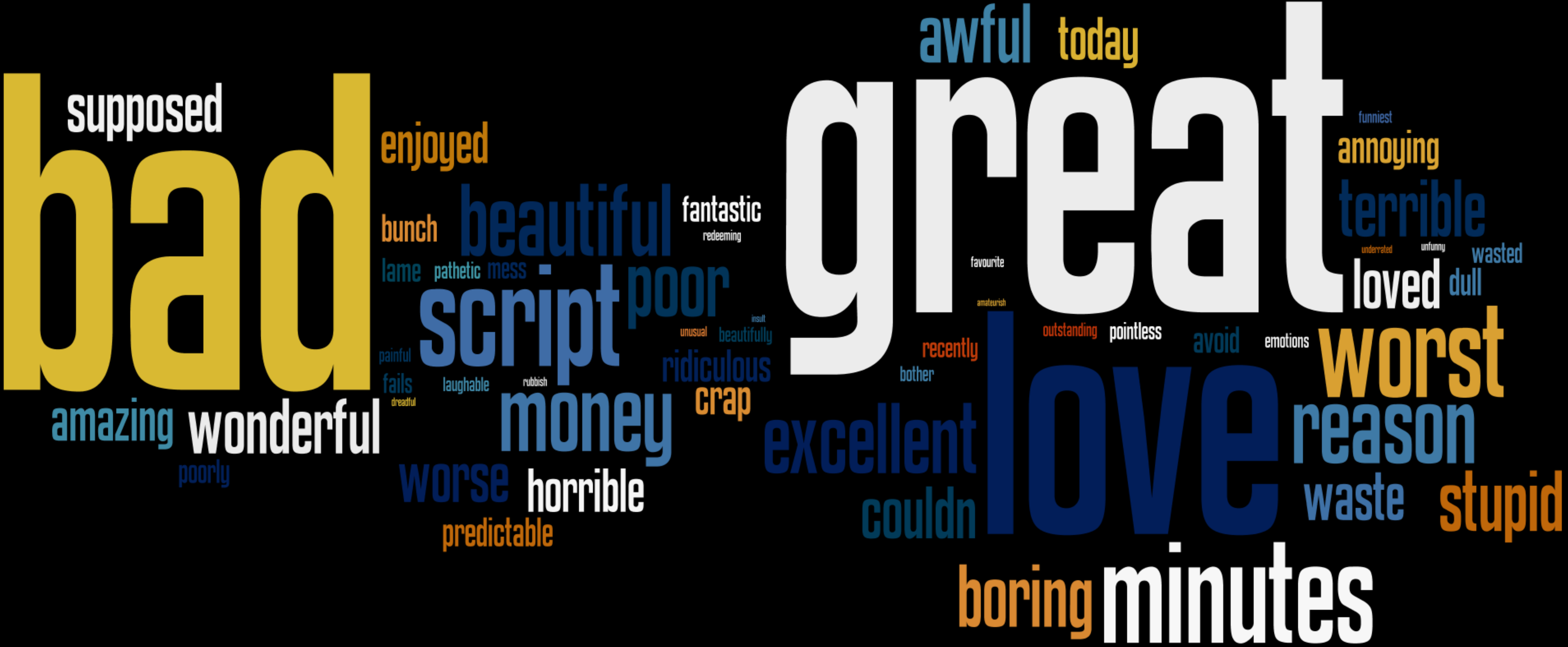
MARCELO PAULON, RODRIGO SILVA, GABRIEL MEDEIROS, RENAN DA FONTE

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


bad
supposed
enjoyed
bunch
beautiful
fantastic
redeeming
lame
pathetic
mess
script
poor
insult
unusual
beautifully
painful
ridiculous
fails
laughable
rubbish
dreadful
crap
amazing
wonderful
money
poorly
worse
horrible
predictable
awful
today
great
funniest
annoying
terrible
underrated
unfunny
wasted
favourite
loved
dull
amateurish
outstanding
pointless
avoid
emotions
recently
bother
worst
love
excellent
reason
couldn
waste
stupid
boring
minutes

