Detectando questões duplicadas: Quora Questions Pairs

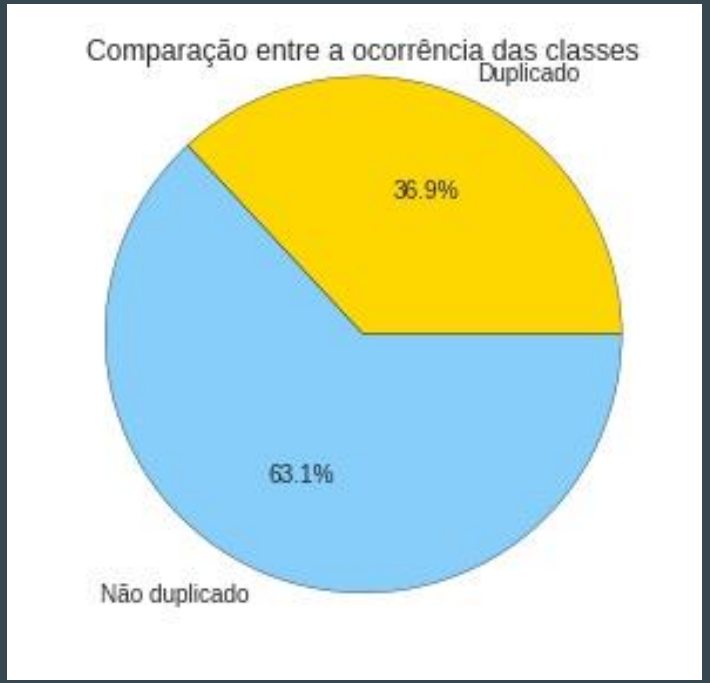
...

