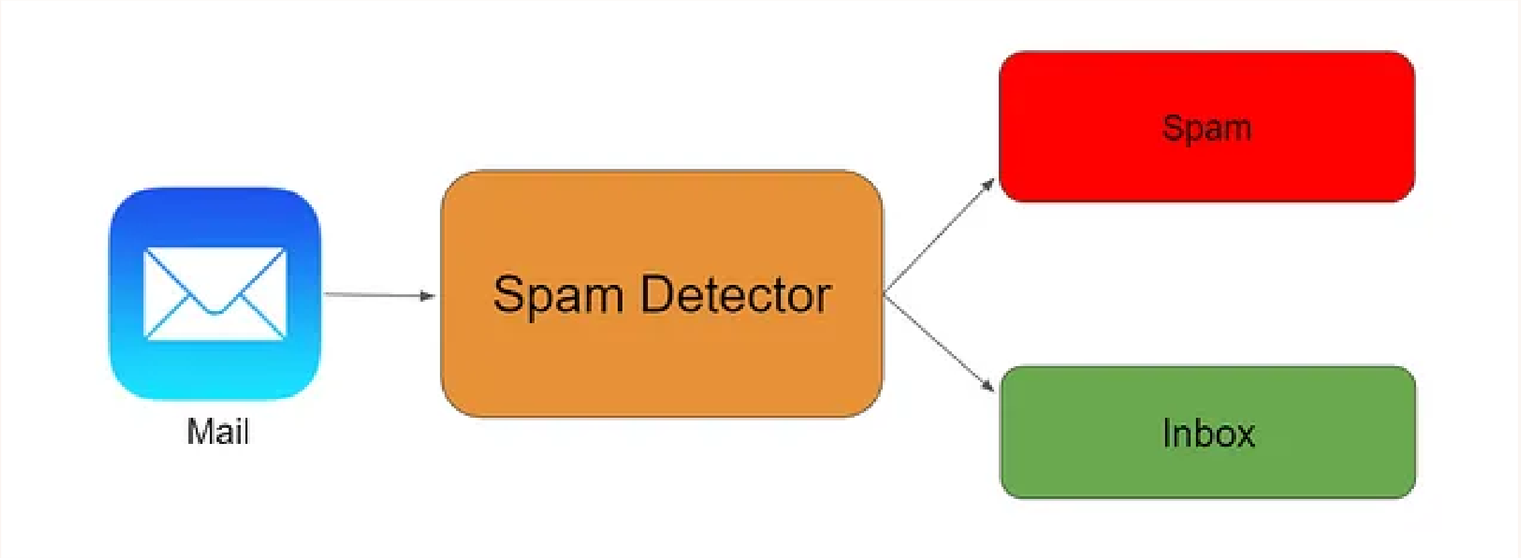
O USO DE APRENDIZADO DE MÁQUINA PARA DETECÇÃO DE SPAM EM EMAILS

por Amanda Reis e Lucas Góes

INTRODUÇÃO

Neste projeto, exploraremos e analisaremos o processo de classificação de e-mails como spam ou não spam. A Detecção de Spam constitui um problema de classificação binária cujo foco reside em identificar e-mails indesejados ou irrelevantes, visando aprimorar a experiência do usuário.



