

Agrupamento de Documentos

Prof. Leandro Alvim, D.Sc.

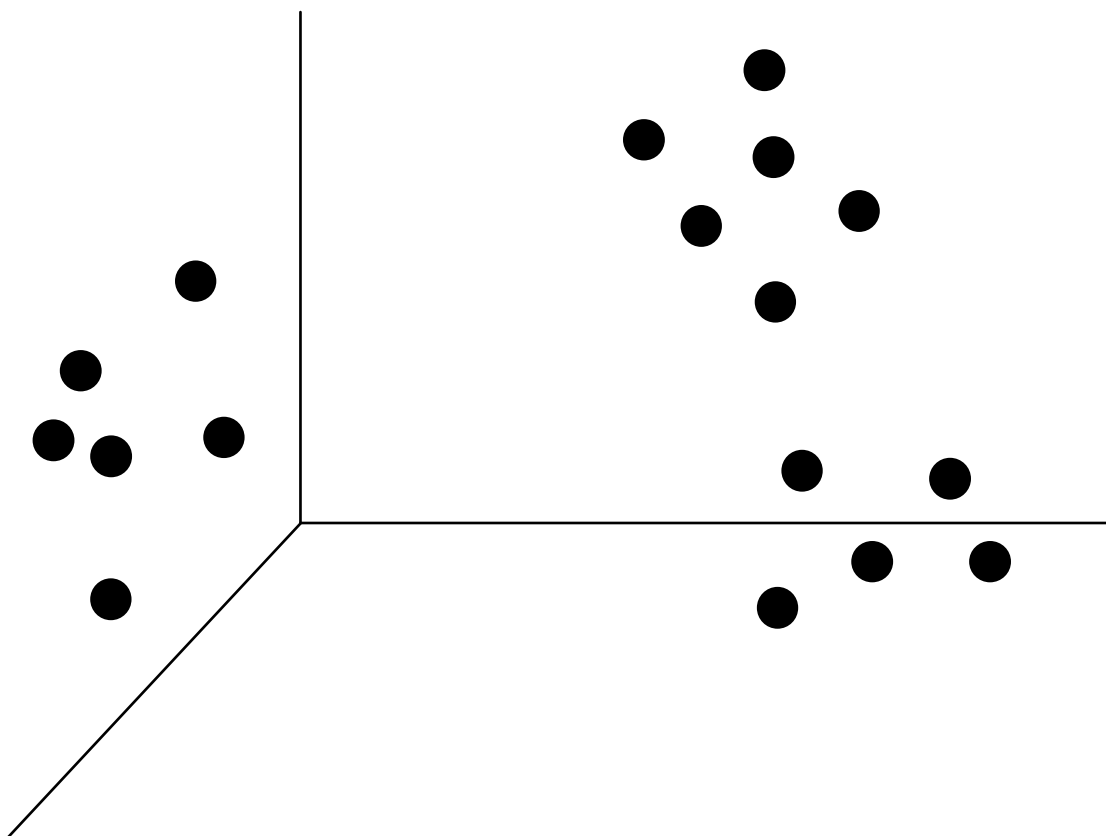
Agenda

- ▶ Motivações
- ▶ O que é ?
- ▶ Metodologia
- ▶ Análise
- ▶ Ferramentas
- ▶ Considerações
- ▶ Exemplos de trabalhos acadêmicos

Motivação

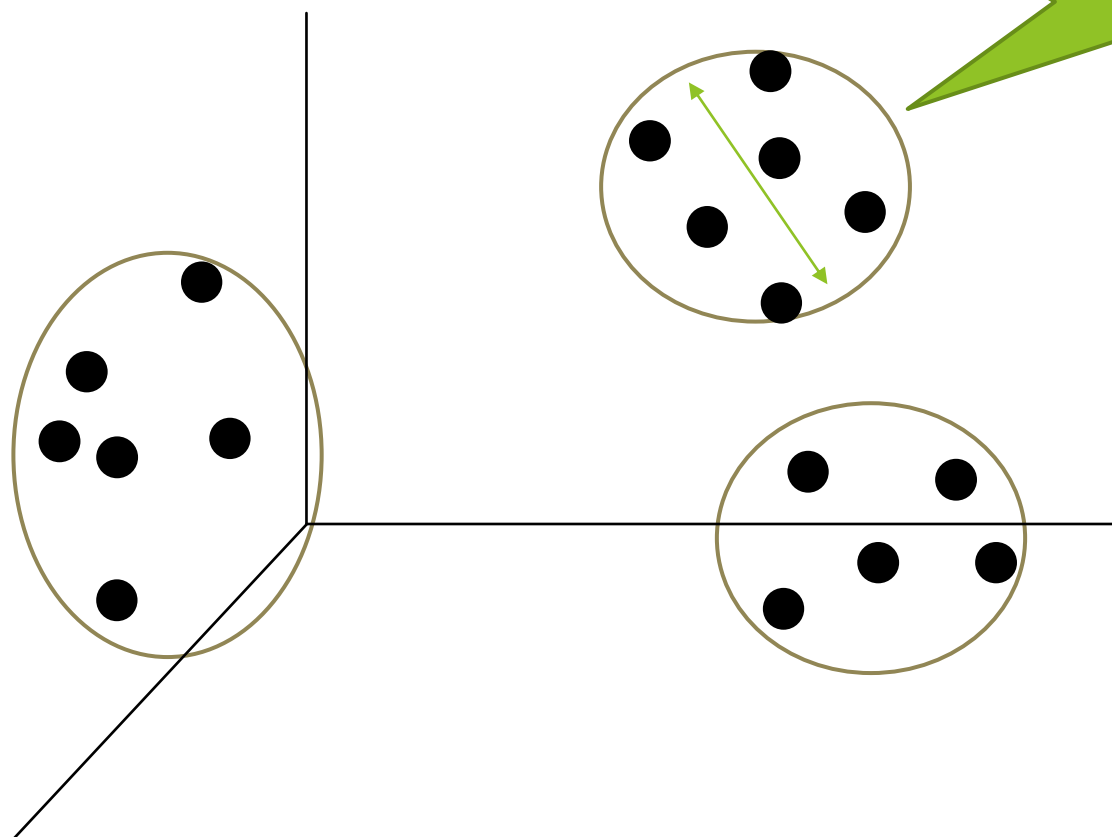
- ▶ Encontrar grupos
 - ▶ Padrões similares dentro do grupo
 - ▶ Padrões distintos entre grupos
- ▶ Mas por que ?
 - ▶ Novas descobertas para grandes volumes de dados
 - ▶ Leitura distante

O que é ?



