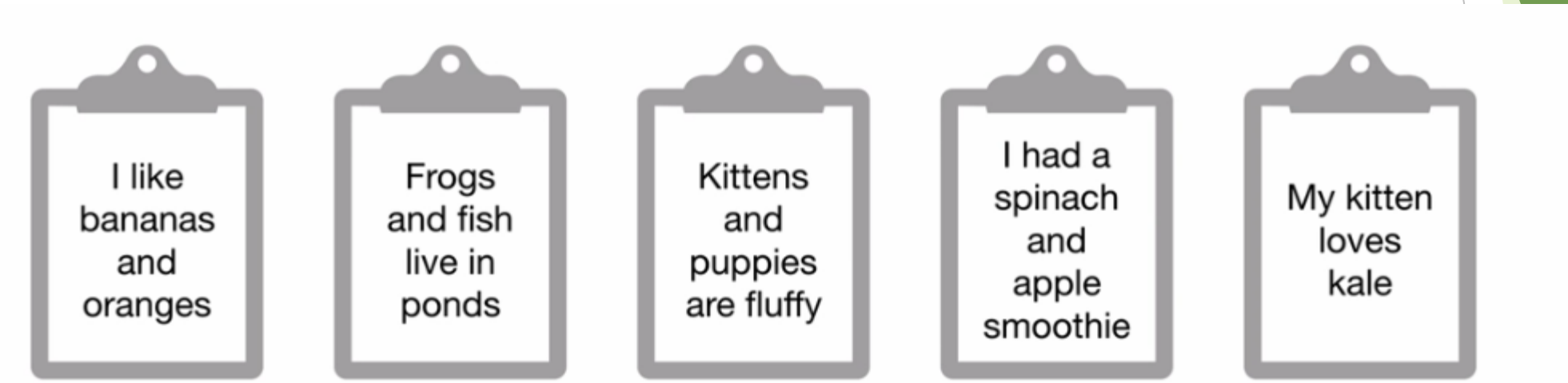
Modelagem de Tópicos

Prof. Leandro Alvim, D. Sc.

