FCT/Unesp – Presidente Prudente
Departamento de Matemática e Computação

Visualização de Dados não Estruturados
Parte 1

Prof. Danilo Medeiros Eler

danilo.eler@unesp.br

Departamento de Matemática e Computação
Faculdade de Ciências e Tecnologia
Unesp – Universidade Estadual Paulista
Presidente Prudente/SP, Brasil

unesp

DEPARTAMENTO DE MATEMÁTICA E COMPUTAÇÃO

Sumário

- Parte 1
  - Dados textuais
  - Coleções de Documentos
- Parte 2
  - Coleções de Imagens

Dados Multivariados

- Dados multivariados são aqueles que possuem mais de uma variável para cada instância dos dados

Country	GDP/capita	Public Debt	Deficit	Inflation	Unemployment
Austria	39.8	72.3	-4.6	1.7	3.9
Belgium	36.3	96.8	-4.1	2.3	6.7
Bulgaria	12.9	16.2	-3.2	3.0	11.9
Cyprus	29.0	60.8	-5.3	2.6	7.8
Czech Republic	25.0	38.5	-4.7	1.2	6.6
Denmark	36.4	43.6	-2.7	2.2	7.1
Estonia	18.5	6.6	0.1	2.7	12.8
Finland	34.9	48.4	-2.5	1.7	7.8
France	33.9	81.7	-7.0	1.7	9.9
Germany	36.1	83.2	-3.3	1.2	5.8
Greece	28.5	142.8	-10.5	4.7	16.7
Hungary	18.8	80.2	-4.2	4.7	9.9

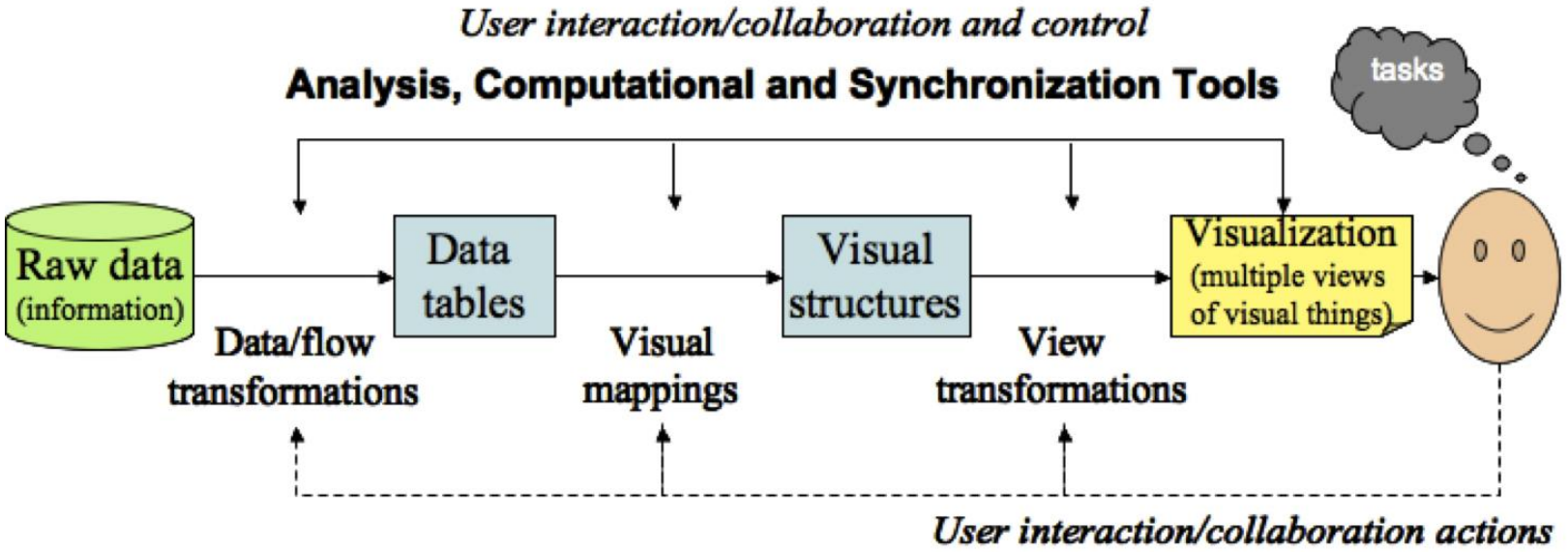
http://www.real-statistics.com/wp-content/uploads/2013/09/eu-data.png

Dados não Estruturados

- Alguns conjuntos de dados não possuem uma estrutura definida, por exemplo
  - Coleções de Documentos e de Imagens
- Por isso, é necessário fazer um processamento das instâncias para extrair dados para serem visualizados ou estruturar o conjunto de dados. Por exemplo,
  - Modelo de espaço vetorial para documentos
  - Espaço de características para imagens

Processo de Visualização

- Pipeline de visualização utilizado pela maioria dos sistemas



Visualização de Coleções de Documentos

- Existem muitas fontes de informação que disponibilizam dados no formato textual
  - Ex.: email, blogs, livros, artigos
- Uma coleção de documentos é definida como um corpus (ou corpora no plural)

Visualização de Coleções de Documentos

- Em uma coleção de documentos, podemos procurar por palavras, frases ou tópicos
- Se a coleção estiver parcialmente estruturado, podemos procurar por relacionamento entre documentos, palavras e tópicos
- Se ela estiver totalmente estruturada, podemos encontrar grupos, padrões e *outliers*

Visualização de Coleções de Documentos

- Podemos definir três níveis de representação de uma coleção de documentos
  - Léxico
    - Transformação em entidades atômicas chamadas de *tokens*
  - Sintático
    - Lida com a rotulação (anotação) dos *tokens*, por exemplo, substantivo, adjetivo
  - Semântico
    - Envolve a extração de significado e relacionamento
    - Define uma interpretação analítica do texto dentro de um contexto

Visualização de Documentos

- Várias técnicas de visualização foram propostas para auxiliar na visualização individual de documentos
  - Algumas delas
    - Tag Clouds ou Word Clouds
    - WordTree
    - TextArc
    - Literature Fingerprinting
    - Visualização baseada em Grafos

Tag Clouds ou Word Clouds

- O tamanho da fonte é proporcional à frequência da palavra no documento

Lista de termos e frequências

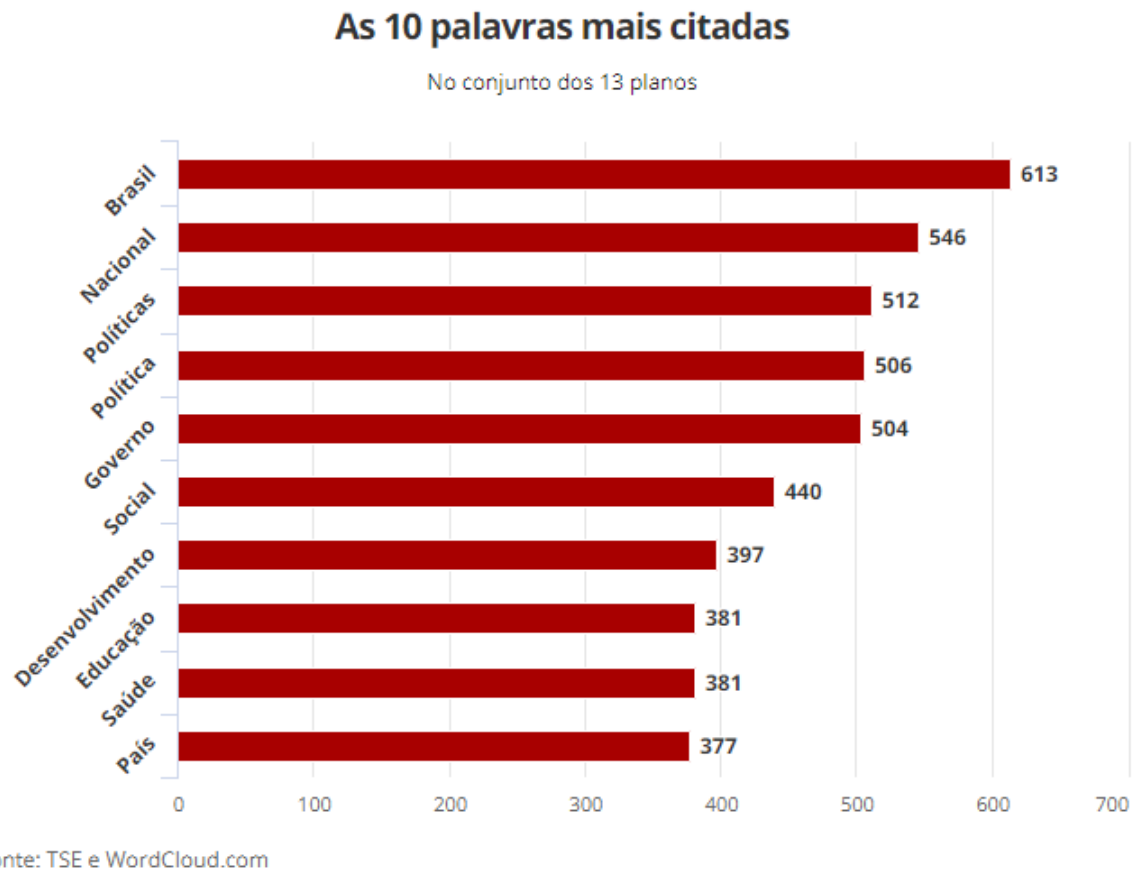
Ngram	Frequency
reasoning	2135
information	1805
retrieval	1544
intelligence	1415
artificial	1293
systems	1162
computer	1129
science	1090
knowledge	1040
system	1037
university	1034
proc	1003
logic	966
machine	937
case	884

