Redes Neurais Recorrentes para Análise de Sentimentos

Trabalho Final
Princípios de Neurocomputação (PSI 5886)
Prof. Dr. Emilio Del Moral Hernandez

Grupo 9

Integrantes

Gabriela de Souza Melo

Taynan Maier Ferreira

Thomaz Calasans dos Santos

Rodrigo Monteiro de Aquino

Rodrigo Amorim Ruiz

Agenda

- Análise de Sentimentos (Taynan M. Ferreira)
- Word Embedding (Thomaz C. Santos)
- Redes Neurais Recorrentes (Rodrigo A. Ruiz)
- Desafio IMDB (Rodrigo Aquino e Gabriela Souza)

1.

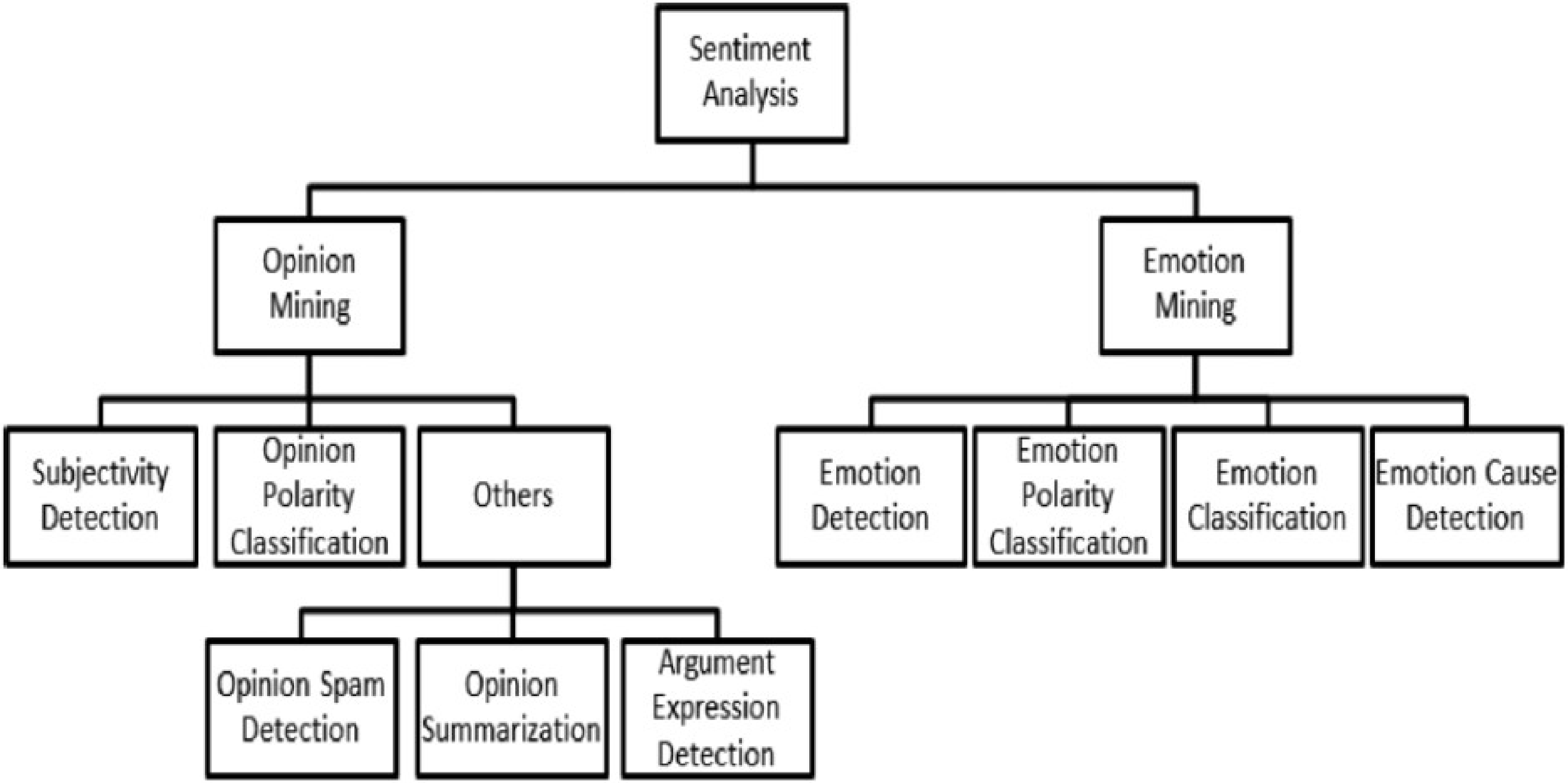
ANÁLISE DE SENTIMENTOS

Análise de Sentimentos

“Tratamento computacional de opiniões, sentimentos e subjetividade”

(Pang & Lee, 2008)

Taxonomia



Taxonomia do campo de Análise de Sentimentos. Retirado de "*Current state of text sentiment analysis from opinion to emotion mining*" (YADOLLAHI, SHAHRAKI, ZAIANE; 2017)

"All the News That's Fit to Print"