TÓPICOS DE ABORDAGEM

- Data Set utilizado
- Pré-processamento
- SVM
- Naive Bayes
- KNN
- Conclusão

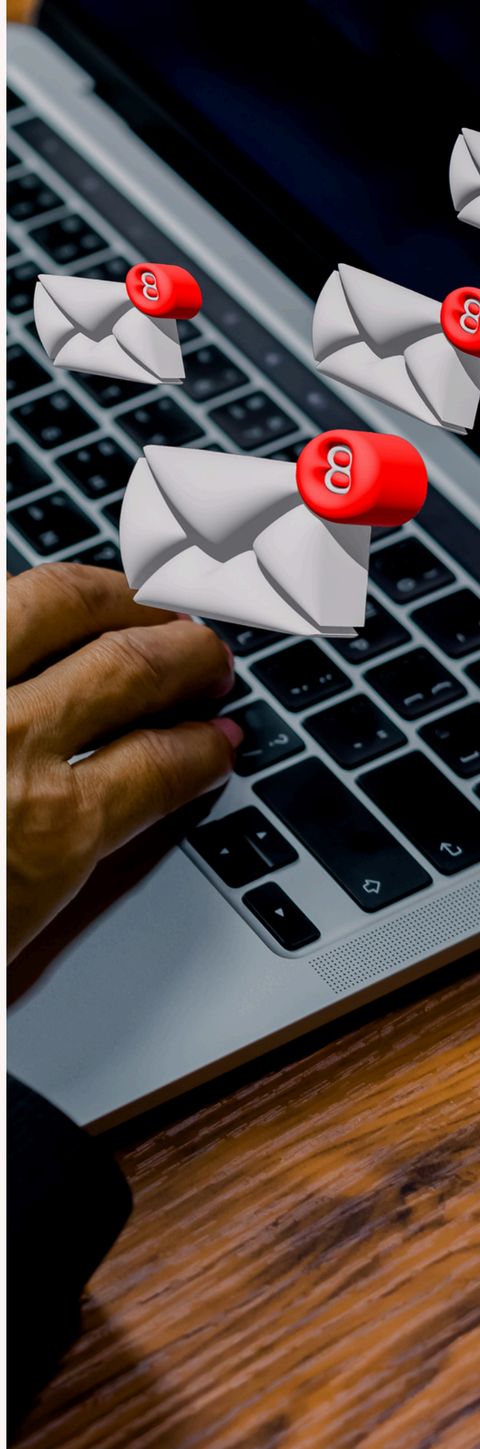
DATA SET UTILIZADO

Com o objetivo de usar o aprendizado de máquina para detecção de spam em emails, faremos o uso do dataset disponível em Ling-Spam Dataset , o qual será dividido em três partes: treinamento, validação e teste.

O conjunto de dados Ling-Spam é uma coleção de 2.893 mensagens de spam e não-spam selecionadas da Linguist List. Essas mensagens focam em interesses linguísticos relacionados a ofertas de emprego, oportunidades de pesquisa e discussões sobre software.

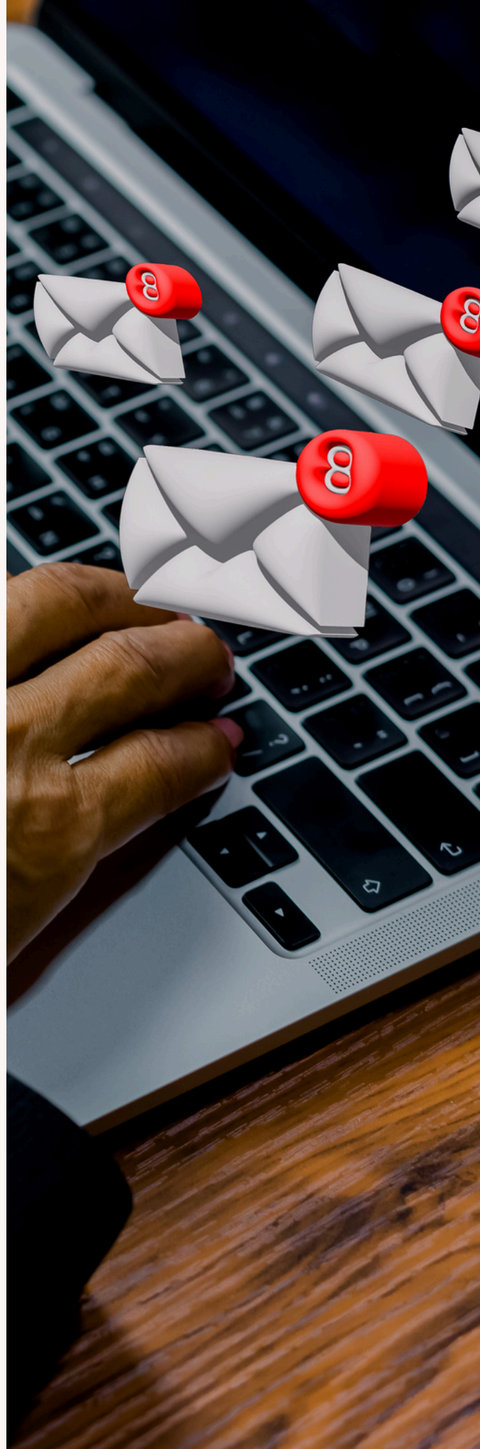
Onde contém:

- 2412 email legítimos
- 481 emails spam



DATA SET UTILIZADO

	subject	message	label
0	job posting - apple-iss research center	content - length : 3386 apple-iss research cen...	0
1	NaN	lang classification grimes , joseph e . and ba...	0
2	query : letter frequencies for text identifica...	i am posting this inquiry for sergei atamas (...	0
3	risk	a colleague and i are researching the differin...	0
4	request book information	earlier this morning i was on the phone with a...	0