Tag Clouds ou Word Clouds

- O tamanho da fonte é proporcional à frequência da palavra no documento
  - No exemplo, a intensidade de negrito também é proporcional à frequência da palavra

Tag Clouds ou Word Clouds

Planos de governo dos candidatos à Presidência



https://g1.globo.com/politica/eleicoes/2018/noticia/2018/09/02/veja-as-palavras-mais-citadas-nos-programas-de-governo-dos-13-candidatos-a-presidencia.ghtml

Tag Clouds ou Word Clouds

- Tag Cloud dos planos de governo dos candidatos à Presidência

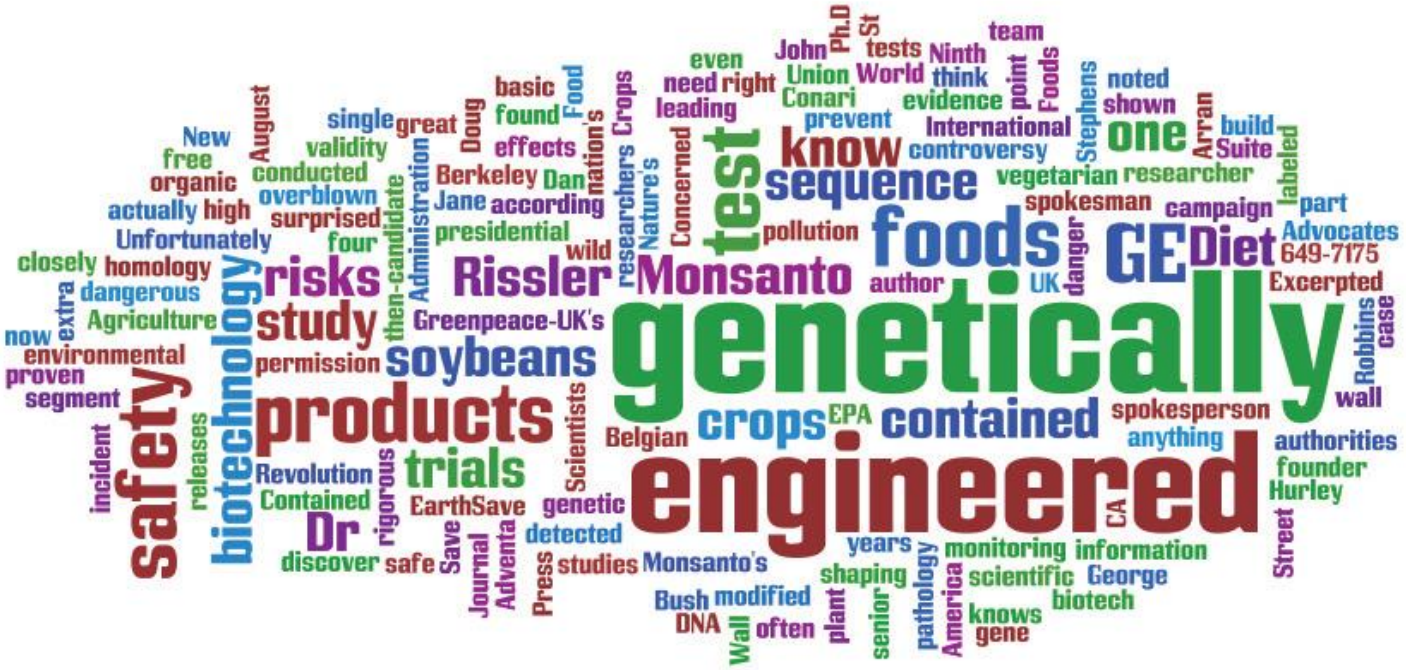
https://g1.globo.com/politica/eleicoes/2018/noticia/2018/09/02/veja-as-palavras-mais-citadas-nos-programas-de-governo-dos-13-candidatos-a-presidencia.ghtml

unesp

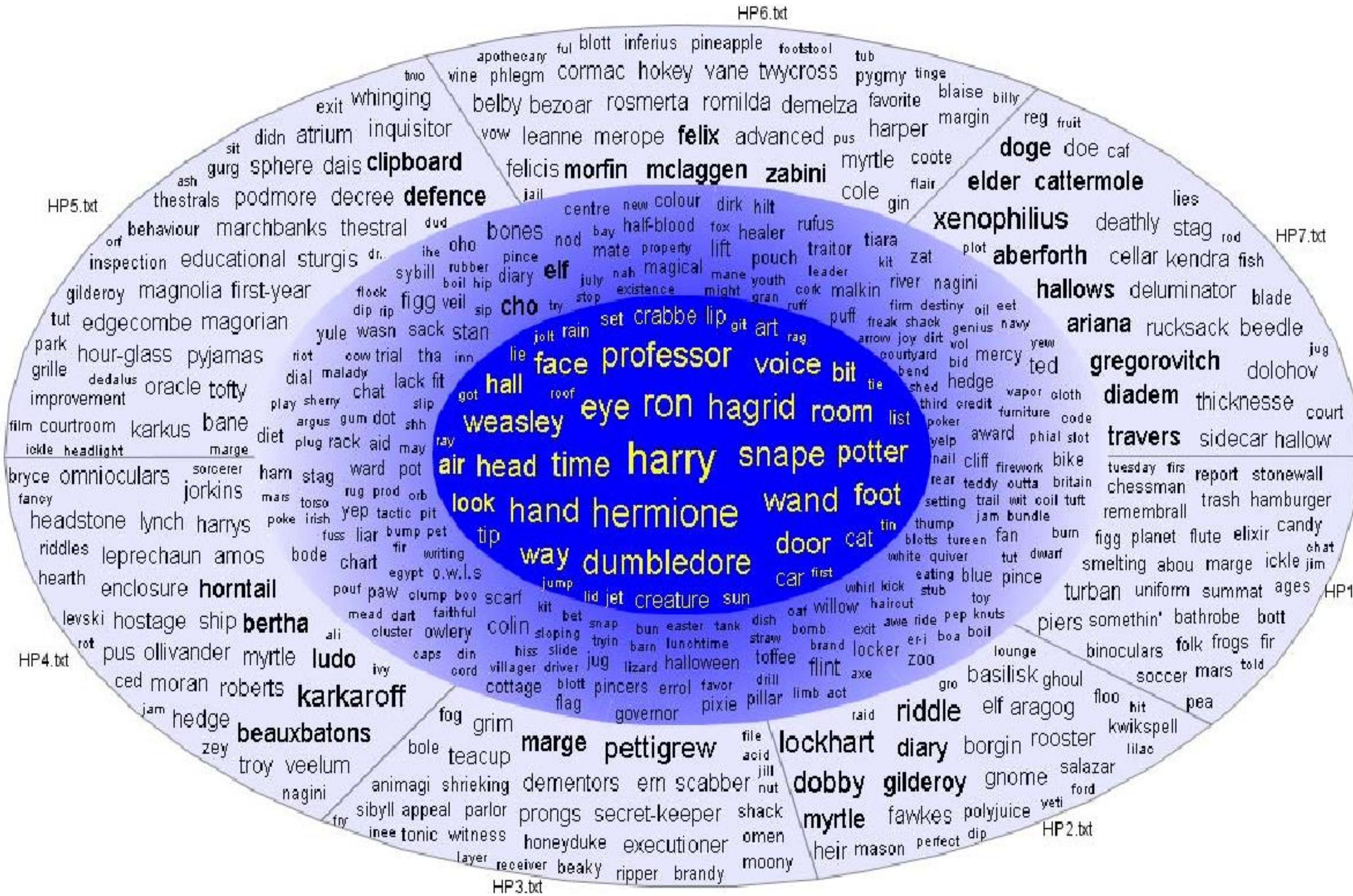
DEPARTAMENTO DE MATEMÁTICA E COMPUTAÇÃO

Tag Clouds ou Word Clouds

- Wordle é uma variação da Tag Cloud, utilizando um layout diferenciado para melhorar a ocupação do espaço
  - Assim, mais palavras podem ser inseridas na visualização



