

# Modelagem de Tópicos

Prof. Leandro Alvim, D. Sc.

# Agenda

- ▶ O que é ?
- ▶ Métodos
- ▶ Metodologia
- ▶ Interpretação
- ▶ Ferramentas
- ▶ Considerações

# Modelagem de Tópicos



# Modelagem de Tópicos



100% Topic A



100% Topic B



100% Topic B



100% Topic A



60% Topic A  
40% Topic B

# Modelagem de Tópicos

**Topic A:** 40% banana, 30% kale, 10% breakfast...

Food

**Topic B:** 30% kitten, 20% puppy, 10% frog, 5% cute...

Animals

# Modelagem de Tópicos



**Food**



**Animals**



**Animals**

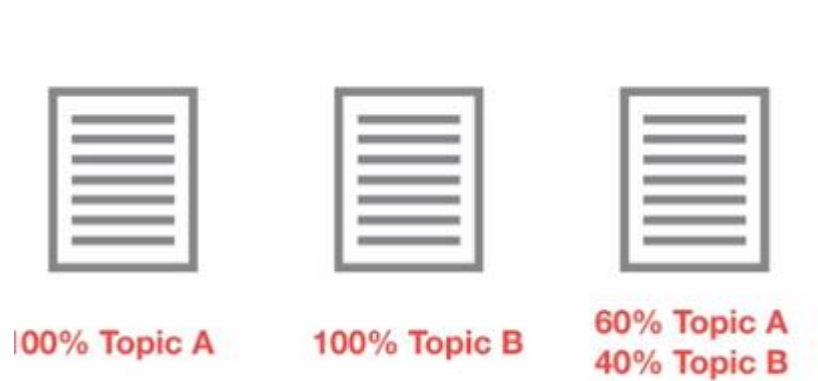


**Food**

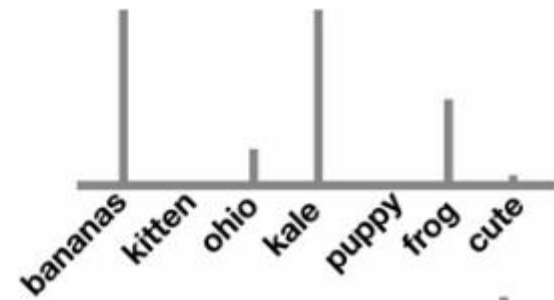


**Food +  
Animals**

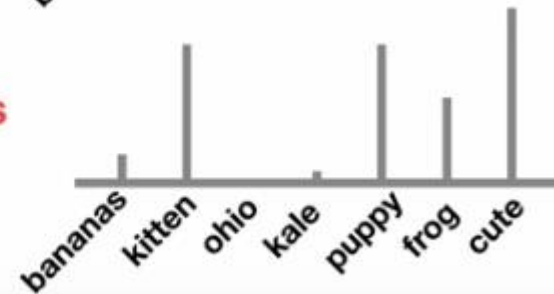
# Modelagem de Tópicos



Topic: Food



Topic: Animals



# Métodos

- ▶ Latent Semantic Analysis
- ▶ Latent Semantic Indexing
- ▶ Probabilistic Semantic Analysis
- ▶ Latent Dirichlet Allocation (LDA)
- ▶ Hierarchical Dirichlet Allocation
- ▶ LDA2Vec
- ▶ Spherical HDP





# Resumo do LDA

## ► Objetivo

- Aprender um modelo gerador de documentos
  - A partir de todos os documentos como exemplos
  - Supõe que **todo documento possui tópicos**
  - Cada tópico é um conjunto de palavras
  - Tópicos não necessariamente disjuntos
- Com o gerador aprendido
  - Você sabe os tópicos que representam a base e cada documento
  - Sabe também como cada tópico é constituído

### Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

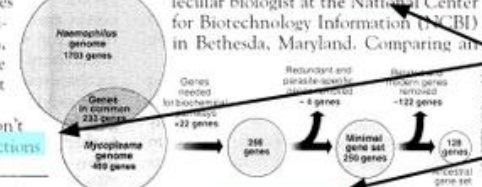
### Documents

#### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **simple** **numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