Agenda

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

Modelagem de Tópicos



Modelagem de Tópicos

Modelagem de Tópicos

Topic A: 40% banana, 30% kale, 10% breakfast…

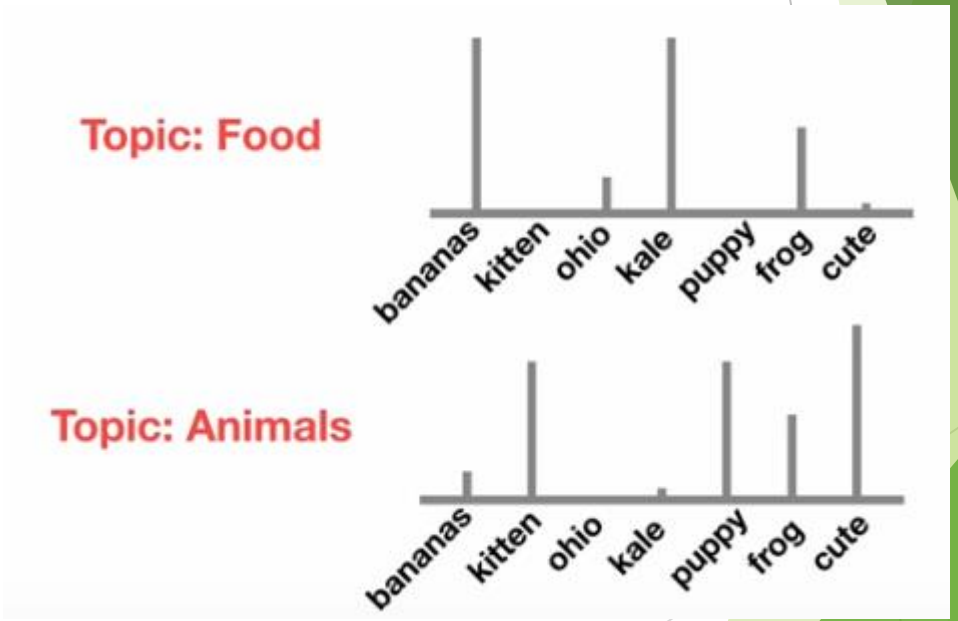
Food

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

Animals

Modelagem de Tópicos

Modelagem de Tópicos

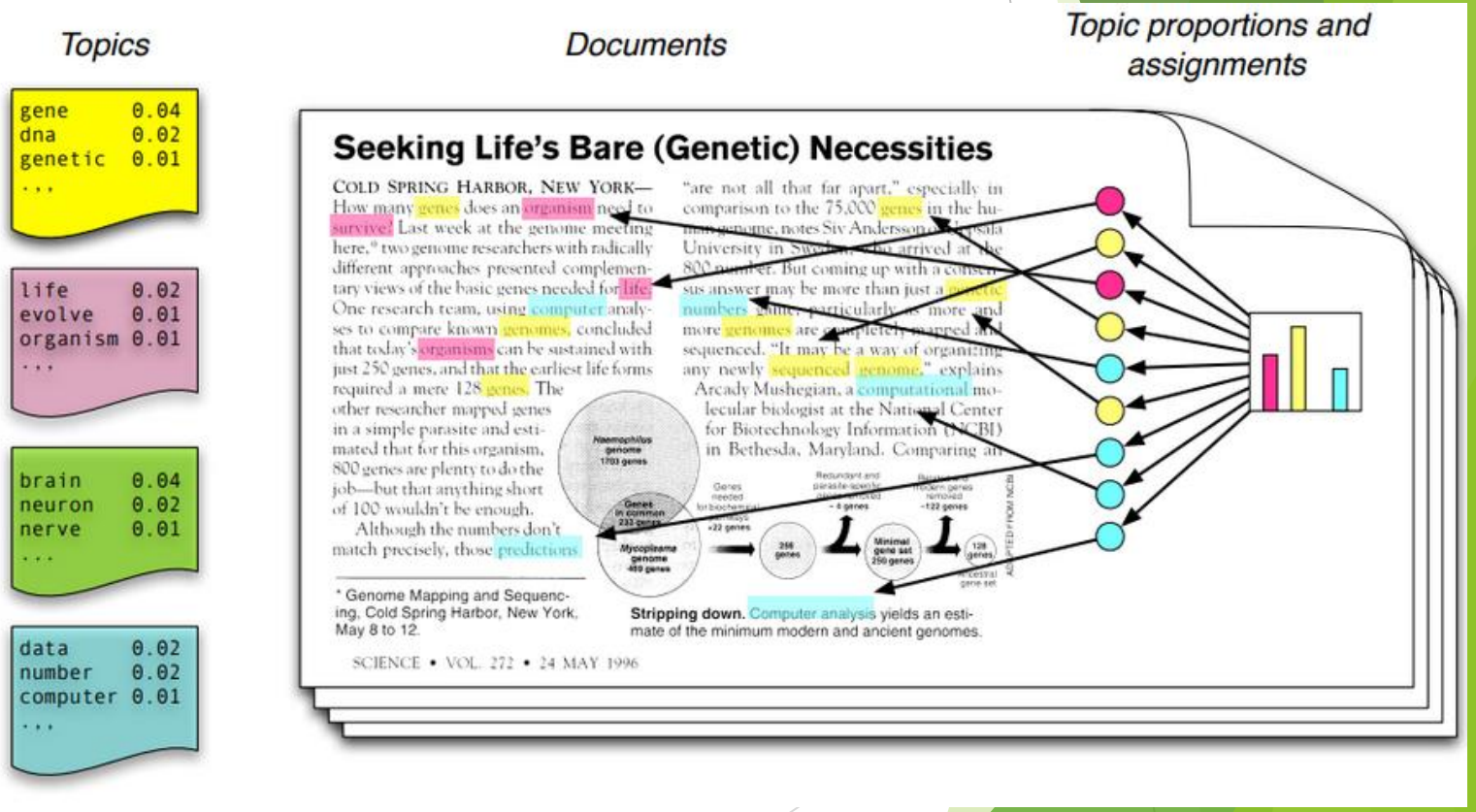


Métodos