Concentri Cloud



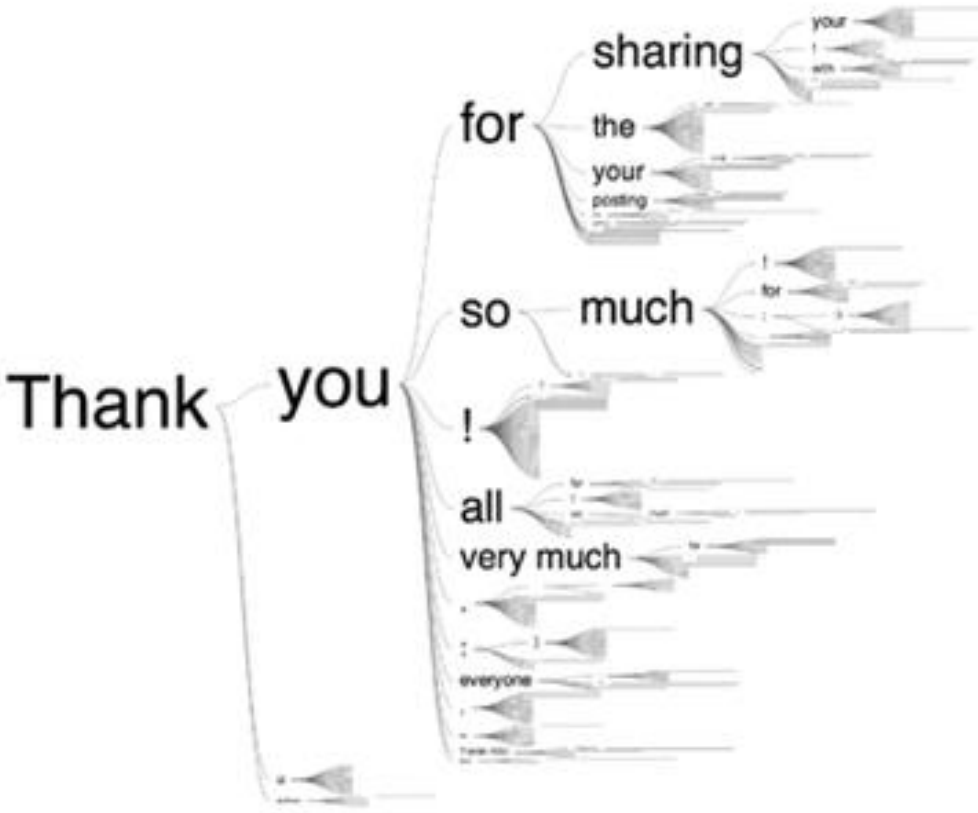
(Lohmann et al, 2015)

unesp

DEPARTAMENTO DE MATEMÁTICA E COMPUTAÇÃO

WordTree

- WordTree é uma representação visual de termos e frequências, bem como seu contexto



https://www.jasondavies.com/wordtree/examples/8.thumb.png

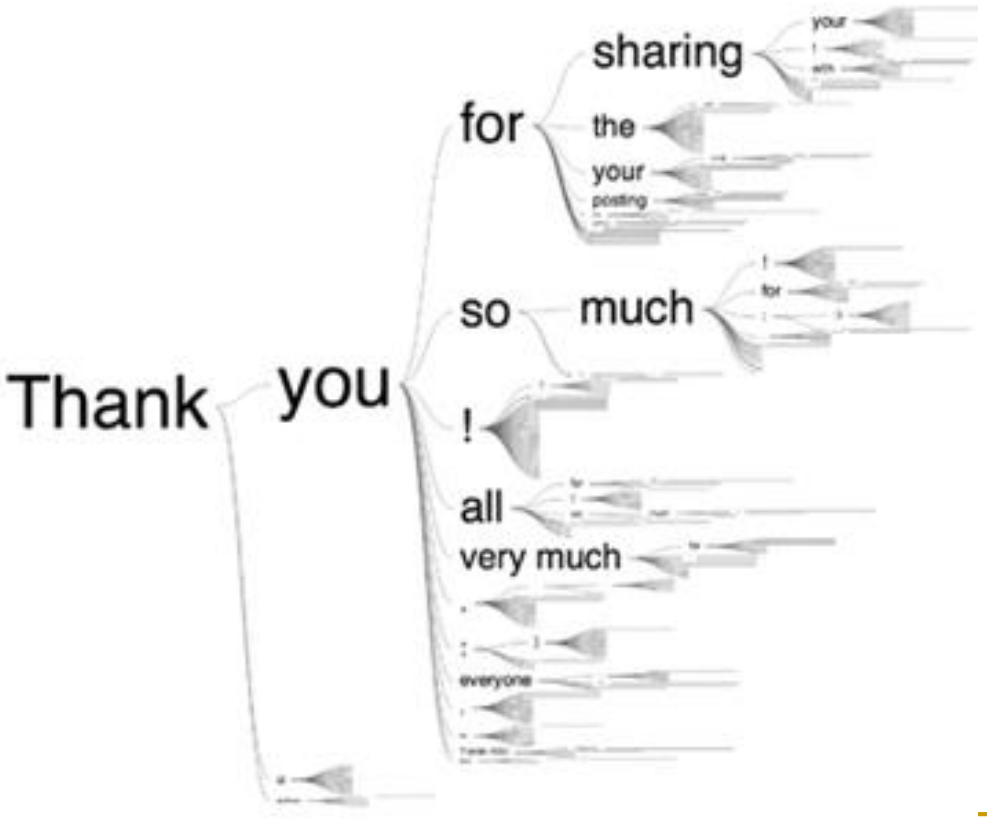
WordTree

- O tamanho do termo é a sua frequência na frase

https://www.jasondavies.com/wordtree/examples/8.thumb.png

WordTree

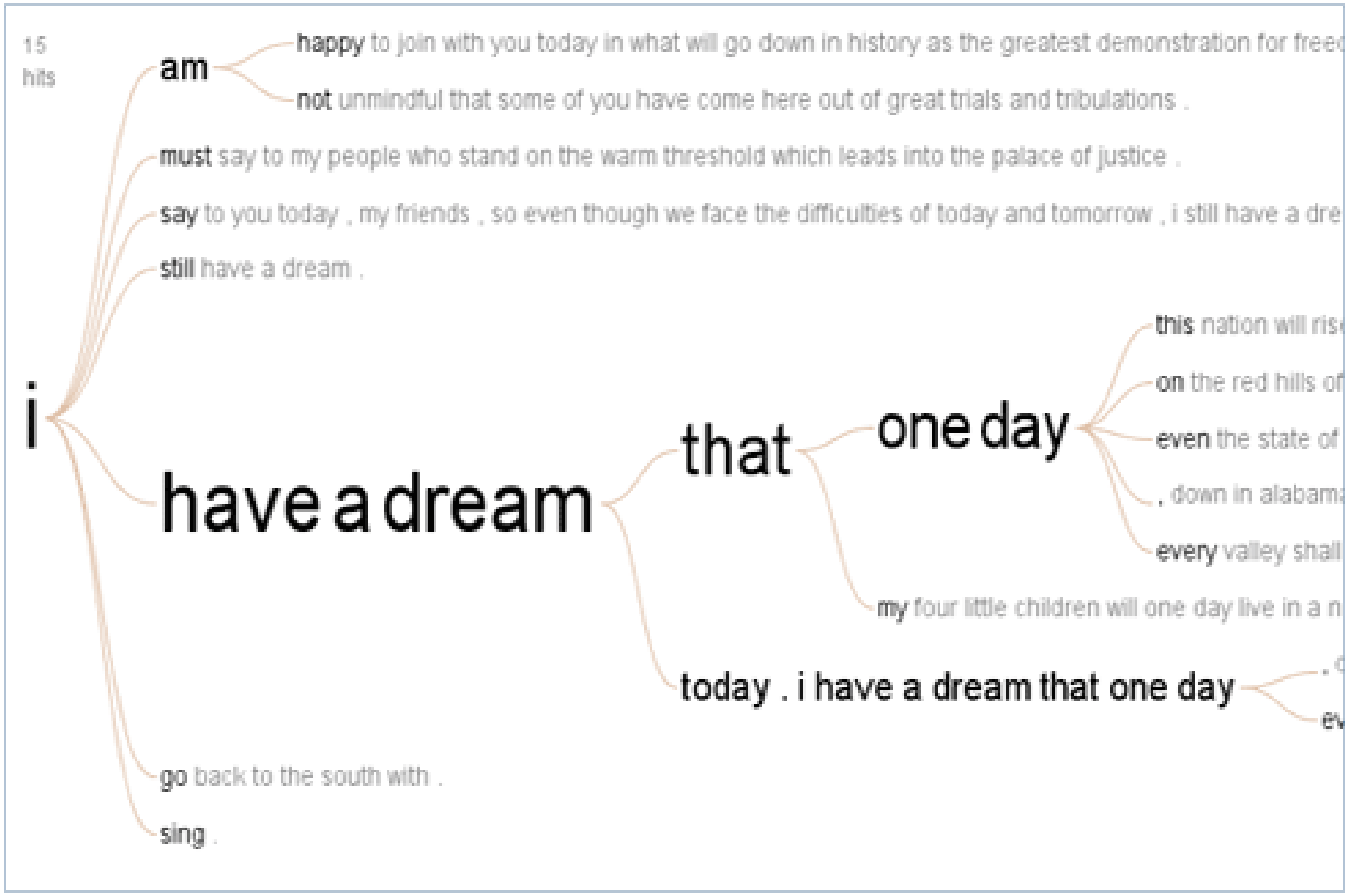
- A raiz da árvore é uma palavra ou frase é escolhida pelo usuário e os ramos representam os diferentes contextos em que são usadas