The New York Times

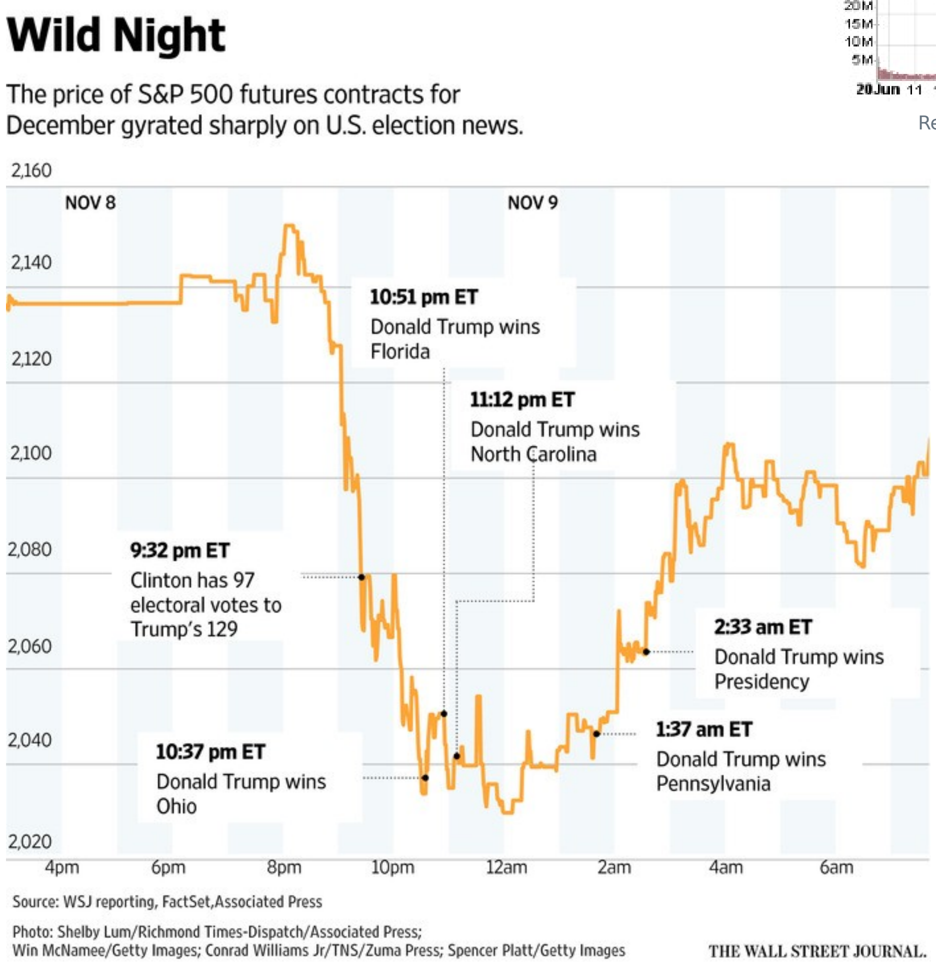
Late Edition

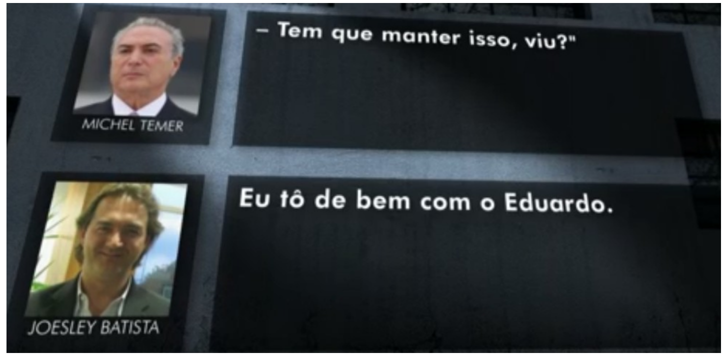
VOL. CLXV ... No. 57,273 — NEW YORK, FRIDAY, JUNE 24, 2016

BRITISH STUN WORLD WITH VOTE TO LEAVE E.U.

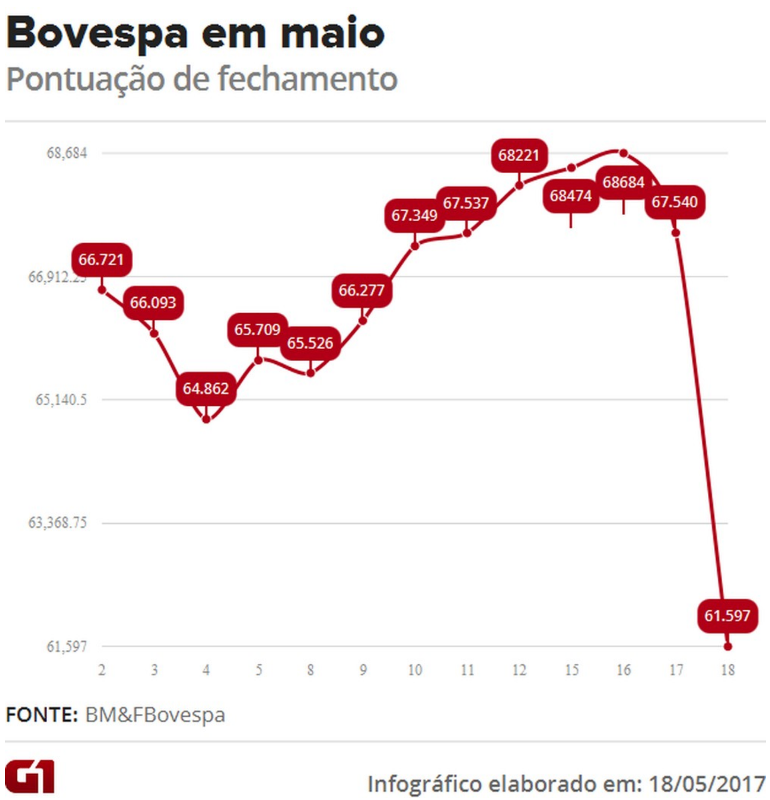
Retirado de https://www.dailymail.co.uk/news/article-3657722/Damn-bad-day-Europe-Shockwaves-spread-continent-politicians-newspapers-react-Britain-quitting-EU.html

Retirado de https://rivemont.ca/wp-content/uploads/2017/04/Lettre-financi%C3%A8re-volume-8-num%C3%A9ro-1.pdf





Retirado de https://g1.globo.com/economia/mercados/noticia/bovespa-fecha-em-forte-queda-de-olho-em-denuncias-sobre-temer.ghtml

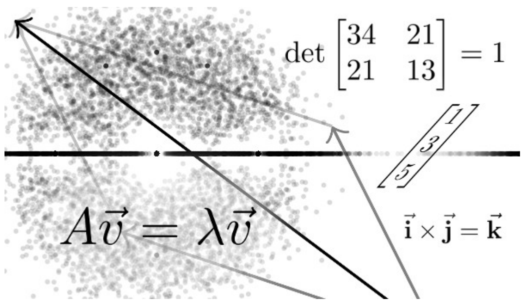


Infográficos retirados de https://g1.globo.com/economia/mercados/noticia/bovespa-fecha-em-forte-queda-de-olho-em-denuncias-sobre-temer.ghtml

2.