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

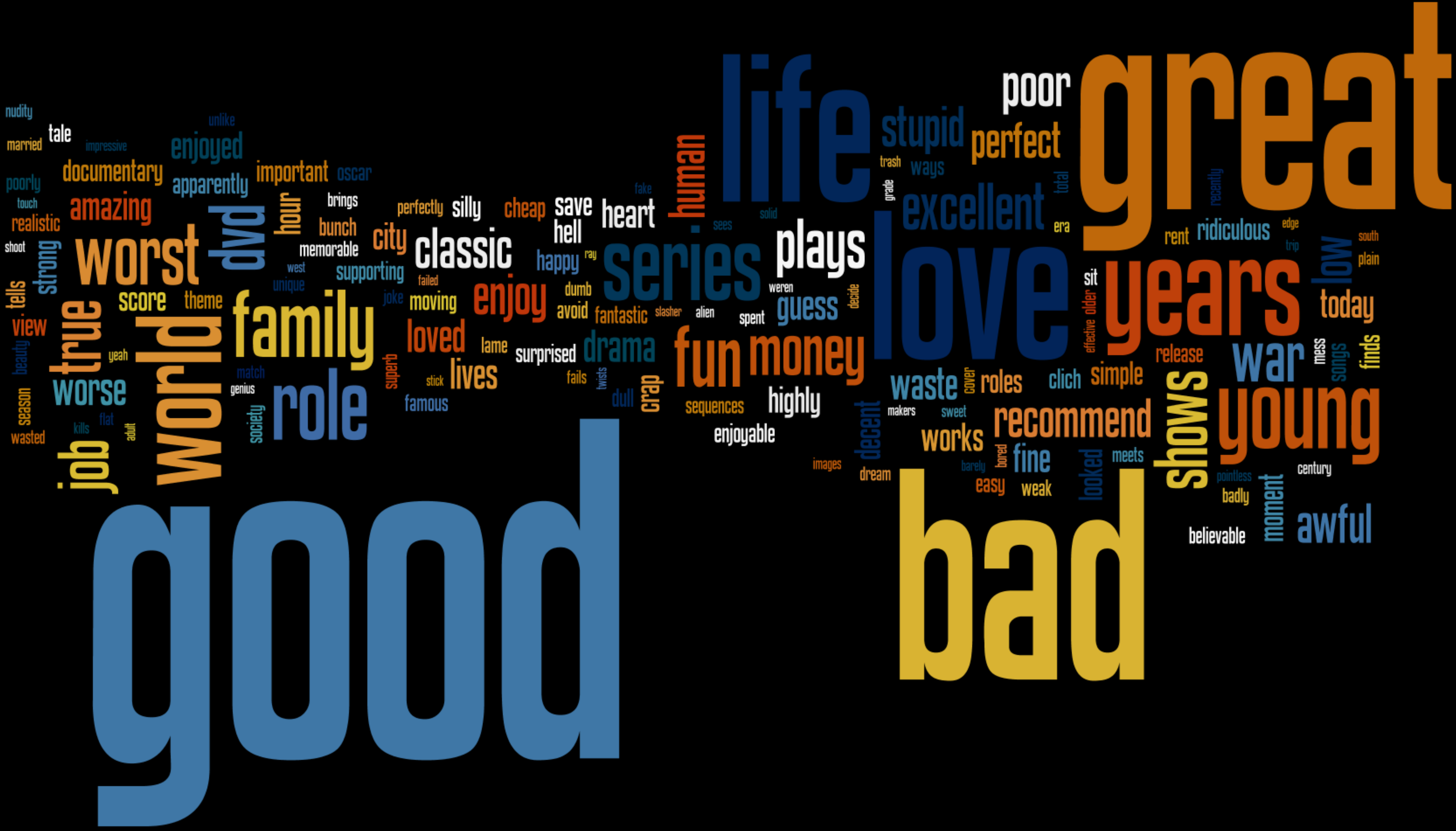
- Classificação:
- SMO - cross validation: 10 folds
- Correctly Classified Instances 39480 78.96 %
- Incorrectly Classified Instances 10520 21.04 %
- Kappa statistic 0.5792
- Mean absolute error 0.2104
- Root mean squared error 0.4587
- Relative absolute error 42.08 %
- Root relative squared error 91.7388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


nudity
unlike
tale
married
impressive
enjoyed
poorly
documentary
apparently
important
oscar
touch
amazing
realistic
hour
brings
perfectly
silly
cheap
save
heart
human
life
stupid
poor
great
perfect
trash
ways
fake
sees
solid
grade
excellent
total
recently
era
edge
rent
ridiculous
south
trip
plain
low
shoot
strong
worst
dvd
bunch
memorable
city
classic
hell
series
plays
love
years
west
unique
supporting
happy
ray
failed
dumb
weren
sit
tells
score
theme
family
joke
moving
enjoy
avoid
fantastic
slasher
alien
spent
guess
decide
older
effective
today
view
true
world
loved
lame
surprised
drama
fun
money
release
war
mess
songs
finds
beauty
yeah
match
superb
stick
lives
fails
twists
crap
waste
cover
roles
clich
simple
shows
young
worse
genius
role
famous
dull
sequences
highly
makers
sweet
recommend
season
kills
flat
adult
society
enjoyable
decent
works
fine
meets
century
wasted
job
images
dream
barely
bored
easy
weak
looked
pointless
badly
moment
awful
believable
good
bad