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

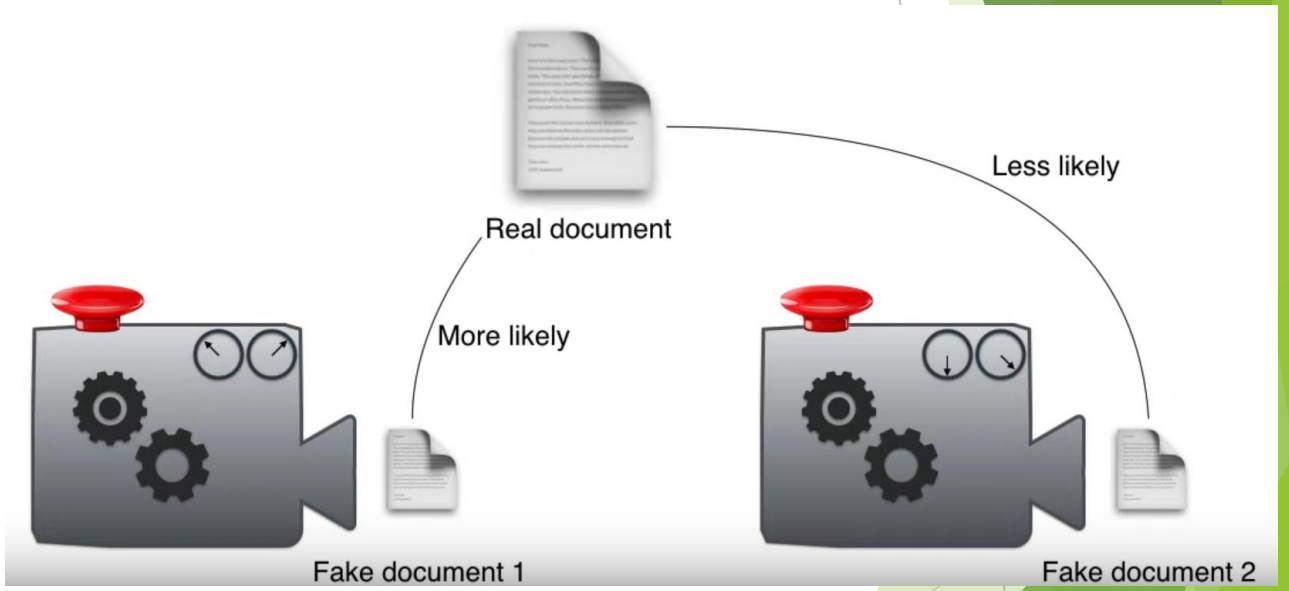
Resumo do LDA

- 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



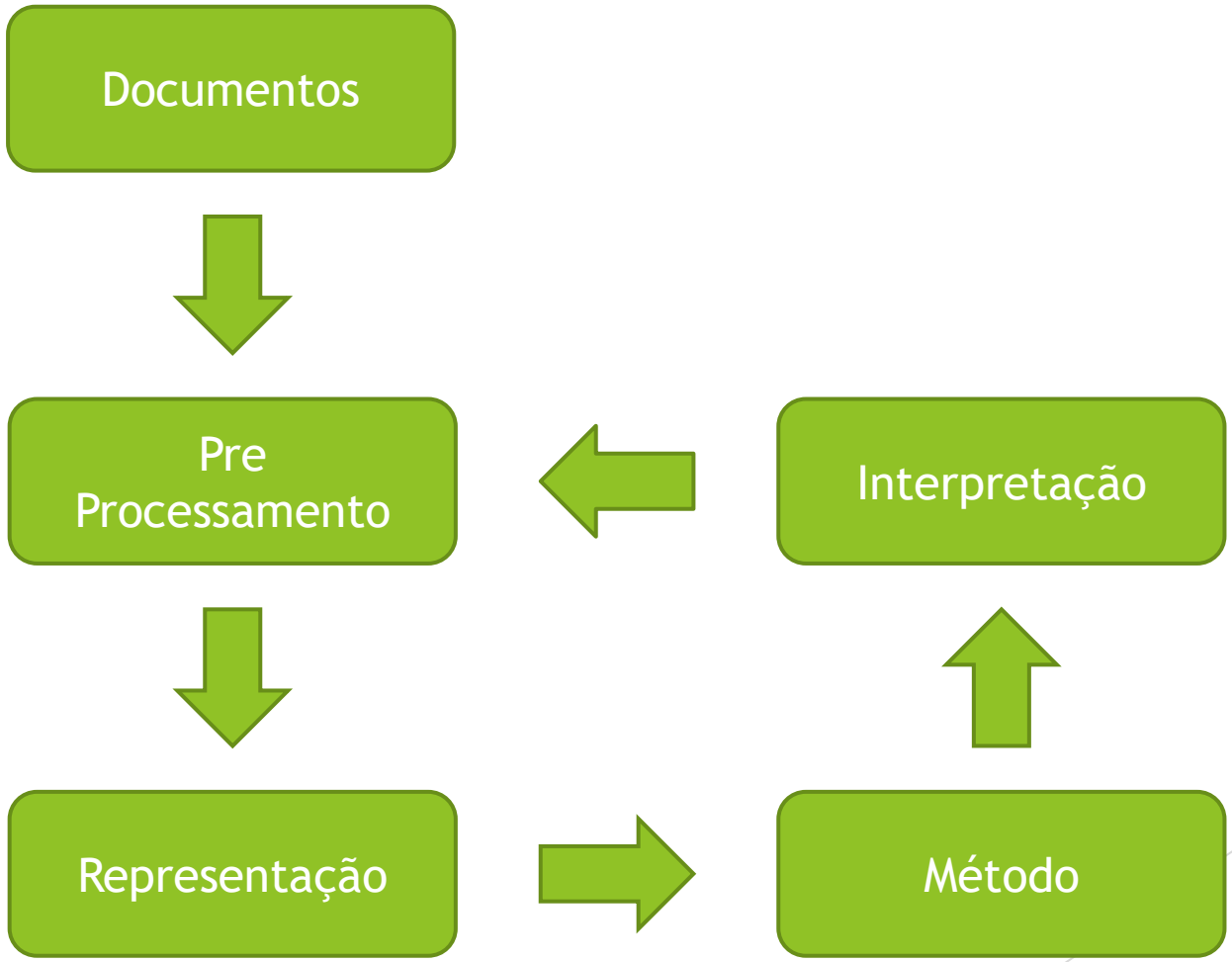
Resumo do LDA

- 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

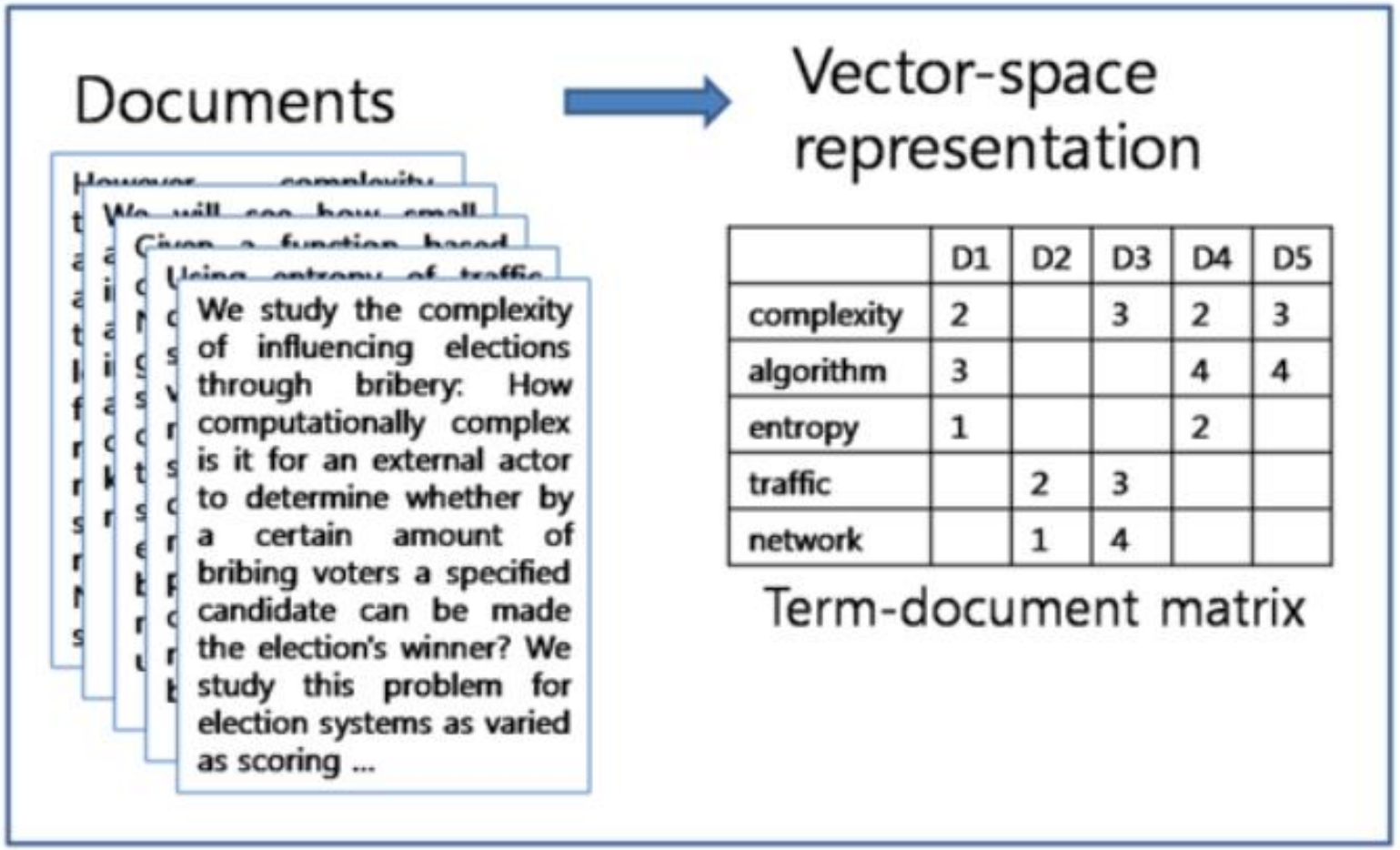


Metodologia – Pre Processamento

- 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



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Interpretação

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

Exemplos de Interpretação

Y Hacker News

Document-Topic Distribution

Topic	P(T|D)
58	0.19
38	0.14
16	0.06
…	…

Sorted Topic-Term Distributions

58	38	16
app	game	language
developer	player	code
mobile	video game	programming
user	gaming	java
app store	developer	programmer