WORD EMBEDDING

Retirado de https://yp.scmp.com/over-to-you/poems-and-short-stories/article/94227/bookaholic



Retirado de https://www.yyshtools.com/image/research-engineers.html

Como representar Linguagem Natural matematicamente?

	Ever tried. Ever failed.	No matter. Try again.	Fail again. Fail better.
Ever	2	0	0
tried	1	0	0
failed	1	0	0
No	0	1	0
matter	0	1	0
Try	0	1	0
again	0	1	1
Fail	0	0	2
better	0	0	1

Retirado de (MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to Information Retrieval, 2008)

Bag of Words

Word Embedding

chair (-0.37, -0.23, 0.33, 0.38, -0.02, -0.37)
on (-0.21, -0.11, -0.10, 0.07, 0.37, 0.15)
dog (0.26, 0.25, -0.39, -0.07, 0.13, -0.17)
...
...
the (-0.43, -0.37, -0.12, 0.13, -0.11, 0.34)
...
...
mouth (-0.32, 0.43, -0.14, 0.50, -0.13, -0.42)
...
...
gone (0.06, -0.21, -0.38, -0.28, -0.16, -0.44)
...

Retirado de (GOLDBERG, Y. A Primer on Neural Network Models for Natural Language Processing, 2016)

GloVe

Global Vector for Word Representation

(Pennington, Socher, Manning; 2014)

```
428  economy -0.12027 -0.72505 0.87014 -0.63944 0.17259 -0.35168 -0.65425 -0.72757 -0.22327 0.132 0.4221 -0.21129 -0.3114 0.47728
     0.31158 0.64071 0.22868 -0.1858 0.80219 0.0069265 0.36053 -0.74774 -0.89363 -0.66631 1.0789 -1.3036 0.0028634 0.36411 0.52839
     1.748 3.713 0.70001 0.92982 -0.17352 -0.83904 -0.42105 -1.4294 0.57824 0.58892 -0.85238 -1.7313 0.045091 0.49483 -1.0151
     0.089959 0.4609 0.0017585 0.62182 1.1893 0.08441
429  press -0.47559 0.59191 -0.0036768 0.59487 -0.203 -0.78632 -0.92919 -0.17557 0.0079159 -0.88082 -0.61542 -0.0094559 -0.62884
     -0.17575 0.78539 0.12645 -0.45861 -0.84154 0.1882 -0.072123 0.93554 0.67691 0.91796 0.22855 0.23493 -1.657 -0.377 0.2731
     -0.687 -0.014572 2.6045 -0.30978 -0.31292 -0.88876 -1.0727 -0.13627 -0.11805 -0.22841 -0.028016 0.40617 0.86576 0.85138
     -0.18297 -0.7882 0.4903 0.1963 -0.567 0.85609 0.31344 -0.077916
430  agency 0.39196 -0.2909 0.63739 0.5607 -0.27845 -0.68137 -0.68517 -0.1992 1.3082 -1.1706 0.77587 0.062061 -0.26642 0.5516
     0.6996 0.24339 -0.46577 0.37241 0.44682 0.71815 0.44093 0.42931 0.32038 -0.2463 -0.29515 -2.1698 -0.17619 0.33691 -0.8596
     0.14617 3.0754 -0.44081 -0.11546 -0.7774 -1.0064 0.24888 0.21279 -0.43476 0.64488 0.3478 0.12427 0.39314 0.92563 -0.6669
     -0.34586 -0.53523 -0.94889 2.1969 0.22165 0.49233
431  water 0.53507 0.5761 -0.054351 -0.208 -0.7882 -0.17592 -0.21255 -0.14388 1.0344 -0.079253 0.27696 0.37951 1.2139 -0.34032
     -0.18118 0.72968 0.89373 0.82912 -0.88932 -1.4071 0.55571 -0.017453 1.2524 -0.57916 0.43 -0.77935 0.4977 1.2746 1.0448
     0.36433 3.7921 0.083653 -0.45044 -0.063996 -0.19866 0.75252 -0.27811 0.42783 1.4755 0.37735 0.079519 0.024462 0.5013 0.33565
     0.051406 0.39879 -0.35603 -0.78654 0.61563 -0.95478
```

