Agrupamento de Documentos

Prof. Leandro Alvim, D.Sc.

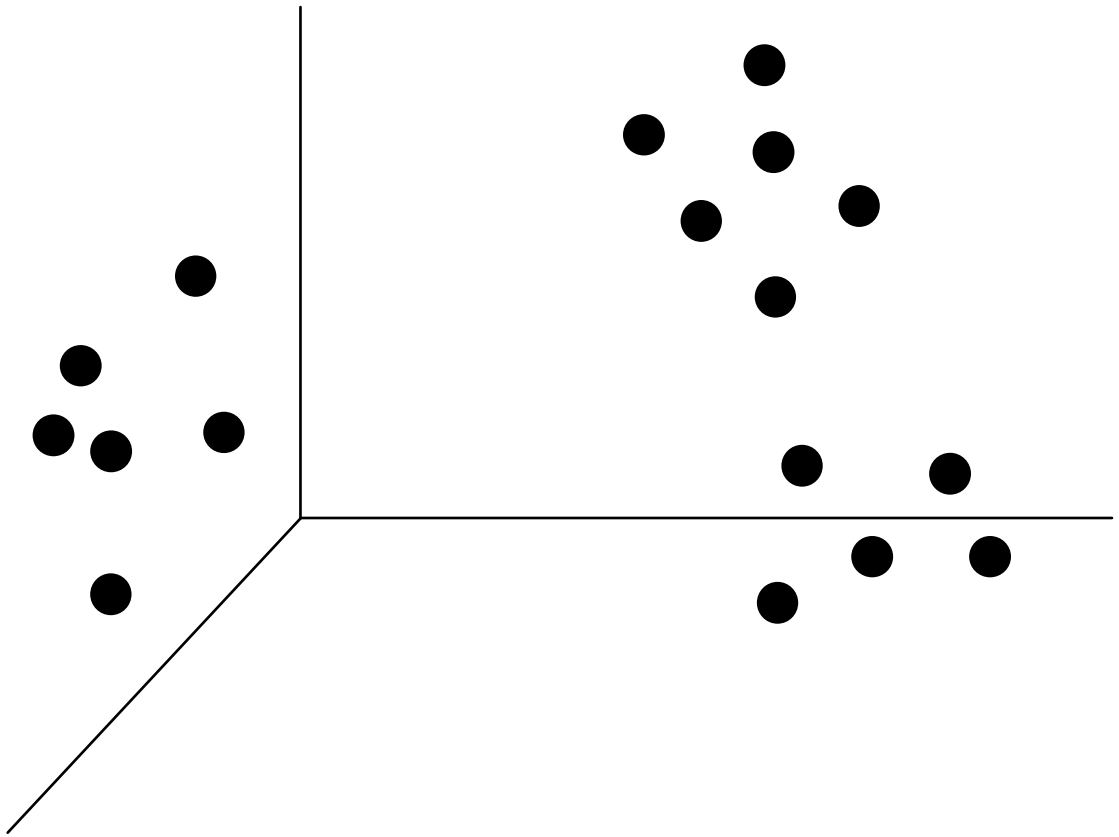
Agenda

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

Motivação

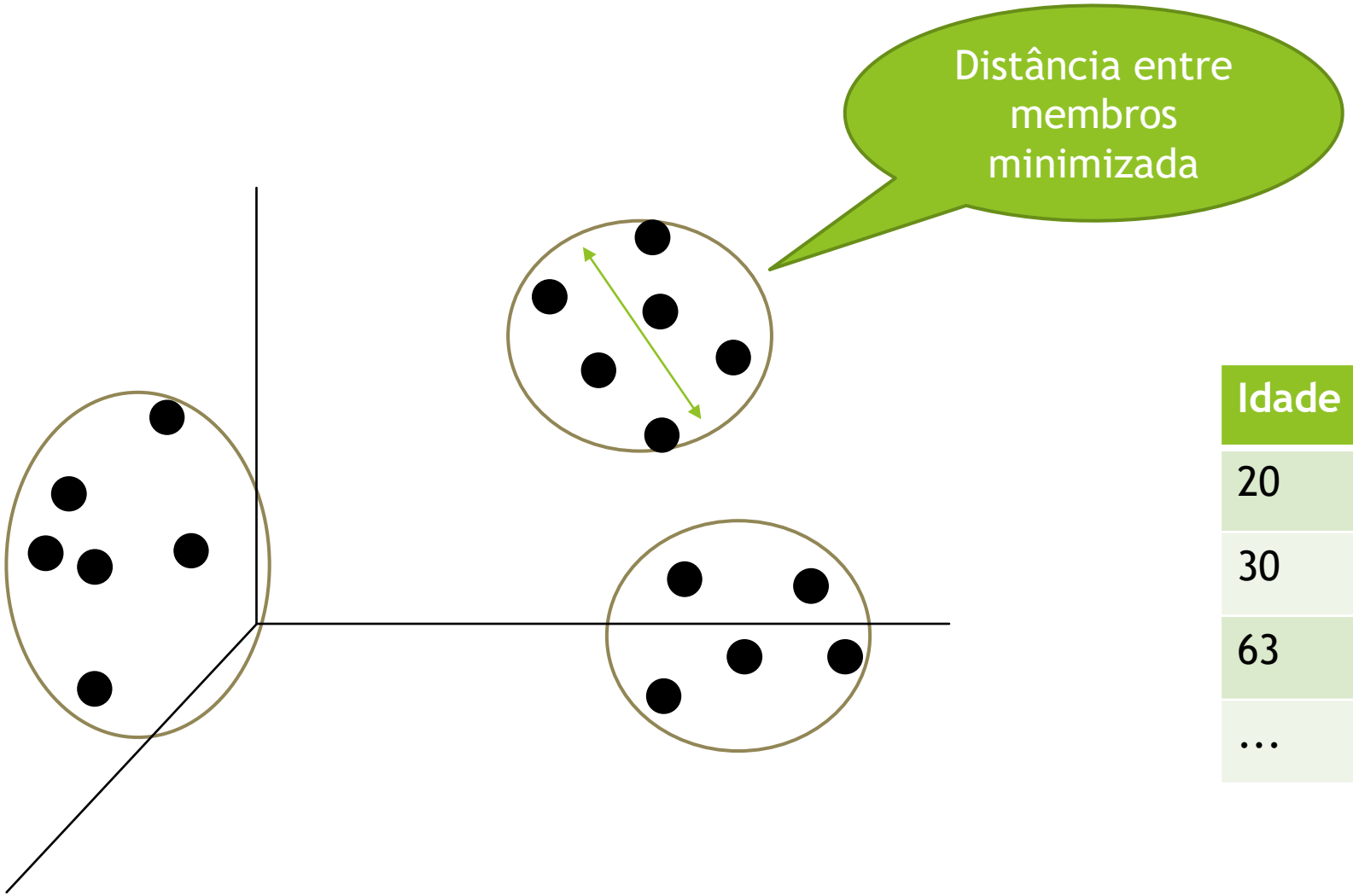
- Encontrar grupos
  - Padrões similares dentro do grupo
  - Padrões distintos entre grupos
- Mas por que ?
  - Novas descobertas para grandes volumes de dados
  - Leitura distante

