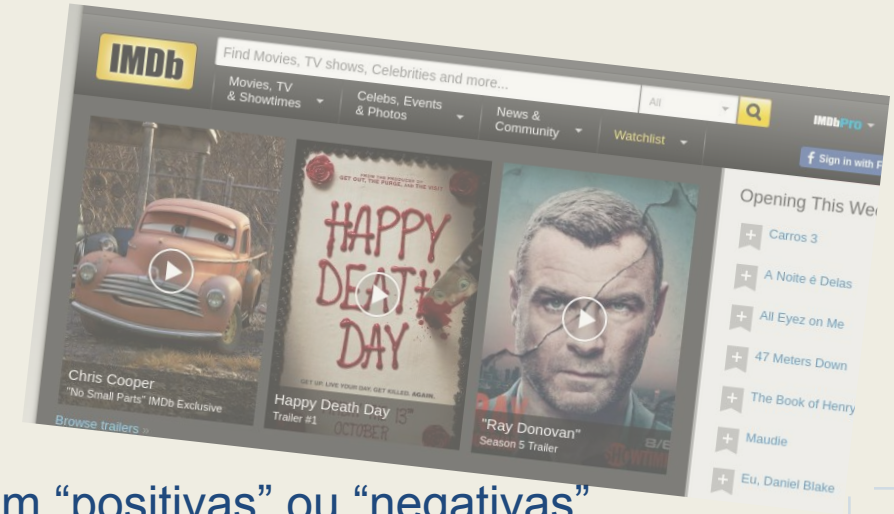
Machine Learning

IA / Redes Neurais

Carlos Cardoso Dias
201510000312

Objetivo

- Análise de sentimentos
- Opiniões sobre filmes
- Classificar opiniões sobre filmes em “positivas” ou “negativas”



Ferramentas

- Python 3.6
- NLTK
- Scikit
- Pandas
- Weka

Dados

- 50 mil avaliações (em inglês) de filmes do IMDb
- 25 mil avaliações “positivas”
- 25 mil avaliações “negativas”

Desenvolvimento

- Fase Exploratória
  - 10% dos dados selecionados aleatoriamente
  - 20% dos dados selecionados aleatoriamente
  - Variações dos parâmetros de treinamento, features extraídas, pré-processamento e algoritmo utilizado
- Fase Analítica
  - Análise dos dados levantados durante a fase exploratória para cada uma das variações
  - Treinamento
  - Seleção e implementação do algoritmo para o melhor resultado utilizando todo o conjunto de dados

Atributos

- N melhores características avaliadas com diferentes medidas estatísticas
- Características N-gram com N variando de 1 à 3
- Exemplos de atributos selecionados: aw, bad, beauti, best, bore, crap, excel, favorit, great, horribl, lame, love, minut, money, perfect, poor, poorli, ridicul, stupid, suck, suppos, terribl, wast, wors, worst
- Pipeline
  - Tokenizing, Stop Words, Stemming, N-Gram, TF-IDF, BestK

Pipeline

- Tokenizing
- Stop Words
- Stemming
- N-Gram
- TF-IDF
- BestK
- Weka

extractor.py

Código….

Árvores de Decisão F: 25, 1

- Tempo para criação do modelo: 4.79 segundos
- Acurácia prevista por cross-validation: 76.7133%
- Acurácia no conjunto de teste: 77.12%
- Ordem do tempo de classificação: milisegundos

Árvores de Decisão F: 25, 2

- Tempo para criação do modelo: 4.44 segundos
- Acurácia prevista por cross-validation: 76.0644%
- Acurácia no conjunto de teste: 76.34%
- Ordem do tempo de classificação: 1 segundo

Árvores de Decisão F: 50, 1

- Tempo para criação do modelo: 13.91 segundos
- Acurácia prevista por cross-validation: 77.7044%
- Acurácia no conjunto de teste: 77.52%
- Ordem do tempo de classificação: 1 segundo

Árvores de Decisão F: 50, 2

- Tempo para criação do modelo: 13.05 segundos
- Acurácia prevista por cross-validation: 77.7844%
- Acurácia no conjunto de teste: 77.4%
- Ordem do tempo de classificação: milisegundos