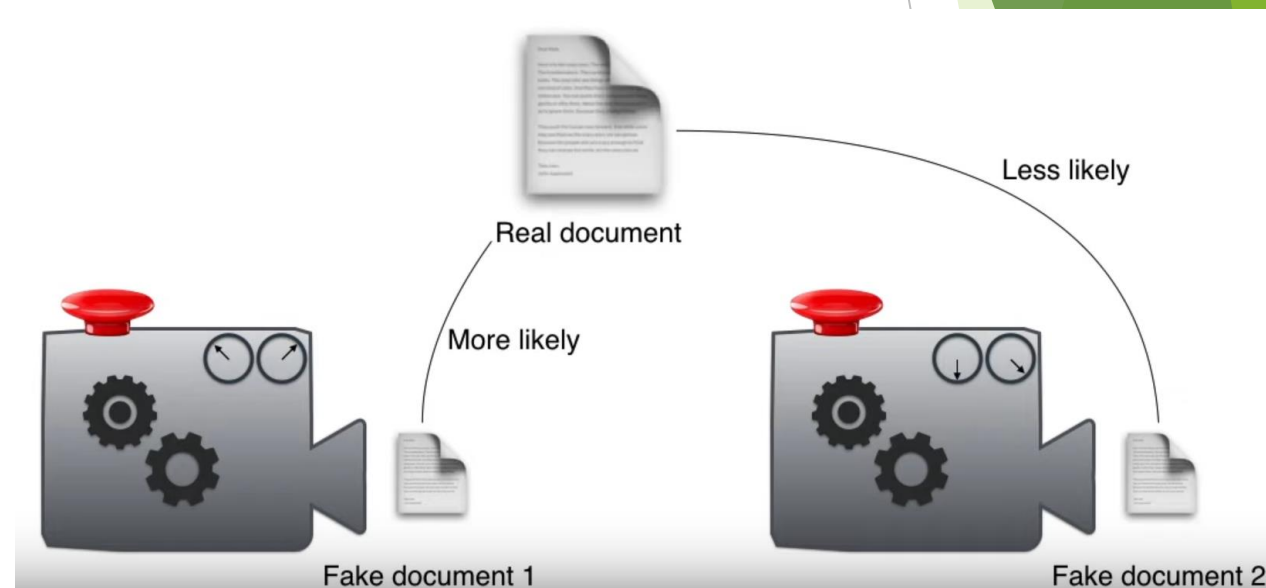
SCIENCE • VOL. 272 • 24 MAY 1996

### Topic proportions and assignments

# Resumo do LDA

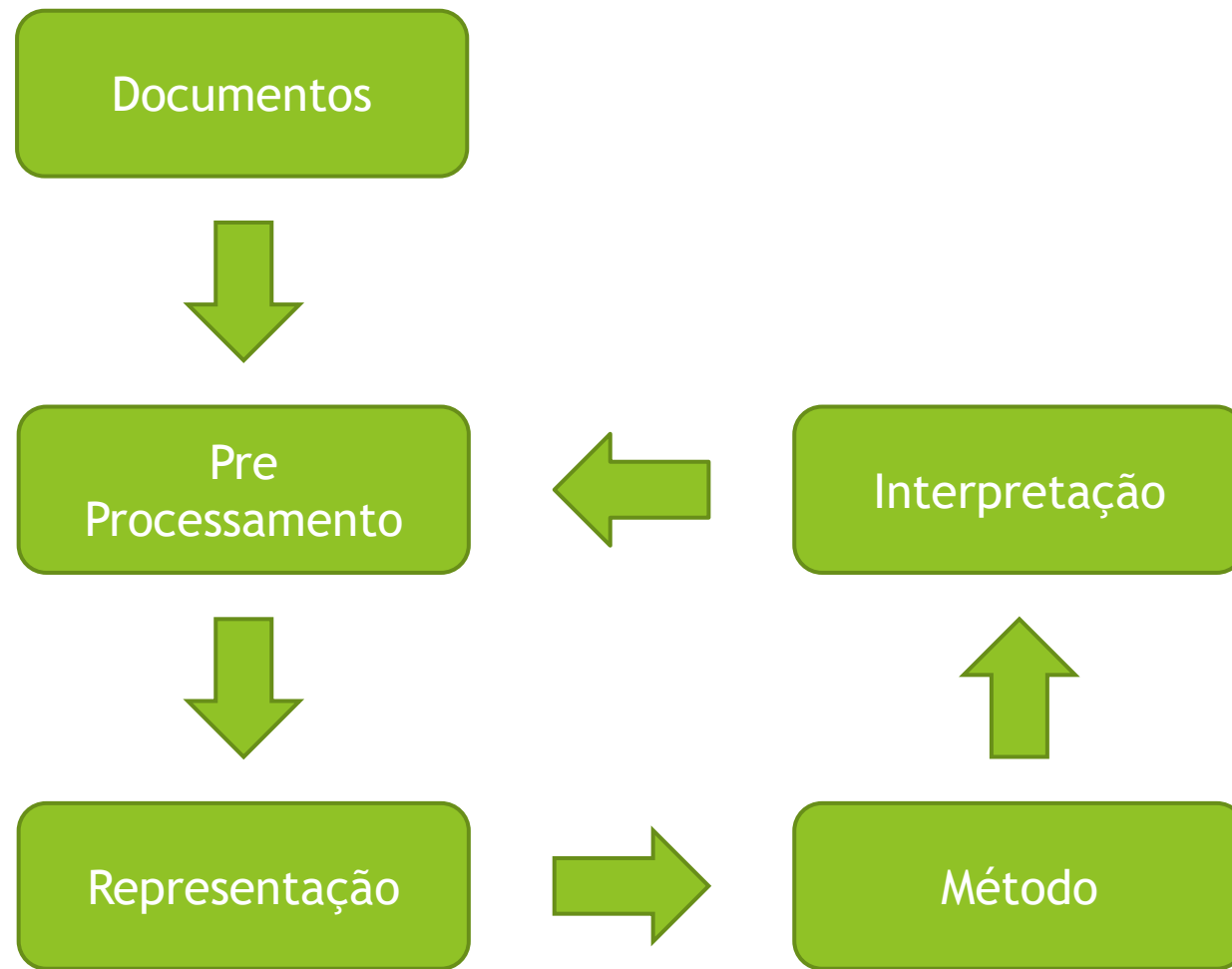
## ► Algoritmo

- Escolha o número de tópicos ( $k$ )
- Associe aleatoriamente cada palavra de cada documento a um dos  $k$  tópicos
- Para cada palavra verifique o quanto o tópico associado a ela ocorre no documento; e também quanto ocorre a palavra no tópico em geral
- Modifique o tópico da palavra dentro do documento
- Repita o processo até ficar bom