O que é ?



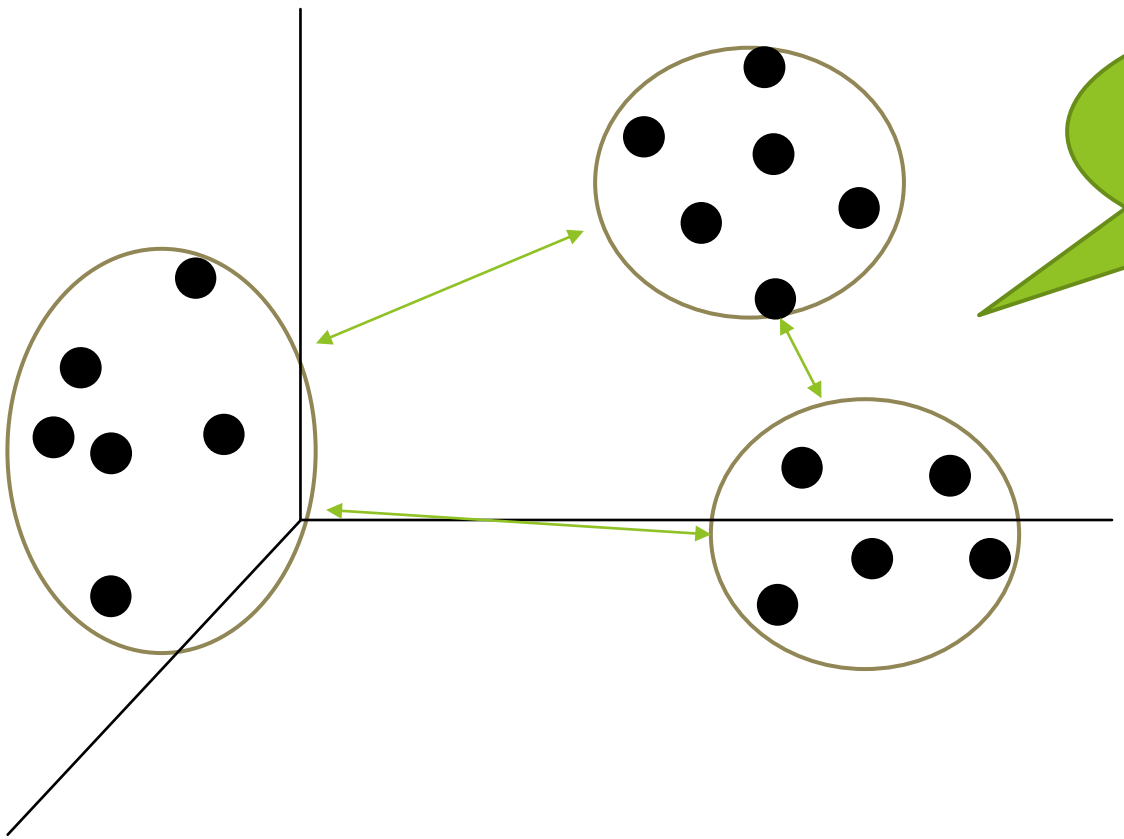
Idade	Peso	Altura
20	80	180
30	98	190
63	90	170
…	…	…

O que é ?