Airine Carmo e Christian Cardozo

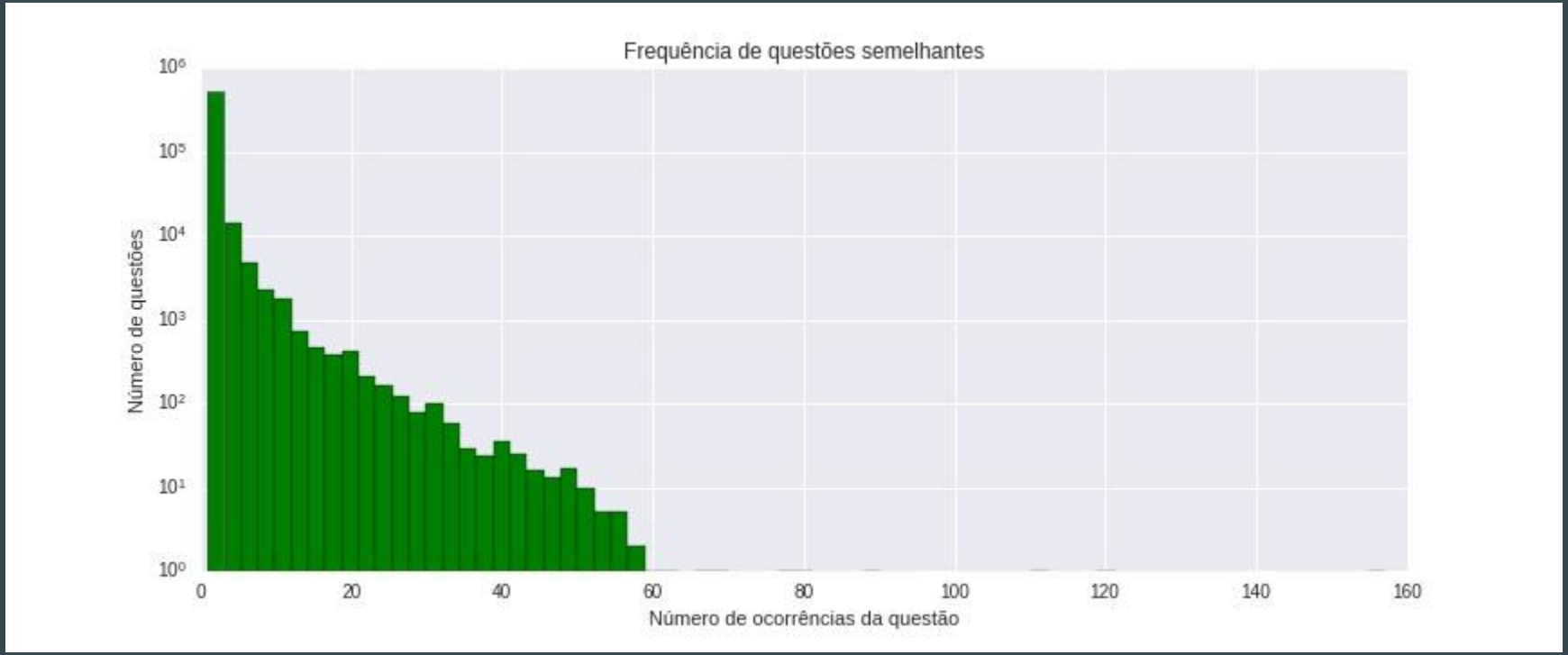
Sumário

Análise dos dados

- Conjunto de dados do treinamento
- 404.290 registros
- 255.027 registros são da classe ‘0’
- 149.263 são da classe ‘1’.



Análise dos dados



Pré-processamento

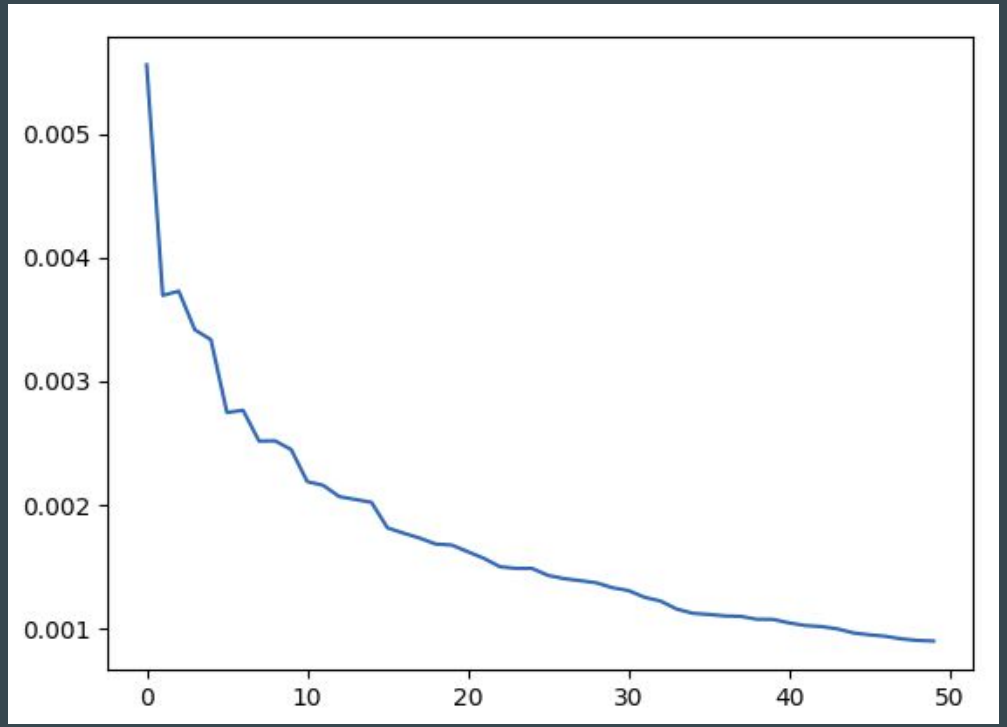
1. Transformação e Tokenização
2. Remoção de Stop Words
3. Stemming
4. Construção da matriz de presença e n-gramas
5. Redução de dimensionalidade
6. Combinando features