https://www.jasondavies.com/wordtree/examples/8.thumb.png

WordTree

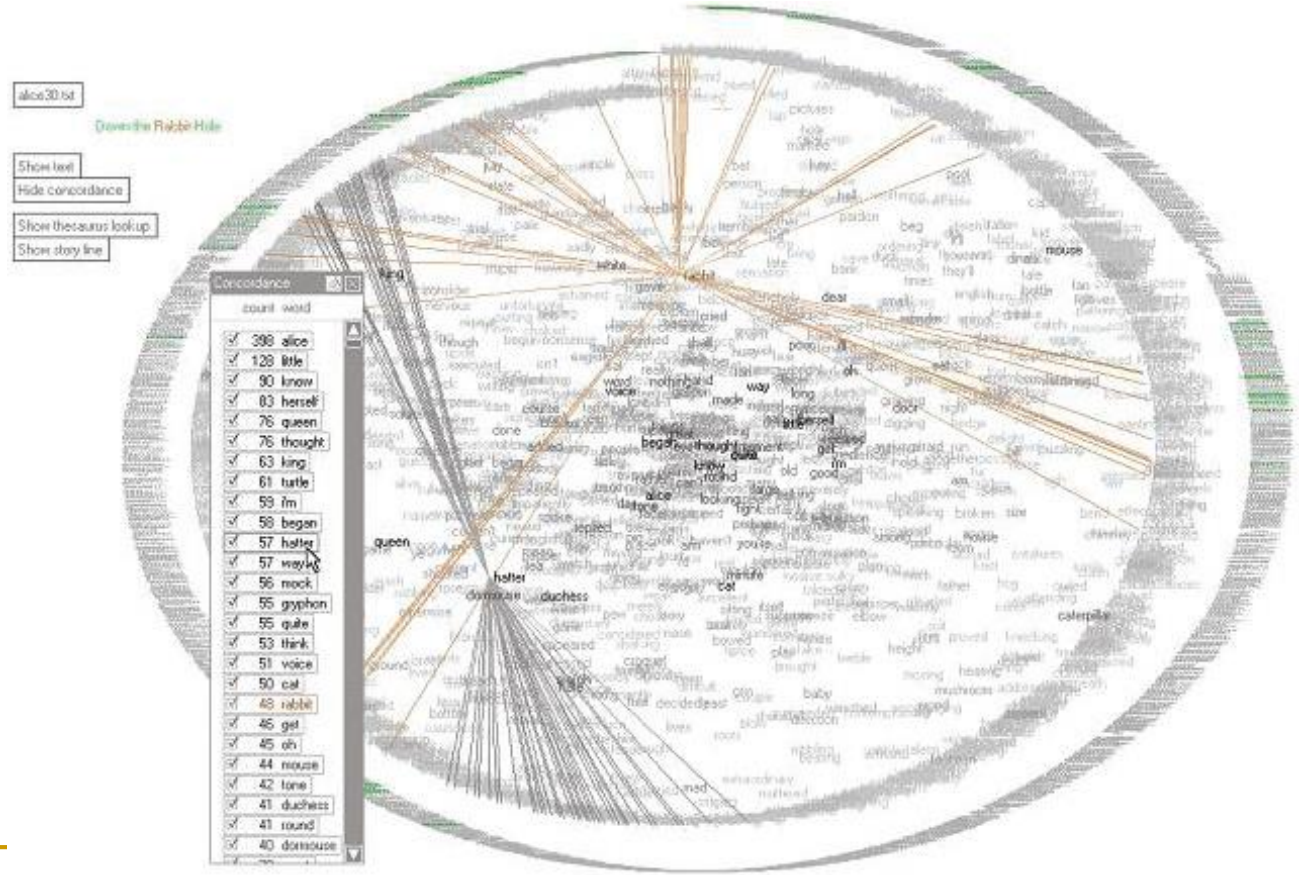
- Parte do discurso de Martin Luther King



http://hint.fm/projects/wordtree/dream.png

TextArc

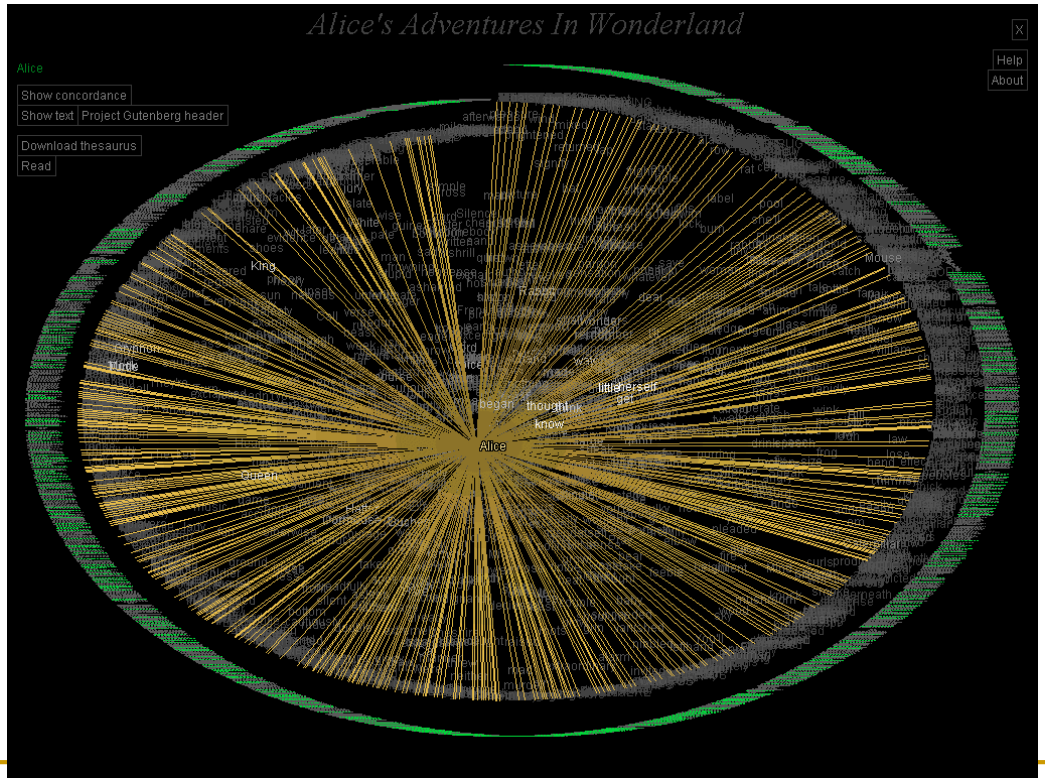
- A técnica TextArc desenha as frases de um texto nas bordas de uma elipse e as palavras mais frequentes em seu interior



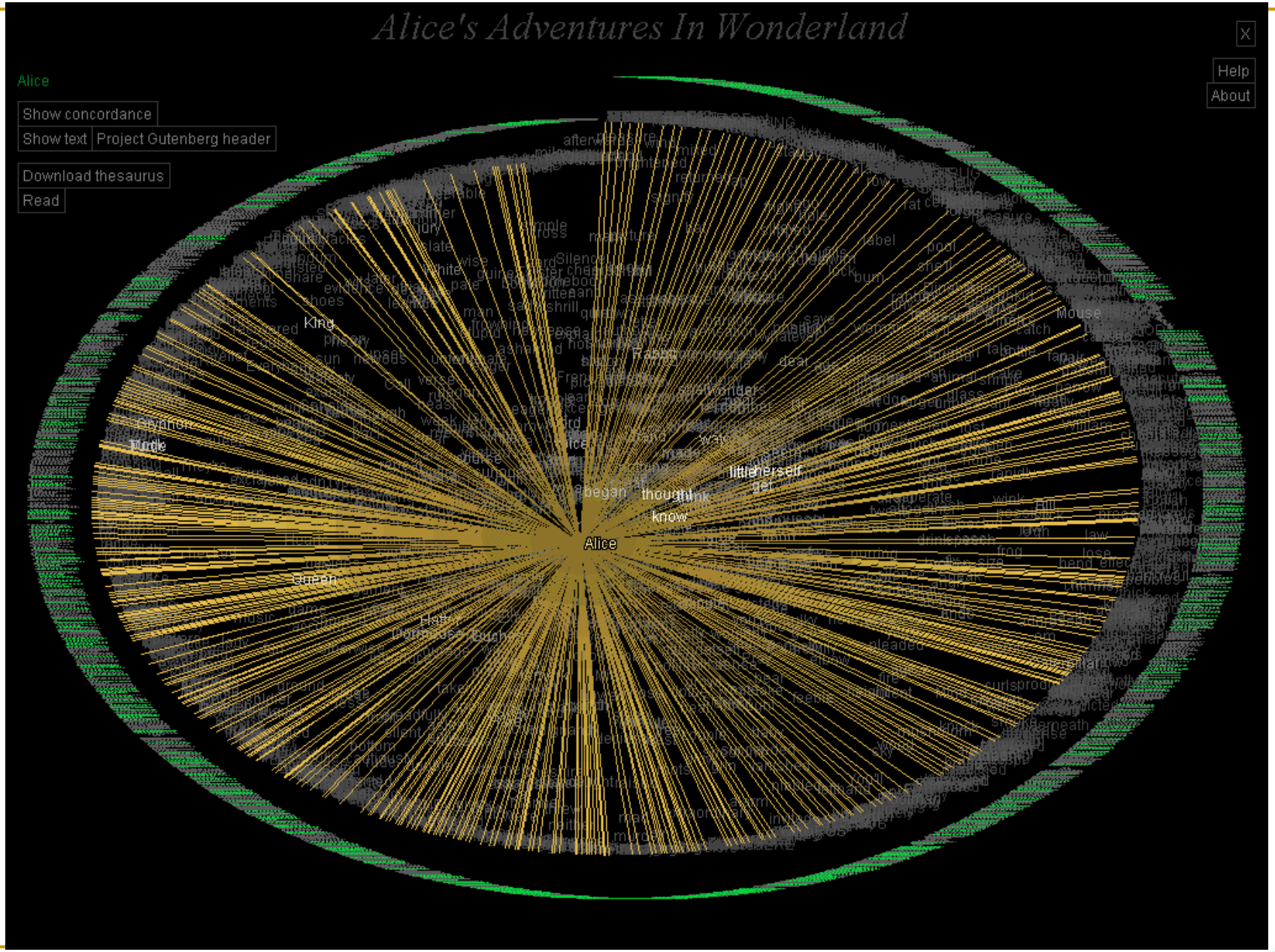
http://textarc.org/

TextArc

- Pode exibir a relação entre as palavras e as frases pela seleção do usuário ou pela simulação da leitura do documento