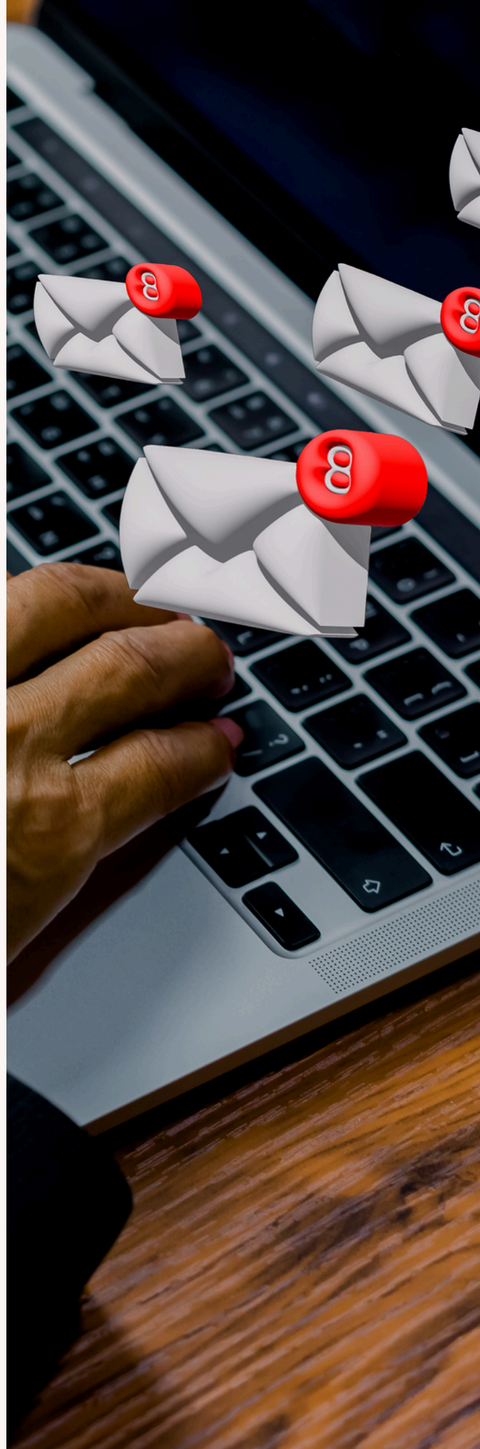


PRÉ-PROCESSAMENTO

Em muitos casos, os algoritmos de aprendizado de máquina ainda não estão devidamente aptos para receber qualquer tipo de dado. No caso deste projeto, será necessário realizar alguns passos de pré-processamento nos textos das colunas

No Processamento de Linguagem Natural (PLN), os textos não seguem sempre a mesma estrutura. Portanto, é necessário padronizá-los ou, ao menos, aproximá-los. Assim, torna-se possível aos algoritmos manipulá-los.

No pré-processamento iremos utilizar técnicas de: remoção de valores nulos, remoção de caracteres especiais, remoção de stop words, lematização, etc



SUBSTITUIÇÃO DO SÍMBOLO '$'

ANTES

```
'win $ 300usd and a cruise ! raquel 's casino , inc . is awarding a cruise + $ 300 to a lucky member . no purchase necessary
to play ! join before month end to participate . ( you will automatically be enteredinto the next drawing . ) you can join at
http : / / www . raquelscasino . com ( casino ) http : / / www . tobet . com ( sportsbook ) raquel 's online casino provides
guests with a state-of - the-art gaming experience : - a chance to win a cruise and $ 300 ! - 10 % sign-up bonus ! - free cas
ino software ! - 25 casino games ! - international sportsbook ! - horse racing ! - play for fun , or for real ! our casino is
secure , audited , private , and insured . thank you for your time and consideration . * * * * * * * * * * * * * * * * * * * *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *\n'
```

DEPOIS

```
'win money 300usd and a cruise ! raquel 's casino , inc . is awarding a cruise + money 300 to a lucky member . no purchase necessary to play ! joi
n before month end to participate . ( you will automatically be enteredinto the next drawing . ) you can join at http : / / www . raquelscasino .
com ( casino ) http : / / www . tobet . com ( sportsbook ) raquel 's online casino provides guests with a state-of - the-art gaming experience : -
a chance to win a cruise and money 300 ! - 10 % sign-up bonus ! - free casino software ! - 25 casino games ! - international sportsbook ! - horse
racing ! - play for fun , or for real ! our casino is secure , audited , private , and insured . thank you for your time and consideration . * * *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *\n'
```

SUBSTITUIÇÃO DOS NÚMEROS DE TELEFONE

ANTES

```
'changes in the journal language the editorial staff and offices of the journal language have been changed as of this m
onth . articles for submission and general correspondence should be sent to the following address : mark aronoff , edit
or language department of linguistics suny stony brook stony brook , ny 11794-4376 , usa book reviews and all correspon
dence concerning reviews should be sent to the following address : edwin battistella , review editor language division
of humanities wayne state college wayne , ne 68787 , usa both offices may be reached by email : main office : language
. eds @ sunysb . edu review office : langrev @ wscgate . wsc . edu the main office may be reached by telephone : phone
: 1-516 - 632-8003 fax : 1-516 - 632-9468\n'
```