Idade	Peso	Altura
20	80	180
30	98	190
63	90	170
...

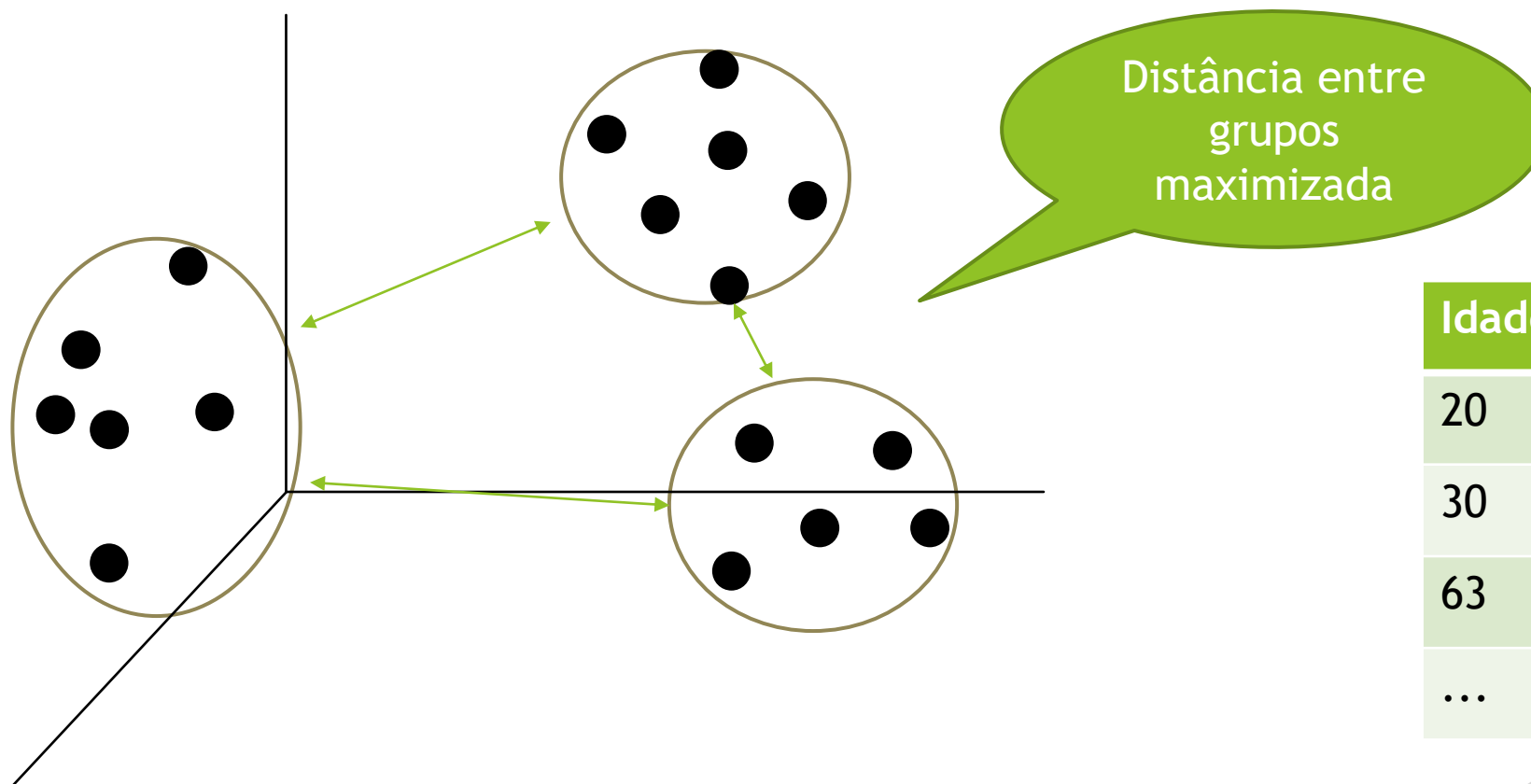
O que é ?