Explicação: <https://www.youtube.com/watch?v=T05t-SqKArY>

# Metodologia

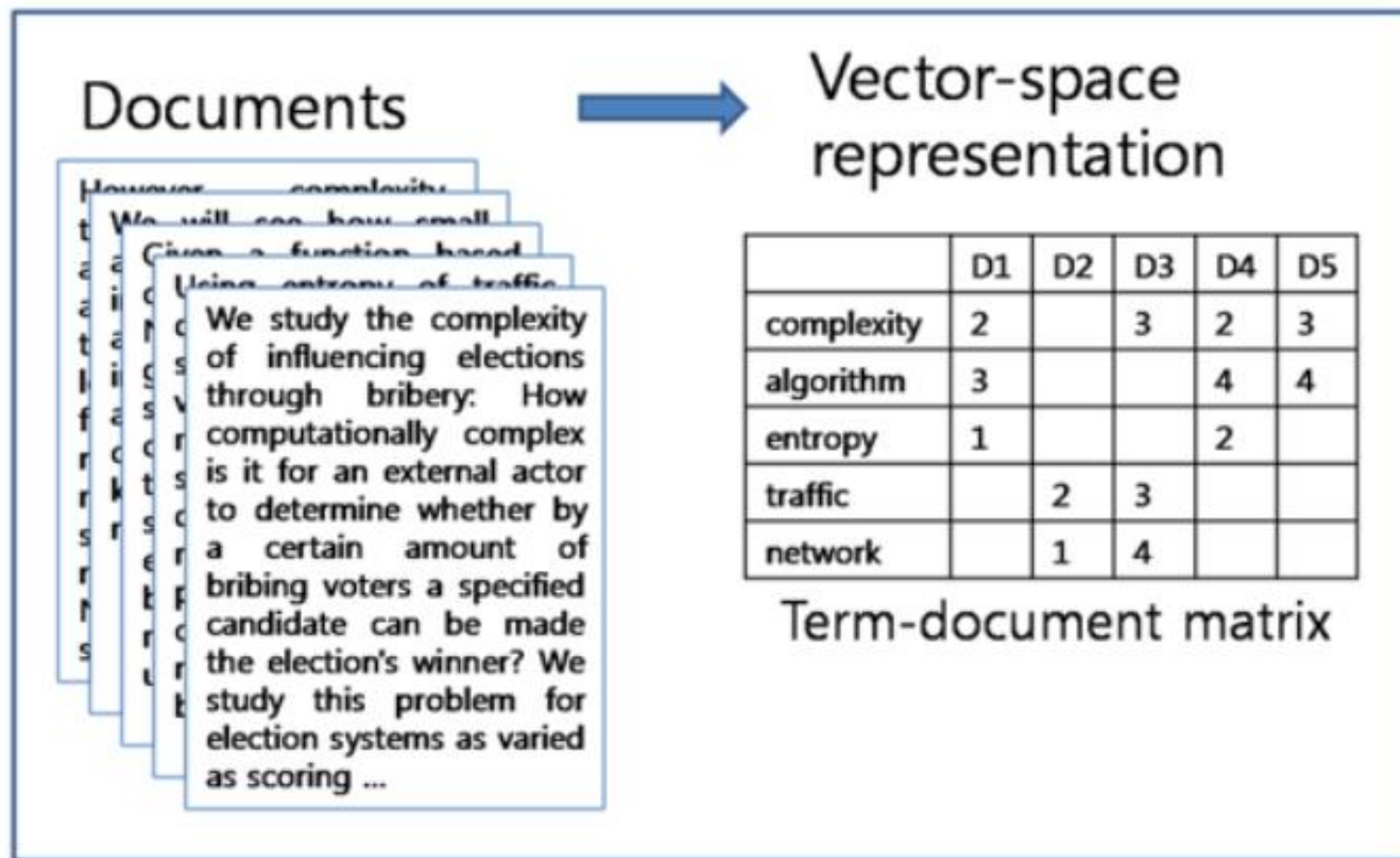


# Metodologia - Pre Processamento

- ▶ Tokenização
- ▶ Normalização
  - ▶ Minúsculas, remoção de acentos, ...
  - ▶ Lemma, ...
- ▶ Filtros
  - ▶ Léxico: Stop Words, Irrelevantes (por tf-idf)
  - ▶ Gramatical

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

# Metodologia - Representação



# Interpretação

- ▶ Os tópicos representam a base ?
- ▶ Os tópicos são inteligíveis ?
- ▶ Os tópicos são de fato coerentes ?
- ▶ Servem para o meu propósito ?

# Exemplos de Interpretação







# Hacker News



Document-Topic Distribution

<i>Topic</i>	<i><math>P(T   D)</math></i>
<b>58</b>	<b>0.19</b>
<b>38</b>	<b>0.14</b>
<b>16</b>	<b>0.06</b>
...	...

Sorted Topic-Term Distributions

<b>58</b>	<b>38</b>	<b>16</b>
<i>app</i>	<i>game</i>	<i>language</i>
<i>developer</i>	<i>player</i>	<i>code</i>
<i>mobile</i>	<i>video game</i>	<i>programming</i>
<i>user</i>	<i>gaming</i>	<i>java</i>
<i>app store</i>	<i>developer</i>	<i>programmer</i>





# Hacker News



Document-Topic Distribution

Topic	$P(T D)$
<i>mobile apps</i>	<i>0.19</i>
<i>38</i>	<i>0.14</i>
<i>16</i>	<i>0.06</i>
...	...