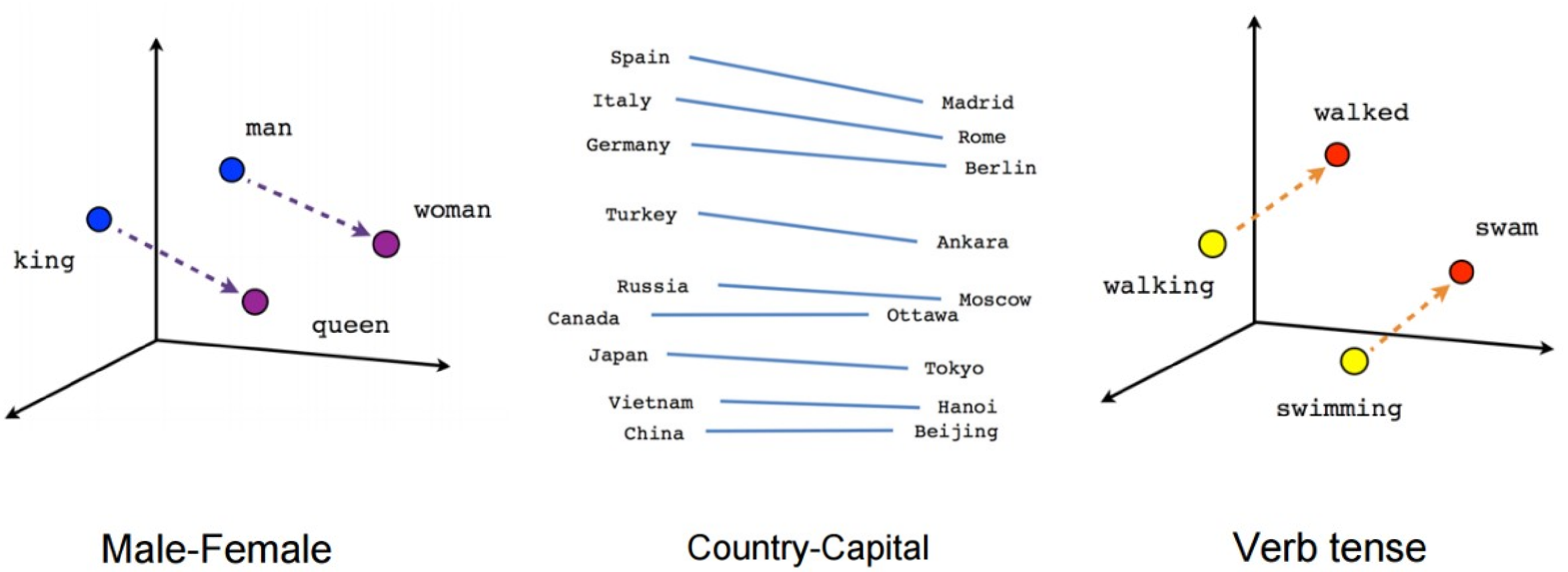
Figura elaborada pelos autores

King – Man + Woman = ?

GloVe

Global Vector for Word Representation

(Pennington, Socher, Manning; 2014)



Retirado de https://www.tensorflow.org/tutorials/representation/word2vec

King – Man + Woman = Queen

Madrid – Spain + Italy = Rome

Walked – Walking + Swimming = Swam

3.

REDES NEURAIS RECORRENTES

Por que não uma MLP?

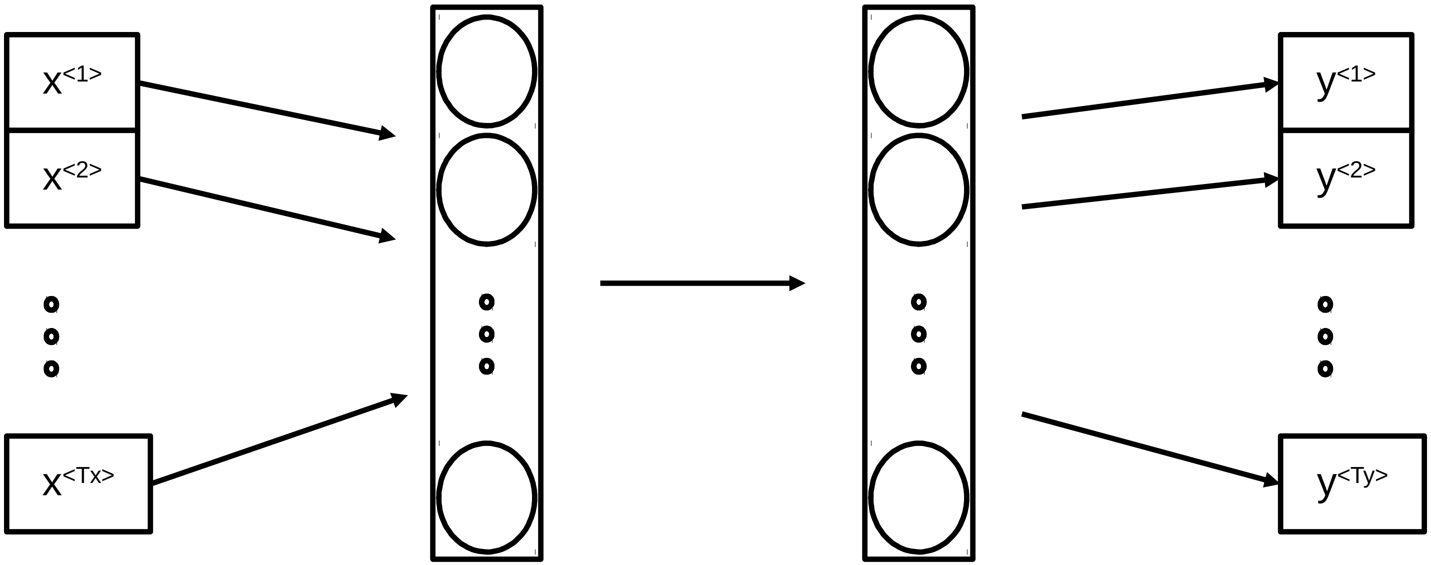


Imagem dos autores

- Entrada e saída têm comprimento diferente em cada exemplo
- Não há correlação para posições diferentes (como a CNN)