Distância entre
membros
minimizada

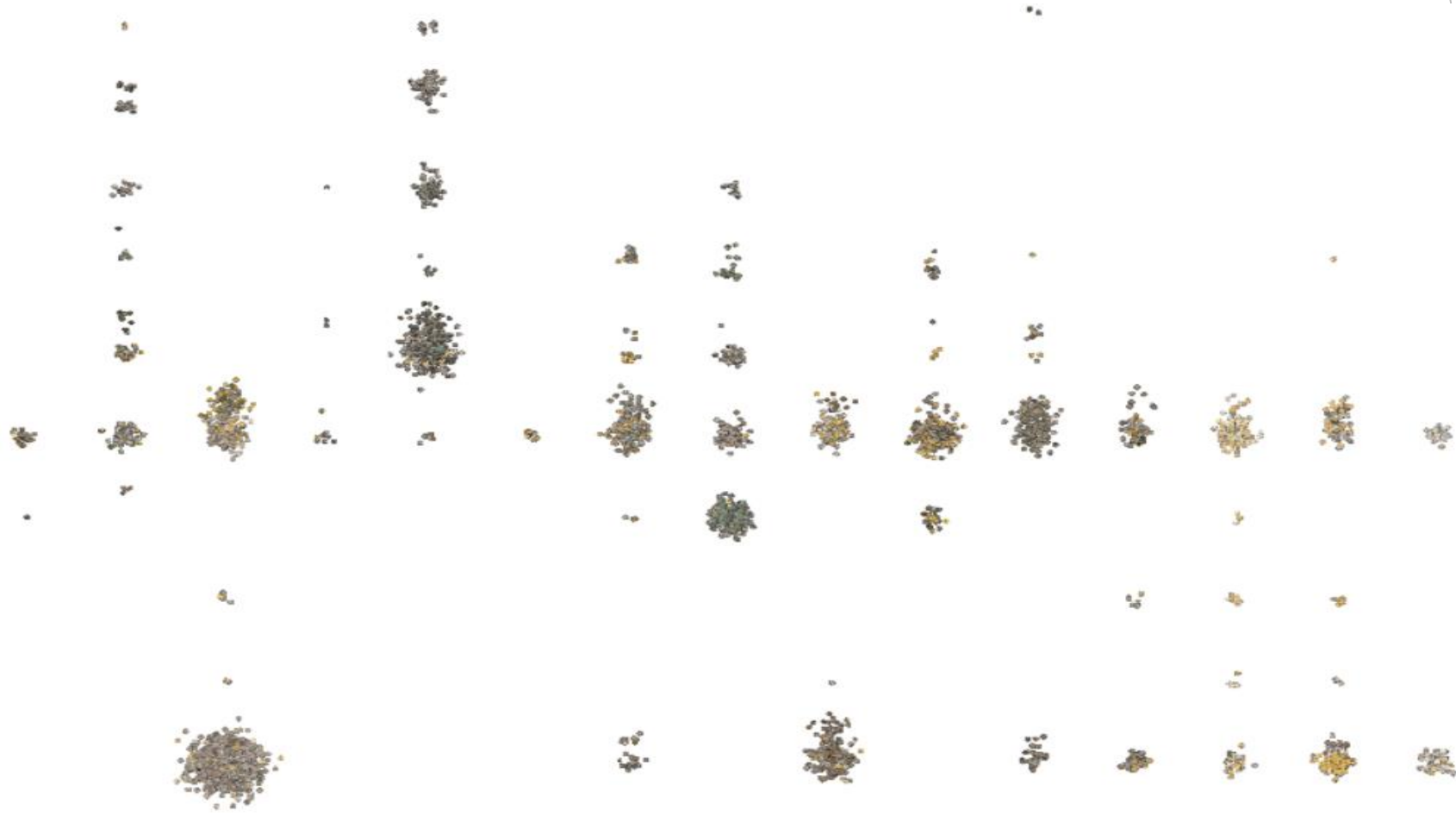
Idade	Peso	Altura	Grupo
20	80	180	2
30	98	190	3
63	90	170	1
...

O que é ?



Idade	Peso	Altura	Grupo
20	80	180	2
30	98	190	3
63	90	170	1
...

Agrupando Moedas



► [COINS – A journey through a rich cultural collection \(fh-potsdam.de\)](http://fh-potsdam.de)

Como agrupar documentos ?



Topic Modeling

T_1	T_2	\dots	T_m
Word ₁₁ , pr ₁₁	Word ₂₁ , pr ₂₁	\dots	Word _{m1} , pr _{m1}
Word ₁₂ , pr ₁₂	Word ₂₂ , pr ₂₂	\dots	Word _{m2} , pr _{m2}
\vdots	\vdots	\vdots	\vdots
Word _{1k} , pr _{1k}	Word _{2k} , pr _{2k}	\dots	Word _{mk} , pr _{mk}

Document Clustering

C_1	C_2	\dots	C_m
D_a	D_i	\dots	D_z
D_b	D_j	\dots	D_a
\vdots	\vdots	\vdots	\vdots
D_c	D_k	\dots	D_s

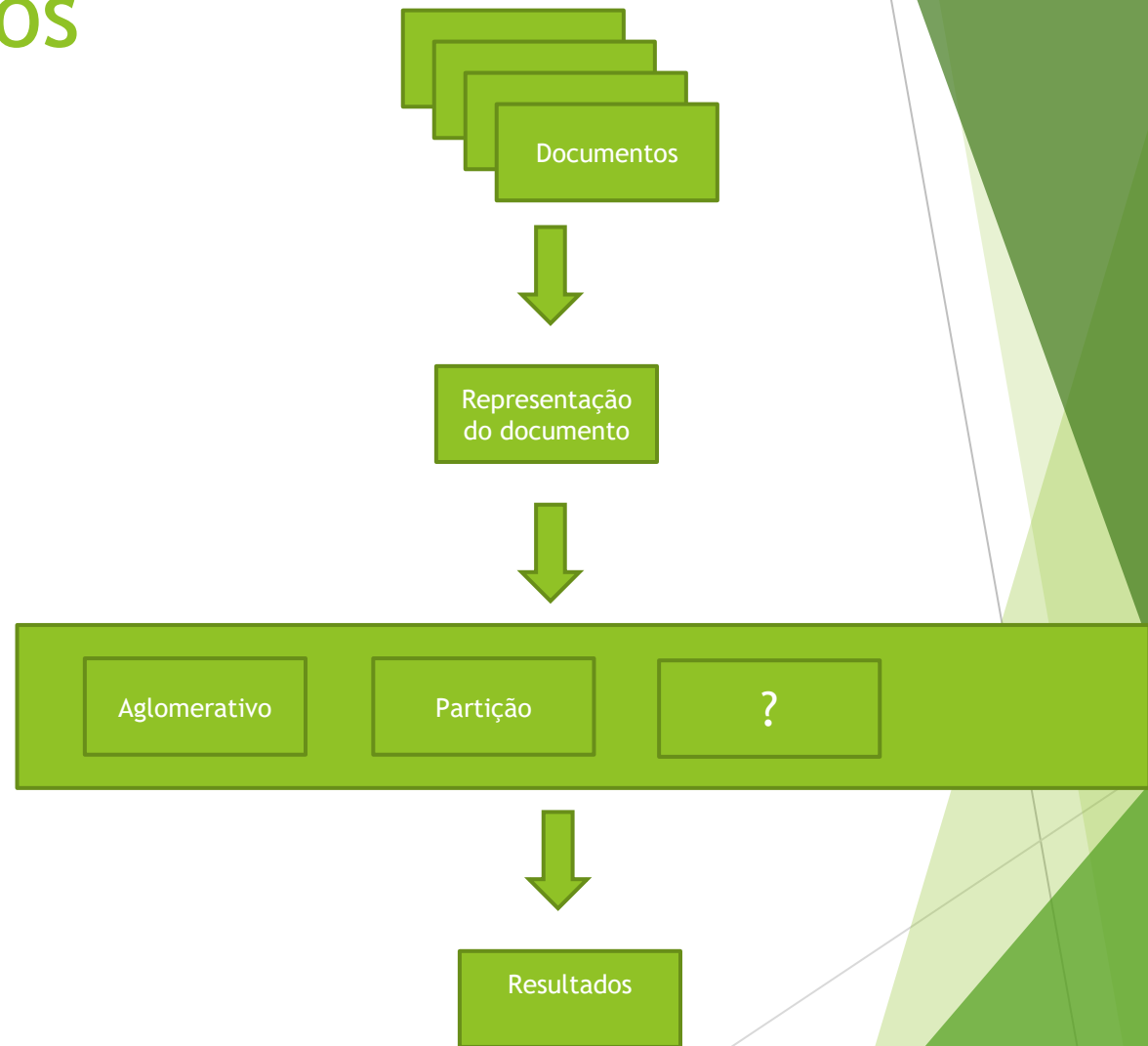
Como agrupar documentos ?