Sorted Topic-Term Distributions

<i>mobile apps</i>	<i>38</i>	<i>16</i>
<i>app</i>	<i>game</i>	<i>language</i>
<i>developer</i>	<i>player</i>	<i>code</i>
<i>mobile</i>	<i>video game</i>	<i>programming</i>
<i>user</i>	<i>gaming</i>	<i>java</i>
<i>app store</i>	<i>developer</i>	<i>programmer</i>



# Hacker News



Document-Topic Distribution

Topic	$P(T   D)$
<i>mobile apps</i>	<i>0.19</i>
<i>video games</i>	<i>0.14</i>
<i>programming</i>	<i>0.06</i>
...	...

Sorted Topic-Term Distributions

<i>mobile apps</i>	<i>video games</i>	<i>programming</i>
<i>app</i>	<i>game</i>	<i>language</i>
<i>developer</i>	<i>player</i>	<i>code</i>
<i>mobile</i>	<i>video game</i>	<i>programming</i>
<i>user</i>	<i>gaming</i>	<i>java</i>
<i>app store</i>	<i>developer</i>	<i>programmer</i>

### TOPIC 32

supply destinations provide programs distance  
efficient measures maintain  
convenient bicycling projects safety information  
improve truck management  
effective land goods travel/safe identify  
choices vehicle systems link freight  
reduce regional transit work  
mobility auto system  
direct trips support facilitate  
mode users movement  
limited pricing parking plan trucks  
network modes  
capacity demand impacts enhance  
calming priority walking passenger  
improvements people efficiency

### TOPIC 33

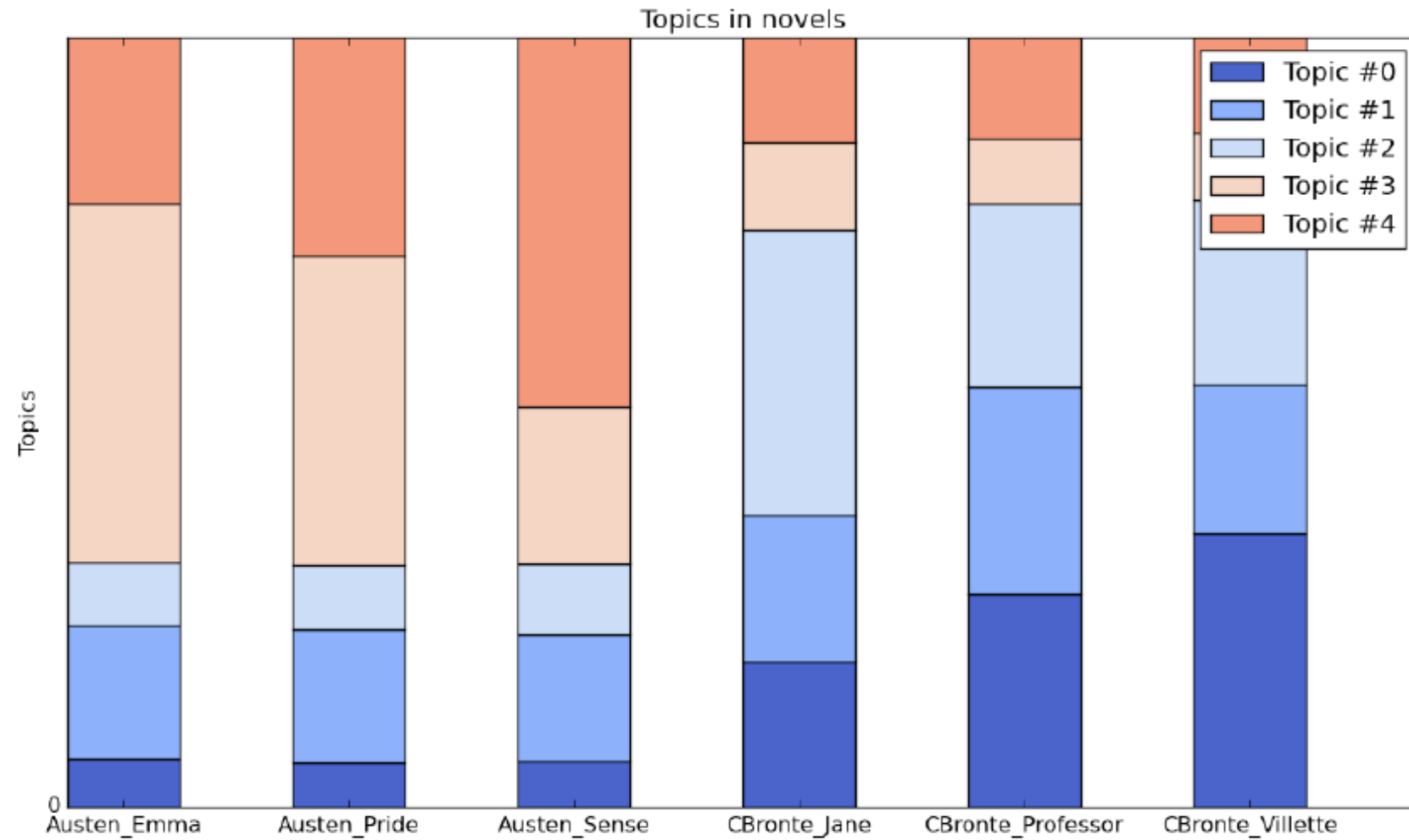
evidence readiness focused  
successfully production trained providers logistics  
experience transfer  
advanced clusters focus tech factors innovations  
good skilled future work role track  
important jobs systems institutions  
technology training venture improving  
research economy  
region lead grow improve  
tracking degrees workforce  
knowledge high innovation industry technologies  
prosperity employers education skills universities  
gain cluster workers world critical  
industries entrepreneurs companies  
talent professionals start leaders model foundations  
philanthropic class