DEPOIS

```
'changes in the journal language the editorial staff and offices of the journal language have been chang
ed as of this month . articles for submission and general correspondence should be sent to the following
address : mark aronoff , editor language department of linguistics suny stony brook stony brook , ny pho
nenumber , usa book reviews and all correspondence concerning reviews should be sent to the following ad
dress : edwin battistella , review editor language division of humanities wayne state college wayne , ne
phonenumber , usa both offices may be reached by email : main office : language . eds @ sunysb . edu rev
iew office : langrev @ wscgate . wsc . edu the main office may be reached by telephone : phone : phonenu
mber fax : phonenumber\n'
```

SUBSTITUIÇÃO DOS NUMEROS POR 'NUMBER'

ANTES

```
'win $ 300usd and a cruise ! raquel 's casino , inc . is awarding a cruise + $ 300 to a lucky member .
no purchase necessary to play ! join before month end to participate . ( you will automatically be ent
eredinto the next drawing . ) you can join at http : / / www . raquelscasino . com ( casino ) http : /
/ www . tobet . com ( sportsbook ) raquel 's online casino provides guests with a state-of - the-art g
aming experience : - a chance to win a cruise and $ 300 ! - 10 % sign-up bonus ! - free casino softwar
e ! - 25 casino games ! - international sportsbook ! - horse racing ! - play for fun , or for real ! o
ur casino is secure , audited , private , and insured . thank you for your time and consideration . *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* * *\n'
```

DEPOIS

```
'win money number usd and a cruise ! raquel 's casino , inc . is awarding a cruise + money number  to a
lucky member . no purchase necessary to play ! join before month end to participate . ( you will automa
tically be enteredinto the next drawing . ) you can join at http : / / www . raquelscasino . com ( casi
no ) http : / / www . tobet . com ( sportsbook ) raquel 's online casino provides guests with a state-o
f - the-art gaming experience : - a chance to win a cruise and money number  ! - number  % sign-up bonu
s ! - free casino software ! - number  casino games ! - international sportsbook ! - horse racing ! - p
lay for fun , or for real ! our casino is secure , audited , private , and insured . thank you for your
time and consideration . * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* * * * * * * * * * * * * * *\n'
```

REMOÇÃO DOS CARACTERES ESPECIAIS

ANTES

```
'sociolinguistics lang classification grimes , joseph e . and barbara f . grimes ; ethnologue language family index ; pb . isbn : 0-88312 - 708 - 3 ; vi , 116 pp . ; $ 14 . 00 . summer institute of linguistics . this companion volume to ethnologue : languages of the world , twelfth edition lists language families of the world with sub-groups shown in a tree arrangement under the broadest classification of language family . the language family index facilitates locating language names in the ethnologue , making the data there more accessible . internet : academic . books @ sil . org languages , reference lang & culture gregerson , marilyn ; ritual , belief , and kinship in sulawesi ; pb . : isbn : 0-88312 - 621 - 4 ; ix , 194 pp . ; $ 25 . 00 . summer institute of linguistics . seven articles discuss five language groups in sulawesi , indonesia ; the primary focus is on cultural matters , with some linguistic content . topics include traditional religion and beliefs , certain ceremonies ...'
```

DEPOIS

```
'sociolinguistics lang classification grimes joseph e and barbara f grimes ethnologue language family index pb isbn phonenumber vi number pp money number number summer institute of linguistics this companion volume to ethnologue languages of the world twelfth edition lists language families of the world with sub groups shown in a tree arrangement under the broadest classification of language family the language family index facilitates locating language names in the ethnologue making the data there more accessible internet academic books @ sil org languages reference lang culture gregerson marilyn ritual belief and kinship in sulawesi pb isbn phonenumber ix number pp money number number summer institute of linguistics seven articles discuss five language groups in sulawesi indonesia the primary focus is on cultural matters with some linguistic content topics include traditional religion and beliefs certain ceremonies and kinship internet academic books @ sil org language and society ind...'
```

REMOÇÃO DE STOP WORDS

Stop words são palavras que ocorrem comumente em um idioma, como "o", "de" e assim por diante. Na maioria das vezes, elas podem ser removidas do texto porque não fornecem informações valiosas.