Redes Neurais Recorrentes

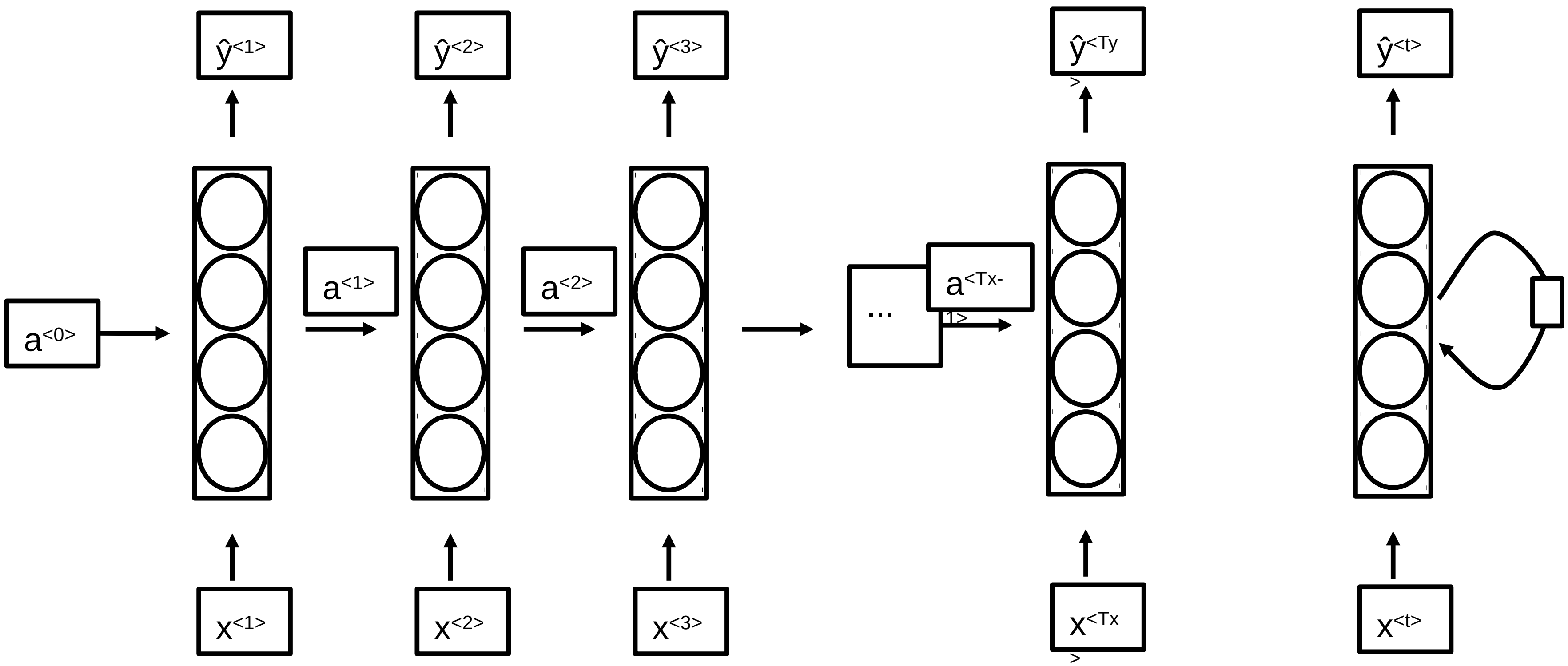


Imagem dos autores

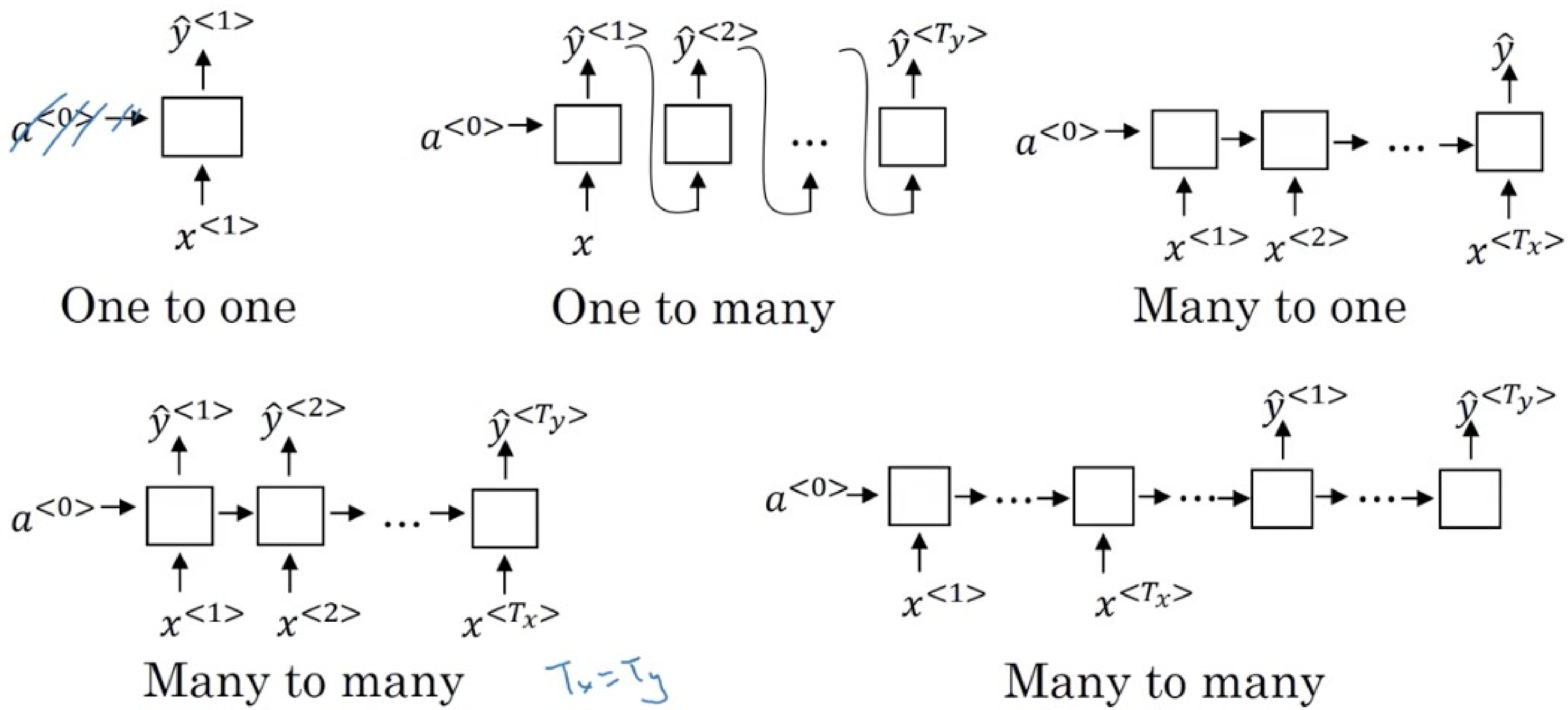
- Considera apenas palavras anteriores

Redes Neurais Recorrentes

- $a^{<t>} = g(W_{aa}\, a^{<t-1>} + W_{ax}\, x^{<t>} + b_a)$
- $\hat{y}^{<t>} = g(W_{ya}\, a^{<t>} + b_y)$

- $L^{<t>}(\hat{y}^{<t>}, y^{<t>}) = - y^{<t>} \log \hat{y}^{<t>} - (1 - y^{<t>}) \log (1 - \hat{y}^{<t>})$
- $L(\hat{y}, y) = \sum L^{<t>}(\hat{y}^{<t>}, y^{<t>})$

Diferentes tipos de RNNs



Retirada do curso Deep Learning Specialization

Long Short-Term Memory (LSTM)

GRU

$\tilde{c}^{<t>} = \tanh(W_c\,[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$

$\Gamma_u = \sigma(W_u\,[c^{<t-1>}, x^{<t>}] + b_u)$

$\Gamma_r = \sigma(w_r\,[c^{<t-1>}, x^{<t>}] + b_r)$

$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$

$a^{<t>} = c^{<t>}$

LSTM

$\tilde{c}^{<t>} = \tanh(W_c\,[a^{<t-1>}, x^{<t>}] + b_c)$

$\Gamma_u = \sigma(W_u\,[a^{<t-1>}, x^{<t>}] + b_u)$