Y Hacker News

Document-Topic Distribution

Topic	P(T | D)
mobile apps	0.19
38	0.14
16	0.06
…	…

Sorted Topic-Term Distributions

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

Y Hacker News

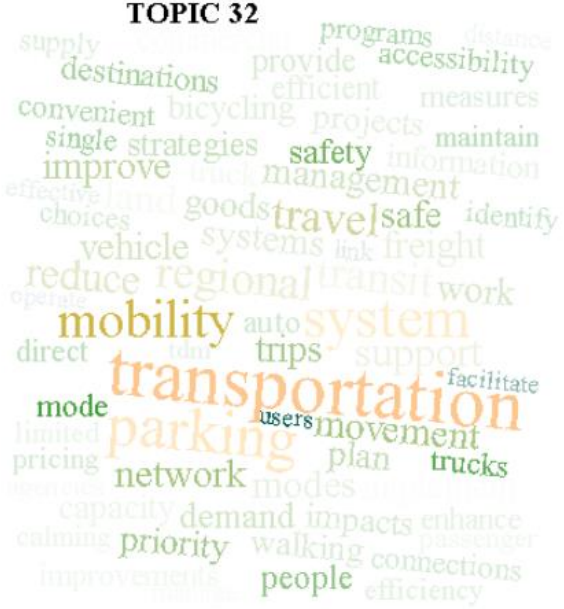
Document-Topic Distribution

Topic	P(T | D)
mobile apps	0.19
video games	0.14
programming	0.06
…	…

Sorted Topic-Term Distributions

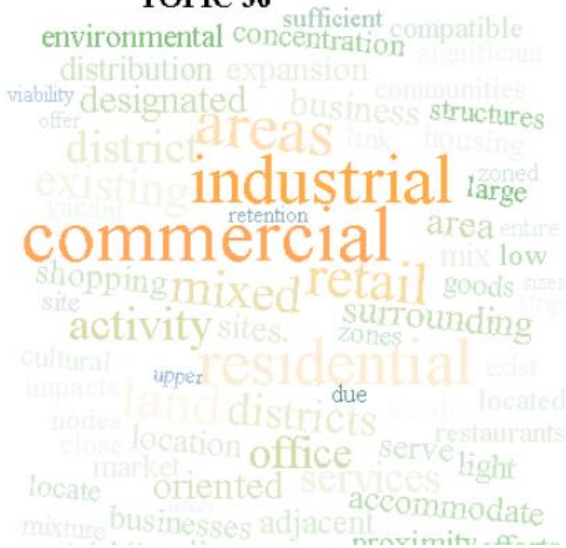
mobile apps	video games	programming
app	game	language
developer	player	code
mobile	video game	programming
user	gaming	java
app store	developer	programmer