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

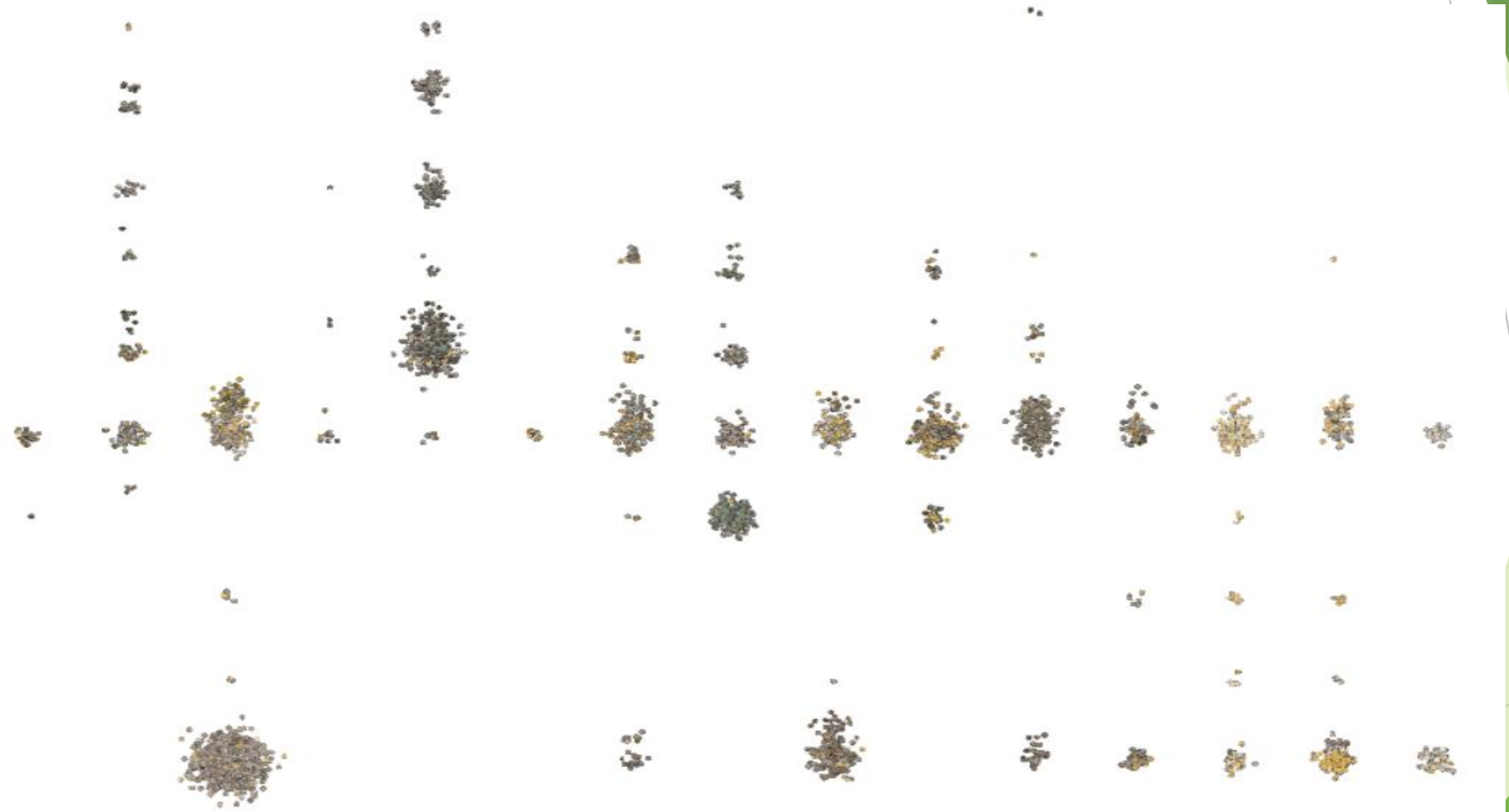
O que é ?



Distância entre grupos maximizada

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)

Como agrupar documentos ?

Topic Modeling

Document Clustering

T_1	T_2	. . .	T_m
$Word_{11}$, pr_{11}	$Word_{21}$, pr_{21}	. . .	$Word_{m1}$, pr_{m1}
$Word_{12}$, pr_{12}	$Word_{22}$, pr_{22}	. . .	$Word_{m2}$, pr_{m2}
.	.	.	.
.	.	.	.
.	.	.	.
$Word_{1k}$, pr_{1k}	$Word_{2k}$, pr_{2k}	. . .	$Word_{mk}$, pr_{mk}

C_1	C_2	. . .	C_m
D_a	D_i	. . .	D_p
D_b	D_j	. . .	D_q
.	.	.	.
.	.	.	.
.	.	.	.
D_c	D_k	. . .	D_s

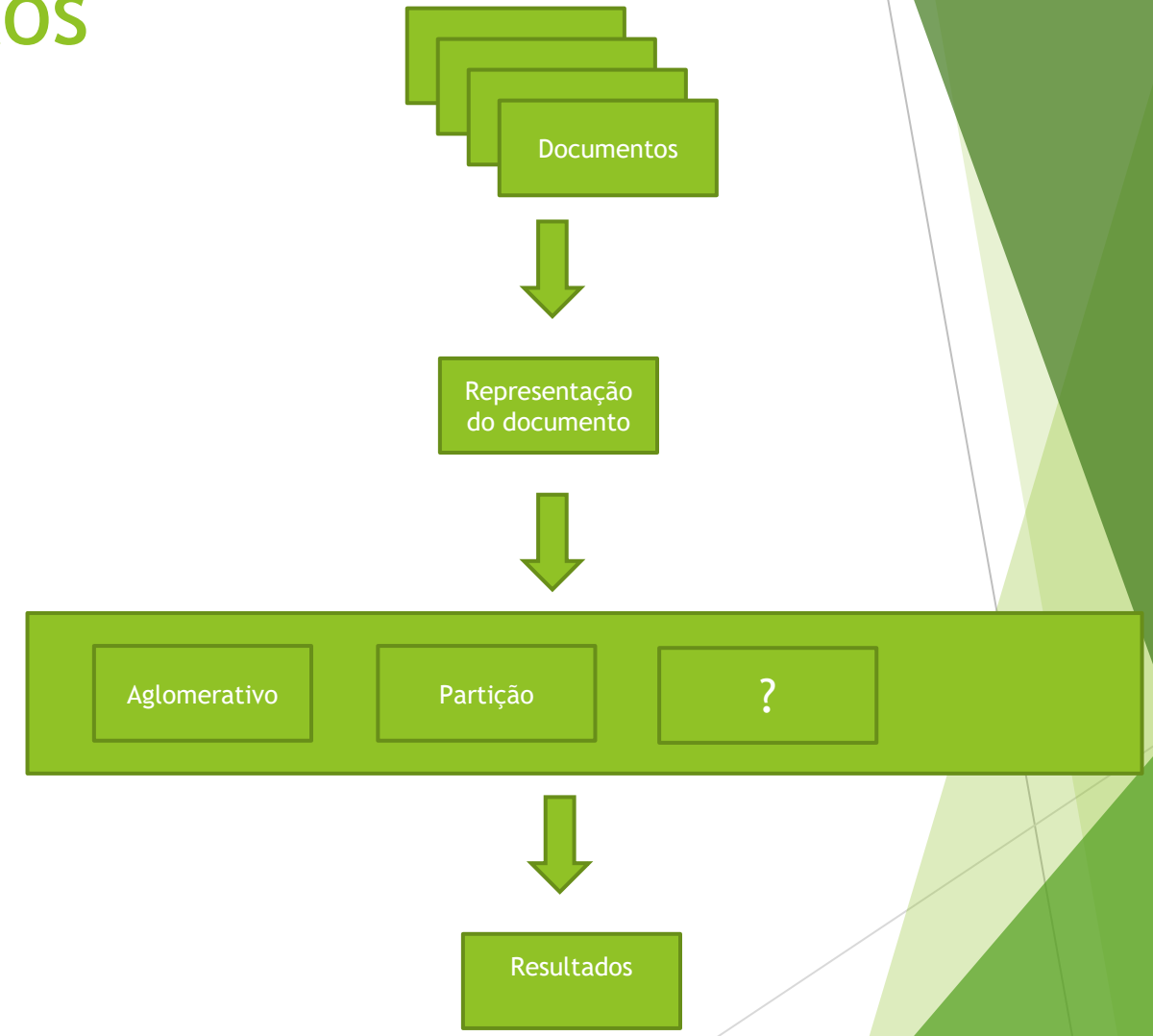
Como agrupar documentos ?