$\Gamma_f = \sigma(W_f\,[a^{<t-1>}, x^{<t>}] + b_f)$

$\Gamma_o = \sigma(W_o\,[a^{<t-1>}, x^{<t>}] + b_o)$

$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$

$a^{<t>} = \Gamma_o * \tanh(c^{<t>})$

Long Short-Term Memory (LSTM)

$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$

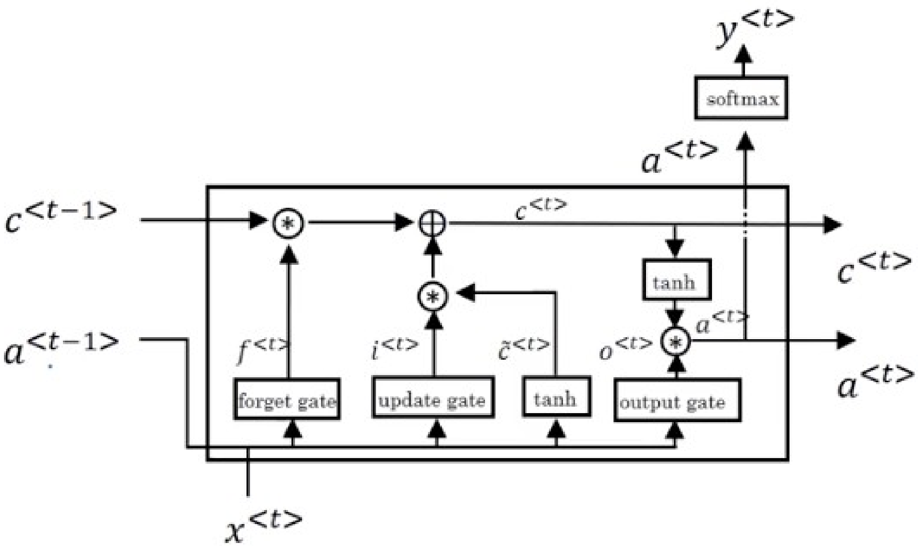
$$\Gamma_u = \sigma(W_u[\, a^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_f = \sigma(W_f[\, a^{<t-1>}, x^{<t>}] + b_f)$$

$$\Gamma_o = \sigma(W_o[\, a^{<t-1>}, x^{<t>}] + b_o)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

$$a^{<t>} = \Gamma_o * \tanh c^{<t>}$$



Retirada do curso Deep Learning Specialization

4.

DESAFIO IMDB

Problema

Retirada de https://www.amazon.jobs/es/teams/imdb

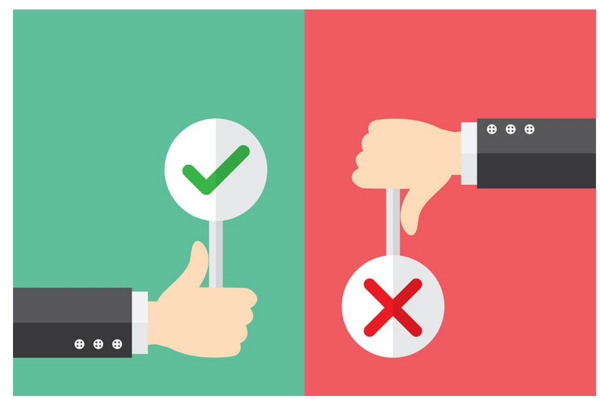
This past weekend, I saw the movie *Harry Potter and the Goblet of Fire*. As part of the popular *Harry Potter* book and movie series, it was definitely one to remember.