### TOPIC 36

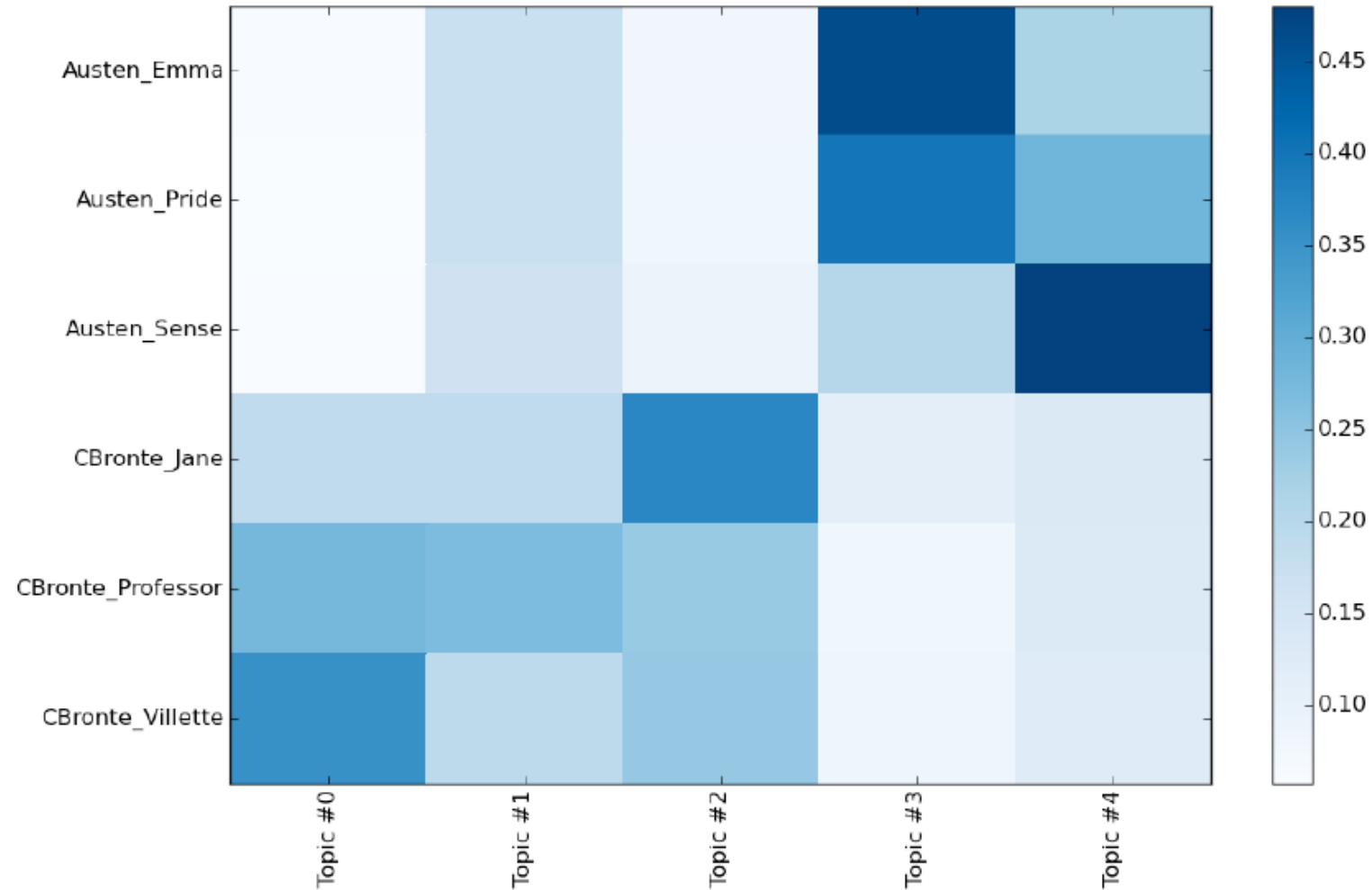
environmental concentration sufficient compatible  
distribution expansion significant  
viability designated business communities  
offer district areas link housing  
existing industrial large  
vacant retention area entire  
commercial mixed retail goods sizes  
site shopping activity sites. surrounding  
cultural upper residential exist  
impacts land districts due scale located  
nodes close location office serve restaurants  
market oriented services light  
locate businesses adjacent accommodate  
mixture proximity effects

### TOPIC 37

unincorporated supplement  
transition streets partially  
include retail station  
order urban subarea west food support  
princess island special existing  
recommendations opportunity studies  
mpsp form action community  
part plan  
space located vision provide  
north street south rail florin  
drainage center study light schenck  
fruitridge figure march  
proposed village housing plans boundary  
support university incorporated  
intensity



[https://de.dariah.eu/tatom/topic\\_model\\_visualization.html](https://de.dariah.eu/tatom/topic_model_visualization.html)



[https://de.dariah.eu/tatom/topic\\_model\\_visualization.html](https://de.dariah.eu/tatom/topic_model_visualization.html)

# Termos importantes com relação aos tópicos (global)

- ▶ Distinção
  - ▶ Mede o quanto o termo é relevante para gerar tópicos
  - ▶ Um termo que aparece muito em todos os tópicos não é um bom gerador de tópico
- ▶ Saliência
  - ▶ Frequência geral x Distinção
  - ▶ Usa-se para filtrar em ordem os termos mais frequentes e informativos da base.