- Aparato de agrupamento
  - Estruturado
  - Numérico

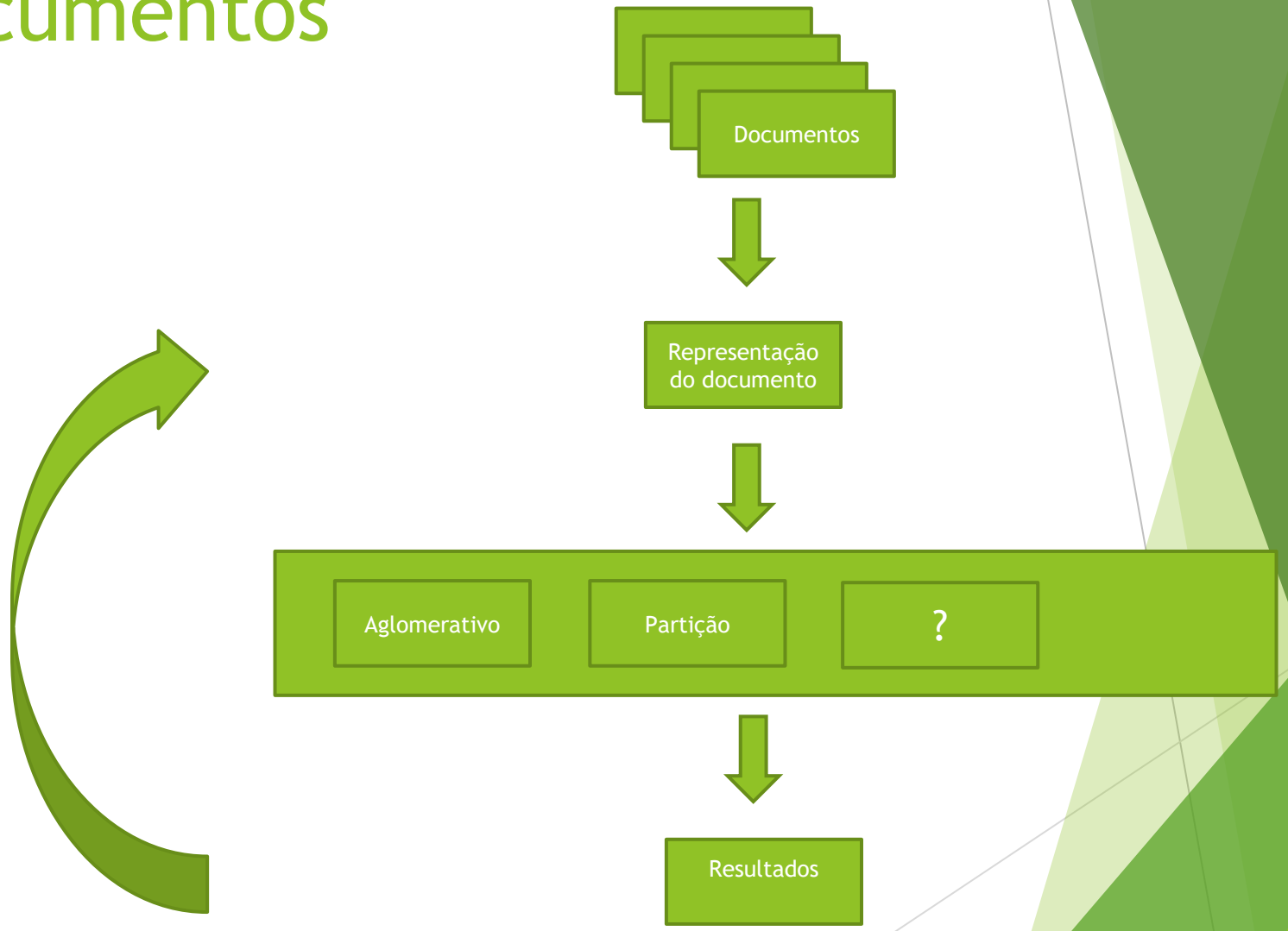
Idade	Peso	Altura	Grupo
20	80	180	2
30	98	190	3
63	90	170	1
…	…	…	…

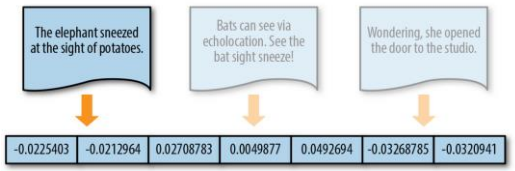
Agrupando Documentos

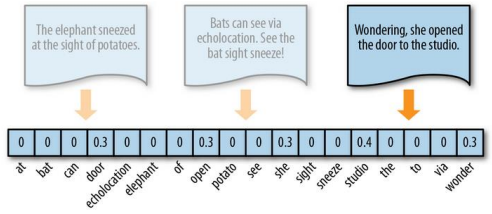
- Aparato de agrupamento
  - Estruturado
  - Numérico
- Documentos
  - Não Estruturado
  - Texto