This movie is mainly about a large-scale competition between wizards that takes place at Hogwarts over the course of the school year. Harry has been entered into the contest, which is known as The Triwizard Tournament, and many twists and turns make for quite the adventurous competition. True to *Harry Potter* form, the characters encounter problems, tragedies, challenges, and mysterious behavior that they are forced to decode.

New characters were introduced in this film, which was an enjoyable aspect and added a new element of intrigue that the previous *Harry Potter* films did not possess. Furthermore, this installment of the series seemed to follow its literary counterpart the most closely and accurately. Also, this film was a good length: not excessively long but long enough to tell the story. Another great aspect of the film was the special effects, particularly with regard to all of the scenes featuring the dragons.

This movie was also different from the rest because it had more story concerning the characters' love lives. In one particular scene, there was a dance that all of the tournament contestants were to attend. The characters went to the ball, and it was so much fun to watch. First of all, the scenery of the ball was exquisite, and it was so cool to actually feel like you were there! This was definitely one of the highlights of the movie.

Overall, this was, in my opinion, the best *Harry Potter* film yet and portrays the fantastical story of wizards and the world of Hogwarts extremely well. With the addition of new characters, the fast-paced nature of the film, and the added romantic aspects, this movie was truly a great one.



Retirada de http://acasignups.net/18/07/15/risk-adjustment-transfer-payment-freeze-how-much-heartburn-your-insurance-carrier

Problema

- Descrição do dataset
  - Dataset aberto, pode ser importado diretamente com a biblioteca Keras [1]
  - 50.000 resenhas de filmes, categorizadas em negativo / positivo
  - Cada filme possui no máximo 30 resenhas
  - Reviews pré-processados (indexados pela sua frequência no dataset)
  - Dataset balanceado (50% de dados de cada classe)

- Treinamento e Testes
  - Divisão do dataset em 80% treino; 10% validação; 10% teste

[1] https://keras.io/datasets/#imdb-movie-reviews-sentiment-classification

Problema

- Exemplo de dado no dataset:

---review---
[1, 194, 1153, 194, 2, 78, 228, 5, 6, 1463, 4369, 2, 134, 26, 4, 715, 8, 118, 1634, 14, 394, 20, 13, 119, 954, 189, 102, 5, 207, 110, 3103, 21, 14, 69, 188, 8, 30, 23, 7, 4, 249, 126, 93, 4, 114, 9, 2300, 1523, 5, 647, 4, 116, 9, 35, 2, 4, 229, 9, 340, 1322, 4, 118, 9, 4, 130, 4901, 19, 4, 1002, 5, 89, 29, 952, 46, 37, 4, 455, 9, 45, 43, 38, 1543, 1905, 398, 4, 1649, 26, 2, 5, 163, 11, 3215, 2, 4, 1153, 9, 194, 775, 7, 2, 2, 349, 2637, 148, 605, 2, 2, 15, 123, 125, 68, 2, 2, 15, 349, 165, 4362, 98, 5, 4, 228, 9, 43, 2, 1157, 15, 299, 120, 5, 120, 174, 11, 220, 175, 136, 50, 9, 4373, 228, 2, 5, 2, 656, 245, 2350, 5, 4, 2, 131, 152, 491, 18, 2, 32, 2, 1212, 14, 9, 6, 371, 78, 22, 625, 64, 1382, 9, 8, 168, 145, 23, 4, 1690, 15, 16, 4, 1355, 5, 28, 6, 52, 154, 462, 33, 89, 78, 285, 16, 145, 95]
---label---
0
---review with words---

big hair big   bad music and a giant safety   these are the words to best describe this terrible movie i love cheesy horror movies and i've seen hundreds but this had got to be on of the worst ever made the plot is paper thin and ridiculous the acting is an   the script is completely laughable the best is the end showdown with the cop and how he worked out who the killer is it's just so damn terribly written the clothes are   …….

Tecnologia

Retirado de https://www.python.org/

Retirado de https://keras.io/

Solução - Teste Inicial

- Tamanho do vocabulário: 5000
  - Palavras desconhecidas são todas marcadas com um mesmo número
- Padding em 80 palavras
  - Possibilita treinamento em batches → diminui tempo de treinamento e testes
- Embeddings de tamanho 32
- Arquitetura utilizada:
  - Uma camada do tipo Embedding, que espera vetores de entradas com tamanho igual ao tamanho do vocabulário (5000) e cuja saída é do tamanho do embedding (32)
  - Uma camada do tipo LSTM, com 100 unidades
  - Uma camada densa de saída, com 1 unidade, e função de ativação do tipo sigmóide

Solução - Teste Inicial

- Função de perda: binary cross-entropy
- Algoritmo de otimização: Adam
- Métrica: acurácia
- Treinamento por 3 épocas com batch size = 64
- Acurácia de treino: 91,96 %
- Acurácia de validação: 83,77 %

Solução - Comparação de Hiperparâmetros

- Variação nos seguintes parâmetros:
  - Tamanho do vocabulário: 2500, 5000, 10000, 20000
  - Tamanho dos vetores de embeddings: 1, 2, 4, 16, 32, 128, 256
  - Tamanho máximo das reviews: 64, 128, 256, 512, 1024
  - Quantidades de neurônios na camada LSTM: 1, 2, 4, 8, 16, 32, 64, 128

