



**MCZA017-13**  
**Processamento de Linguagem Natural**

**NLTK através de exemplos:**

- Análise de sentimentos**
- Similaridade entre ementas UFABC**

Prof. Jesús P. Mena-Chalco  
jesus.mena@ufabc.edu.br

2Q-2019

# Algumas avaliações de um livro

Trata-se de um livro da mais alta excelência: bem escrito, objetivo, em ótima tradução e, principalmente, com um personagem principal capaz de segurar a nossa atenção e prender o fôlego durante toda a leitura. **[Rev 1]**

Li o livro em dois dias. Livro muito bom e que conta a história de um empreendedor que não tem medo do "não". Confesso que após ler o livro, me sinto ainda mais inspirado a trabalhar mais e conquistar meus objetivos.

Logo no início do livro, há uma definição clara: "Ele é do tipo, faça ou morra, mas não desista." **[Rev 2]**

Otimo livro... bem embasado..leitura facil...fatos e narrativas aparentemente imparciais. Recomendo a leitura. **[Rev 3]**



## Elon Musk: Como o CEO bilionário da SpaceX e da Tesla está moldando nosso futuro

**Capa comum:** 416 páginas

**Editora:** Intrínseca; Edição: 1ª (26 de setembro de 2015)

**Idioma:** Português

**ISBN-10:** 8580578280

**ISBN-13:** 978-8580578287

**Dimensões do produto:** 22,9 x 15,5 x 2,3 cm

**Peso de envio:** 585 g

**Avaliação média:** ★★★★★ ∨ 315 avaliações de clientes

[https://www.amazon.com.br/Elon-Musk-Ashlee-Vance/dp/8580578280/ref=pd\\_sbs\\_14\\_3/130-2356547-5609000?encoding=UTF8&pd\\_rd\\_i=8580578280&pd\\_rd\\_r=046dbeeb-a714-4156-90c7-e29fd22a20b6&pd\\_rd\\_w=vtstc&pd\\_rd\\_wg=25dAz&pf\\_rd\\_p=80c6065d-57d3-41bf-b15e-ee01dd80424f&pf\\_rd\\_r=94SD2J87VCZD2D36SSWA&pvc=1&refRID=94SD2J87VCZD2D36SSWA](https://www.amazon.com.br/Elon-Musk-Ashlee-Vance/dp/8580578280/ref=pd_sbs_14_3/130-2356547-5609000?encoding=UTF8&pd_rd_i=8580578280&pd_rd_r=046dbeeb-a714-4156-90c7-e29fd22a20b6&pd_rd_w=vtstc&pd_rd_wg=25dAz&pf_rd_p=80c6065d-57d3-41bf-b15e-ee01dd80424f&pf_rd_r=94SD2J87VCZD2D36SSWA&pvc=1&refRID=94SD2J87VCZD2D36SSWA)

# Algumas avaliações de um livro

Ao parecer as 3 avaliações são positivas.

Provavelmente, baseados nessas avaliações, poderíamos decidir comprar e ler livro.

**Mas, o que não leva a determinar que as avaliações sejam positivas, ou negativas ou neutras?**

# Algumas avaliações de um livro

Trata-se de um livro da mais **alta excelência**: bem escrito, objetivo, em **ótima tradução** e, principalmente, com um personagem principal capaz de segurar a nossa atenção e prender o fôlego durante toda a leitura. **[Rev 1]**

Li o livro em dois dias. Livro **muito bom** e que conta a história de um empreendedor que não tem medo do "não". Confesso que após ler o livro, me sinto ainda **mais inspirado** a trabalhar mais e **conquistar** meus objetivos.

Logo no início do livro, há uma definição clara: "Ele é do tipo, faça ou morra, mas **não desista.**" **[Rev 2]**

**Otimo livro**... bem embasado.. **leitura fácil**... fatos e narrativas aparentemente imparciais. **Recomendo a leitura.** **[Rev 3]**

# Algumas avaliações de um livro

- O autor do livro é **muito parcial** a favor de Elon.. Considero um livro **muito mal escrito**. Não gostei e não recomendo.
- **Cansativo**. Começo interessante mas depois **fica cansativo**, não flui. Muitos detalhes técnicos dos carros, foguetes etc. **poderia ser mais resumida** essa parte
- O **problema do livro** é se ater de mais aos **fatos mundanos** na vida do biografado, falando sobre as relações conjugais dele. Imagino que para tirar um pouco o brilho de herói messiânico infalível que ele mesmo cria no livro, mas **infelizmente é chato** e eu **não quero saber** quantas vezes ele se casou, inclusive não tira nem um mérito dos trabalhos dele e sim da moral dele, por assim dizer.

# Algumas avaliações de um livro

Fanfarrão, mas um visionário. O livro é bem escrito, a história é bacana, não muito mais do que isso.

Livro muito bem escrito, mas abaixo das minhas expectativas poi me parecia mais uma propaganda do que uma biografia propriamente dita da pessoa Elon Musk.

Biografia chapa branca.. leitura agradável, mas decepcionante.

# Análise de sentimento

- É uma subárea **muito popular de PLN**.
- De fato, os primeiros artigos científicos foram publicados **nos finais dos 90s**.
- Nos últimos anos foram realizadas grandes contribuições para essa subárea de pesquisa.  
Em 2018 foram publicados 8668 artigos (base Dimensions.ia)

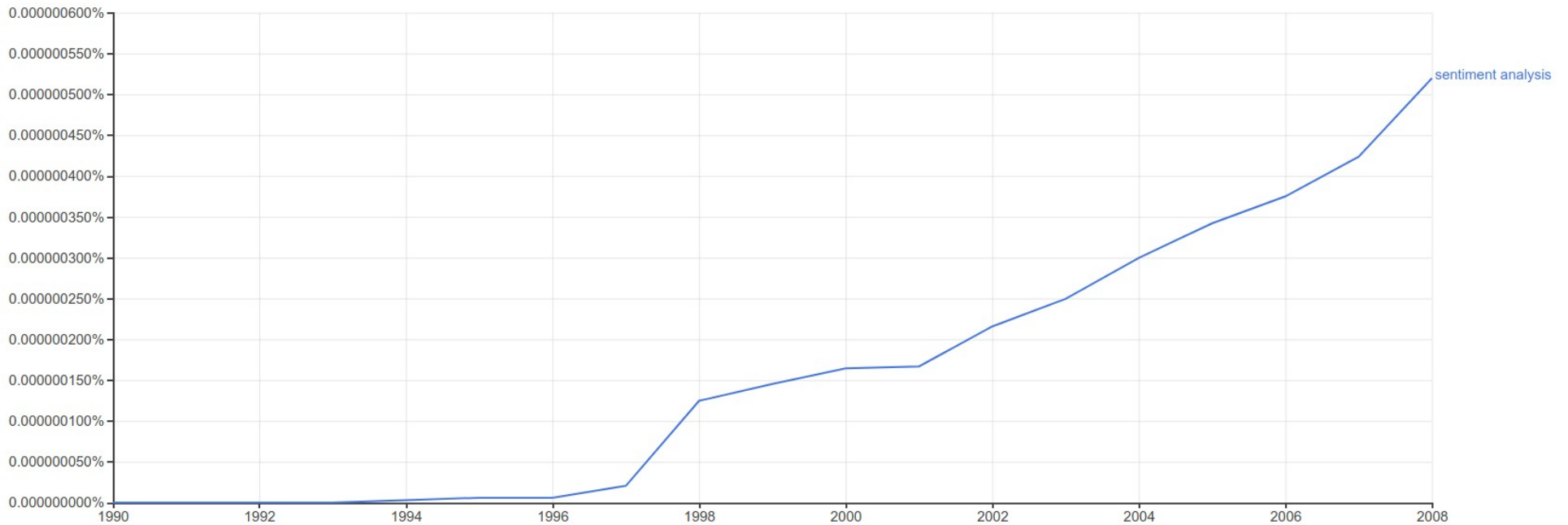
PUBLICATION YEAR							
<input type="radio"/> 2019	4,547	<input type="radio"/> 2010	478	<input type="radio"/> 2001	4	<input type="radio"/> 1980	1
<input type="radio"/> 2018	8,668	<input type="radio"/> 2009	463	<input type="radio"/> 2000	3	<input type="radio"/> 1979	1
<input type="radio"/> 2017	5,532	<input type="radio"/> 2008	170	<input type="radio"/> 1999	1	<input type="radio"/> 1977	1
<input type="radio"/> 2016	4,044	<input type="radio"/> 2007	88	<input type="radio"/> 1998	1	<input type="radio"/> 1974	1
<input type="radio"/> 2015	3,105	<input type="radio"/> 2006	53	<input type="radio"/> 1997	1	<input type="radio"/> 1972	1
<input type="radio"/> 2014	2,527	<input type="radio"/> 2005	33	<input type="radio"/> 1996	2	<input type="radio"/> 1961	1
<input type="radio"/> 2013	1,805	<input type="radio"/> 2004	10	<input type="radio"/> 1990	1		
<input type="radio"/> 2012	1,246	<input type="radio"/> 2003	6	<input type="radio"/> 1989	1		
<input type="radio"/> 2011	750	<input type="radio"/> 2002	4	<input type="radio"/> 1986	1		