ANTES

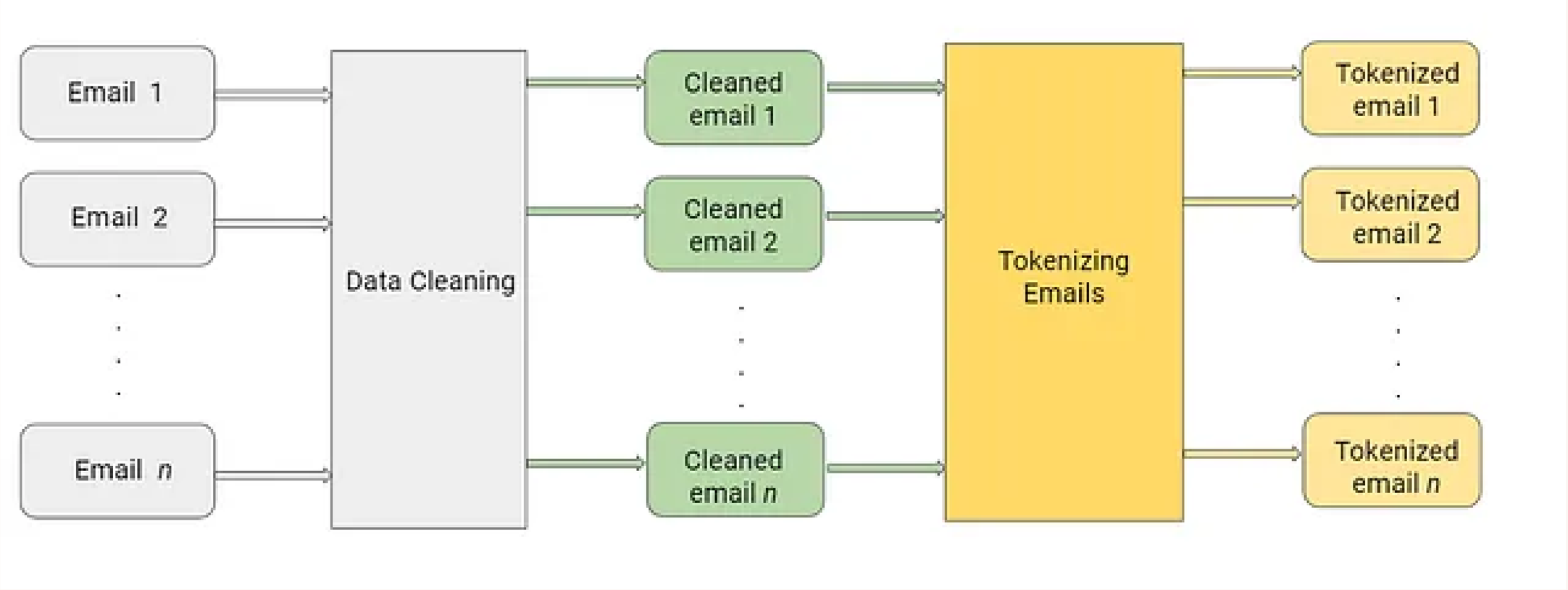
'sociolinguistics lang classification grimes joseph e and barbara f grimes ethnologue language family index pb isbn phonenumber vi number pp money number number summer institute of linguistics this companion volume to ethnologue languages of the world twelfth edition lists language families of the world with sub groups shown in a tree arrangement under the broadest classification of language mily the language family index facilitates locating language names in the ethnologue making the data there more accessible internet academic books @ sil org languages reference lang culture gregerson marilyn ritual belief and kinship in sulawesi pb isbn phonenumber ix number pp money number number summer institute of linguistics seven articles discuss five language groups in sulawesi indonesia the primary focus is on cultural matters with some linguistic content topics include traditional religion and beliefs certain ceremonies and kinship internet academic books @ sil org language and society indonesia computers ling weber david j stephen r mcconnel diana d weber and beth j bryson primer a tool for developing early reading materials pb isbn phonenumber xvi number pp ms dos software money number number summer institute of linguistics the authors present a computer program and instructions for developing reading materials in languages with little or no background in literacy the book is structured as a how to manual with step by step procedures to establish an appropriate primer sequence and to organize words phrases and sentences that correlate with the sequence it presupposes a thorough knowledge of linguistics internet academic books @ sil org literacy computer <'

DEPOIS

'sociolinguistics lang classification grimes joseph e barbara f grimes ethnologue language family index pb isbn phonenumber vi number pp money number number summer institute linguistics companion volume ethnologue languages world twelfth edition lists language families world sub groups shown tree arrangement broadest classification language family language family index facilitates locating language names ethnologue making data accessible internet academic books @ sil org languages reference lang culture gregerson marilyn ritual belief kinship sulawesi pb isbn phonenumber ix number pp money number number summer institute linguistics seven articles discuss language groups sulawesi indonesia primary focus cultural matters linguistic content topics include traditional religion beliefs certain ceremonies kinship internet academic books @ sil org language society indonesia computers ling weber david j stephen r mcconnel diana weber beth j bryson primer tool developing early reading materials pb isbn phonenumber xvi number pp ms dos software money number number summer institute linguistics authors present computer program instructions developing reading materials languages little background literacy book structured manual step step procedures establish appropriate primer sequence organize words phrases sentences correlate sequence presupposes thorough knowledge linguisti