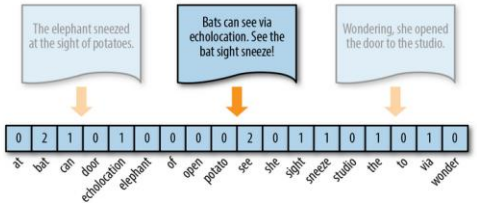


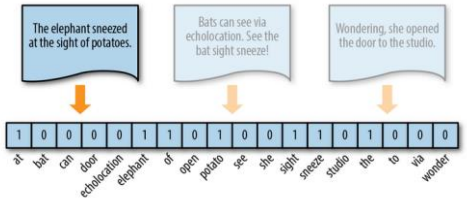
Agrupando Documentos











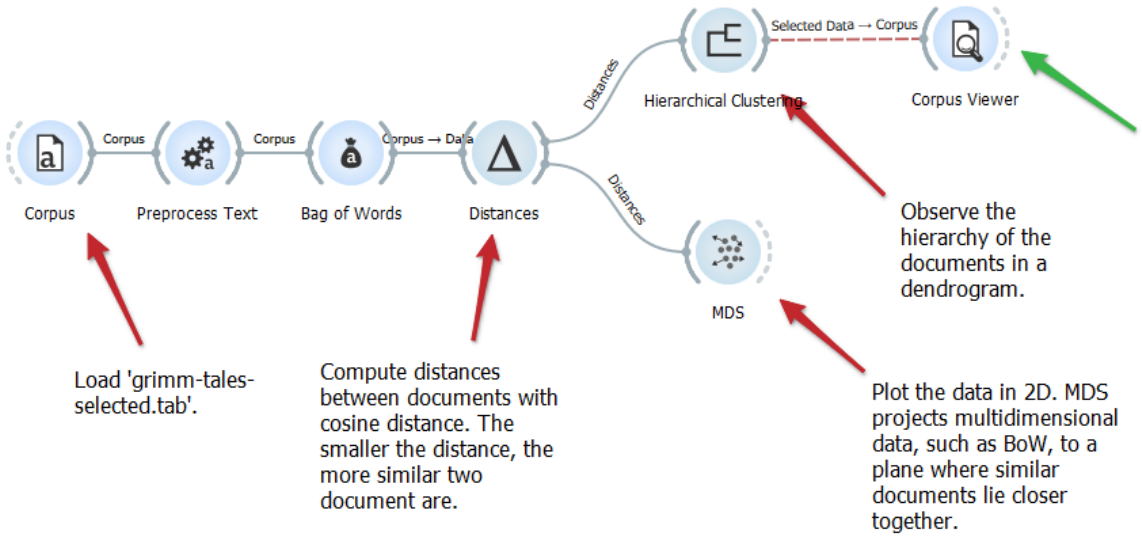
Formas de Representação

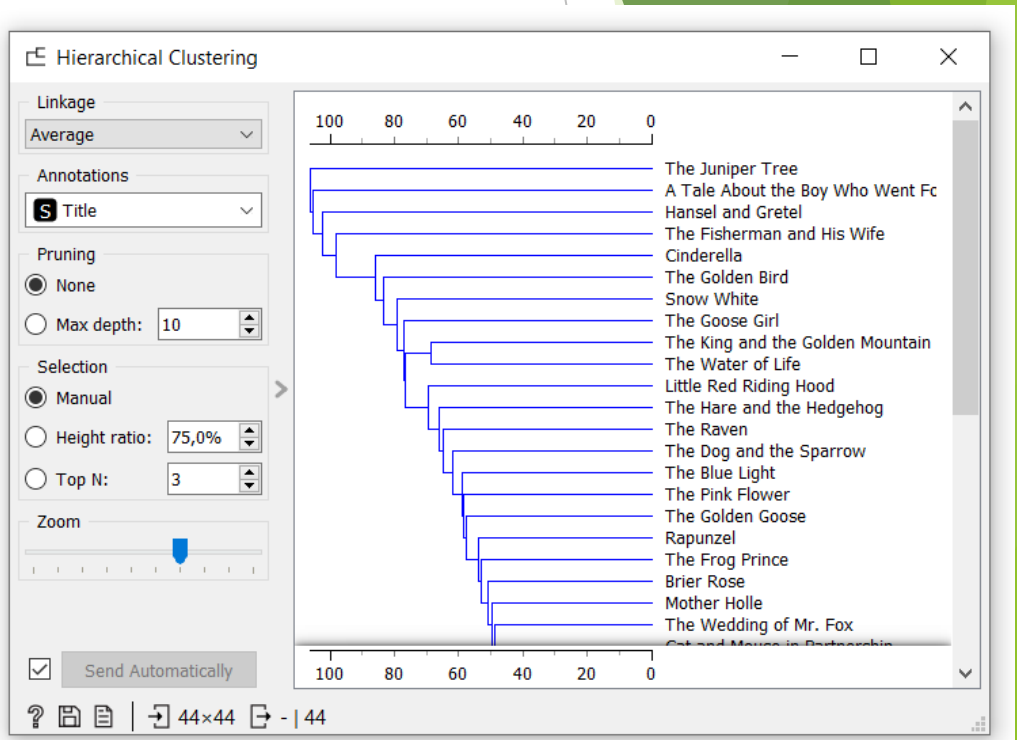
- Bag of Words
- One-Hot-Encoding
- TF-IDF
- Embeddings