# Google Books Ngram Viewer

Graph these comma-separated phrases:   case-insensitive

between  and  from the corpus  with smoothing of



[https://books.google.com/ngrams/graph?content=sentiment+analysis&year\\_start=1990&year\\_end=2008&corpus=15&smoothing=3&share=&direct\\_url=t1%3B%2Csentiment%20analysis%3B%2Cc0](https://books.google.com/ngrams/graph?content=sentiment+analysis&year_start=1990&year_end=2008&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Csentiment%20analysis%3B%2Cc0)

FILTERS FAVORITES

PUBLICATION YEAR

- 2019 4,547
- 2018 8,668
- 2017 5,532
- 2016 4,044
- 2015 3,105
- 2014 2,527
- 2013 1,805
- 2012 1,246
- 2011 750
- 2010 478

[More](#)

- RESEARCHER
- FIELDS OF RESEARCH
- PUBLICATION TYPE
- SOURCE TITLE
- JOURNAL LIST
- OPEN ACCESS

**PUBLICATIONS**  
33,553

GRANTS  
330

PATENTS  
3,613

CLINICAL TRIALS  
3

POLICY DOCUMENTS  
162

Sort by: Relevance

Title, Author(s), Bibliographic reference - About the metrics

**Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm**

Ahmed Al-Saffar, Suryanti Awang, Hai Tao, Nazlia Omar, Wafaa Al-Saiagh, Mohammed Al-bared  
2018, PLoS ONE - Article

Citations 2 Altmetric 1 View PDF Add to Library

**Microblog sentiment analysis using social and topic context**

Xiaomei Zou, Jing Yang, Jianpei Zhang  
2018, PLoS ONE - Article

Citations 4 Altmetric 2 View PDF Add to Library

**Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias**

Hongyu Han, Yongshi Zhang, Jianpei Zhang, Jing Yang, Xiaomei Zou  
2018, PLoS ONE - Article

View PDF Add to Library

**Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach**

María Del Pilar Salas-Zárate, José Medina-Moreira, Katty Lagos-Ortiz, Harry Luna-Aveiga, Miguel Ángel Rodríguez-García, Rafa...  
2017, Computational and Mathematical Methods in Medicine - Article

Citations 18 Altmetric 1 View PDF Add to Library

**Feature engineering for sentiment analysis in e-health forums**

Jorge Carrillo-de-Albornoz, Javier Rodríguez Vidal, Laura Plaza  
2018, PLoS ONE - Article

Altmetric 3 View PDF Add to Library

**Extracting features with medical sentiment lexicon and position encoding for drug reviews**

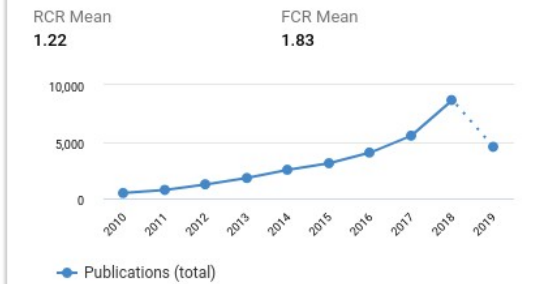
Sisi Liu, Ickjai Lee

ANALYTICAL VIEWS

FIELDS OF RESEARCH

0801 Artificial Intelligence and Image Processing	12,737
0806 Information Systems	8,958
1701 Psychology	3,728
2004 Linguistics	1,801
1117 Public Health and Health Services	1,072

OVERVIEW



RESEARCHERS

Erik Cambria Nanyang Technological University, Singapore	135
Iztok Podbregar University of Maribor, Slovenia	134
Polona Sprajc Polona Sprajc University of Maribor, Slovenia	101
Pascal Ravesteijn	98
Roger W H Bons	97

# Um exemplo para o mercado de ações

StockFluence

Search

Data has been delayed by three days. Subscribe to StockFluence in order to view real-time data and predictions.

## We predict the stock market for the next five days!

About StockFluence

### FINANCIAL SENTIMENT ANALYSIS

StockFluence.com provides financial sentiment analysis for investors to discover, react and respond to market opinions. We monitor (social) media channels and analyze the overall sentiment with our algorithms. Based on the sentiment, we make predictions with an accuracy level of 70%.

More about StockFluence

▲ +3.33% 18

Buffalo Wild Wings Inc.

▲ +0.54% 37

BJ's Restaurants, Inc.

▲ +0% 20

Cambium Learning Group, Inc.

KONINKLIJKE PHILIPS ELECTRONICS NV >

35

3.43%

THE GOLDMAN SACHS GROUP, INC. >

31

1.29%

APPLE INC. >

33

1.21%

## Popular

Most popular funds

Apollo Group Inc. (APOL)	29 (+2.76%)
Apple Inc. (AAPL)	33 (+1.21%)
Marvell Technology Group Ltd. (MRVL)	34 (+-2.94%)

## Gainers

Gaining funds

Existe, inclusive, sites especializados para a Análise de Sentimentos para auxiliar investidores com o humor dos investidores.

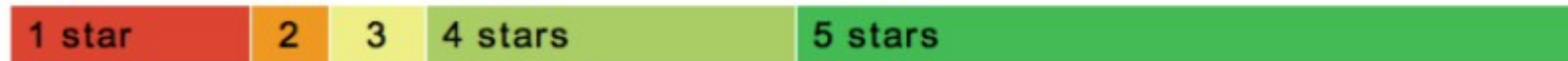
# Google product search



**HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner**  
**\$89 online, \$100 nearby** ★★★★★ 377 reviews  
September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sheets