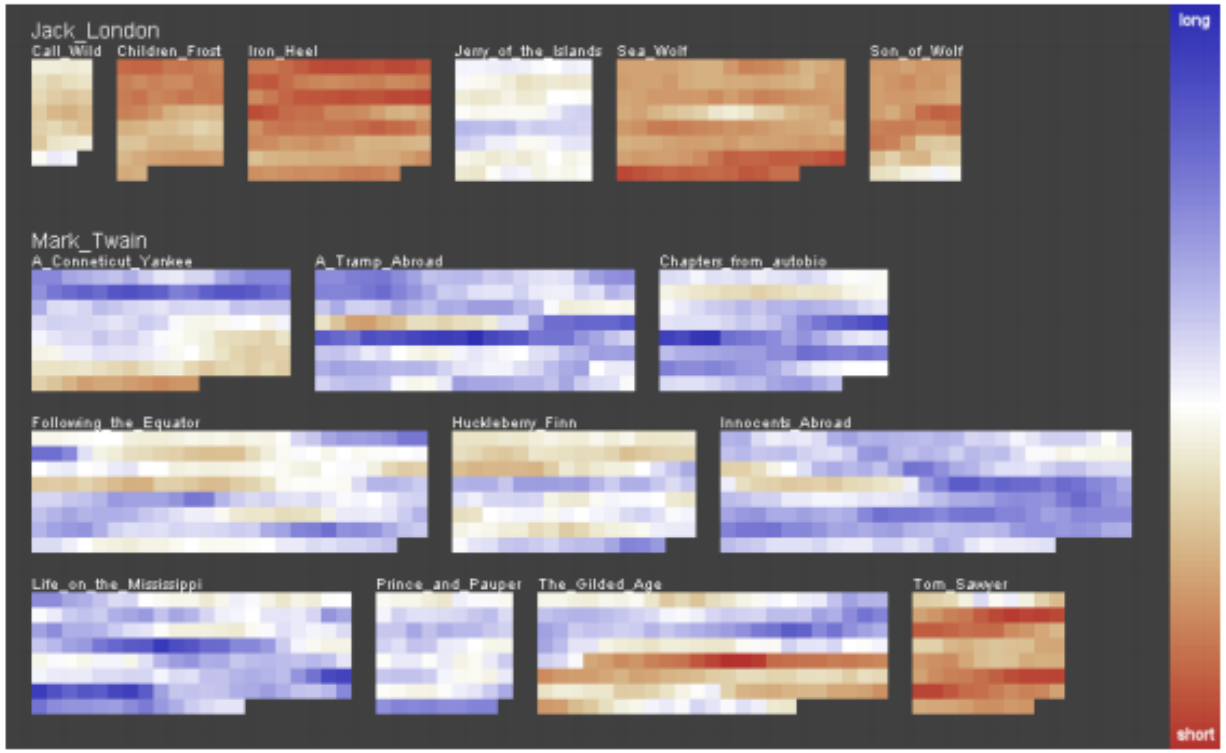
http://textarc.org/images/alice2.gif



http://textarc.org/images/alice2.gif

Literature Fingerprinting

- É um método de visualização de características textuais
  - Várias características são extraídas do texto e apresentadas como impressão digital do documento

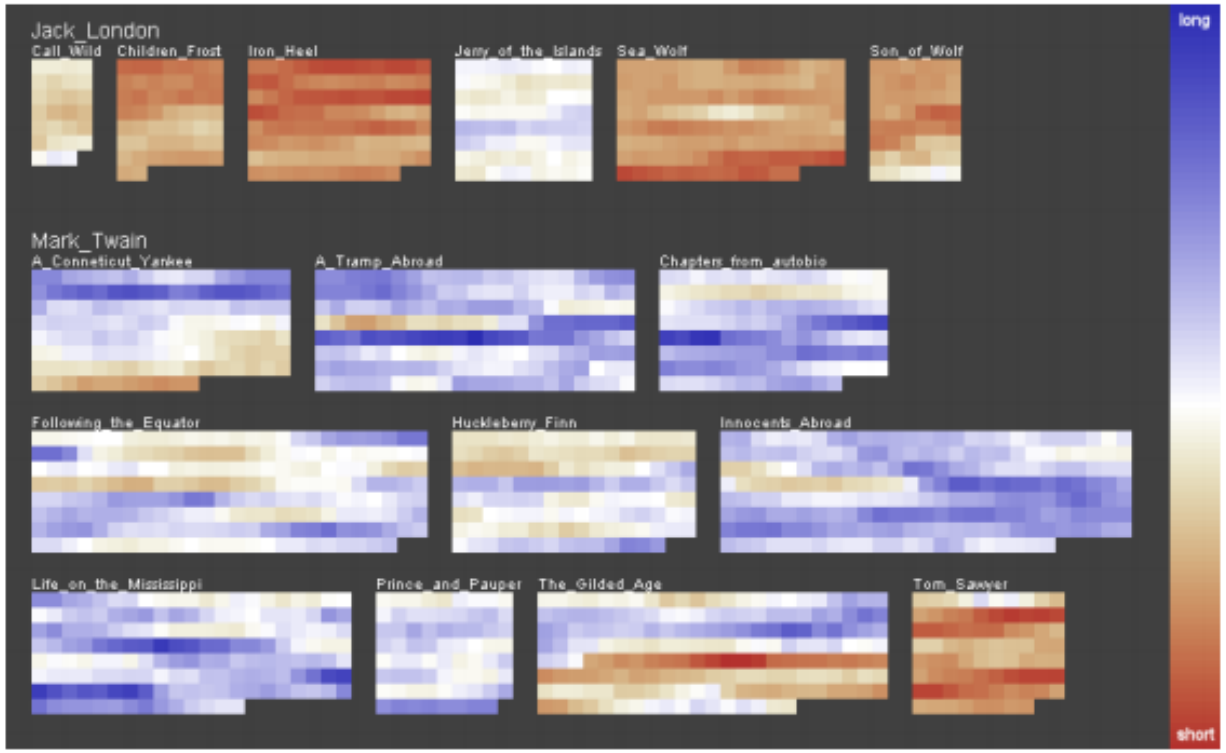


Average sentence length

(D. A. Keim and D. Oelke, 2007)

Literature Fingerprinting

- Medidas foram calculadas para analisar o estilo literário de dois autores diferentes
  - Tamanho de sentenças de diferentes obras de Mark Twain e de Jack London

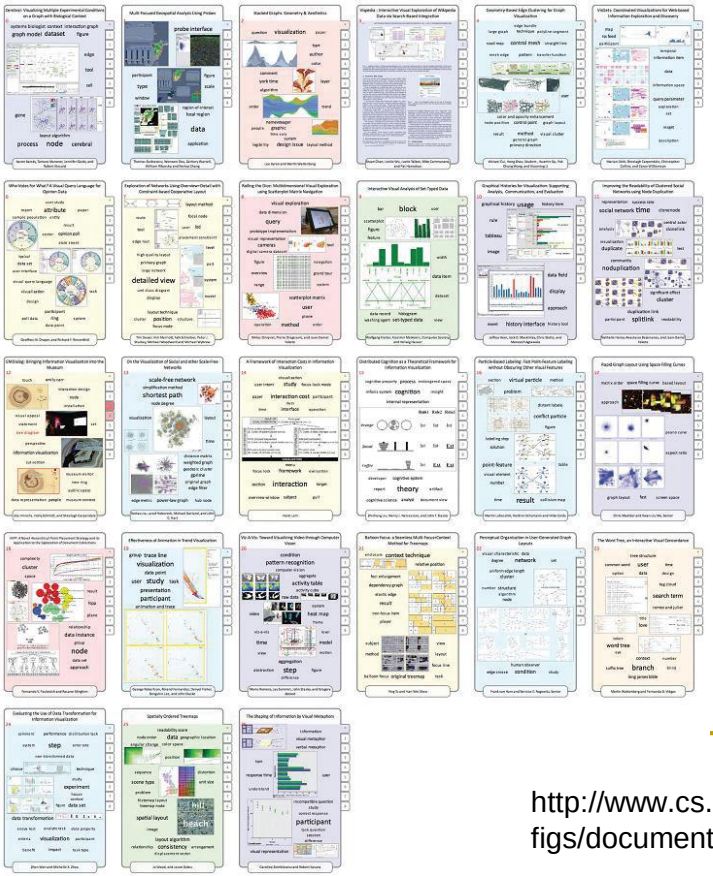


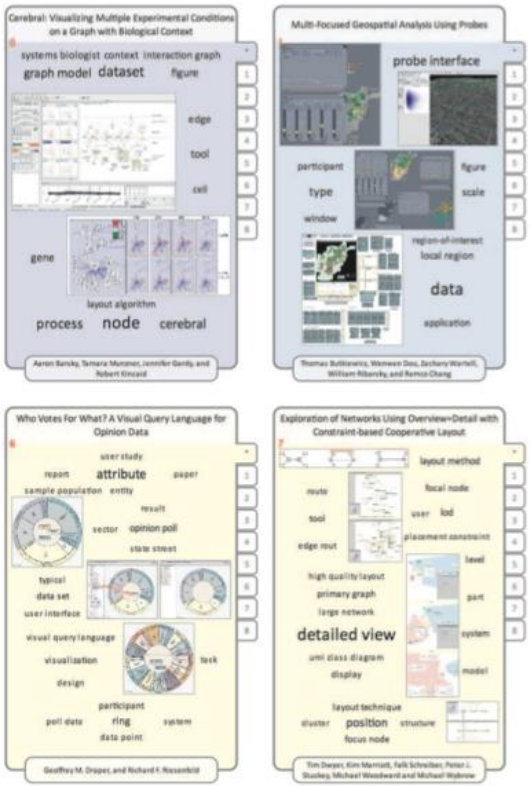
Average sentence length

(D. A. Keim and D. Oelke, 2007)