Tipos de Agrupamento

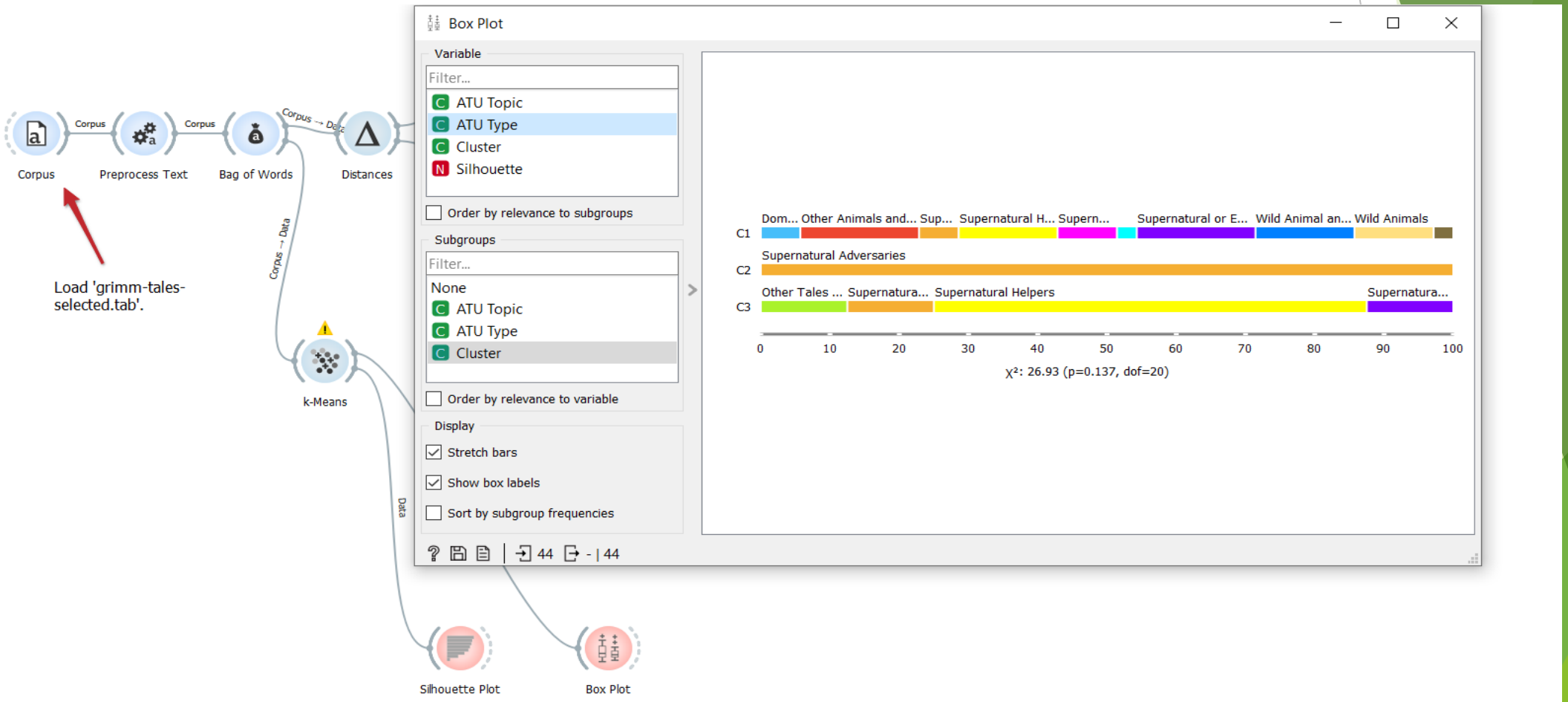
- Hierárquico
  - Gera uma hierarquia de subconjuntos
  - Aglomerativo ou Divisivo
  - Quando é útil extrair uma taxonomia
- Partição
  - Constrói k (k<n) partições
  - Não há hierarquia
- Densidade
  - Cada ponto deve ter um número mínimo de vizinhos
  - Consegue achar grupos com formas não convexas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