## Reviews

Summary - Based on 377 reviews



### What people are saying

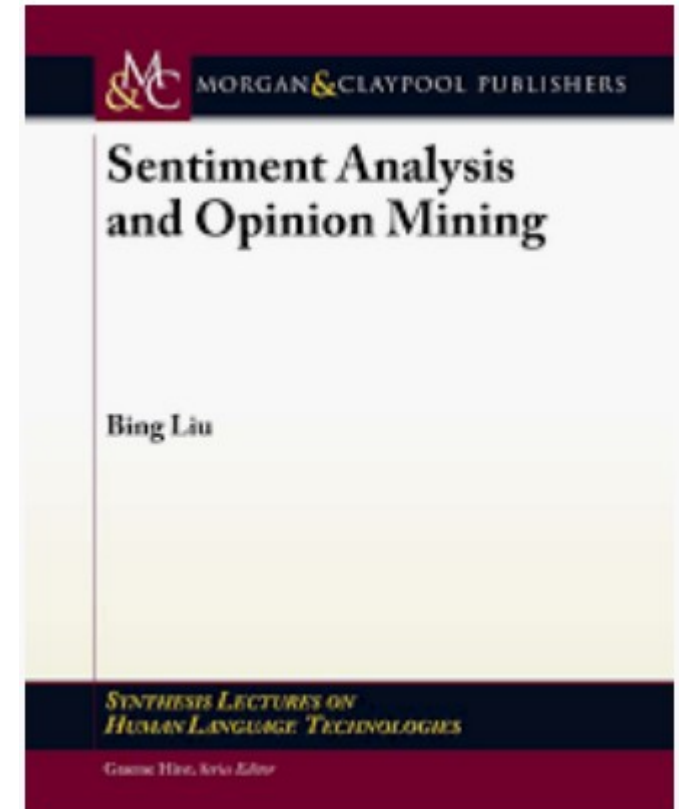
ease of use		"This was very easy to setup to four computers."
value		"Appreciate good quality at a fair price."
setup		"Overall pretty easy setup."
customer service		"I DO like honest tech support people."
size		"Pretty Paper weight."
mode		"Photos were fair on the high quality mode."
colors		"Full color prints came out with great quality."

↑  
Aspectos/  
atributos

↑  
Sentimento

# Termos correlatos

- Subjectivity Analysis.
- Sentiment Analysis.
- Opinion Analysis.
  
- Sentiment Mining.
- Opinion Mining.





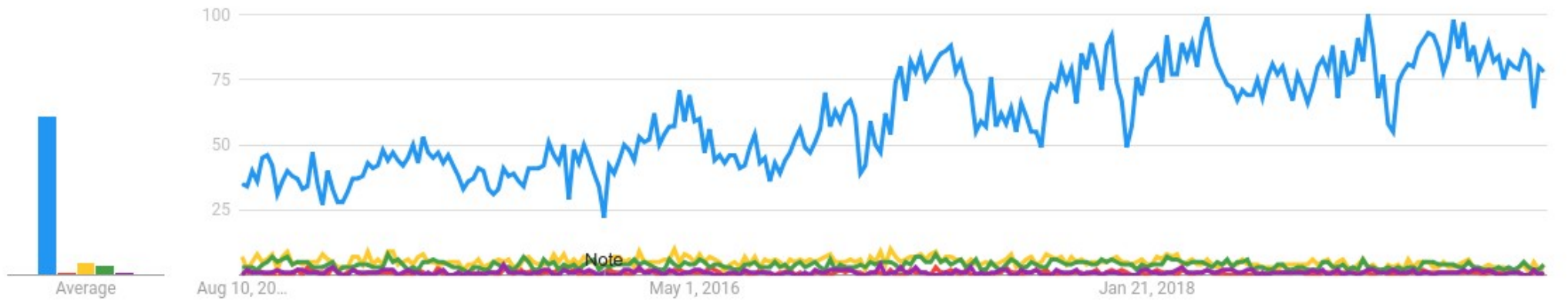


<http://search.carrot2.org/stable/search?query=sentiment+analysis&results=100&source=web&algorithm=lingo&view=foamtree&skin=fancy-compact&EToolsDocumentSource.country=ALL&EToolsDocumentSource.language=ALL&EToolsDocumentSource.safeSearch=false>

- sentiment analysis Search term
- opinion extraction Search term
- opinion mining Search term
- sentiment mining Search term
- subjectivity analysis Search term

Worldwide ▼ Past 5 years ▼ All categories ▼ Web Search ▼

Interest over time ?



<https://trends.google.com/trends/explore?date=today%20-5-y&q=sentiment%20analysis,opinion%20extraction,opinion%20mining,sentiment%20mining,subjectivity%20analysis>

# Análise de sentimentos (AS)

- A **Análise de sentimentos (AS)** ou **Mineração de Opinião (MO)** é a aplicação de um conjunto de tarefas sobre análise de textos para evidenciar:
  - Opiniões,
  - Emoções,
  - Julgamentos, ou
  - Pareceres.
- A AS tornou-se o “Santo Graal” da:
  - Pesquisa de mercado,
  - Pesquisa de opinião, e
  - Gerenciamento de reputação online.



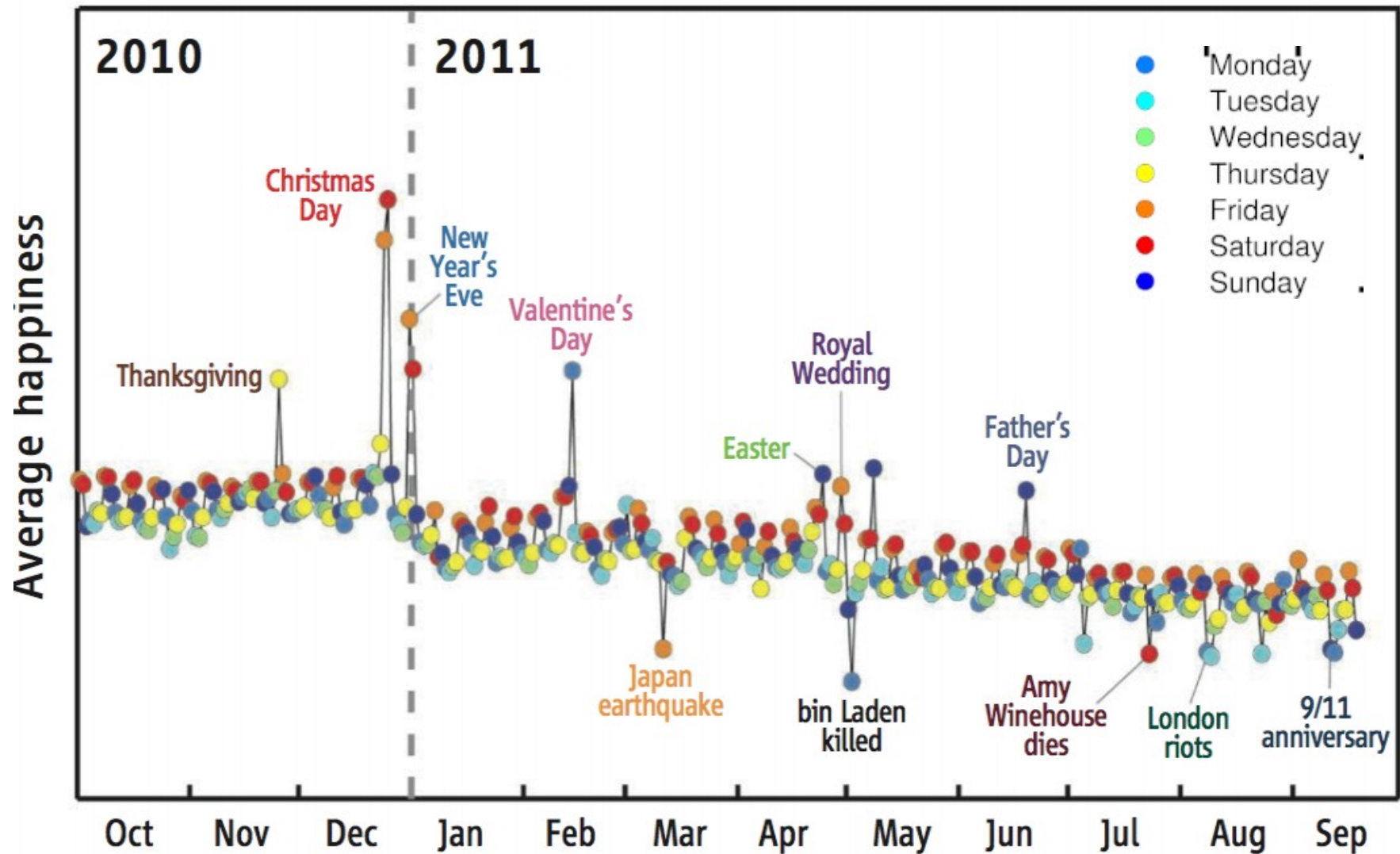
# Aplicações sobre "AS" ou "MO"

São inúmeras as aplicações.