Document Cards

- A *Document Cards* apresenta uma visualização compacta de uma coleção
  - Apresentando elementos chaves e mais importantes, tais como, texto e figuras

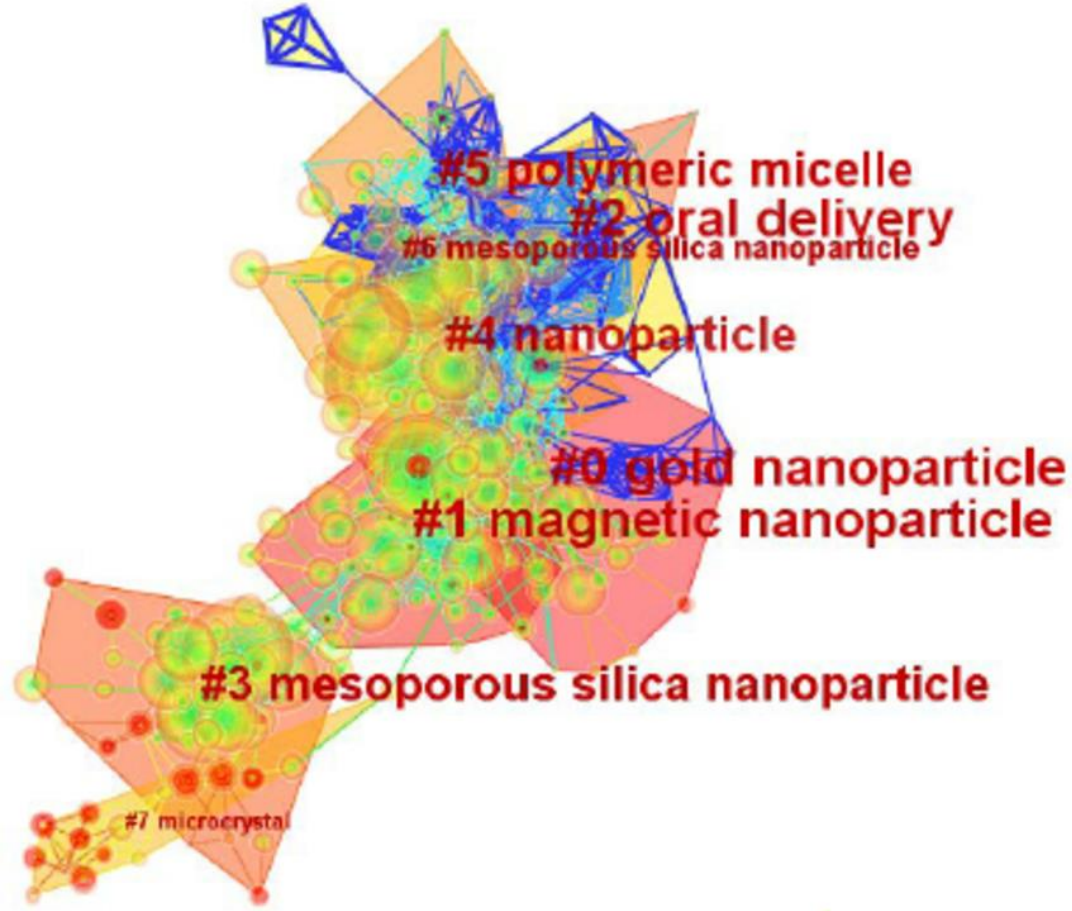




http://www.cs.odu.edu/~mweigle/research/archive-vis/2013-figs/documentcards.png

Visualização Baseada em Grafos

- Visualização da rede de co-citação entre artigos



https://www.researchgate.net/profile/Chaomei_Chen/publication/288827232/figure/fig1/AS:315264722849792@1452176375584/Figure-2-Clusters-visualization-based-on-a-document-co-citation-network-Figure-2.png

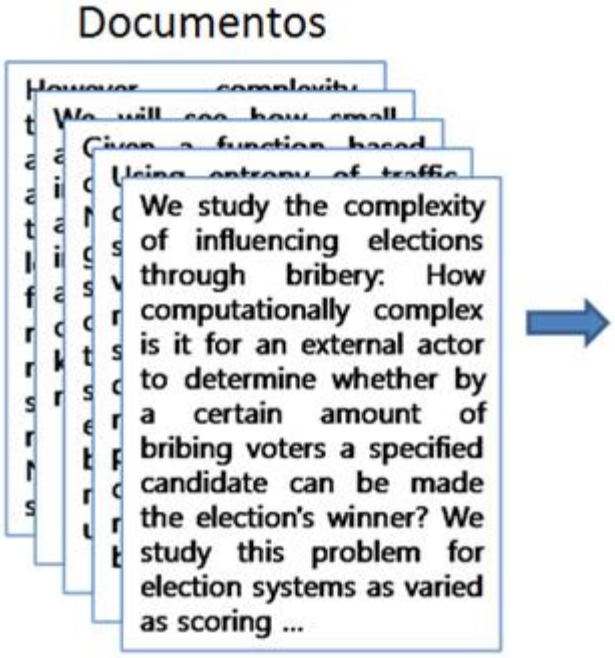
Visualização de Coleções de Documentos

Visualização de Coleções de Documentos

- Quando visualizações são aplicadas para criar representações visuais de **coleções de documentos**, geralmente, o foco é apresentar a relação de similaridade desses documentos
  - A similaridade entre cada par de documentos é calculada para gerar o layout da visualização
  - Para tanto, é necessário extrair dados dos documentos e estruturar a coleção para calcular as similaridades

Modelo de Espaço Vetorial

- Geralmente, as coleções de documentos são estruturadas em um matriz de termos por documentos, também conhecida como Modelo de Espaço Vetorial



Modelo de Espaço Vetorial

	Complexity	Algorithm	Entropy	Traffic	Network
Doc1	2	3	1	0	0
Doc2	0	0	0	2	1
Doc3	3	0	0	3	4
Doc4	2	4	2	0	0
Doc5	3	4	0	0	0

Matriz de Termos x Documentos

http://mlg.postech.ac.kr/static/research/nmf_cluster1.PNG

Pré-processamento de Documentos

- A construção desse modelo pode seguir as seguintes etapas
  - Identificação de termos
  - Eliminação de stopwords
  - Stemming
  - Contagem de Frequência
  - Ponderação dos termos

Identificação dos Termos

A frase do ex-presidente Fernando Henrique Cardoso, que escorregou no português formal ao criticar indiretamente o presidente Luiz Inácio Lula da Silva, foi considerada "politicamente incorreta" pela professora de português Thaís Nicoleti. No 3º Congresso do PSDB, anteontem em Brasília, o ex-presidente disse que quer "brasileiros melhor educados, e não brasileiros liderados por gente que despreza a educação, a começar pela própria".

http://www1.folha.uol.com.br/folha/brasil/ult96u348430.shtml

Eliminação de Stopwords

A frase do ex-presidente Fernando Henrique Cardoso, que escorregou no português formal ao criticar indiretamente o presidente Luiz Inácio Lula da Silva, foi considerada "politicamente incorreta" pela professora de português Thaís Nicoleti. No 3º Congresso do PSDB, anteontem em Brasília, o ex-presidente disse que quer "brasileiros melhor educados, e não brasileiros liderados por gente que despreza a educação, a começar pela própria".

http://www1.folha.uol.com.br/folha/brasil/ult96u348430.shtml

Stemming

A frase do ex-presidente Fernando Henrique Cardoso, que escorreg no português formal ao critic indireta o presidente Luiz Inácio Lula da Silva, foi consider "politic incorret" pela professor de português Thaís Nicoleti. No 3º Congresso do PSDB, anteontem em Brasília, o ex-presidente disse que quer "brasileir melhor educ, e não brasileir liderad por gente que desprez a educ, a começ pela própria".

http://www1.folha.uol.com.br/folha/brasil/ult96u348430.shtml

Termos resultants – n-grams

- Os termos resultantes são agrupados em n-grams, que são a combinação dos termos, conforme aparecem no texto
  - Ex. 1-grams:
    - frase, ex-presidente, Fernando, Henrique, Cardoso, escorreg, português, formal, critic, indireta, presidente,Luiz, Inácio, Lula, Silva, foi, consider, politic, incorreta, professor, Thaís, Nicoleti, congresso, PSDB, anteontem, Brasília, disse, quer, brasileir, melhor, educ, não, liderad, gente, desprez, educ, começ, própria

Termos resultants – n-grams

- Os termos resultantes são agrupados em n-grams, que são a combinação dos termos, conforme aparecem no texto
  - Ex. 2-grams:
    - frase<>ex-presidente, ex-presidente<>Fernando, Fernando<>Henrique, Henrique<>Cardoso, Cardoso<>escorreg, escorreg<>português, português<>formal, formal<>critic, critic<>indireta, indireta<>presidente, presidente<>Luiz, Luiz<>Inácio.....

http://www1.folha.uol.com.br/folha/brasil/ult96u348430.shtml

Contagem de Frequência

- A contagem de frequência consiste em verificar a ocorrência dos termos (n-grams) na lista de termos resultantes
  - Exemplo