*Termite: Visualization Techniques for Assessing Textual Topic Models*, Jason Chuang, Christopher D. Manning and Jeffrey Heer. 2012

# Distinção

	<i>coding</i>	<i>tech news</i>	<i>video games</i>
<i>game</i>	<b>10</b>	<b>10</b>	<b>50</b>
<i>apple</i>	<b>20</b>	<b>40</b>	<b>20</b>
<i>angry birds</i>	<b>1</b>	<b>1</b>	<b>30</b>
<i>python</i>	<b>50</b>	<b>5</b>	<b>10</b>

# Distinção

	<i>coding</i>	<i>tech news</i>	<i>video games</i>
<i>game</i>	<b>10</b>	<b>10</b>	<b>50</b>
<i>apple</i>	<b>20</b>	<b>40</b>	<b>20</b>
<i>angry birds</i>	<b>1</b>	<b>1</b>	<b>30</b>
<i>python</i>	<b>50</b>	<b>5</b>	<b>10</b>

$P(T   \text{game})$	<b>0.14</b>	<b>0.14</b>	<b>0.71</b>
$P(T   \text{apple})$	<b>0.25</b>	<b>0.50</b>	<b>0.25</b>
$P(T   \text{angry birds})$	<b>0.03</b>	<b>0.03</b>	<b>0.94</b>
$P(T   \text{pyhton})$	<b>0.77</b>	<b>0.08</b>	<b>0.15</b>
$P(T)$	<b>0.33</b>	<b>0.23</b>	<b>0.45</b>

$$\text{distinctiveness}(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$



# Distinção

	<i>coding</i>	<i>tech news</i>	<i>video games</i>
<i>game</i>	10	10	50
<i>apple</i>	20	40	20
<i>angry birds</i>	1	1	30
<i>python</i>	50	5	10

<i>P(T   game)</i>	0.14	0.14	0.71
<i>P(T   apple)</i>	0.25	0.50	0.25
<i>P(T   angry birds)</i>	0.03	0.03	0.94
<i>P(T   python)</i>	0.77	0.08	0.15
<i>P(T)</i>	0.33	0.23	0.45

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

# Distinção

	<i>coding</i>	<i>tech news</i>	<i>video games</i>	<i>distinctiveness</i>
<i>game</i>	<b>10</b>	<b>10</b>	<b>50</b>	
<i>apple</i>	<b>20</b>	<b>40</b>	<b>20</b>	
<i>angry birds</i>	<b>1</b>	<b>1</b>	<b>30</b>	<b>0.56</b>
<i>python</i>	<b>50</b>	<b>5</b>	<b>10</b>	

<i>P(T   game)</i>	<i>0.14</i>	<i>0.14</i>	<i>0.71</i>	
<i>P(T   apple)</i>	<i>0.25</i>	<i>0.50</i>	<i>0.25</i>	
<b><i>P(T   angry birds)</i></b>	<b>0.03</b>	<b>0.03</b>	<b>0.94</b>	
<i>P(T   pyhton)</i>	<i>0.77</i>	<i>0.08</i>	<i>0.15</i>	
<b><i>P(T)</i></b>	<b>0.33</b>	<b>0.23</b>	<b>0.45</b>	

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

# Distinção

	<i>coding</i>	<i>tech news</i>	<i>video games</i>	<i>distinctiveness</i>
<i>game</i>	<b>10</b>	<b>10</b>	<b>50</b>	<b>0.15</b>
<i>apple</i>	<b>20</b>	<b>40</b>	<b>20</b>	<b>0.18</b>
<i>angry birds</i>	<b>1</b>	<b>1</b>	<b>30</b>	<b>0.56</b>
<i>python</i>	<b>50</b>	<b>5</b>	<b>10</b>	<b>0.41</b>

<i>P(T   game)</i>	<b>0.14</b>	<b>0.14</b>	<b>0.71</b>
<i>P(T   apple)</i>	<b>0.25</b>	<b>0.50</b>	<b>0.25</b>
<i>P(T   angry birds)</i>	<b>0.03</b>	<b>0.03</b>	<b>0.94</b>
<i>P(T   python)</i>	<b>0.77</b>	<b>0.08</b>	<b>0.15</b>
<i>P(T)</i>	<b>0.33</b>	<b>0.23</b>	<b>0.45</b>