As aplicações mais comuns:

- **Opinião** sobre filmes.
- **Avaliação** de produtos/serviços.
- **Polarização** política.
- **Predição** do mercado de ações.
- **Sentimento público**.

# Sentimiento público



DODDS, Peter Sheridan et al. **Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter.** PloS one, v. 6, n. 12, p. e26752, 2011.

# Subtarefas na "AS" ou "MO"

- **Classificação de sentimento:**

Classifica se um trecho de texto é pos., neg. ou neu.

- **Geração de lexicon:**


Determina se uma palavra tende a ser pos., neg. ou neu.

- **Quantificação de sentimento:**

Estima a prevalência de sentimentos pos., neg. ou neu.

- **Extração de Opinião:**

Dada uma opinião em uma sentença, identificar o titular da opinião, seu objeto, sua polaridade, a força dessa polaridade, e o tipo de opinião.



**Classificação de sentimento:  
positivo (1) e negativo (0)**

**function** TRAIN NAIVE BAYES(D, C)

**for each** class  $c \in C$

$N_{doc}$  = number of documents in D

$N_c$  = number of documents from D in class  $c$

$logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$

$V \leftarrow$  vocabulary of D

$bigdoc[c] \leftarrow$  **append**(d) **for**  $d \in D$  **with** class  $c$

**for each** word  $w$  in  $V$

$count(w, c) \leftarrow$  # of occurrences of  $w$  in  $bigdoc[c]$

$loglikelihood[w, c] \leftarrow \log \frac{count(w, c) + 1}{\sum_{w' \text{ in } V} (count(w', c) + 1)}$

**return**  $logprior, loglikelihood, V$

**function** TEST NAIVE BAYES(*testdoc*, *logprior*, *loglikelihood*, *C*, *V*)

**for each** class  $c \in C$

$sum[c] \leftarrow logprior[c]$

**for each** position  $i$  in *testdoc*

$word \leftarrow testdoc[i]$

**if**  $word \in V$

$sum[c] \leftarrow sum[c] + loglikelihood[word, c]$

**return**  $argmax_c sum[c]$

Se uma palavra é desconhecida  
no treinamento, então será  
desconsiderada (solução padrão)

# Conjunto de dados



KOTZIAS, Dimitrios et al.

**From group to individual labels using deep features.**

## Sentiment Labelled Sentences Data Set

Download: [Data Folder](#), [Data Set Description](#)

**Abstract:** The dataset contains sentences labelled with positive or negative sentiment.

Data Set Characteristics:	Text	Number of Instances:	3000	Area:	N/A
Attribute Characteristics:	N/A	Number of Attributes:	N/A	Date Donated	2015-05-30
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	134808

In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015. p. 597-606.

Para cada frase: A pontuação é 1 (para positivo) ou 0 (para negativo)

## Bases:

- imdb.com
- amazon.com
- yelp.com

Para cada site, existem 500 frases positivas e 500 frases negativas.



# Conjunto de datos: Ejemplos

Avoid, avoid, avoid! 0

My experience was terrible.....

This was my fourth bluetooth headset, and while it was much more comfortable than my last Jabra (which I HATED!!! 0

In conclusion, I will not bother with this movie because a volcano in Los Angeles is nothing but nonsense. 0

For a product that costs as much as this one does, I expect it to work far better and with greater ease than this thing does.0

Their network coverage in Los Angeles is horrible. 0

The characters were all funny and had the peculiarity of not having a true lead character. 1

Nice headset priced right.1

Works fine. 1

cool phone. 1

Great Product. 1



# analiseDeSentimentosNB1.py

```
regex = r"[-'a-zA-ZÀ-ÖØ-öø-ÿ]+"
```

```
class NBClassifier:
```

```
def __init__(self, training_file=None):
    self.Data          = []
    self.Classes       = dict([])
    self.V             = set([])
    self.bigdoc        = dict([])

    self.logprior      = dict([])
    self.loglikelihood = dict([])

    if training_file is not None:
        self.load_data(training_file)
```

```
def load_data(self, training_file):
    training_document = open(training_file, 'r')

    for line in training_document.readlines():
        d, c = tuple(line.strip().split("\t"))
        self.Data.append((c,d))

        if c not in self.Classes:
            self.Classes[c] = 0
            self.bigdoc[c] = []
            self.Classes[c] += 1

        for w in re.findall(regex, d):
            self.V.add(w)
            self.bigdoc[c].append(w)

    print("Total: classes={} documentos={} vocabulario={}".format(len(self.Classes), len(self.Data), len(self.V) ) )
```

# analiseDeSentimentosNB1.py

```
def train(self):
    for c in self.Classes:
        Ndoc = len(self.Data)
        Nc = self.Classes[c]

        self.logprior[c] = math.log(Nc/Ndoc)
        #self.logprior[c] = Nc/Ndoc

        count_wc = 0
        for w in self.V:
            count_wc += self.bigdoc[c].count(w)

        for w in self.V:
            self.loglikelihood[(w,c)] = math.log( (self.bigdoc[c].count(w) + 1) / (count_wc + len(self.V) ) )
            #self.loglikelihood[(w,c)] = (self.bigdoc[c].count(w) + 1) / (count_wc + len(self.V) )

    print("\n", self.logprior)
```

```
def test(self, testdoc):
    s = dict([])
    for c in self.Classes.keys():
        s[c] = self.logprior[c]
        for w in re.findall(regex, testdoc):
            if w in self.V:
                s[c] += self.loglikelihood[(w,c)]
                #s[c] *= self.loglikelihood[(w,c)]

    return max(s, key=s.get)
```

# analiseDeSentimentosNB1.py

```
if __name__ == '__main__':  
    fileName = sys.argv[1]  
  
    NBC = NBClassifier(fileName)  
    NBC.train()  
  
    while True:  
        phrase = input("\nDigite uma frase: ")  
        print("Resposta: {}".format( NBC.test(phrase) ) )
```

# analiseDeSentimentosNB1.py

```
python analiseDeSentimentosNB1.py datasets/all_datasets.txt
```

```
Total: classes=2 documentos=3000 vocabulario=6013
```

```
{'0': -0.6931471805599453, '1': -0.6931471805599453}
```

```
Digite uma frase: cool phone
```

```
Resposta: 1
```

```
Digite uma frase: probably never coming back
```

```
Resposta: 0
```

```
Digite uma frase: waste of time
```

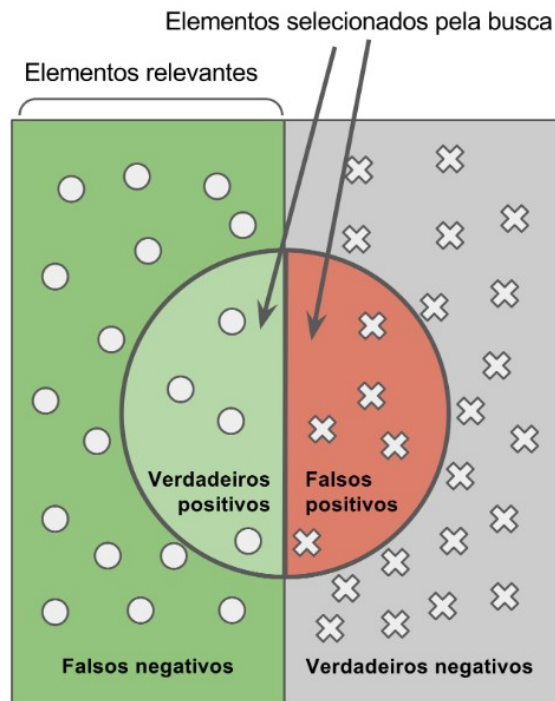
```
Resposta: 0
```

```
Digite uma frase: Damn good steak
```

```
Resposta: 1
```

# Avaliando o desempenho

		Actual	
		Positive	Negative
Predicted	Positive	<b>True Positive</b>	<b>False Positive</b>
	Negative	<b>False Negative</b>	<b>True Negative</b>



Precisão =  $\frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos positivos}}$

"Quantos elementos selecionados são relevantes?"