TOPIC 32

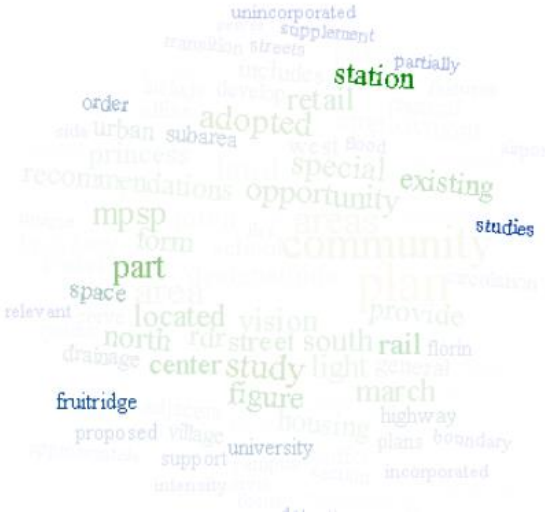


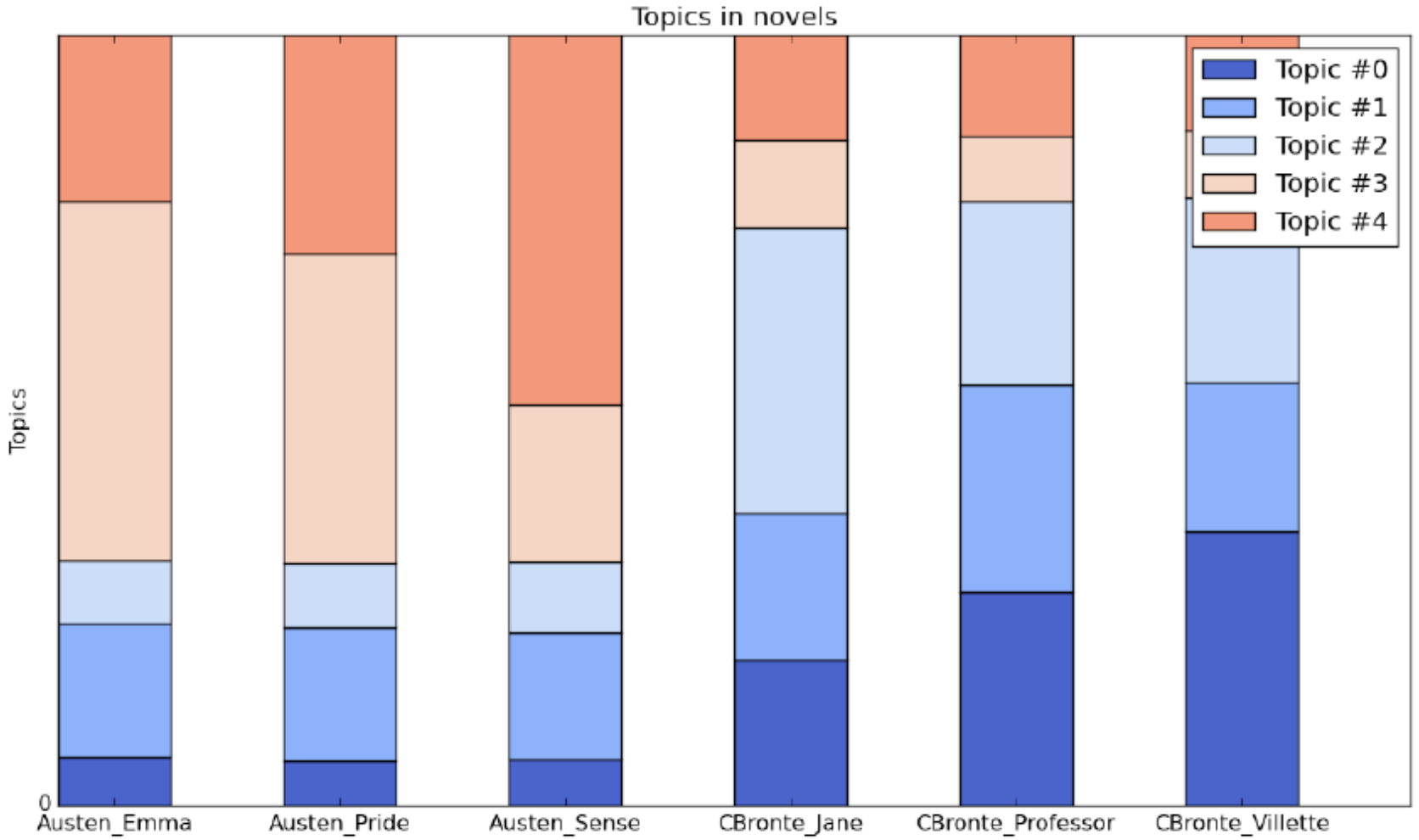
TOPIC 33

TOPIC 36

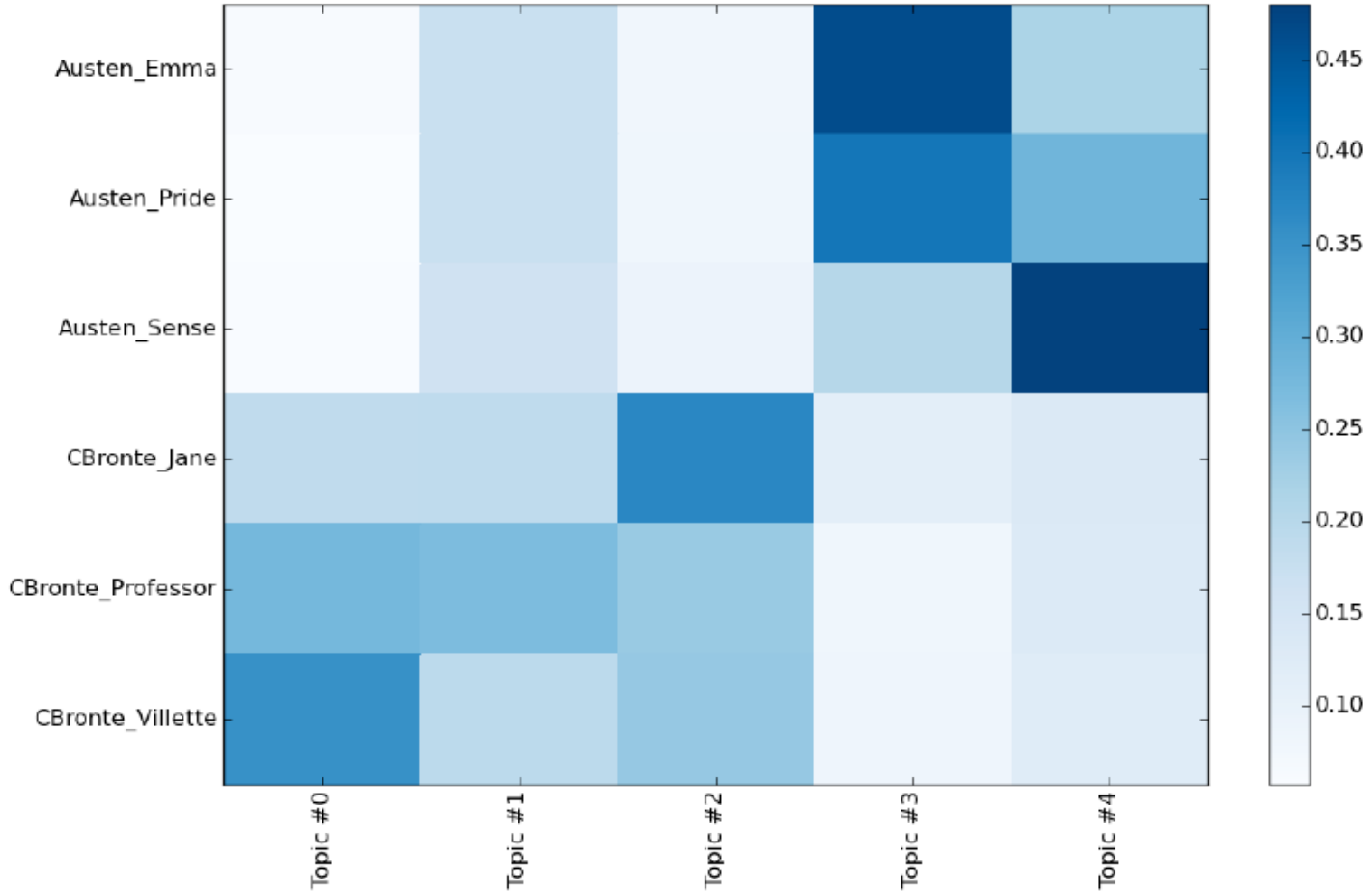


TOPIC 37





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

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

Distinção

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

P(T | game)	0.14	0.14	0.71
P(T | apple)	0.25	0.50	0.25
P(T | angry birds)	0.03	0.03	0.94
P(T | pyhton)	0.77	0.08	0.15
P(T)	0.33	0.23	0.45