TOKENIZAÇÃO

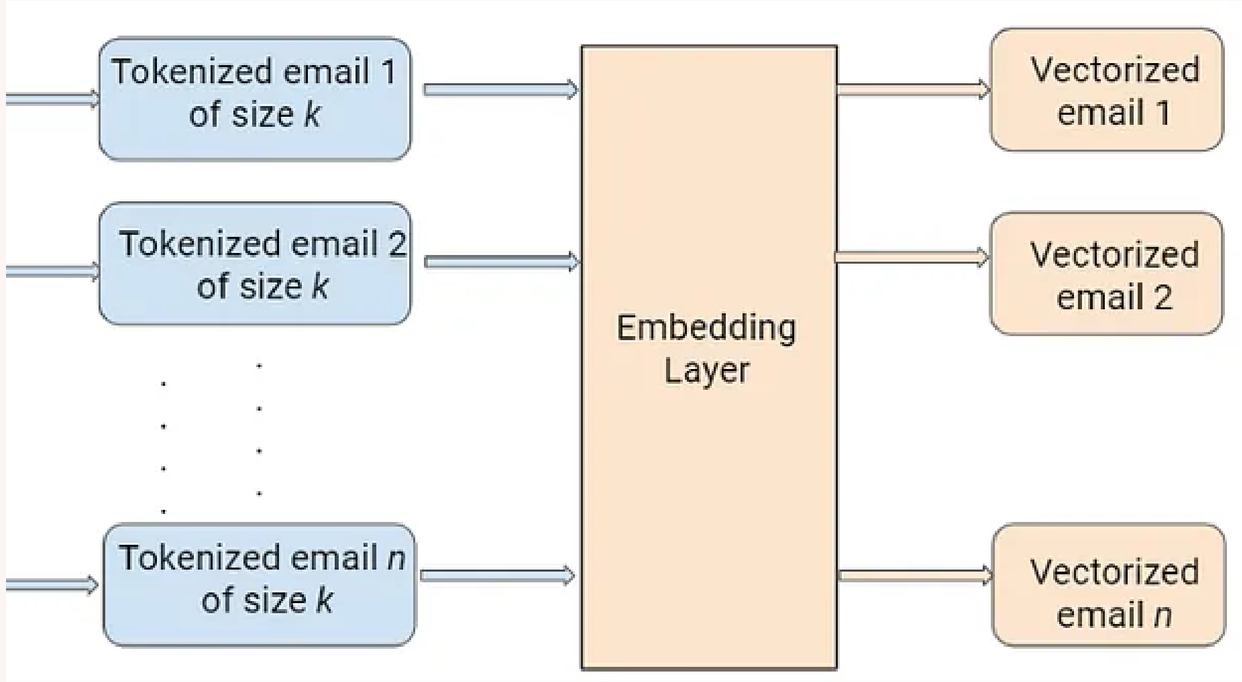
Tokenização é o processo de dividir o texto em partes menores, chamadas tokens. Cada token é uma entrada para o algoritmo de aprendizado de máquina como uma característica.



EMBEDDING

Os dados de texto são facilmente interpretados por humanos, mas complexos para máquinas. Precisamos converter nosso texto em um formato compreensível para as máquinas.

O Embedding é o processo de converter dados de texto formatados em valores/vetores numéricos que uma máquina pode interpretar.



	aa	aaa	aaai	aachen	aalborg	aamt	aan	aarhus	aaron	aart	...	zubizarreta	zum	zur	zurich	zwart	zwei	zweigenbaum	zwicky	zwischen	zygmunt
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2020	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2021	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2022	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2023	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2024	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

025 rows × 12382 columns

SVM

O SVM (Support Vector Machine) ou Máquina de Vetores de Suporte é um algoritmo de aprendizado de máquina supervisionado utilizado para tarefas de classificação e regressão. No contexto de aprendizado de máquina, o SVM funciona buscando um hiperplano que melhor separa as classes de dados.

VANTAGENS

- Efetividade em Alta Dimensionalidade
- Uso de Subconjunto de Pontos de Treinamento
- Flexibilidade com Kernels

DESVANTAGENS

- Complexidade Computacional
- Escolha do Kernel e Parâmetros
- Escalabilidade

NAIVE BAYES

O classificador Naive Bayes é um algoritmo probabilístico de Machine Learning baseado no Teorema de Bayes — uma fórmula matemática usada para calcular probabilidades condicionais. Trata-se de uma ferramenta muito usada em uma ampla variedade de tarefas de classificação no campo da estatística.