- ▶ Aparato de agrupamento
 - ▶ Estruturado
 - ▶ Numérico

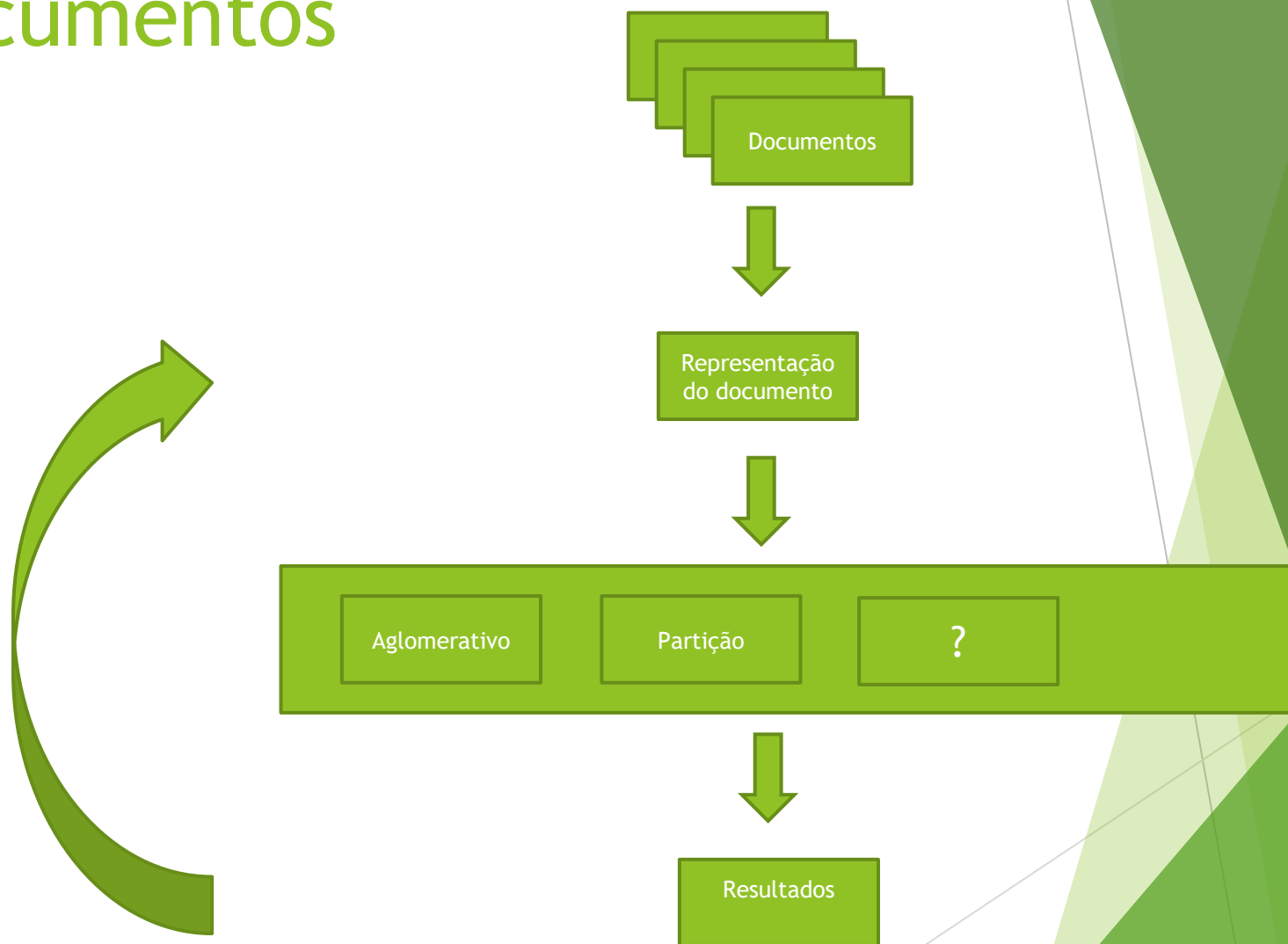
Idade	Peso	Altura	Grupo
20	80	180	2
30	98	190	3
63	90	170	1
...