*Precisión*

Revocação =  $\frac{\text{Verdadeiros positivos} + \text{Falsos positivos}}{\text{Verdadeiros positivos} + \text{Verdadeiros negativos}}$

"Quantos elementos relevantes foram selecionados?"

*Recall (sensibilidade)*

# analiseDeSentimentosNB2.py

```
def test_batch(self, testing_file):
    testing_document = open(testing_file, 'r')
    correct = 0
    total = 0
    (tp, tn, fp, fn) = (0,0,0,0)

    for line in testing_document.readlines():
        total += 1
        d, c = tuple(line.strip().split("\t"))
        result = NBC.test(d)
        print ("Classe_Verdadeira={} Classe_Identificada={}\t{}".format(c, result, d))
        if c==result:
            correct += 1
        if c=='1' and result=='1':
            tp += 1
        if c=='0' and result=='1':
            fp += 1
        if c=='1' and result=='0':
            fn += 1
        if c=='0' and result=='0':
            tn += 1

    print ("Corretos={}/{}\tAcurácia={}".format(correct, total, correct/float(total) ))
    print ("Precisão = {}".format(float(tp)/(tp+fp)))
    print ("Revocação = {}".format(float(tp)/(tp+fn)))
```

# analiseDeSentimentosNB2.py

**C\_Verdadeira=1 C\_Identicada=1:**

Also were served hot bread and butter, and home made potato chips with bacon bits on top.... very original and very good.

**C\_Verdadeira=1 C\_Identicada=0:**

Logitech Bluetooth Headset is a 10!.

**C\_Verdadeira=0 C\_Identicada=1:**

The budget was evidently very limited.

**C\_Verdadeira=0 C\_Identicada=0:**

This does not fit the Palm Tungsten E2 and it broke the first time I tried to plug it in.

**C\_Verdadeira=0 C\_Identicada=0:**

The soundtrack sucked.

**Corretos=594/750      Acurácia=0.792**

**Precisão = 0.759**

**Revocação = 0.811**

# analiseDeSentimentosNB2.py

**C\_Verdadeira=1 C\_Identicada=1:**

Also were served hot bread and butter, and home made potato chips with bacon bits on top.... very original and very good.

*Verdadeiro positivo*

**C\_Verdadeira=1 C\_Identicada=0:**

Logitech Bluetooth Headset is a 10!

*Falso negativo*

**C\_Verdadeira=0 C\_Identicada=1:**

The budget was evidently very limited.

*Falso positivo*

**C\_Verdadeira=0 C\_Identicada=0:**

This does not fit the Palm Tungsten E2 and it broke the first time I tried to plug it in.

**C\_Verdadeira=0 C\_Identicada=0:**

The soundtrack sucked.

*Verdadeiro negativo*

**Corretos=594/750      Acurácia=0.792**

**Precisão = 0.759**

**Revocação = 0.811**



# Como lidar com a Negação?

- Eu **não** gostei muito desse filme, mas ...
- Eu **realmente** gostei muito desse filme.

# Como lidar com a Negação?

- Eu **não** gostei muito desse filme, mas...
- -> Eu **não\_**gostei **não\_**muito **não\_**desse **não\_**filme,

Uma solução de Das & Chen:

Adicionar o prefixo **não\_** a cada palavra entre a negação e o sinal de pontuação.

DAS, Sanjiv; CHEN, Mike. **Yahoo! for Amazon: Extracting market sentiment from stock message boards**. In: Proceedings of the Asia Pacific finance association annual conference (APFA). 2001. p. 43.

# Como lidar com a Negação?

- there is no sign of improvement; the most expected ...
- there is not\_sign not\_of not\_improvement; the most expected...]
  
- but this movie is not funny, considering the ridiculousness of it.
- but this movie is not\_funny, considering the ridiculousness of it .
  
- this place is not quality sushi, it is not a quality restaurant.
- this place is not\_quality not\_sushi , it is not\_a not\_quality not\_restaurant .

# analiseDeSentimentosNB3.py

```
def negate_sequence(text):
    text2 = ""
    prefix = ""
    for w in re.findall(r"[-'a-zA-ZÀ-Öø-ÿ]+[.,;!]", text):
        if w in ["not", "didn't", "no"]:
            prefix = "not_"
            continue
        if w in ".,;!":
            prefix = ""
        text2 += " "+prefix+w
    return text2
```

# analiseDeSentimentosNB3.py

```
// versão 2  
Corretos=594/750      Acurácia=0.792  
Precisão   = 0.759  
Revocação  = 0.811
```

```
// versão 3  
Corretos=597/750      Acurácia=0.796  
Precisão   = 0.730  
Revocação  = 0.891
```

# Complexidade das tarefas em AS

- **Tarefa simples:**
  - > O sentimento é positivo ou negativo?
- **Tarefa mais complexa:**
  - > Classificar o sentimento de 1 a 5.
- **Tarefa que é um grande desafio:**
  - > Identificar tipos complexos de sentimento.
  - > Identificar o alvo.
  - > Identificar a fonte.

# Ainda é uma tarefa difícil

“A beautiful hotel in a horrible town!” vs.

“A horrible hotel in a beautiful town!”

“Tive vitória esmagadora” vs.

“Minha reputação foi esmagada”