Construção da matriz de presença e n-gramas

- Foram geradas todas as combinações presentes no texto de 1-gramas, 2-gramas e 3-gramas
- Foram geradas 3 matrizes de presença onde:
  - A primeira matriz se refere à presença de n-gramas na primeira questão do par.
  - A segunda, se refere à presença de n-gramas na segunda questão do par.
  - A terceira, é feita utilizando a presença de n-gramas que aparecem nas duas perguntas.
- São concatenadas horizontalmente, formando uma matriz com total de 3.315.693 colunas

Redução de dimensionalidade (SVD)

- Mostra os primeiros 50 autovalores da matriz diagonal Sigma gerada pelo SVD.
- Foi definido um ponto de corte K=10
- SVD gera como saída uma nova matriz com apenas 10 colunas.



Combinando features

Além das 10 geradas no SVD, são geradas 6 features adicionais sendo:

1. Diferença de quantidade de tokens entre as questões dividido pela quantidade de tokens na questão 1
2. A diferença de quantidade de tokens entre as questões dividido pela quantidade de tokens na questão 2
3. Distância de Jaccard entre os conjuntos de tokens das questões
4. Distância de Leveinshtein entre os conjuntos de tokens das questões
5. Quantidade de tokens que aparecem nas duas perguntas dividido pela quantidade de tokens na questão 1
6. Quantidade de tokens que aparecem nas duas perguntas dividido pela quantidade de tokens na questão 2

Totalizando 16 features para a matriz final

Experimentos

Foram utilizados três modelos de aprendizado de máquina sendo eles:

- Naive Bayes do scikit-learn
- Redes neurais do scikit-learn
- Xgboost

Seguindo estes parâmetros, para todos:

- Utilizando o conjunto de exemplos dividido em 10% para validação e 90% para treinamento.
- Método de K-Fold Cross Validation com K=3
- Duas métricas de avaliação sendo: acurácia e log loss.

Naive Bayes

Naive Bayes Gaussiano		
	Acurácia	Log loss
Conjunto de treinamento	68.2416% (+/- 0.0422%)	1.3230 (+/- 0.0097)
Conjunto de validação	68.1317%	1.3590

Naive Bayes Bernoulli		
	Acurácia	Log loss
Conjunto de treinamento	62.0856% (+/- 0.0993%)	0.6463 (+/- 0.0014)
Conjunto de validação	66.7095%	0.5987

Naive Bayes Multinomial		
	Acurácia	Log loss
Conjunto de treinamento	66.6351% (+/- 0.0734%)	0.5994 (+/- 0.0001)
Conjunto de validação	62.1954%	0.6483

Rede Neural

Foram testadas as seguintes variações:

- Quantidade de camadas: 1 e 2
- Quantidade de neurônios por camada: 10, 30 e 50
- Função de ativação: relu, tangente hiperbólica e logística
- Ajuste de pesos: adam, lbfgs e sgd

Melhor desempenho

Quantidade de camadas	2
Quantidade de neurônios por camada	50
Função de ativação	relu
Ajuste de pesos	adam

	Acurácia	Log loss
Conjunto de treinamento	75.6323% (+/- 0.0554%)	0.4633 (+/- 0.0012)
Conjunto de validação	75.6511%	0.4629

Xgboost

XGBoost é uma biblioteca de gradiet boosting otimizada e muito eficiente

Parâmetros iniciais:

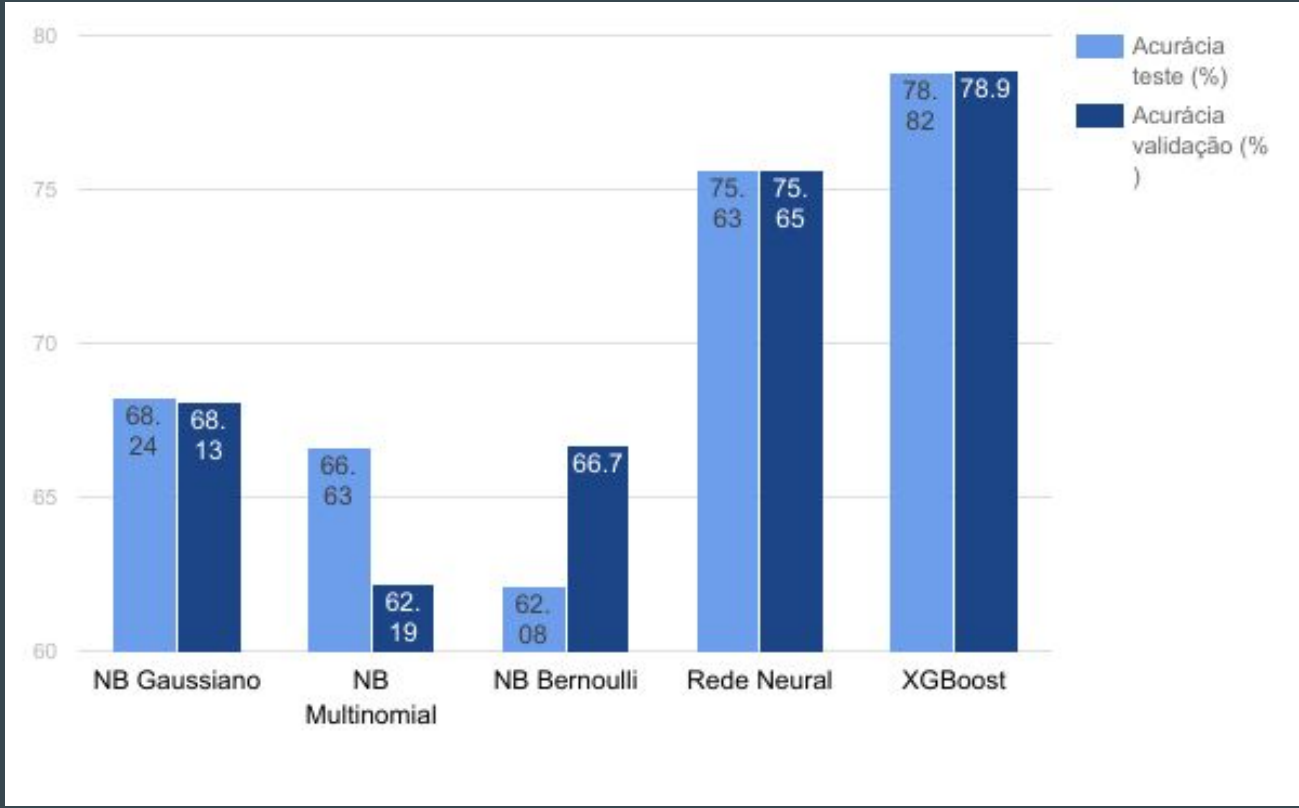
- learning_rate: 0.15
- n_estimators: 170
- max_depth: 6