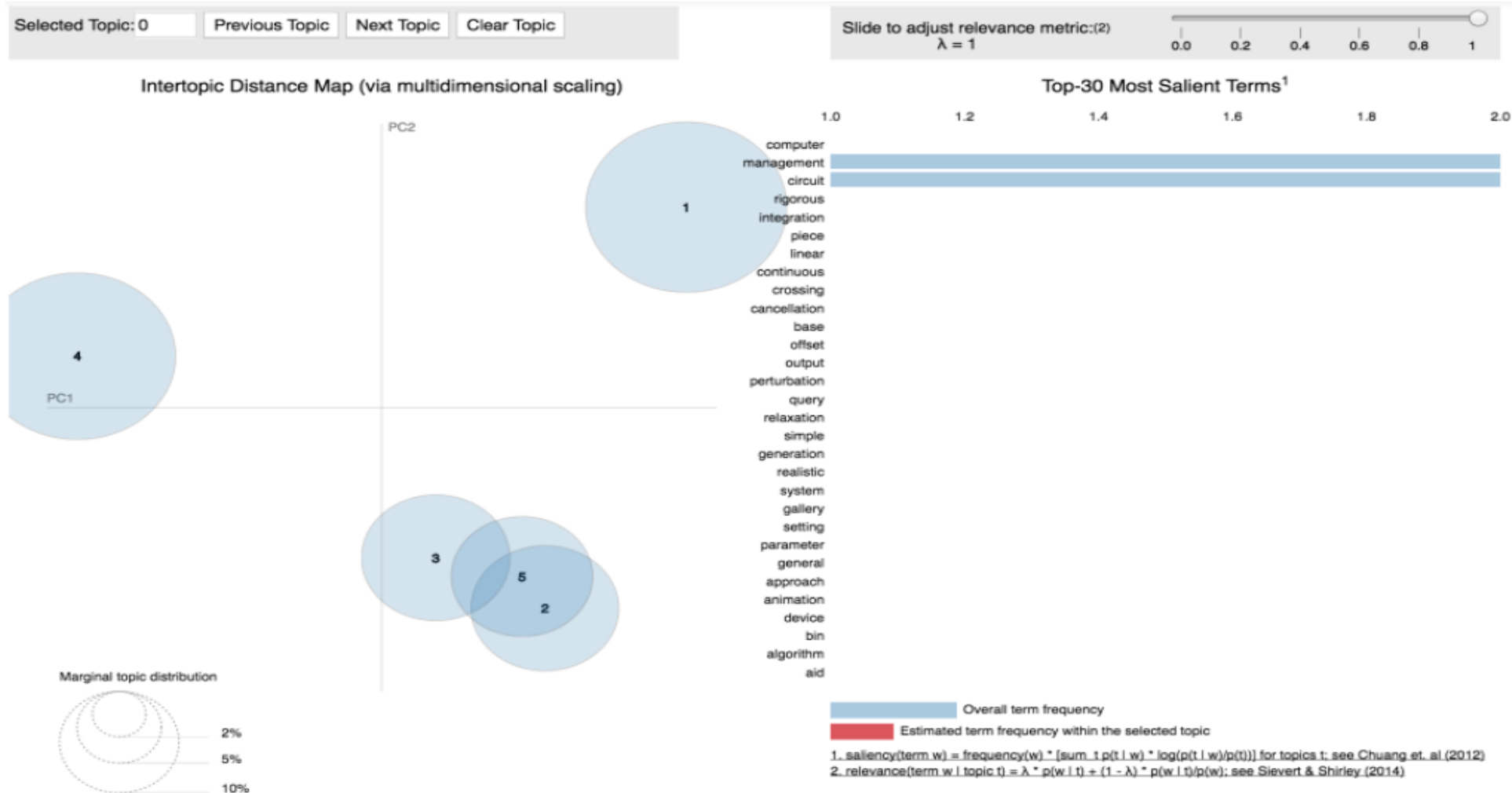
$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

# Saliência

	<i><b>coding</b></i>	<i><b>tech news</b></i>	<i><b>video games</b></i>	<i><b>distinctiveness</b></i>	<i><b>P(w)</b></i>	<i><b>saliency</b></i>
<i><b>game</b></i>	<i><b>10</b></i>	<i><b>10</b></i>	<i><b>50</b></i>	<i><b>0.15</b></i>	<i><b>0.28</b></i>	<i><b>0.04</b></i>
<i><b>apple</b></i>	<i><b>20</b></i>	<i><b>40</b></i>	<i><b>20</b></i>	<i><b>0.18</b></i>	<i><b>0.32</b></i>	<i><b>0.06</b></i>
<i><b>angry birds</b></i>	<i><b>1</b></i>	<i><b>1</b></i>	<i><b>30</b></i>	<i><b>0.56</b></i>	<i><b>0.13</b></i>	<i><b>0.07</b></i>
<i><b>python</b></i>	<i><b>50</b></i>	<i><b>5</b></i>	<i><b>10</b></i>	<i><b>0.41</b></i>	<i><b>0.26</b></i>	<i><b>0.11</b></i>

$$saliency(w) = P(w) \times distinctiveness(w)$$

# Exemplo de ordenação por saliência (outra base)



# Análise local do Termo

- ▶ Relevância
  - ▶ Mede o quão relevante o termo é dentro do tópico
  - ▶  $R(W|T,a) = a \cdot P(W|T) + (1-a) \cdot P(W|T)/P(W)$