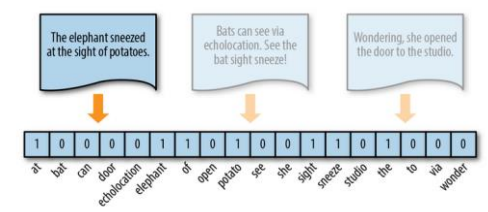
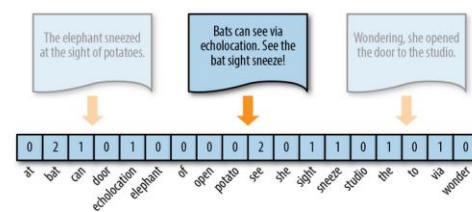
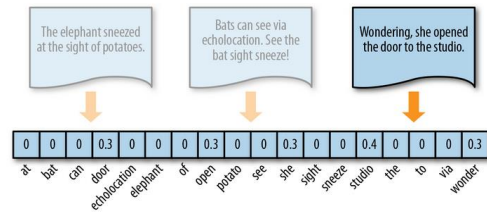
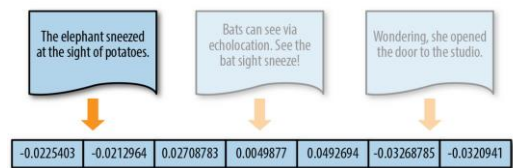
Agrupando Documentos

- ▶ Aparato de agrupamento
 - ▶ Estruturado
 - ▶ Numérico
- ▶ Documentos
 - ▶ Não Estruturado
 - ▶ Texto



Agrupando Documentos





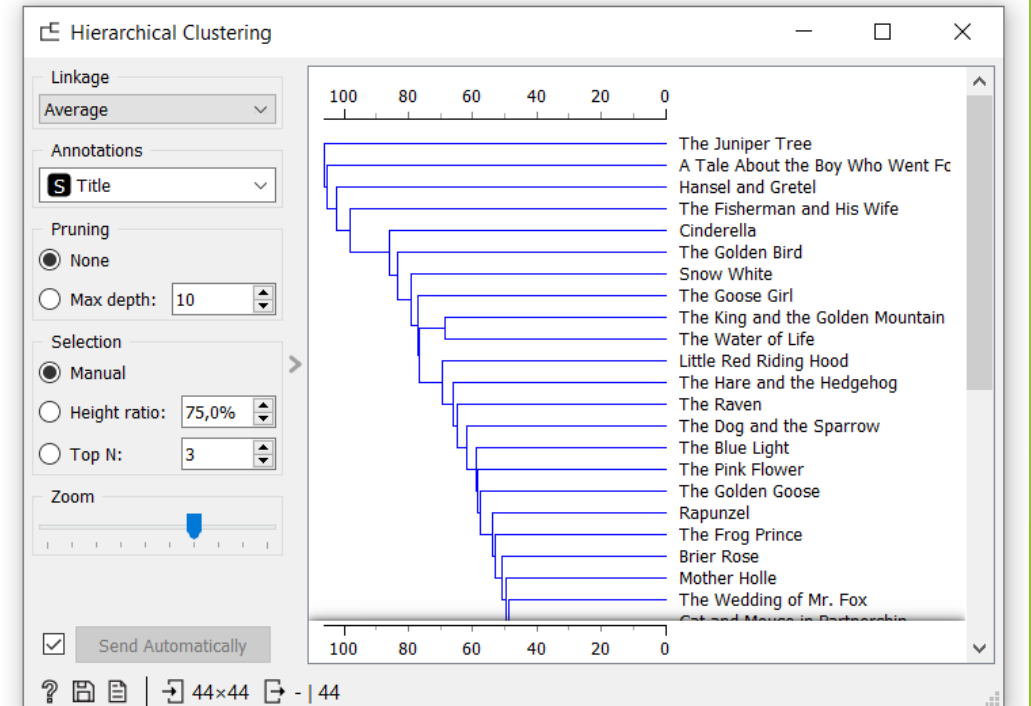
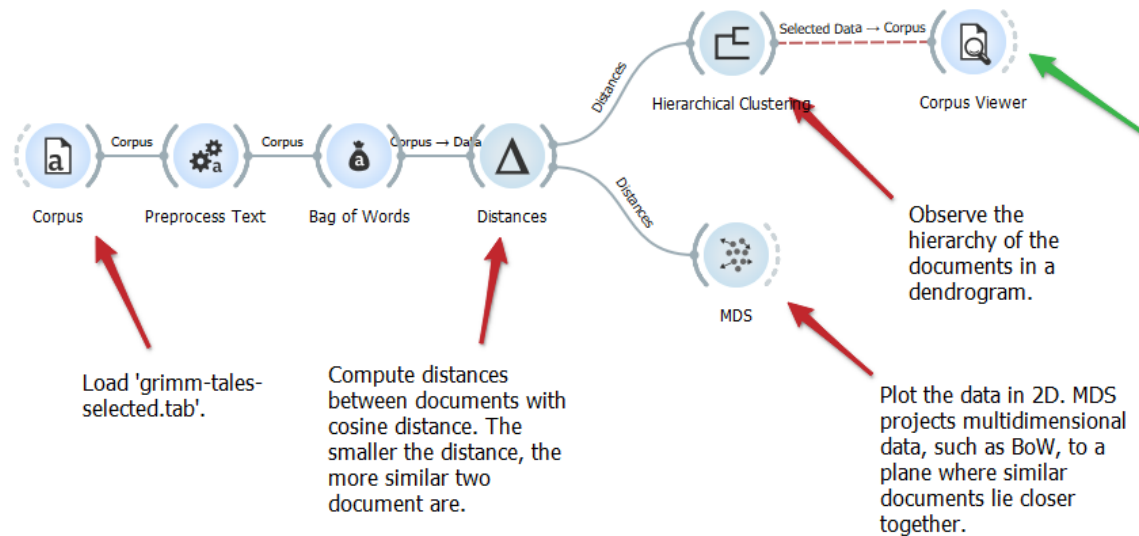
Formas de Representação