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

Distinção

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

P(T | game)	0.14	0.14	0.71
P(T | apple)	0.25	0.50	0.25
P(T | angry birds)	0.03	0.03	0.94
P(T | pyhton)	0.77	0.08	0.15
P(T)	0.33	0.23	0.45

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

Distinção

	coding	tech news	video games	distinctiveness
game	10	10	50	
apple	20	40	20	
angry birds	1	1	30	0.56
python	50	5	10	
P(T | game)	0.14	0.14	0.71	
P(T | apple)	0.25	0.50	0.25	
P(T | angry birds)	0.03	0.03	0.94	
P(T | pyhton)	0.77	0.08	0.15	
P(T)	0.33	0.23	0.45	

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

Distinção

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

P(T | game)	0.14	0.14	0.71
P(T | apple)	0.25	0.50	0.25
P(T | angry birds)	0.03	0.03	0.94
P(T | pyhton)	0.77	0.08	0.15
P(T)	0.33	0.23	0.45

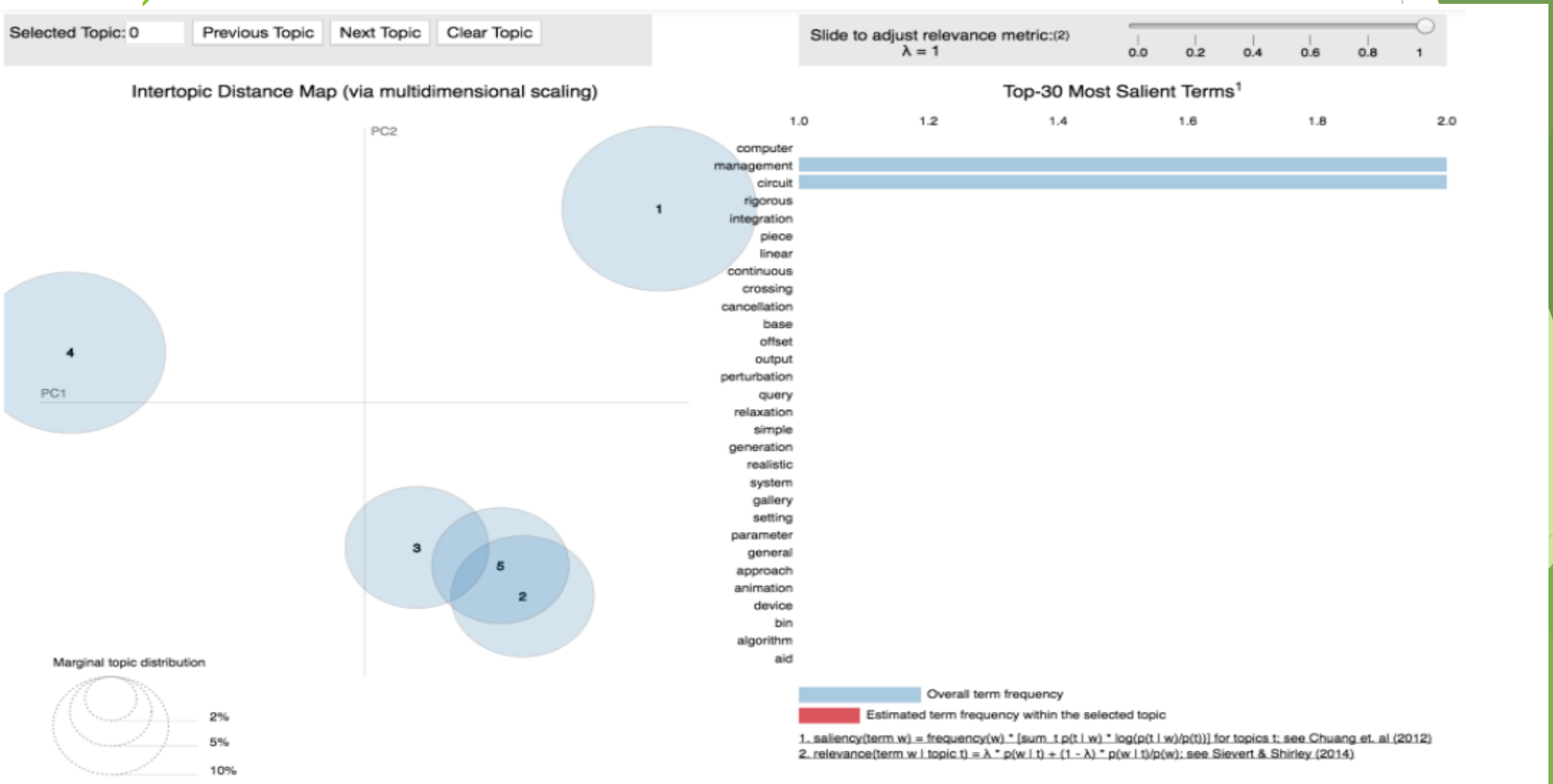
$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