# **Similaridade entre ementas: O caso da UFABC**



# Objetivo

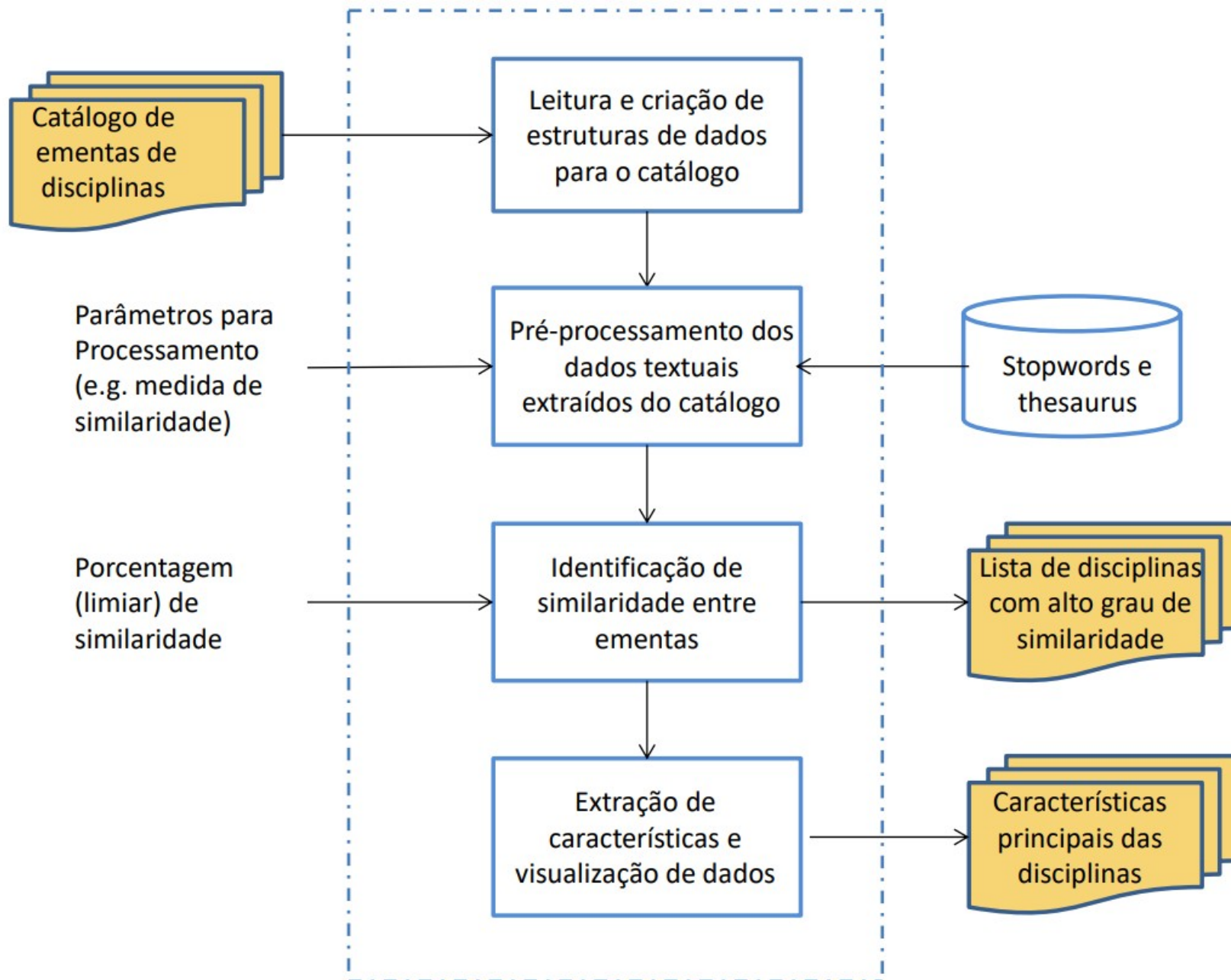
Projetar e desenvolver **algoritmos de detecção de similaridade entre ementas de disciplinas** associadas ao catálogo da UFABC, usando Processamento da Linguagem Natural e diversos coeficientes de similaridade.

## Identificação de similaridade em disciplinas do catálogo da UFABC

**Ana Laura Belotto Claudio**, Marcelo Bussotti Reyes, Jesús P. Mena-Chalco

BC&T, BCC, Universidade Federal do ABC

# Processos



# Medidas de similaridade

Documento	Texto	Termos
Ementa 1	História do petróleo. Exploração do petróleo.	história do petróleo exploração
Ementa 2	Classificação e composição do petróleo.	classificação e composição do petróleo
Ementa 3	Processamento primário do petróleo.	processamento primário do petróleo

	história	do	petróleo	explor.	classif.	e	compos.	proces.	primário	Tamanho
Ementa 1	1	1	1	1	0	0	0	0	0	4
Ementa 2	0	1	1	0	1	1	1	0	0	5
Ementa 3	0	1	1	0	0	0	0	1	1	4

	Ementa 1	Ementa 2	Ementa 3
Ementa 1	1	0,45	0,50
Ementa 2	0,45	1	0,45
Ementa 3	0,50	0,45	1

$$\text{coef. do Cosseno} = \frac{A \cap B}{|A| \cdot |B|}$$

# Medidas de similaridade

Documento	Texto	Termos
Ementa 1	História do petróleo. Exploração do petróleo.	história do petróleo exploração
Ementa 2	Classificação e composição do petróleo.	classificação e composição do petróleo
Ementa 3	Processamento primário do petróleo.	processamento primário do petróleo

	história	do	petróleo	explor.	classif.	e	compos.	proces.	primário	Tamanho
Ementa 1	1	1	1	1	0	0	0	0	0	4
Ementa 2	0	1	1	0	1	1	1	0	0	5
Ementa 3	0	1	1	0	0	0	0	1	1	4

	Ementa 1	Ementa 2	Ementa 3
Ementa 1	1	0,45	0,50
Ementa 2	0,45	1	0,45
Ementa 3	0,50	0,45	1

$$\text{coef. do Cosseno} = \frac{|A \cap B|}{|A| \cdot |B|}$$



# Palavras com pouca expressão

Tabela 8: Palavras mais frequentes no catálogo de disciplinas 2016/17 da UFABC.

nº	Palavra	Frequência	nº	Palavra	Frequência
1	sistemas	433	51	básicos	87
2	análise	290	52	estrutura	85
3	desenvolvimento	221	53	método	85
4	introdução	218	54	dados	83
5	aplicações	185	55	fundamentos	81
6	políticas	175	56	educação	80
7	teoria	173	57	modelo	79
8	modelos	172	58	integração	79
9	trabalho	172	59	social	76
10	conceitos	169	60	-	76
11	métodos	162	61	matemática	75
12	energia	162	62	história	74
13	processos	163	63	formação	73
14	técnicas	156	64	noções	72
15	projeto	151	65	dinâmica	71
16	aparece	146	66	processamento	70
17	planejamento	140	67	evolução	70
18	sistema	139	68	conhecimento	69
19	curso	137	69	tempo	67
20	principais	130	70	temas	67
21	materiais	128	71	informação	66
22	política	119	72	cultura	66
23	engenharia	118	73	lei	66
24	controle	117	74	pesquisa	65
25	equações	116	75	segurança	65
26	redes	115	76	relação	64

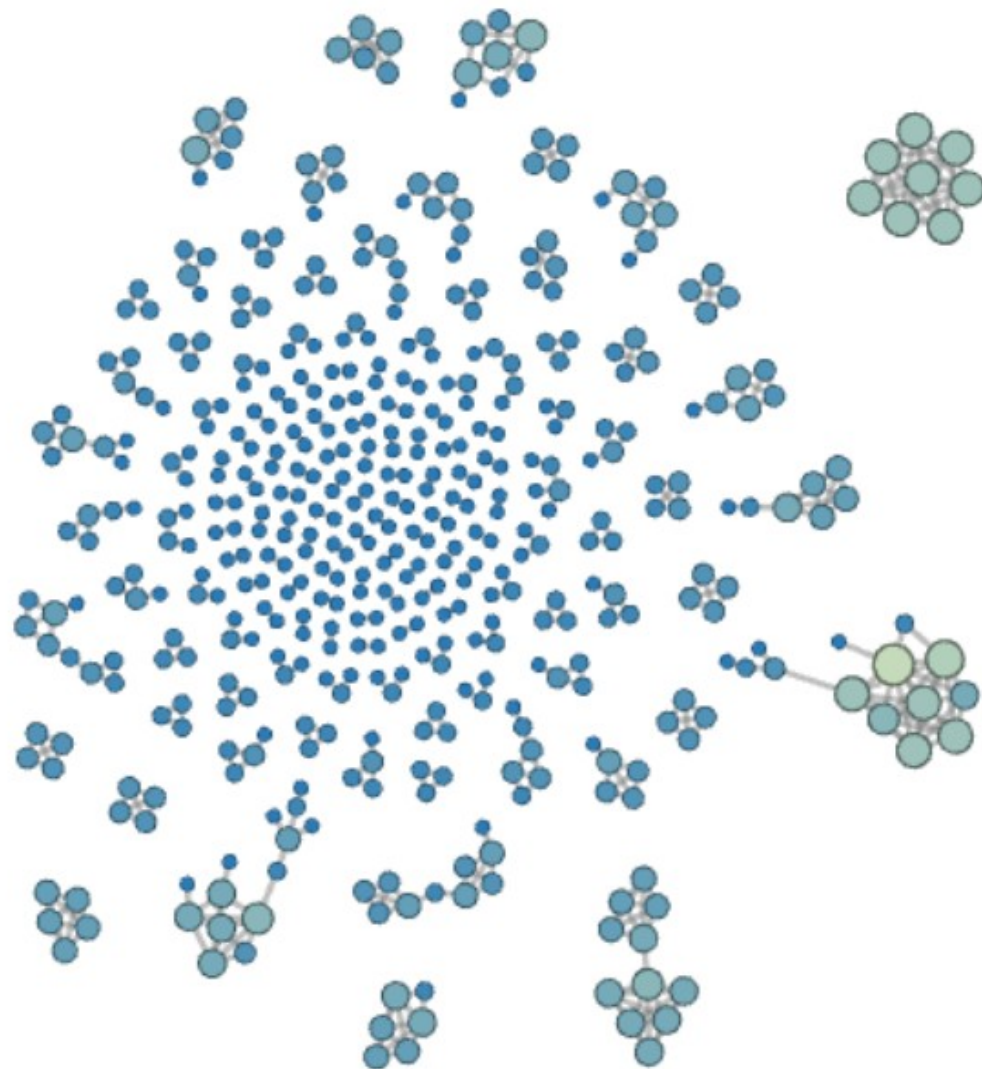
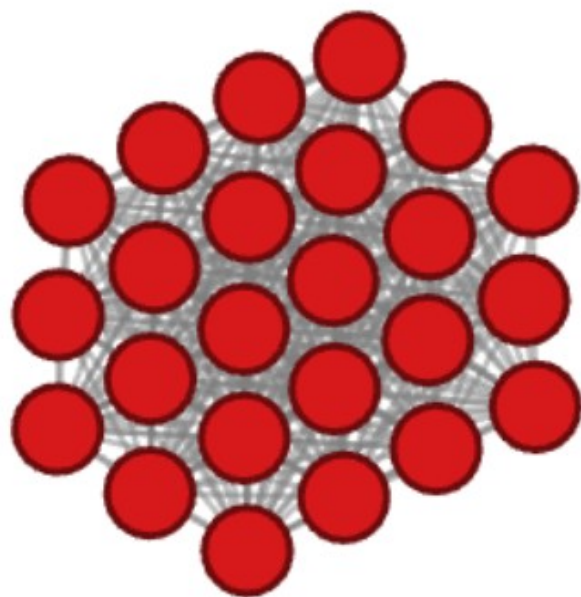
# Algoritmo