- Bag of Words
- One-Hot-Encoding
- TF-IDF
- Embeddings

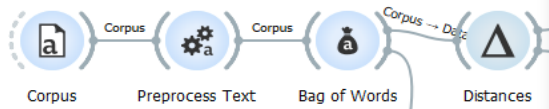
Tipos de Agrupamento

- ▶ Hierárquico
 - ▶ Gera uma hierarquia de subconjuntos
 - ▶ Aglomerativo ou Divisivo
 - ▶ Quando é útil extrair uma taxonomia
- ▶ Partição
 - ▶ Constrói k ($k < n$) partições
 - ▶ Não há hierarquia
- ▶ Densidade
 - ▶ Cada ponto deve ter um número mínimo de vizinhos
 - ▶ Consegue achar grupos com formas não convexas
 - ▶ Útil para identificação de anomalias

Hierárquico



Partição



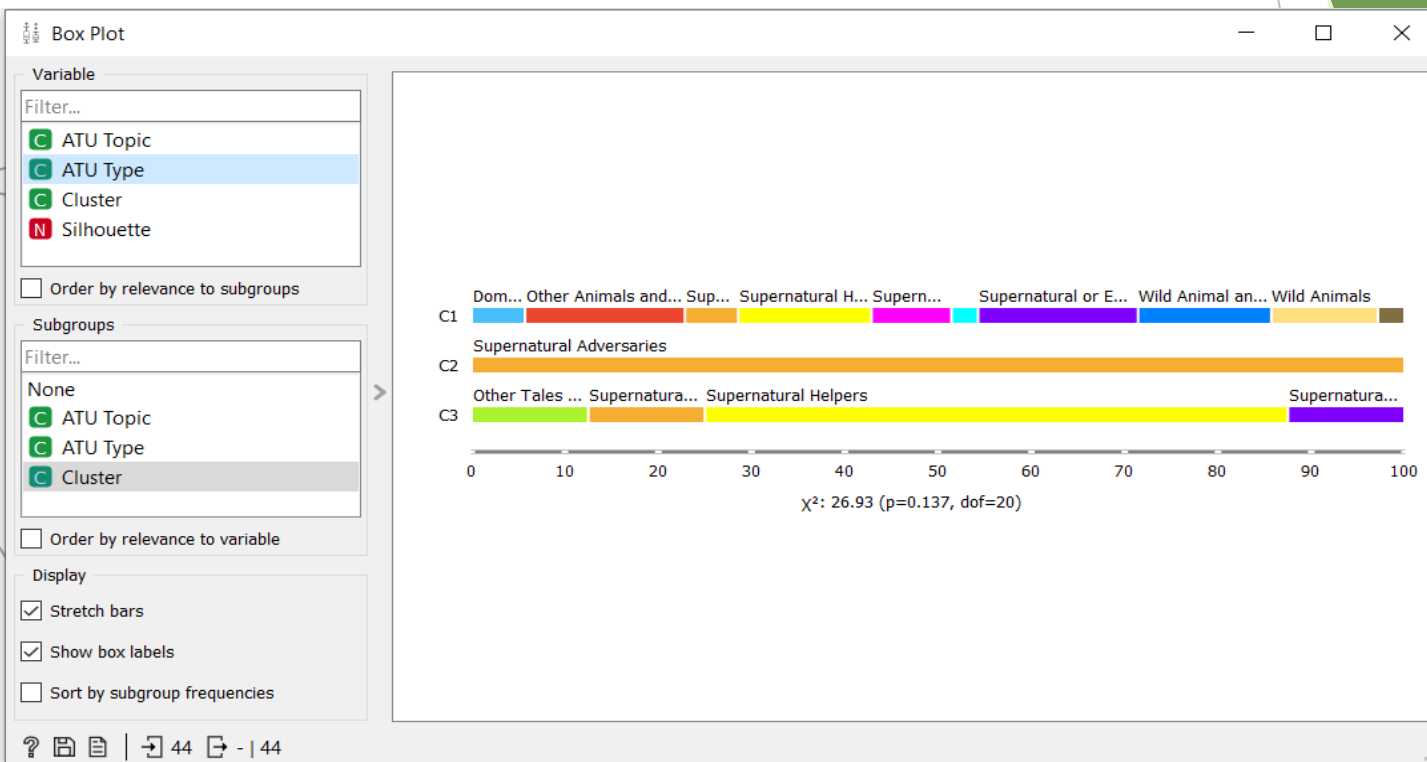
Load 'grimm-tales-selected.tab'.