VANTAGENS

- Bom desempenho em tarefas de classificação de texto
- Altamente escalável e relativamente simples de implementar
- Escalonar linearmente na complexidade do tempo com o número de recursos

DESVANTAGENS

- Ignora Interações Entre Características
- Dificuldade com Dados Numéricos
- Sensibilidade a Dados de Treinamento

KNN

O KNN (K-nearest neighbors) é um algoritmo que pode ser utilizado para problemas de classificação e regressão. Em resumo, dado um elemento sem rótulo conhecido, seu rótulo será definido por uma análise de seus k vizinhos mais próximos.

VANTAGENS

- Simplicidade e Facilidade de Implementação
- Adaptabilidade a Dados com Diferentes Tipos de Características
- Capacidade de Aprendizado Incremental

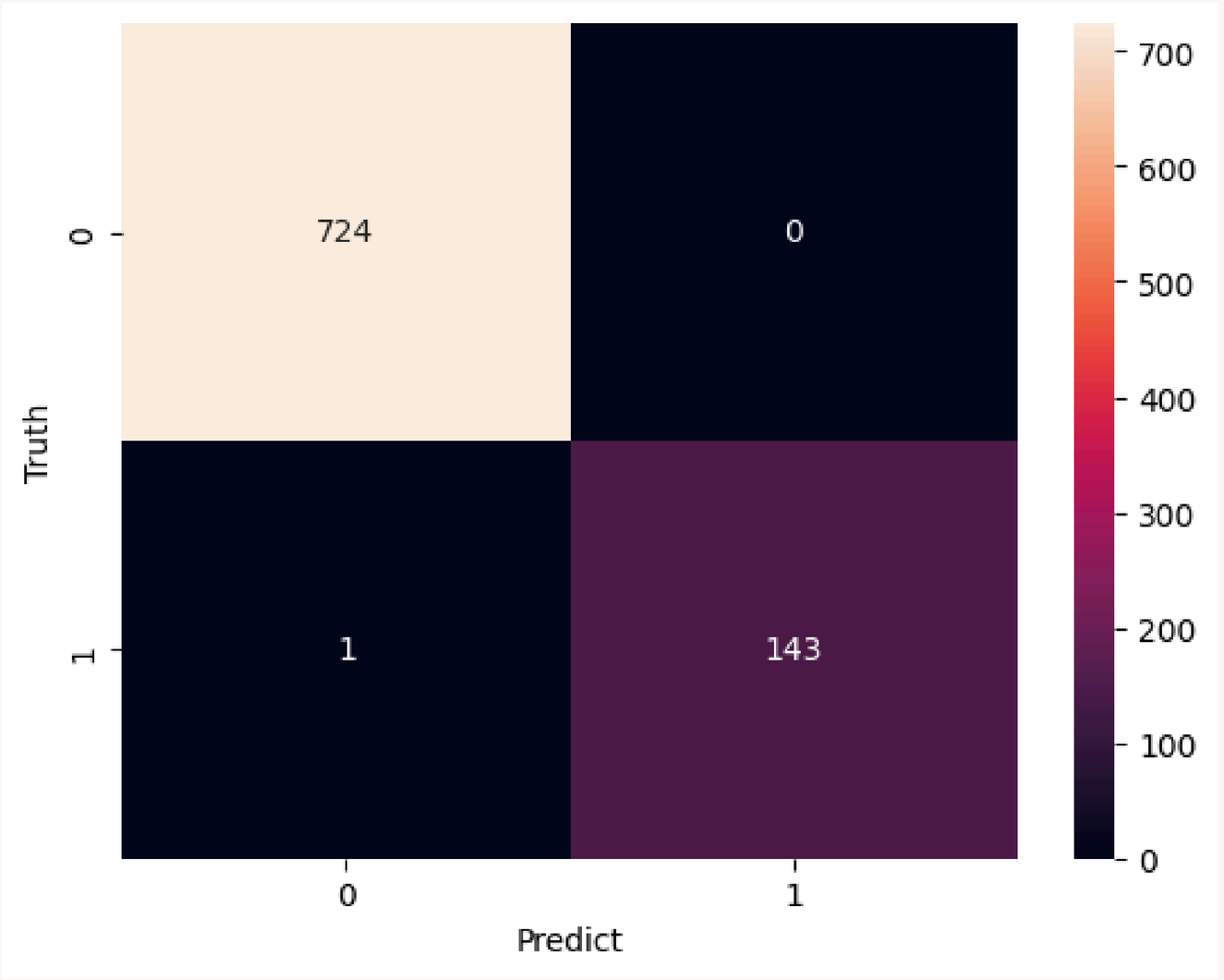
DESVANTAGENS

- Alta Complexidade Computacional
- Sensibilidade à Escala dos Dados
- Influência de Outliers