## ALGORITMO 1: IDENTIFICACAO-DE-SIMILARIDADE-DE-EMENTAS ( *Catalogo*, *medida\_similaridade* )

- *Catalogo*: Estrutura que contém o código das disciplinas, o nome, a ementa e as bibliografias.
- *medida\_similaridade*: Medida escolhida para o cálculo do coeficiente de similaridade.

```
1   for each ementa in Catalogo
2       (ementa) ← LEITURA_DE_EMENTA(disciplina)
3       (ementa) ← NORMALIZADOR(ementa, Stopwords)
4       (ementa) ← STEMMING(ementa)
5       (palavras) ← DICIONARIO(ementa)
6   end for
7   matriz_distancia ← DISTANCIA(palavras, frequencia)
8   matriz_distancia ← REDIMENCIONA(matriz_distancia)
9   matriz_similaridade ← SIMILARIDADE(matriz_distancia, medida_similaridade)
10  arquivo_grafo ← CRIAGRAFO(matriz_similaridade)
```

# Agrupamento de disciplinas similares

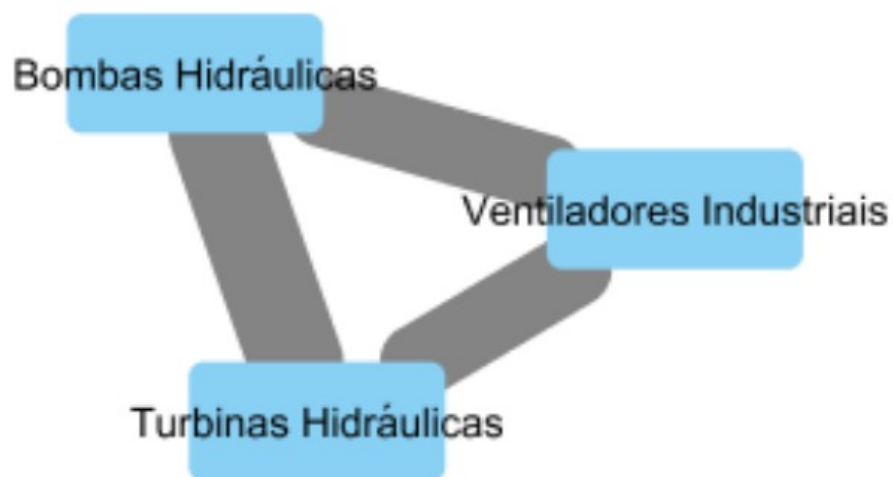
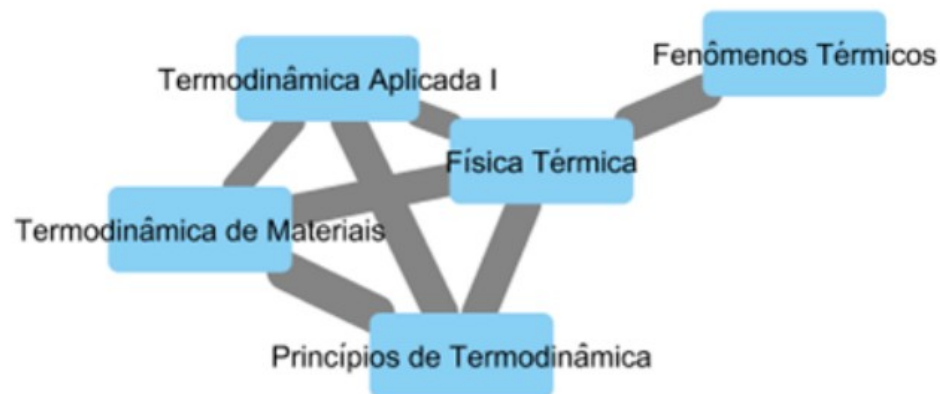
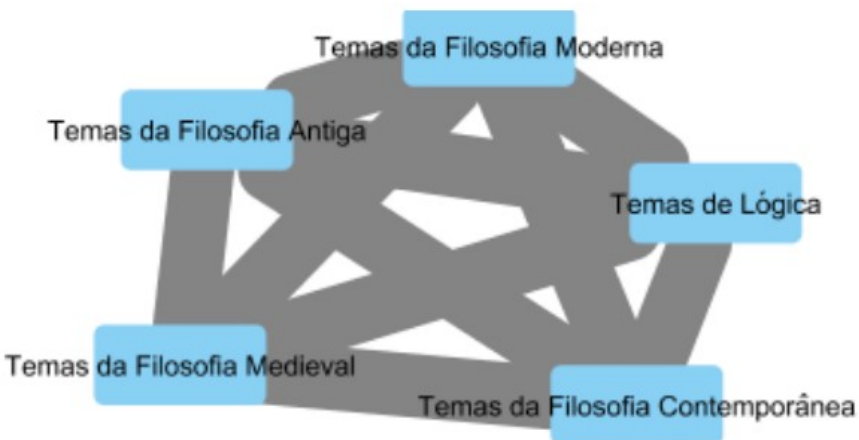