1-gram

Ngram	Frequency
statistical	179
queries	179
david	177
framework	177
present	176
wess	176
publishers	174
task	172
phd	171
time	168
show	168
examples	168
multiple	167
concept	167
understanding	166

2-gram

Ngram	Frequency
international<>works...	184
the<>acm	182
for<>information	182
on<>inductive	173
conf<>on	173
intelligence<>pages	170
system<>for	163
national<>conference	163
pages<>springer	161
international<>joint	159
san<>mateo	159
volume<>of	159
phd<>thesis	158
joint<>conference	155
natural<>language	155

3-gram

Ngram	Frequency
european<>workshop<>on	146
and<>development<>in	141
on<>inductive<>logic	140
acm<>sigir<>conference	140
this<>paper<>we	139
on<>research<>and	136
of<>lecture<>notes	136
science<>university<>of	129
international<>acm<>sigir	127
development<>in<>information	126
conference<>on<>case-based	125
annual<>international<>acm	123
the<>use<>of	122
computer<>science<>university	120
san<>mateo<>ca	120

Contagem de Frequência

- Nesta etapa é comum a eliminação de termos que não estejam dentro de uma frequência desejada

1-gram

Ngram	Frequency
statistical	179
queries	179
david	177
framework	177
present	176
wess	176
publishers	174
task	172
phd	171
time	168
show	168
examples	168
multiple	167
concept	167
understanding	166

2-gram

Ngram	Frequency
international<>works...	184
the<>acm	182
for<>information	182
on<>inductive	173
conf<>on	173
intelligence<>pages	170
system<>for	163
national<>conference	163
pages<>springer	161
international<>joint	159
san<>mateo	159
volume<>of	159
phd<>thesis	158
joint<>conference	155
natural<>language	155

3-gram

Ngram	Frequency
european<>workshop<>on	146
and<>development<>in	141
on<>inductive<>logic	140
acm<>sigir<>conference	140
this<>paper<>we	139
on<>research<>and	136
of<>lecture<>notes	136
science<>university<>of	129
international<>acm<>sigir	127
development<>in<>information	126
conference<>on<>case-based	125
annual<>international<>acm	123
the<>use<>of	122
computer<>science<>university	120
san<>mateo<>ca	120

Representação Vetorial

- Por fim, temos a matriz termos por documentos ou Modelo de Espaço Vetorial

	$term_1$	$term_2$	$term_3$	$term_4$	...	$term_m$
Doc_1	10	1	3	0	...	1
Doc_2	3	11	100	3	...	33
Doc_3	2	0	0	44	...	77
...	...	...	...	...	...	...
Doc_n	2	12	2	92	...	0

Ponderação de Termos

- Como podemos medir a importância de um termo no documento?
  - Como podemos atribuir pesos para os termos?

	$term_1$	$term_2$	$term_3$	$term_4$	...	$term_m$
Doc_1	10	1	3	0	...	1
Doc_2	3	11	100	3	...	33
Doc_3	2	0	0	44	...	77
...	...	...	...	...	...	...
Doc_n	2	12	2	92	...	0

Ponderação de Termos

- Como podemos medir a importância de um termo no documento?
  - Como podemos atribuir pesos para os termos?
- Uma das maneira mais utilizadas é conhecida como TF-IDF (term frequency–inverse document frequency)
  - TF * IDF

Ponderação de Termos

- TF-IDF
  - A TF é a frequência do termo *i* em um documento *j*
  - A IDF de um termo *i* é dado por $\log\left(\frac{N}{df_i}\right)$
  - Em que
    - *N* é a quantidade de documentos da coleção
    - *dfi* é a quantidade de documentos em que o termo *i* aparece
  - O novo valor do termo *i* para o documento *j* é calculado como

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Ponderação dos Termos

- Calculo da IDF (N = 10)
  - Termo ‘example’
    - IDF = log(N/df) = log(10/10) = log(1) = 0
  - Termo ‘visualization’
    - IDF = log(N/df) = log(10/6) = log(1,66) = 0,22

Doc/Term	example	visualization	computer	book	artificial
Doc1	10	5	6	8	11
Doc2	15	8	7	4	12
Doc3	2	7	0	6	10
Doc4	9	0	8	1	13
Doc5	8	5	3	7	0
Doc6	13	3	12	10	14
Doc7	17	0	0	5	18
Doc8	5	0	9	6	9
Doc9	4	0	8	11	7
Doc10	1	1	0	0	6

Ponderação dos Termos

- Calculo da IDF (N = 10)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

  - Termo ‘example’
    - IDF = log(N/df) = log(10/10) = log(1) = 0
  - Termo ‘visualization’
    - IDF = log(N/df) = log(10/6) = log(1,66) = 0,22

Doc/Term	example	visualization	computer	book	artificial
Doc1	10 * 0	5 * 0,22	6	8	11
Doc2	15 * 0	8 * 0,22	7	4	12
Doc3	2 * 0	7 * 0,22	0	6	10
Doc4	9 * 0	0 * 0,22	8	1	13
Doc5	8 * 0	5 * 0,22	3	7	0
Doc6	13 * 0	3 * 0,22	12	10	14
Doc7	17 * 0	0 * 0,22	0	5	18
Doc8	5 * 0	0 * 0,22	9	6	9
Doc9	4 * 0	0 * 0,22	8	11	7
Doc10	1 * 0	1 * 0,22	0	0	6

Ponderação dos Termos

- Calculo da IDF (N = 10)

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

  - Termo ‘example’
    - IDF = log(N/df) = log(10/10) = log(1) = 0
  - Termo ‘visualization’
    - IDF = log(N/df) = log(10/6) = log(1,66) = 0,22

Doc/Term	example	visualization	computer	book	artificial
Doc1	0	1,10	6	8	11
Doc2	0	1,76	7	4	12
Doc3	0	1,54	0	6	10
Doc4	0	0	8	1	13
Doc5	0	1,10	3	7	0
Doc6	0	0,66	12	10	14
Doc7	0	0	0	5	18
Doc8	0	0	9	6	9
Doc9	0	0	8	11	7
Doc10	0	0,22	0	0	6

Visualização de Coleções de Documentos

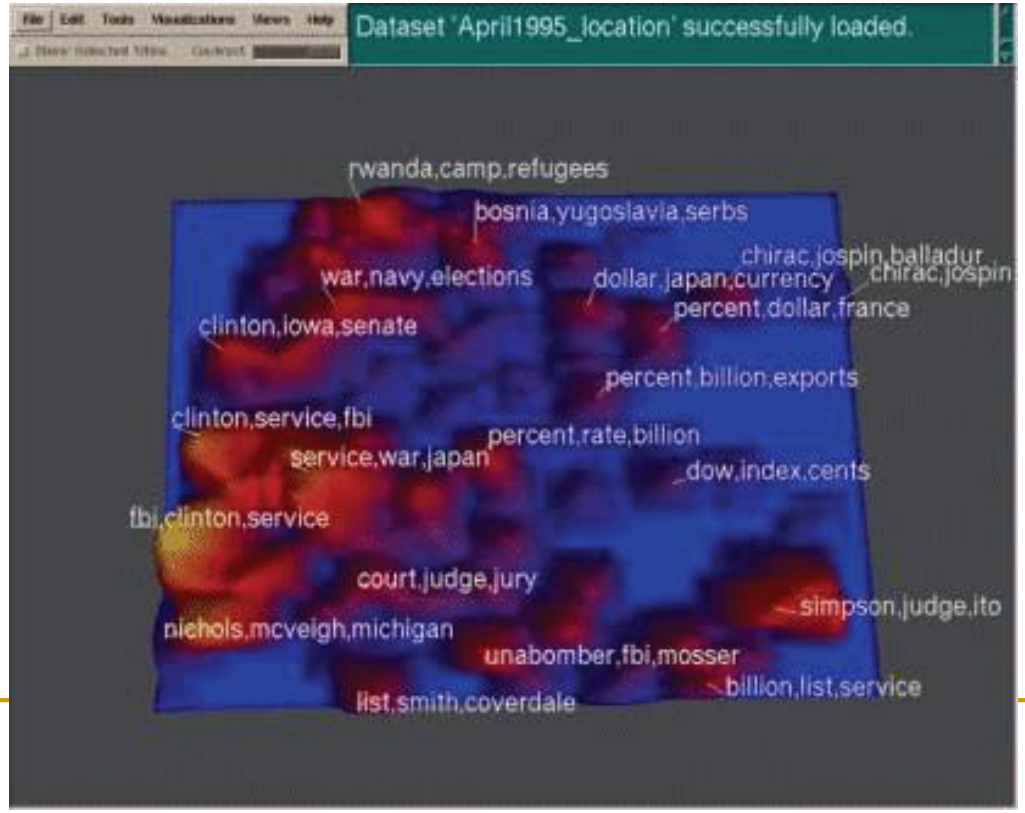
- Uma vez que o modelo de espaço vetorial é construído podemos calcular a similaridade entre os documentos
  - Por exemplo, distância Euclidiana entre os vetores
- Além da visualização, técnicas de mineração de dados podem ser utilizadas para classificar ou agrupar a coleção de documentos

Visualização de Coleções de Documentos

- Algumas das técnicas mais utilizadas para visualizar coleções de documentos são
  - Desenho de grafos baseados em força
  - Self-organizing maps
  - Projeções Multidimensionais