Hierárquico

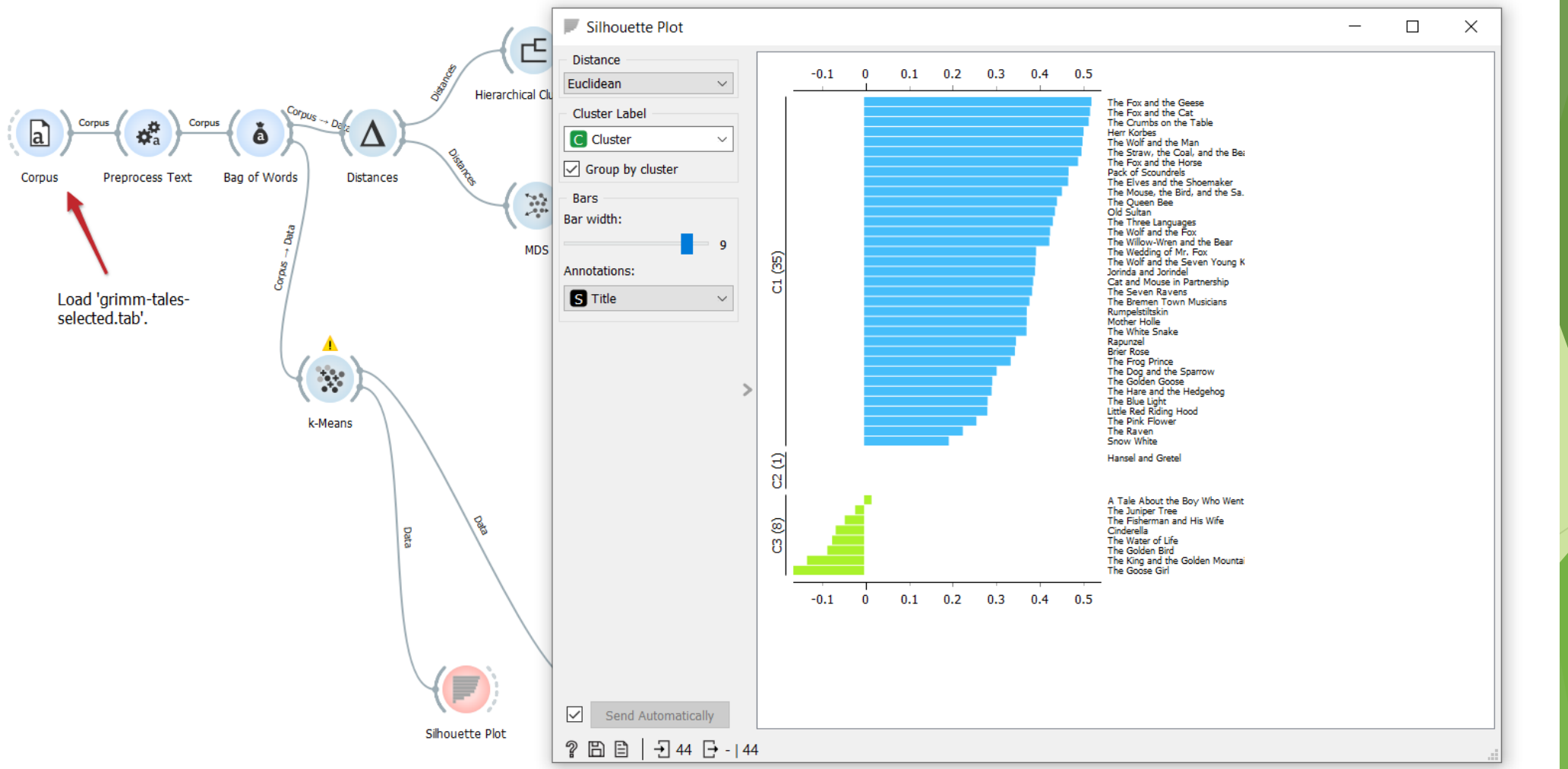




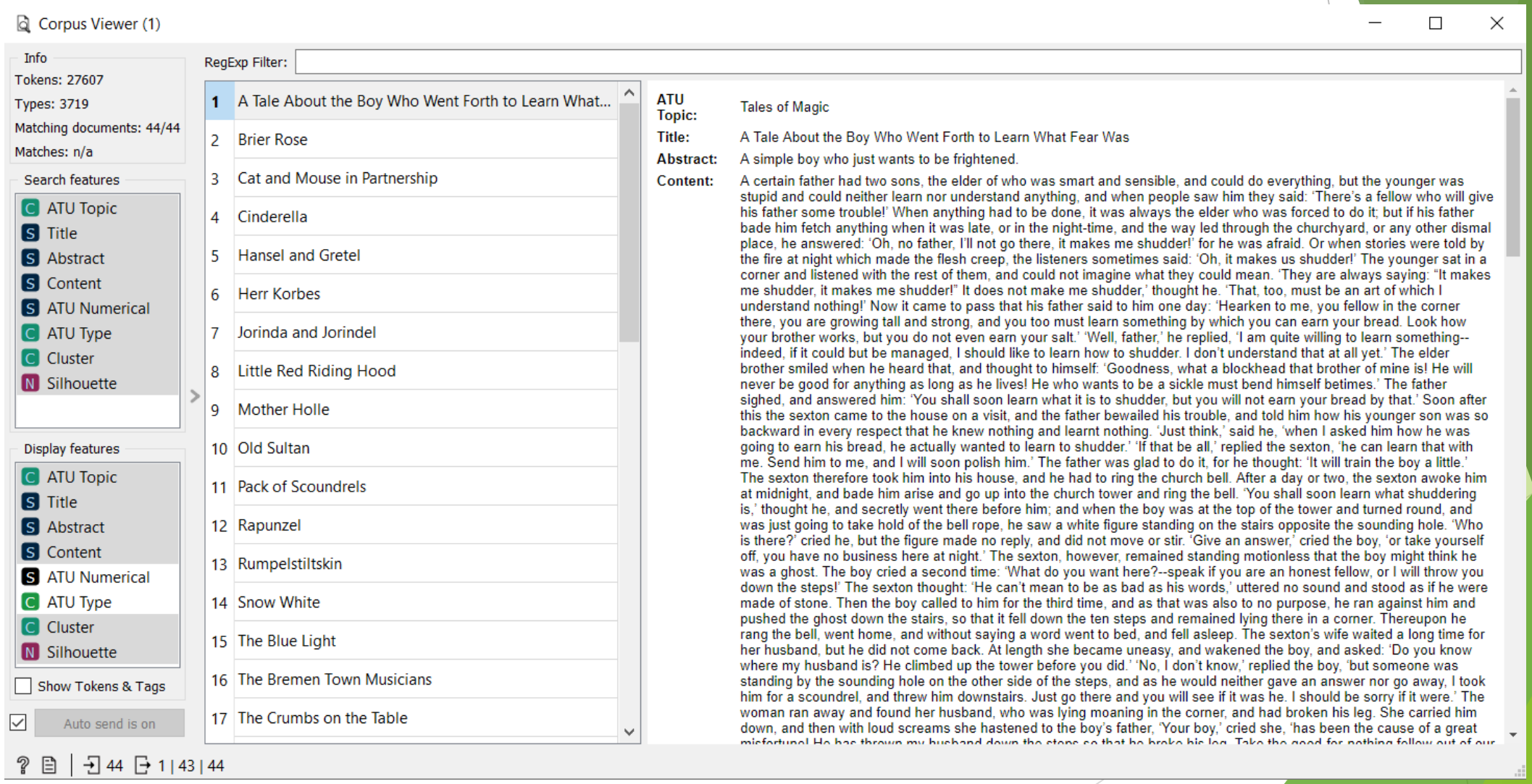
Partição


Box Plot
Variable
Filter...
ATU Topic
ATU Type
Cluster
Silhouette
Order by relevance to subgroups
Subgroups
Filter...
None
ATU Topic
ATU Type
Cluster
Order by relevance to variable
Display
Stretch bars
Show box labels
Sort by subgroup frequencies
44
- | 44
C1
Dom... Other Animals and... Sup... Supernatural H... Supern... Supernatural or E... Wild Animal an... Wild Animals
C2
Supernatural Adversaries
C3
Other Tales ... Supernatura... Supernatural Helpers Supernatura...
0 10 20 30 40 50 60 70 80 90 100
X²: 26.93 (p=0.137, dof=20)
Corpus
Corpus
Corpus
Preprocess Text
Bag of Words
Distances
Load 'grimm-tales-selected.tab'.
Corpus → Data
k-Means
Data
Silhouette Plot
Box Plot