Corpus → Data



k-Means

Data

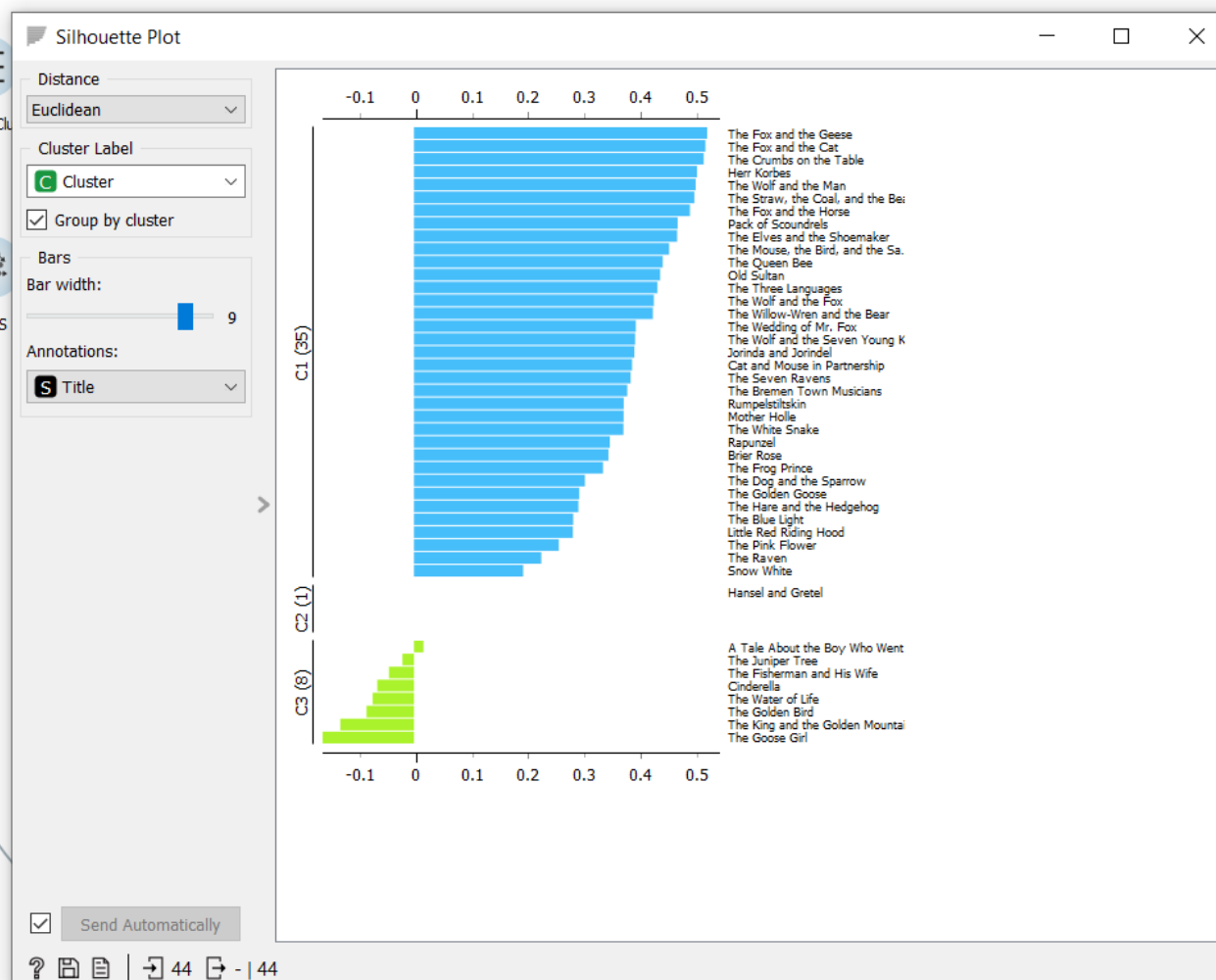
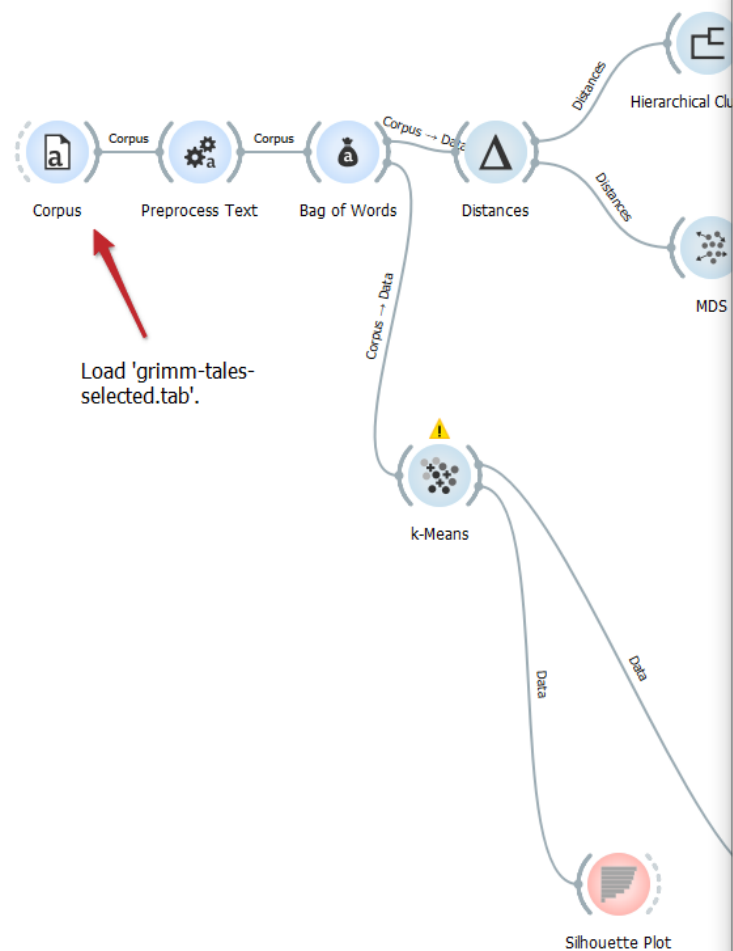


Silhouette Plot



Box Plot

Partição



Análise Qualitativa

Corpus Viewer (1)

Info

Tokens: 27607
Types: 3719
Matching documents: 44/44
Matches: n/a

Search features

- ATU Topic
- Title
- Abstract
- Content
- ATU Numerical
- ATU Type
- Cluster
- Silhouette

Display features

- ATU Topic
- Title
- Abstract
- Content
- ATU Numerical
- ATU Type
- Cluster
- Silhouette

☐ Show Tokens & Tags

☒ Auto send is on

RegExp Filter:

1	A Tale About the Boy Who Went Forth to Learn What...
2	Brier Rose
3	Cat and Mouse in Partnership
4	Cinderella
5	Hansel and Gretel
6	Herr Korbes
7	Jorinda and Jorindel
8	Little Red Riding Hood
9	Mother Holle
10	Old Sultan
11	Pack of Scoundrels
12	Rapunzel
13	Rumpelstiltskin
14	Snow White
15	The Blue Light
16	The Bremen Town Musicians
17	The Crumbs on the Table

ATU Topic: Tales of Magic

Title: A Tale About the Boy Who Went Forth to Learn What Fear Was

Abstract: A simple boy who just wants to be frightened.