Resultados - Comparação de Hiperparâmetros

tam. vocab	tam. max. review	tam. embedding	neuronios camada LSTM	acuracia validacao (%)
5000	512	32	128	89.2
5000	512	32	128	89.2
5000	256	32	128	88.62
5000	1024	32	128	88.74
20000	512	32	128	88.46
5000	128	128	128	87.7
5000	128	256	128	87.38
2500	128	32	128	86.98
5000	128	32	16	86.94
5000	128	32	32	86.72
5000	128	32	128	86.6
10000	128	32	128	86.62
5000	128	32	8	86.52
20000	128	32	128	86.36
5000	128	16	8	86.46
5000	128	2	2	86.08
5000	128	4	4	86.06
5000	128	32	64	85.98
5000	128	32	2	85.8
5000	128	1	1	84.7
5000	64	32	128	82.86

Figura elaborada pelos autores

Solução Final

Acurácia de teste:

```
Parameters
- vocabulary_size: 5000
- sequence_maximum_length: 128
- embedding_size: 32
- lstm_neurons: [32]

Accuracies
- Training:   0.94815
- Validation: 0.8672
```

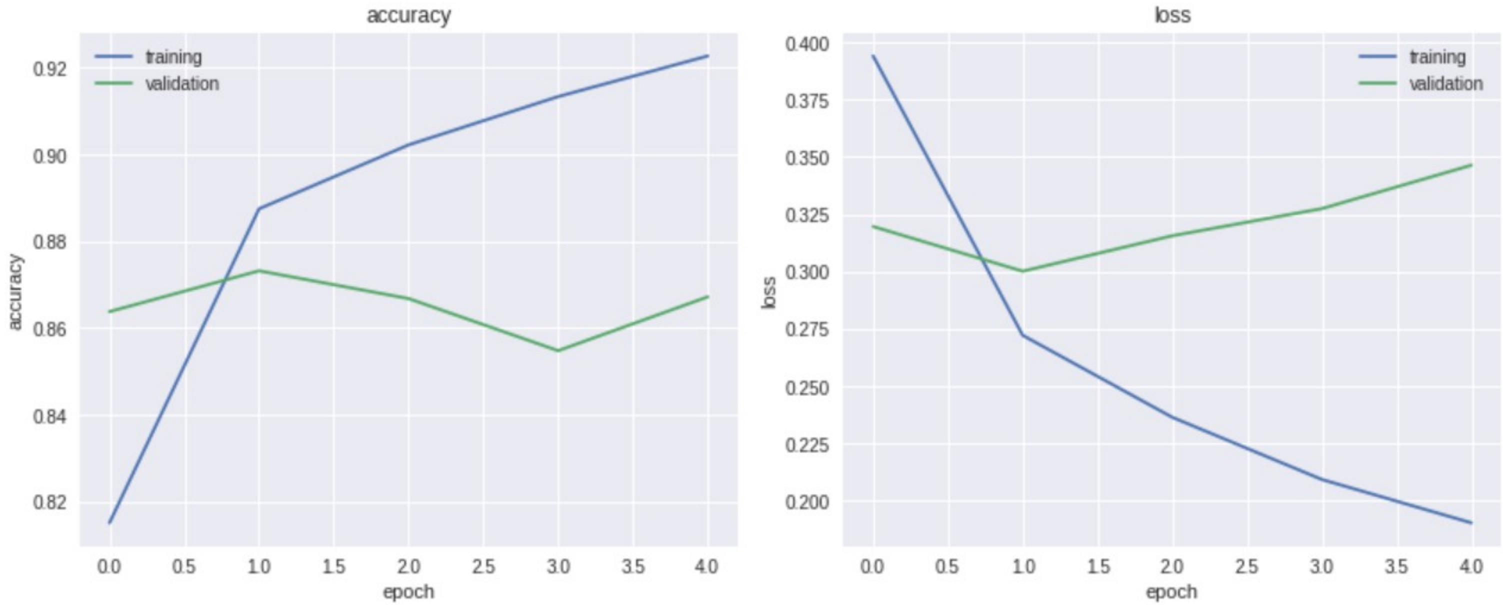


Figura elaborada pelos autores

Exemplo de acerto

Review: please give this one a miss br br and the rest of the cast terrible performances the show is flat flat flat br br i don't know how michael could have allowed this one on his he almost seemed to know this wasn't going to work out and his performance was quite so all you fans give this a miss

Label esperada: 0

Saída do modelo: 0.01855629

Exemplo de erro

Review: this film requires a lot of because it focuses on mood and character development the plot is very simple and many of the scenes take place on the same set in the dennis character apartment but the film builds to a disturbing climax br br the characters create an atmosphere with sexual tension and psychological it's very interesting that robert directed this considering the style and structure of his other films still the ...

Label esperada: 0

Saída do modelo: 0.84447217

Trabalhos Futuros

- Algumas outras coisas poderiam ser testadas:
  - Utilizar embeddings pré-treinados
  - Arquiteturas diferentes (quantidade e tipos de camadas, quantidade de unidades por camadas)
  - Utilização de técnicas de regularização, para evitar sobrejuste, como Dropout
  - Variação de outros hiperparâmetros
    - Quantidade de épocas
    - Tamanho dos lotes (batches)
    - Função de otimização
    - Taxa de aprendizagem (learning rate)

8.
REFERÊNCIAS

Referências

1. YADOLLAHI, A.; SHAHRAKI, A.G; ZAIANE, O. R. 2017. Current state of text sentiment analysis from opinion to emotion mining. ACM Comput. Surv. 50, 2, Article 25, 2017.

2. CAMBRIA, E. et al. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, v. 28, n. 2, p. 15–21, 2013.

3. PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr., v. 2, p. 1–135, 2008.

4. https://keras.io/datasets/#imdb-movie-reviews-sentiment-classification. Website consultado em 03/12/2018.

Obrigado!
Dúvidas?

Gabriela de Souza Melo
Rodrigo Monteiro de Aquino
Rodrigo Amorim Ruiz
Taynan Maier Ferreira
Thomaz Calasans dos Santos