Selected Topic: 2

Previous Topic

Next Topic

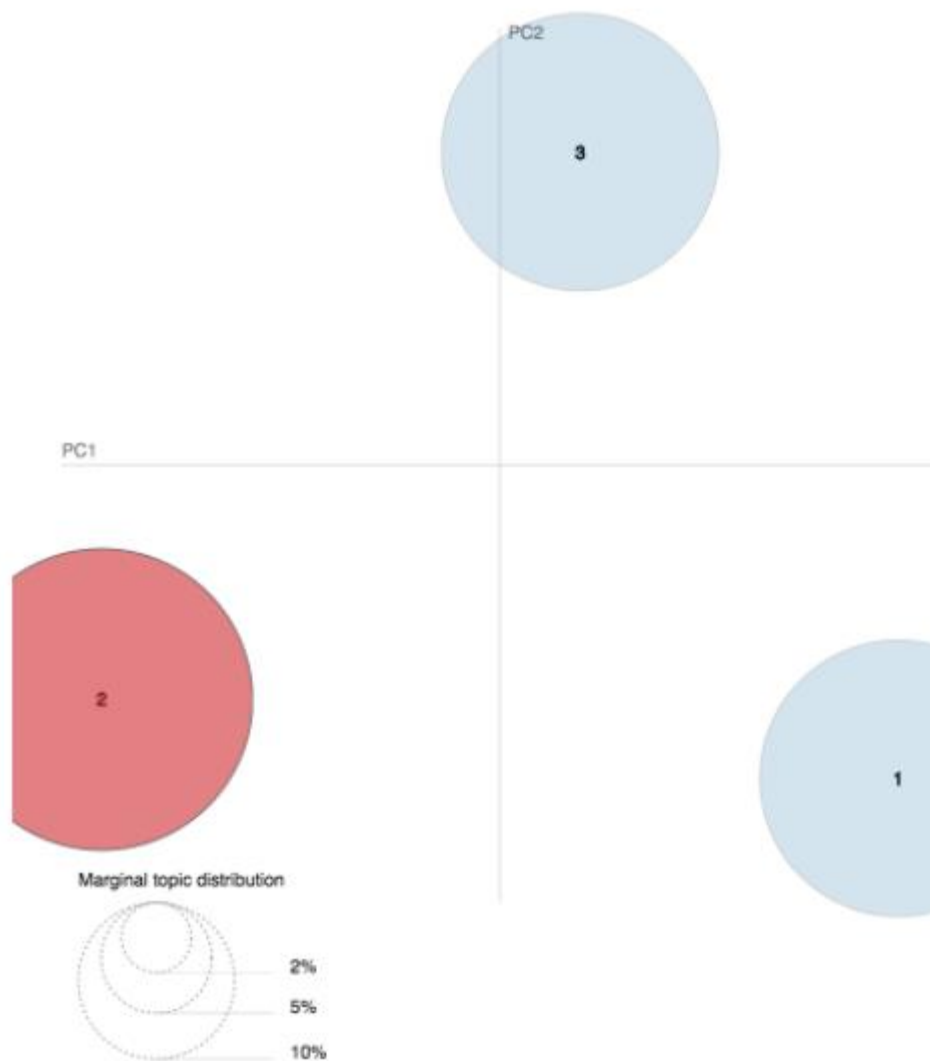
Clear Topic

Slide to adjust relevance metric:(2)

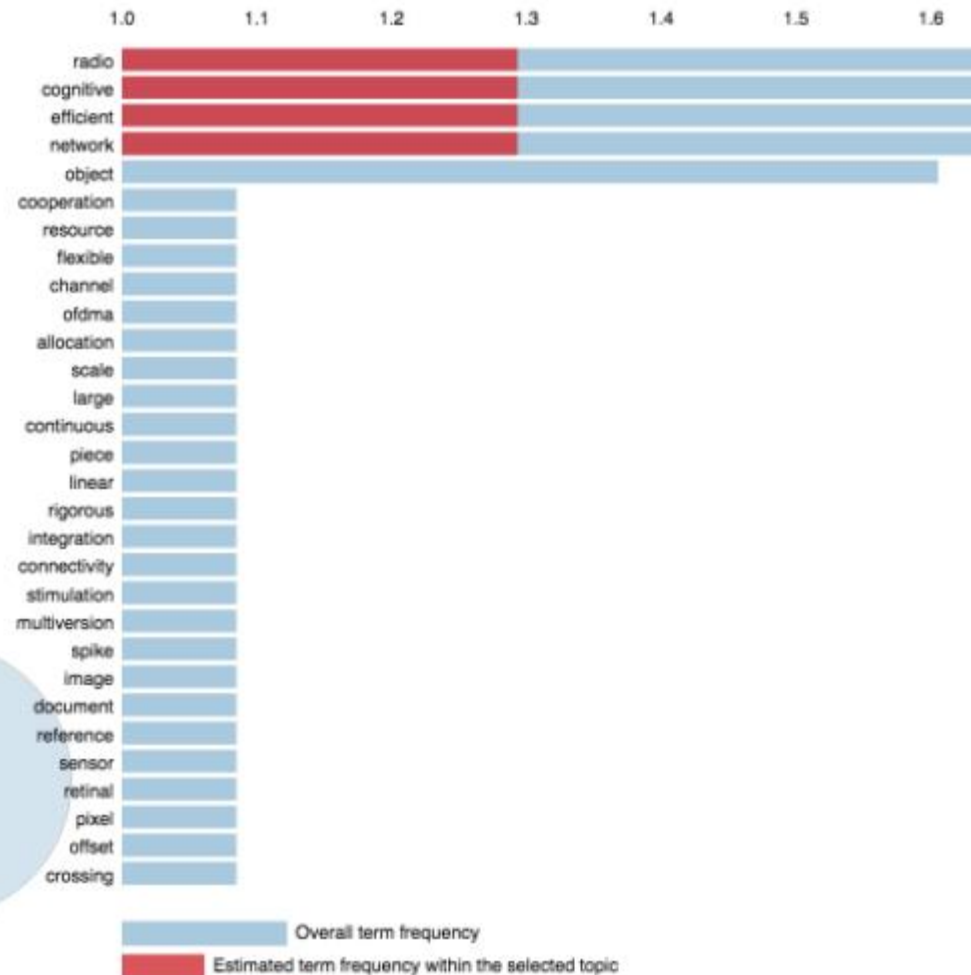
$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 2 (37.2% of tokens)



1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t))]; for topics t; see Chuang et al. (2012)

2. relevance(term w | topic t) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)

# Ferramentas

- ▶ Orange
- ▶ Mallet
- ▶ Python
  - ▶ Gensim + PyLDAvis
- ▶ Stanford Topic Modeling
- ▶ Serviços
  - ▶ jsLDA





# Considerações

- ▶ Ajuste dos tópicos
  - ▶ Não há método universal
  - ▶ Métricas que quantificam a qualidade não são suficientes
  - ▶ Análise qualitativa dos tópicos
- ▶ Quanto aos métodos
  - ▶ Recomenda-se o uso do LDA ou HLDA
- ▶ Pre processamento
  - ▶ Um grande diferencial muitas vezes ignorado por preguiça