Escala





# Agrupamento de disciplinas similares

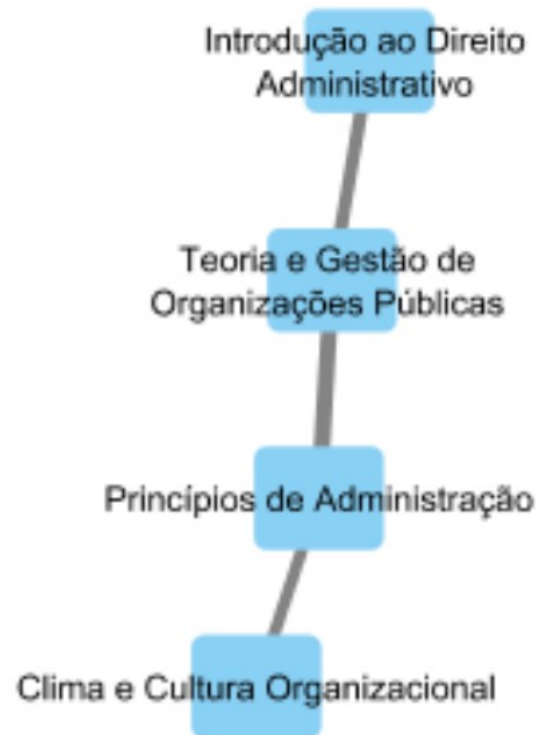




# Disciplinas acima de 70% similares



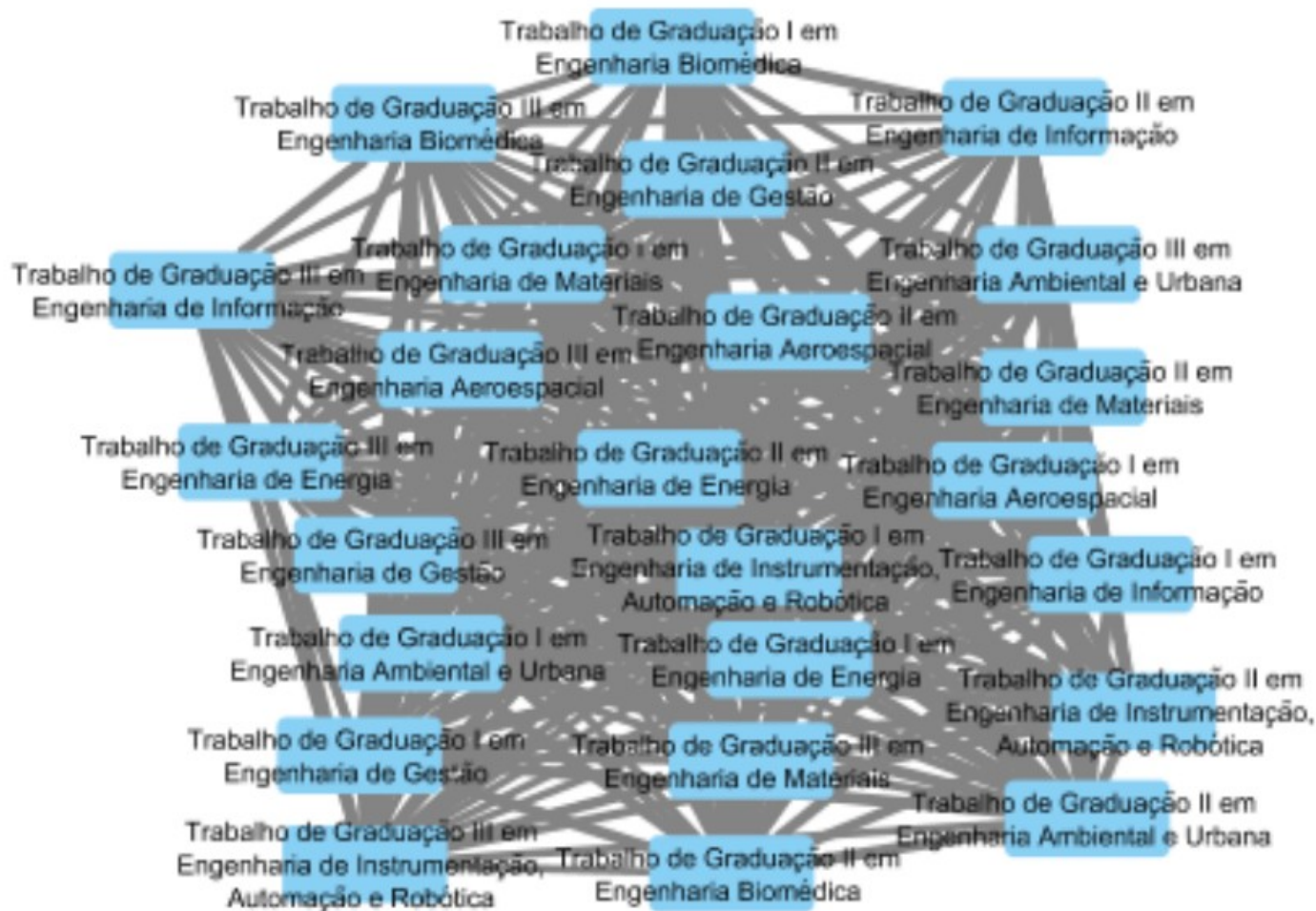
(a) Agrupamento de disciplinas similares com a temática mecânica.



(b) Agrupamento de disciplinas similares com a temática administração.

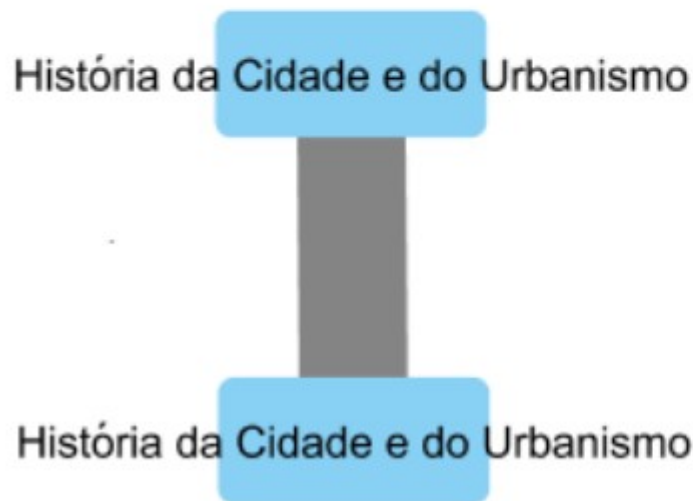
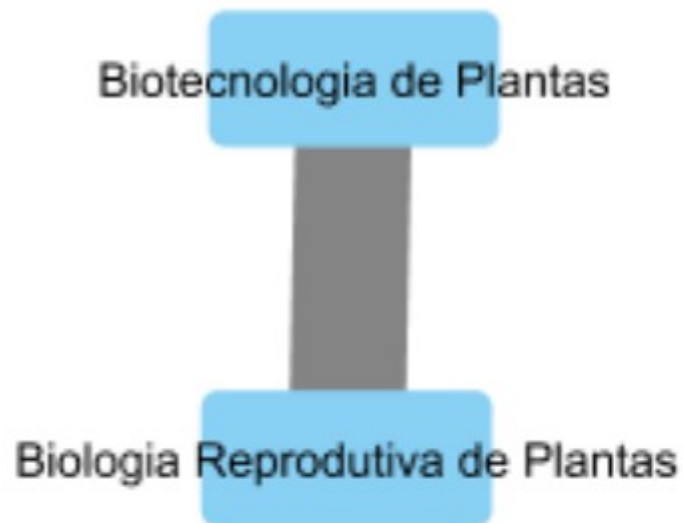
**Figura 17:** Exemplos de agrupamentos de disciplinas com similaridade acima de 70% utilizando o coeficiente de Sobreposição.

# Disciplinas 100% similares



**Figura 10:** Agrupamento de disciplinas de Trabalho de Graduação encontrado no grafo de similaridade entre disciplinas do catálogo da UFABC com o coeficiente de Jaccard.

# Disciplinas 100% similares



# Considerações finais

Utilizando medidas de similaridade, encontramos algumas disciplinas, no catálogo de disciplinas da UFABC, que são similares entre si.

Desde disciplinas completamente idênticas até disciplinas com parte da ementa similar.

Acreditamos que estes resultados encontrados podem auxiliar na tomada de decisões relacionadas com a logística de ofertas de disciplinas na Universidade.