Content: A certain father had two sons, the elder of who was smart and sensible, and could do everything, but the younger was stupid and could neither learn nor understand anything, and when people saw him they said: 'There's a fellow who will give his father some trouble!' When anything had to be done, it was always the elder who was forced to do it; but if his father bade him fetch anything when it was late, or in the night-time, and the way led through the churchyard, or any other dismal place, he answered: 'Oh, no father, I'll not go there, it makes me shudder!' for he was afraid. Or when stories were told by the fire at night which made the flesh creep, the listeners sometimes said: 'Oh, it makes us shudder!' The younger sat in a corner and listened with the rest of them, and could not imagine what they could mean. 'They are always saying: "It makes me shudder, it makes me shudder!" It does not make me shudder,' thought he. 'That, too, must be an art of which I understand nothing!' Now it came to pass that his father said to him one day: 'Hearken to me, you fellow in the corner there, you are growing tall and strong, and you too must learn something by which you can earn your bread. Look how your brother works, but you do not even earn your salt.' 'Well, father,' he replied, 'I am quite willing to learn something--indeed, if it could but be managed, I should like to learn how to shudder. I don't understand that at all yet.' The elder brother smiled when he heard that, and thought to himself: 'Goodness, what a blockhead that brother of mine is! He will never be good for anything as long as he lives! He who wants to be a sickle must bend himself betimes.' The father sighed, and answered him: 'You shall soon learn what it is to shudder, but you will not earn your bread by that.' Soon after this the sexton came to the house on a visit, and the father bewailed his trouble, and told him how his younger son was so backward in every respect that he knew nothing and learnt nothing. 'Just think,' said he, 'when I asked him how he was going to earn his bread, he actually wanted to learn to shudder.' 'If that be all,' replied the sexton, 'he can learn that with me. Send him to me, and I will soon polish him.' The father was glad to do it, for he thought: 'It will train the boy a little.' The sexton therefore took him into his house, and he had to ring the church bell. After a day or two, the sexton awoke him at midnight, and bade him arise and go up into the church tower and ring the bell. 'You shall soon learn what shuddering is,' thought he, and secretly went there before him; and when the boy was at the top of the tower and turned round, and was just going to take hold of the bell rope, he saw a white figure standing on the stairs opposite the sounding hole. 'Who is there?' cried he, but the figure made no reply, and did not move or stir. 'Give an answer,' cried the boy, 'or take yourself off, you have no business here at night.' The sexton, however, remained standing motionless that the boy might think he was a ghost. The boy cried a second time: 'What do you want here?--speak if you are an honest fellow, or I will throw you down the steps!' The sexton thought: 'He can't mean to be as bad as his words,' uttered no sound and stood as if he were made of stone. Then the boy called to him for the third time, and as that was also to no purpose, he ran against him and pushed the ghost down the stairs, so that it fell down the ten steps and remained lying there in a corner. Thereupon he rang the bell, went home, and without saying a word went to bed, and fell asleep. The sexton's wife waited a long time for her husband, but he did not come back. At length she became uneasy, and awakened the boy, and asked: 'Do you know where my husband is? He climbed up the tower before you did.' 'No, I don't know,' replied the boy, 'but someone was standing by the sounding hole on the other side of the steps, and as he would neither gave an answer nor go away, I took him for a scoundrel, and threw him downstairs. Just go there and you will see if it was he. I should be sorry if it were.' The woman ran away and found her husband, who was lying moaning in the corner, and had broken his leg. She carried him down, and then with loud screams she hastened to the boy's father, 'Your boy,' cried she, 'has been the cause of a great misfortune! He has thrown my husband down the steps so that he broke his leg. Take the good for nothing fellow out of our

Análise Qualitativa

Corpus Viewer (1)

Info
Tokens: 27607
Types: 3719
Matching documents: 44/44
Matches: n/a

Search features

- ATU Topic
- Title
- Abstract
- Content
- ATU Numerical
- ATU Type
- Cluster
- Silhouette

Display features

- ATU Topic
- Title
- Abstract
- Content
- ATU Numerical
- ATU Type
- Cluster
- Silhouette

☐ Show Tokens & Tags

☒ Auto send is on

RegExp Filter:

1	A Tale About the Boy Who Went Forth to Learn What...	ATU Topic: Tales of Magic
2	Brier Rose	Title: A Tale About the Boy Who Went Forth to Learn What Fear Was
3	Cat and Mouse in Partnership	Abstract: A simple boy who just wants to be frightened.
4	Cinderella	Cluster: C3
5	Hansel and Gretel	Silhouette: 0.505589
6	Herr Korbes	
7	Jorinda and Jorindel	
8	Little Red Riding Hood	
9	Mother Holle	
10	Old Sultan	
11	Pack of Scoundrels	
12	Rapunzel	
13	Rumpelstiltskin	
14	Snow White	
15	The Blue Light	
16	The Bremen Town Musicians	
17	The Crumbs on the Table	

? | 44 | 1 | 43 | 44

Ferramentas

- ▶ Orange
 - ▶ Provê um workflow completo e simples
- ▶ Python e Scikit
 - ▶ Exige codificação
 - ▶ Bem documentado
 - ▶ Muitos exemplos

Considerações

- ▶ Pre processamento é importante!
 - ▶ Stemming ou lema
 - ▶ Stop words
 - ▶ Lower case
 - ▶ ...
- ▶ Tente mudar a representação do documento
 - ▶ Embeddings são muito bons e modernos

Considerações

- ▶ Tente diferentes formas de visualização
 - ▶ Leitura distante
 - ▶ Leitura próxima
- ▶ Identifique o que você quer fazer
 - ▶ Taxonomia automática ?
 - ▶ Grupos disjuntos ?
 - ▶ Anomalias ?
- ▶ Se o algoritmo tiver parâmetros
 - ▶ Usem algo que ajuste pela silhueta ou alguma outra função