Classificador - Sites de Jogos

Italo Soares

Dificuldades

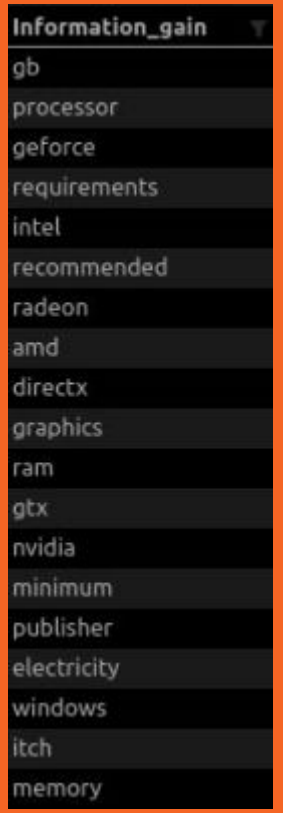
- **Origin sempre baixava o mesmo HTML -> Tive que baixar manualmente**
- **Nem todas as páginas eram em UTF-8 -> Só converter antes de pré processar**

Classificação

- **Scikit-learn**
- **Naive Bayes -> Gaussian NB**
- **SVM -> SVC**
- **Logistic Regression -> LogisticRegression**
- **Decision Tree -> DecisionTreeClassifier**
- **Multilayer Perceptron -> MLPClassifier**
- **Validação cruzada estratificada em 10 segmentos**

- Desempenho
  - Acurácia
  - Recall
  - Precisão
  - Tempo de Treinamento
- 9 bags of words
  - Contagem e frequência
  - Stemming, stopwords e information gain

- Information Gain
- Melhores palavras são as de requisitos do sistema



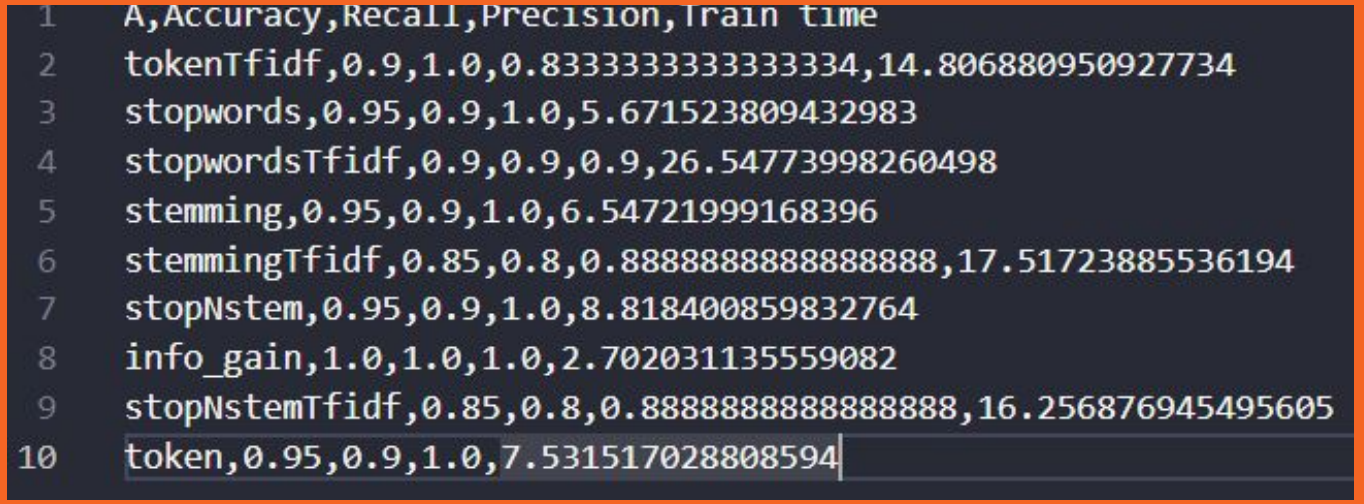
Decision Tree

```
A,Accuracy,Recall,Precision,Train time
tokenTfidf,1.0,1.0,1.0,0.09793806076049805
stopwords,0.9,1.0,0.8333333333333334,0.08618307113647461
stopwordsTfidf,1.0,1.0,1.0,0.09923601150512695
stemming,1.0,1.0,1.0,0.06911587715148926
stemmingTfidf,0.9,0.8,1.0,0.09247183799743652
stopNstem,1.0,1.0,1.0,0.07476687431335449
info_gain,0.85,1.0,0.7692307692307693,0.020652055740356445
stopNstemTfidf,0.9,0.9,0.9,0.09745097160339355
token,0.85,0.8,0.8888888888888888,0.09821200370788574
```

Logistic Regression

```
A,Accuracy,Recall,Precision,Train time
tokenTfidf,0.95,0.9,1.0,0.030242919921875
stopwords,0.9,1.0,0.8333333333333334,0.08200979232788086
stopwordsTfidf,0.95,1.0,0.9090909090909091,0.032950878143310555
stemming,0.9,1.0,0.8333333333333334,0.090000992774963379
stemmingTfidf,0.95,0.9,1.0,0.027146100997924805
stopNstem,0.9,1.0,0.8333333333333334,0.0693368911743164
info_gain,0.9,1.0,0.8333333333333334,0.014688014984130866
stopNstemTfidf,0.95,0.9,1.0,0.024194955825805664
token,0.9,1.0,0.8333333333333334,0.11249399185180664
```

MLP



```
A,Accuracy,Recall,Precision,Train time
tokenTfidf,0.9,1.0,0.8333333333333334,14.806880950927734
stopwords,0.95,0.9,1.0,5.671523809432983
stopwordsTfidf,0.9,0.9,0.9,26.54773998260498
stemming,0.95,0.9,1.0,6.54721999168396
stemmingTfidf,0.85,0.8,0.8888888888888888,17.51723885536194
stopNstem,0.95,0.9,1.0,8.818400859832764
info_gain,1.0,1.0,1.0,2.702031135559082
stopNstemTfidf,0.85,0.8,0.8888888888888888,16.256876945495605
token,0.95,0.9,1.0,7.531517028808594
```

Naive Bayes

```
A,Accuracy,Recall,Precision,Train time
tokenTfidf,0.7,0.4,1.0,0.05216789245605469
stopwords,0.7,0.6,0.75,0.0586240291595459
stopwordsTfidf,0.7,0.4,1.0,0.05155611038208008
stemming,0.85,0.8,0.8888888888888888,0.0479788780212402 34
stemmingTfidf,0.85,0.8,0.8888888888888888,0.05680203437805176
stopNstem,0.85,0.8,0.8888888888888888,0.05696916580200195
info_gain,0.75,0.7,0.7777777777777778,0.008568048477172852
stopNstemTfidf,0.85,0.8,0.8888888888888888,0.04703497886657715
token,0.7,0.6,0.75,0.06105995178222656
```

SVM

```
A,Accuracy,Recall,Precision,Train time
tokenTfidf,0.9,1.0,0.8333333333333334,0.5715620517730713
stopwords,0.9,0.8,1.0,0.5496160984039307
stopwordsTfidf,0.95,0.9,1.0,0.5635490417480469
stemming,0.9,0.9,0.9,0.4180440902709961
stemmingTfidf,0.9,1.0,0.8333333333333334,0.5406520366668701
stopNstem,0.9,0.9,0.9,0.480798959732055 66
info_gain,0.85,0.7,1.0,0.09015393257141113
stopNstemTfidf,0.9,1.0,0.8333333333333334,0.5049870014190674
token,0.9,0.9,0.9,0.626694917678833
```