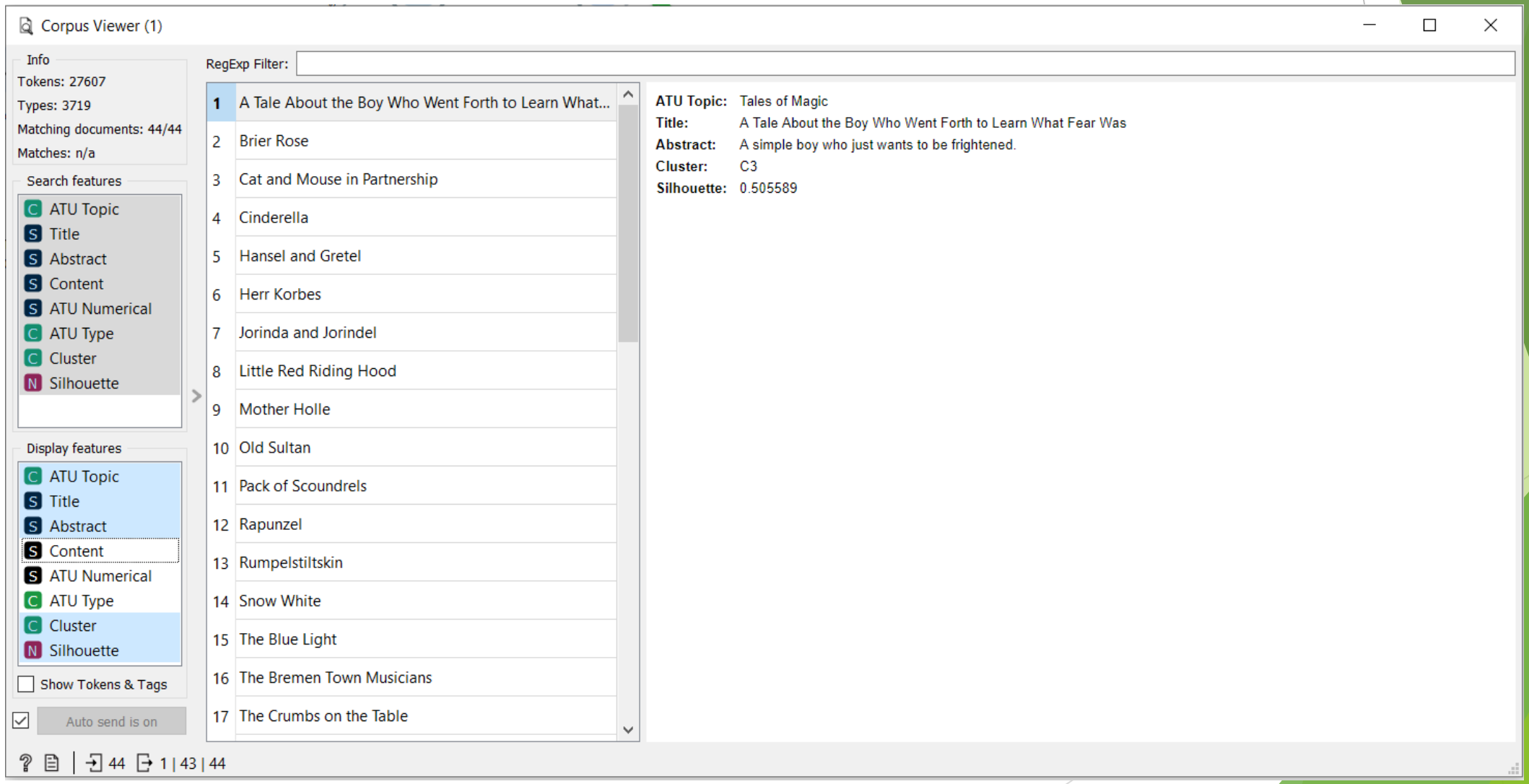
Partição


Corpus
Preprocess Text
Bag of Words
Distances
Hierarchical Cl
MDS
k-Means
Silhouette Plot
Load 'grimm-tales-selected.tab'.
Silhouette Plot
Distance
Euclidean
Cluster Label
Cluster
Group by cluster
Bars
Bar width:
9
Annotations:
Title
Send Automatically
C1 (35)
C2 (1)
C3 (8)
The Fox and the Geese
The Fox and the Cat
The Crumbs on the Table
Herr Korbes
The Wolf and the Man
The Straw, the Coal, and the Bea
The Fox and the Horse
Pack of Scoundrels
The Elves and the Shoemaker
The Mouse, the Bird, and the Sa.
The Queen Bee
Old Sultan
The Three Languages
The Wolf and the Fox
The Willow-Wren and the Bear
The Wedding of Mr. Fox
The Wolf and the Seven Young K
Jorinda and Jorindel
Cat and Mouse in Partnership
The Seven Ravens
The Bremen Town Musicians
Rumpelstiltskin
Mother Holle
The White Snake
Rapunzel
Brier Rose
The Frog Prince
The Dog and the Sparrow
The Golden Goose
The Hare and the Hedgehog
The Blue Light
Little Red Riding Hood
The Pink Flower
The Raven
Snow White
Hansel and Gretel
A Tale About the Boy Who Went
The Juniper Tree
The Fisherman and His Wife
Cinderella
The Water of Life
The Golden Bird
The King and the Golden Mountai
The Goose Girl
-0.1
0
0.1
0.2
0.3
0.4
0.5
44
- | 44

Análise Qualitativa



Análise Qualitativa



Ferramentas

- Orange
  - Provê um workflow completo e simples
- Python e Scikit
  - Exige codifição
  - 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