RESULTADOS

SVM

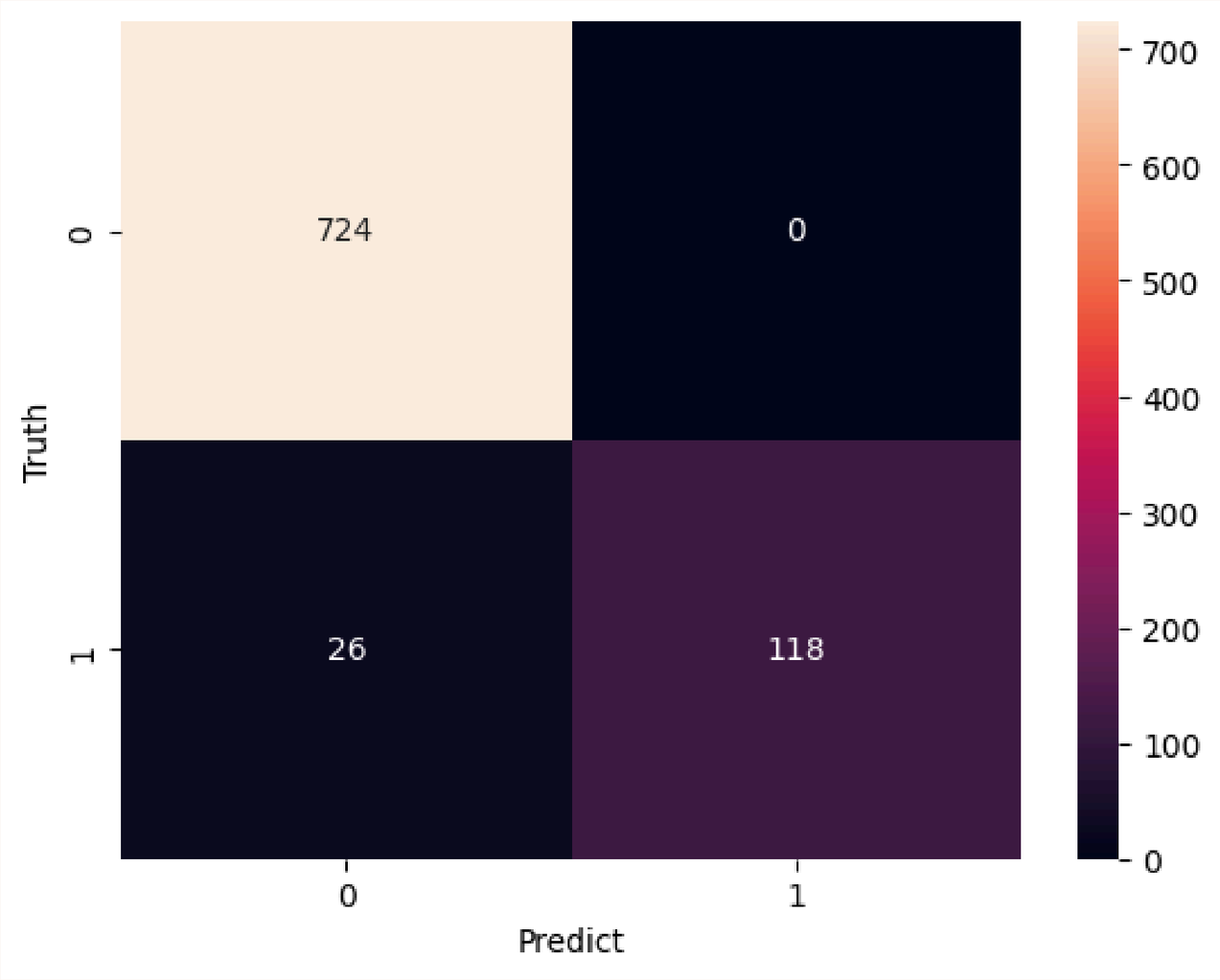
```
Accuracy:  0.9988479262672811
Precision:  0.9988559267793139
Recall:  0.9988479262672811
F1 Score:  0.9988495358234 06
```



RESULTADOS

NAIVE BAYES

```
Accuracy:  0.9700460829493087
Precision:  0.9754544290834614
Recall:  0.9700460829493087
F1 Score:  0.9712681662793208
```



RESULTADOS

KNN

```
Accuracy:  0.9389400921658986
Precision:  0.9421326058790415
Recall:  0.9389400921658986
F1 Score:  0.9355703309292046
```