Melhor desempenho

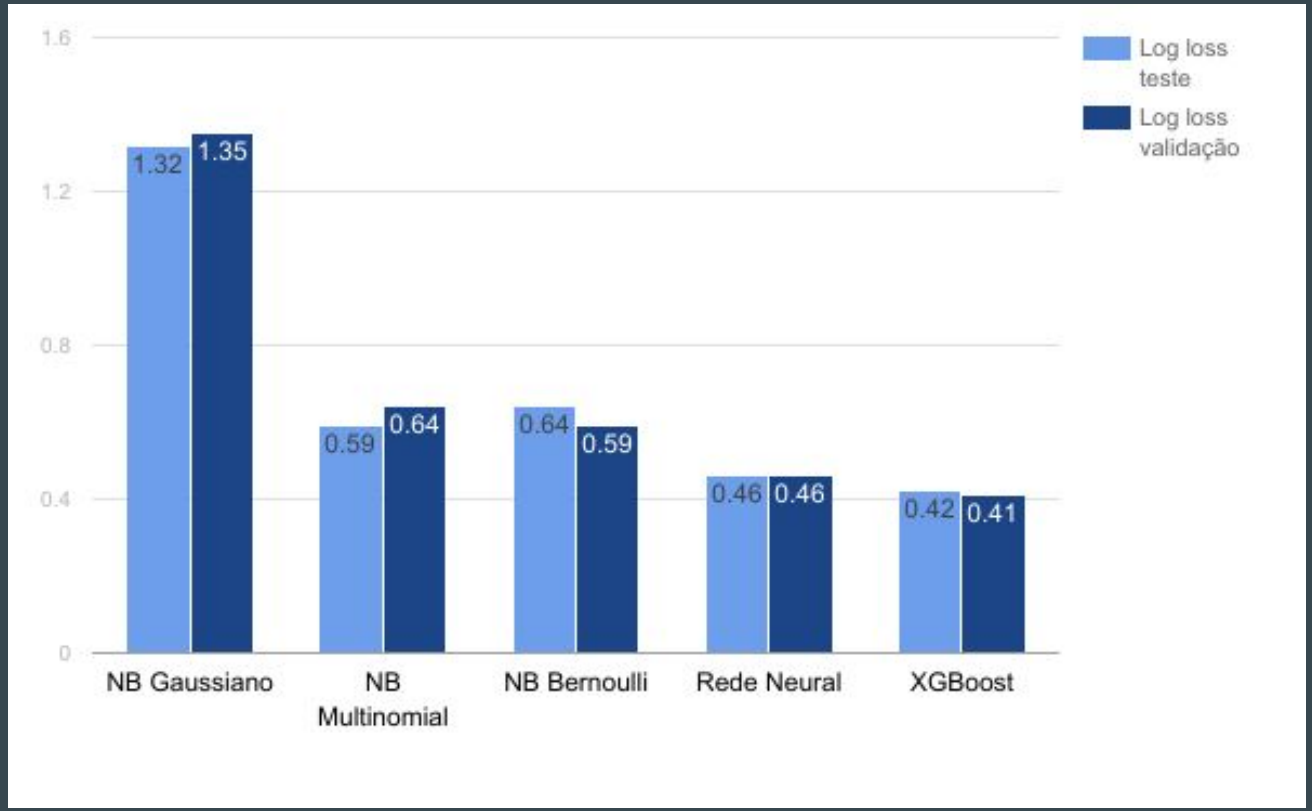
learning_rate	0.15
n_estimators	200
max_depth	12

	Acurácia	Log loss
Conjunto de treinamento	78.8276% (+/- 0.05%)	0.422606 (+/- 0.0008)
Conjunto de validação	78.9012%	0.4196

Resultados e considerações finais



Resultados e considerações finais



Referências

Quora Questions Pairs, 2017. Disponível em: https://www.kaggle.com/c/quora-question-pairs. Acesso em: 15 de mai. 2017.

Kaggle, 2017. Disponível em: https://www.kaggle.com/. Acesso em: 15 de mai. 2017.

Quora, 2017. Disponível em: https://www.quora.com/. Acesso em: 15 de mai. 2017.

Python Software Foundation, 2017. Disponível em: https://www.python.org/. Acesso em: 15 de mai. 2017.

scikit-learn: Machine Learning in Python, 2017. Disponível em: http://scikit-learn.org/stable/. Acesso em: 15 de mai. 2017.

Natural Language Toolkit, 2017. Disponível em: http://www.nltk.org/. Acesso em: 15 de mai. 2017.

GitHub, 2017. Disponível em: https://github.com/chriiscardozo/QuoraQuestionPair/. Acesso em: 15 de mai. 2017.

Stemmers, 2017. Disponível em: http://www.nltk.org/howto/stem.html. Acesso em: 15 de mai. 2017.

XGBoost Documents, 2017. Disponível em: https://xgboost.readthedocs.io/en/latest/. Acesso em: 15 de mai. 2017.

Complete Guide to Parameter Tuning in XGBoost (with codes in Python), 2017. Disponível em: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/, 2017. Acesso em: 15 de mai. 2017.