K-Nearest Neighbor (KNN) F: 25, 1

- Tempo para criação do modelo: 0.01 segundos
- Acurácia prevista por cross-validation: 71.1489%
- Acurácia no conjunto de teste: 72.1%
- Ordem do tempo de classificação: 24 segundos

K-Nearest Neighbor (KNN) F: 25, 2

- Tempo para criação do modelo: 0.01 segundos
- Acurácia prevista por cross-validation: 70.92%
- Acurácia no conjunto de teste: 71.06%
- Ordem do tempo de classificação: 25 segundos

K-Nearest Neighbor (KNN) F: 50, 1

- Tempo para criação do modelo: 0.01 segundos
- Acurácia prevista por cross-validation: 71.5778%
- Acurácia no conjunto de teste: 71.52%
- Ordem do tempo de classificação: 44 segundos

K-Nearest Neighbor (KNN) F: 50, 2

- Tempo para criação do modelo: 0.01 segundos
- Acurácia prevista por cross-validation: 72.7644%
- Acurácia no conjunto de teste: 73.24%
- Ordem do tempo de classificação: 45 segundos

Rede Neural (backpropagation) F: 25, 1

- Tempo para criação do modelo: 128.75 segundos
- Acurácia prevista por cross-validation: 76.8756%
- Acurácia no conjunto de teste: 78.2%
- Ordem do tempo de classificação: milisegundos

Rede Neural (backpropagation) F: 25, 2

- Tempo para criação do modelo: 131.57 segundos
- Acurácia prevista por cross-validation: 76.1711%
- Acurácia no conjunto de teste: 77.66%
- Ordem do tempo de classificação: milisegundos

Rede Neural (backpropagation) F: 50, 1

- Tempo para criação do modelo: 419.04 segundos
- Acurácia prevista por cross-validation: 77.1356%
- Acurácia no conjunto de teste: 77.48%
- Ordem do tempo de classificação: 1 segundo

Rede Neural (backpropagation) F: 50, 2

- Tempo para criação do modelo: 420.19 segundos
- Acurácia prevista por cross-validation: 79.1933%
- Acurácia no conjunto de teste: 77.46%
- Ordem do tempo de classificação: 1 segundo

Support Vector Machine (SVM) F: 25, 1

- Tempo para criação do modelo: 67.17 segundos
- Acurácia prevista por cross-validation: 78.0689%
- Acurácia no conjunto de teste: 78.22%
- Ordem do tempo de classificação: 3 segundos

Support Vector Machine (SVM) F: 25, 2

- Tempo para criação do modelo: 64.68 segundos
- Acurácia prevista por cross-validation: 77.0267%
- Acurácia no conjunto de teste: 77.72%
- Ordem do tempo de classificação: 3 segundos

Support Vector Machine (SVM) F: 50, 1

- Tempo para criação do modelo: 126.78 segundos
- Acurácia prevista por cross-validation: 80.8733%
- Acurácia no conjunto de teste: 81.02%
- Ordem do tempo de classificação: 5 segundos

Support Vector Machine (SVM) F: 50, 2

- Tempo para criação do modelo: 108.5 segundos
- Acurácia prevista por cross-validation: 80.3489%
- Acurácia no conjunto de teste: 80.52%
- Ordem do tempo de classificação: 4 segundos

Support Vector Machine (SVM) F: 200, 1

- Tempo para criação do modelo: 354.93 segundos
- Acurácia prevista por cross-validation: 84.5867%
- Acurácia no conjunto de teste: 84.16%
- Ordem do tempo de classificação: 16 segundos

Análise - Taxa de Reconhecimento

1. SVM
2. Rede Neural
3. Decision Tree
4. KNN

Análise - Tempo de Treinamento

1. KNN
2. Decision Tree
3. SVM
4. Rede Neural

Análise - Tempo de Classificação

1. Decision Tree
2. Rede Neural
3. SVM
4. KNN

analyser.py

Código….

Conclusões

- A qualidade do classificador está diretamente associada à qualidade dos dados
- A variação dos parâmetros para cada classificador pode produzir mudanças significativas
- O aumento do número de características não significa necessariamente um aumento na qualidade da classificação