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

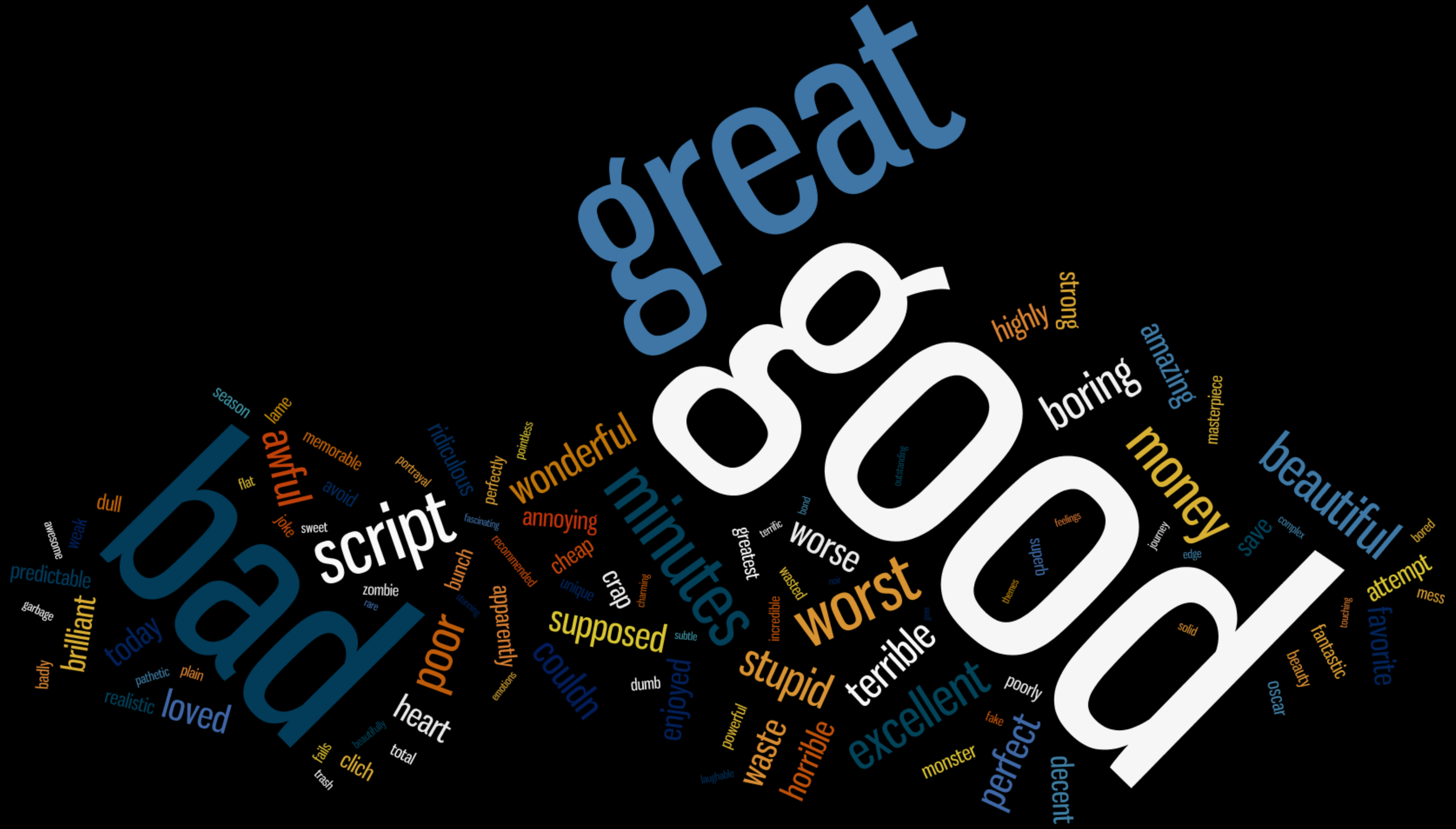
- Classificação:
- SMO - cross validation: 10 folds
- Correctly Classified Instances 41569 83.5743 %
- Incorrectly Classified Instances 8170 16.4257 %
- Kappa statistic 0.6715
- Mean absolute error 0.1643
- Root mean squared error 0.4053
- Relative absolute error 32.8515 %
- Root relative squared error 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
  - 100 palavras – parâmetro: TF-IDF
  - 95,25% do dataset