Saliência

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

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

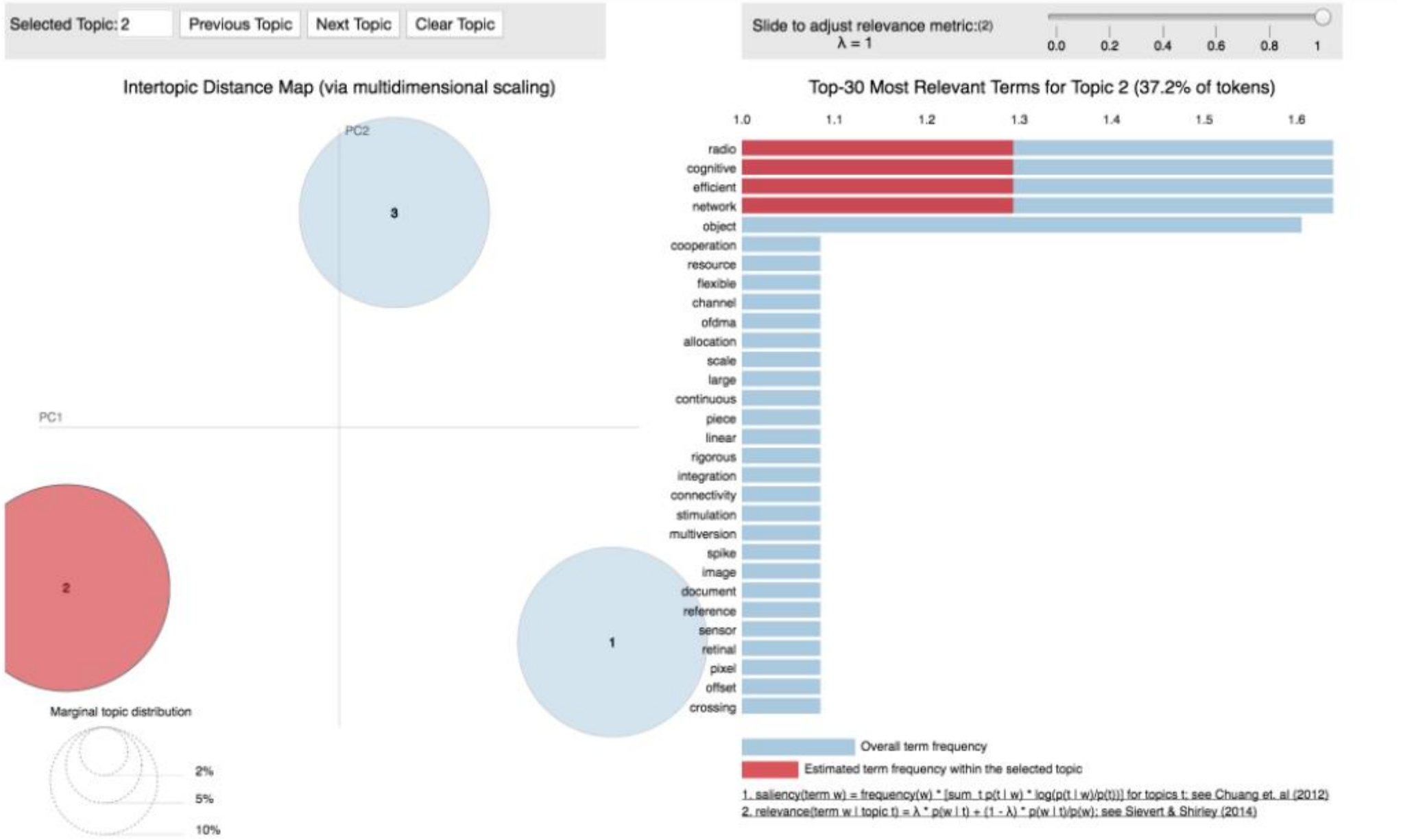
Exemplo de ordenação por saliência (outra base)



Análise local do Termo

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

LDAvis: A method for visualizing and interpreting topics, C. Sievert, K. E. Shirley


Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
2
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (37.2% of tokens)
1.0
1.1
1.2
1.3
1.4
1.5
1.6
radio
cognitive
efficient
network
object
cooperation
resource
flexible
channel
ofdma
allocation
scale
large
continuous
piece
linear
rigorous
integration
connectivity
stimulation
multiversion
spike
image
document
reference
sensor
retinal
pixel
offset
crossing
Overall term frequency
Estimated term frequency within the selected topic
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) = λ * p(w | t) + (1 - λ) * 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