good
great
bad
script
minutes
worst
boring
money
beautiful
excellent
wonderful
stupid
terrible
worse
poor
awful
perfect
waste
horrible
supposed
couldn
annoying
amazing
highly
strong
loved
heart
brilliant
today
ridiculous
memorable
apparently
enjoyed
attempt
favorite
fantastic
decent
save
masterpiece
predictable
realistic
avoid
cheap
crap
bunch
dull
lame
season
flat
portrayal
perfectly
pointless
joke
sweet
weak
awesome
garbage
badly
pathetic
plain
zombie
rare
fascinating
recommended
unique
stunning
charming
emotions
beautifully
fails
clich
trash
total
dumb
subtle
greatest
terrific
wasted
incredible
noir
gem
powerful
laughable
monster
fake
poorly
outstanding
bond
themes
superb
feelings
journey
edge
complex
solid
touching
beauty
oscar
bored
mess

MODELO 3

- Classificação:
- IBk (k-nearest neighbor) – cross validation: 10 folds
- Correctly Classified Instances 33965 71.3161 %
- Incorrectly Classified Instances 13661 28.6839 %
- Kappa statistic 0.4267
- Mean absolute error 0.2887
- Root mean squared error 0.5322
- Relative absolute error 57.7551 %
- Root relative squared error 106.4392 %
- Total Number of Instances 47626

MODELO 3

- Classificação:
- J48 (C45) – cross validation: 10 folds
- Correctly Classified Instances 37357 78.4382 %
- Incorrectly Classified Instances 10269 21.5618 %
- Kappa statistic 0.5688
- Mean absolute error 0.2894
- Root mean squared error 0.4036
- Relative absolute error 57.8959 %
- Root relative squared error 80.7329 %
- Total Number of Instances 47626

MODELO 3

- Classificação:
- LibSVM - cross validation: 10 folds
- Correctly Classified Instances 38825 81.5206 %
- Incorrectly Classified Instances 8801 18.4794 %
- Kappa statistic 0.6307
- Mean absolute error 0.1848
- Root mean squared error 0.4299
- Relative absolute error 36.9628 %
- Root relative squared error 85.9801 %
- Total Number of Instances 47626

MODELO 3

- Classificação:
- SMO - cross validation: 10 folds

Correctly Classified Instances	39363	82.6502 %
Incorrectly Classified Instances	8263	17.3498 %
Kappa statistic	0.6532	
Mean absolute error	0.1735	
Root mean squared error	0.4165	
Relative absolute error	34.7033 %	
Root relative squared error	83.3107 %	
Total Number of Instances	47626	

MODELO 3

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

Correctly Classified Instances	37377	78.4802 %
Incorrectly Classified Instances	10249	21.5198 %
Kappa statistic	0.5703	
Mean absolute error	0.2672	
Root mean squared error	0.3886	
Relative absolute error	53.4488 %	
Root relative squared error	77.7282 %	
Total Number of Instances	47626	

MODELO 4 – 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

good
bad
great
love
life
world
years
family
young
series
worst
shows
excellent
money
fun
awful
true
classic
stupid
human
poor
role
job
enjoy
plays
worse
drama
war
dvd
loved
highly
lives
hour
view
amazing
perfect
recommend
documentary
guess
strong
ridiculous
cheap
simple
fine
waste
low
today
looked
decent
heart
enjoyed
city
moment
finds
happy
poorly
meets
total
believable
score
easy
wasted
tale
oscar
moving
dream
effective
dull
release
season
songs
works
tells
ways
important
crap
silly
apparently
fantastic
save
century
cover
mess
hell
theme
brings
roles
enjoyable
lame
recently
supporting
superb
bunch
sequences
older
unique
famous
match
weak
society
realistic
sit
perfectly
beauty
joke
rent
spent
avoid
trash
clich
plain
dumb
surprised
memorable
married
west
yeah
slasher
barely
touch
shoot
decent
adult
flat
decide
alien
era
sweet
makers
grade
solid
images
bored
edge
sees
impressive
stick
badly
pointless
unlike
genius
twists
fails
kills
fake
warren
failed
south
mostly

MODELO 4

- 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 4

- 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 4

- 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 4

- 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 4

- Classificação:
- MultiLayerPerceptron (Rede Neural) - cross validation: 10 folds
- Correctly Classified Instances 26830 54.0459 %
- Incorrectly Classified Instances 22813 45.9541 %
- Kappa statistic 0.0809
- Mean absolute error 0.4589
- Root mean squared error 0.6292
- Relative absolute error 91.7876 %
- Root relative squared error 125.832 %
- Total Number of Instances 49643

OBRIGADO