Themescape

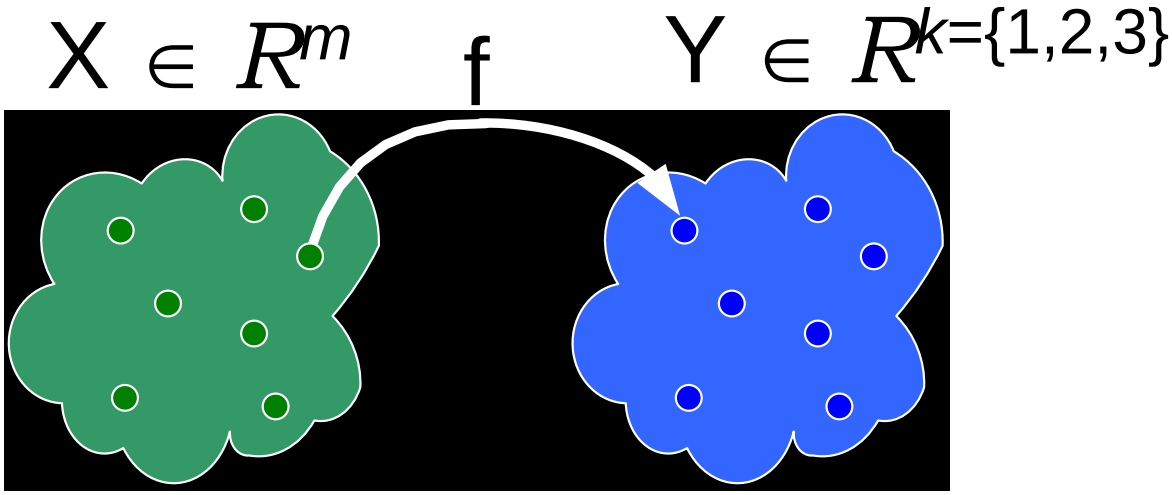
- A Themescape utiliza uma paisagem 3D abstrata com alturas e cores para representar a densidade de documentos similares
  - Abaixo é apresentada uma visualização de artigos de notícias



Projeções Multidimensionais

- As Projeções Multidimensionais reduzem a dimensionalidade do conjunto de dados para um espaço de menor dimensão
- No caso de coleções de documentos, a redução é aplicada no modelo de espaço vetorial
  - Reduzindo o espaço para duas ou três dimensões

Projeção Multidimensional



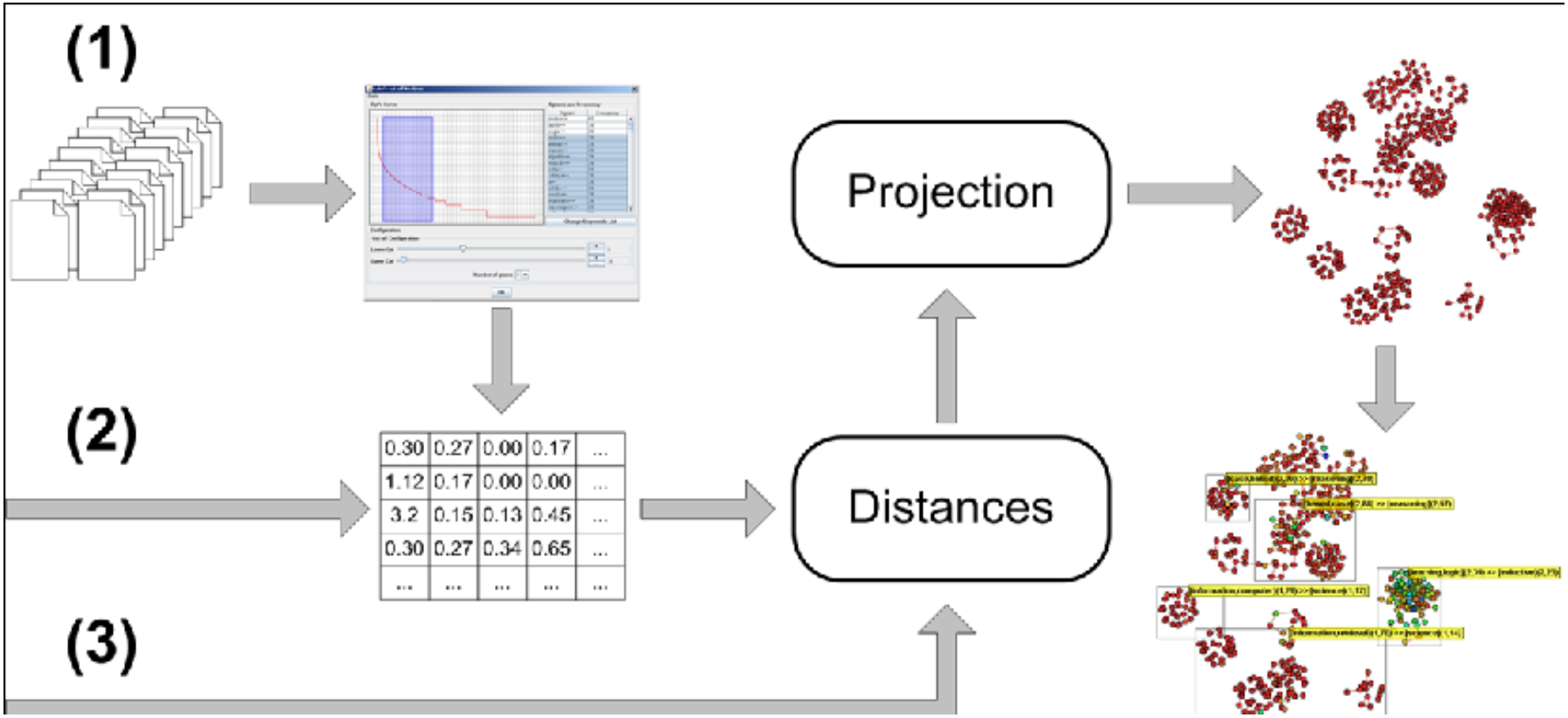
δ: $x_i, x_j \rightarrow \mathbb{R}$, $x_i, x_j \in X$

d: $y_i, y_j \rightarrow \mathbb{R}$, $y_i, y_j \in Y$

f: $X \rightarrow Y, |\delta(x_i, x_j) - d(f(x_i), f(x_j))| \approx 0, \forall\ x_i, x_j \in X$

Projeção Multidimensional

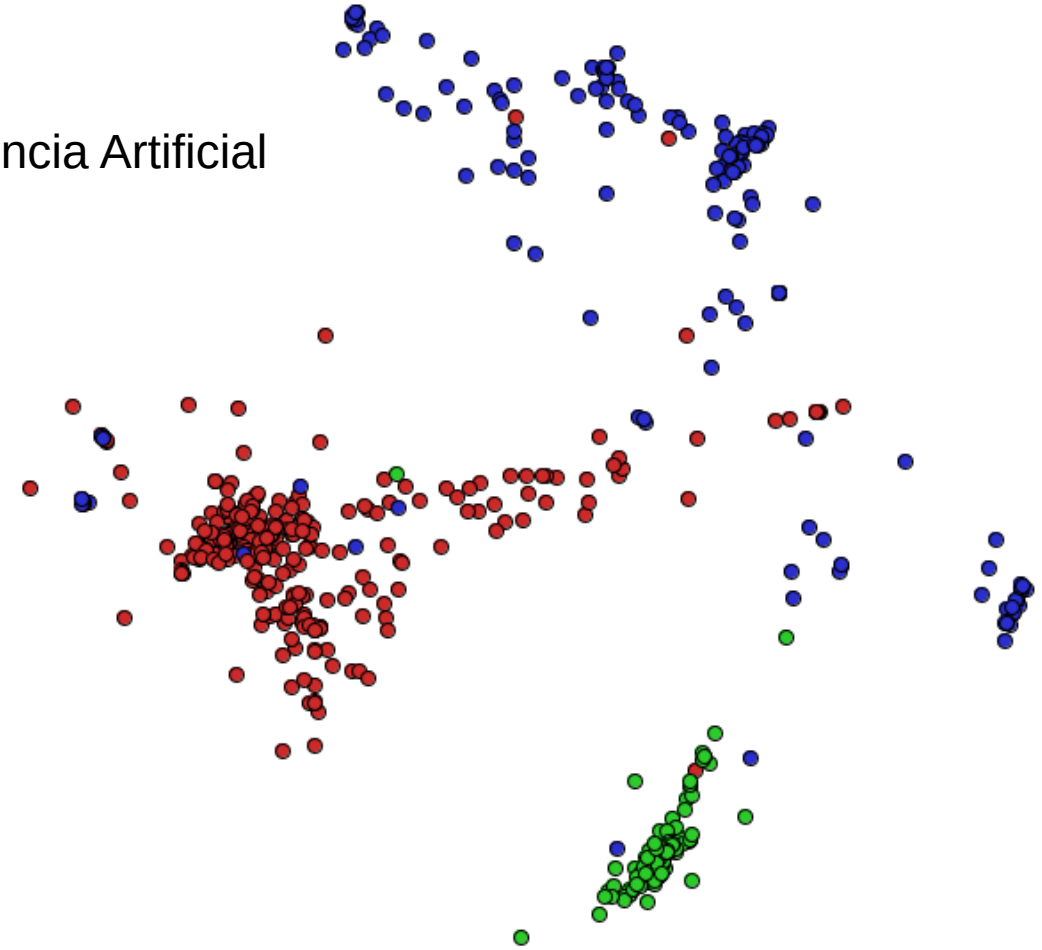
- Pipeline



Projeção Multidimensional

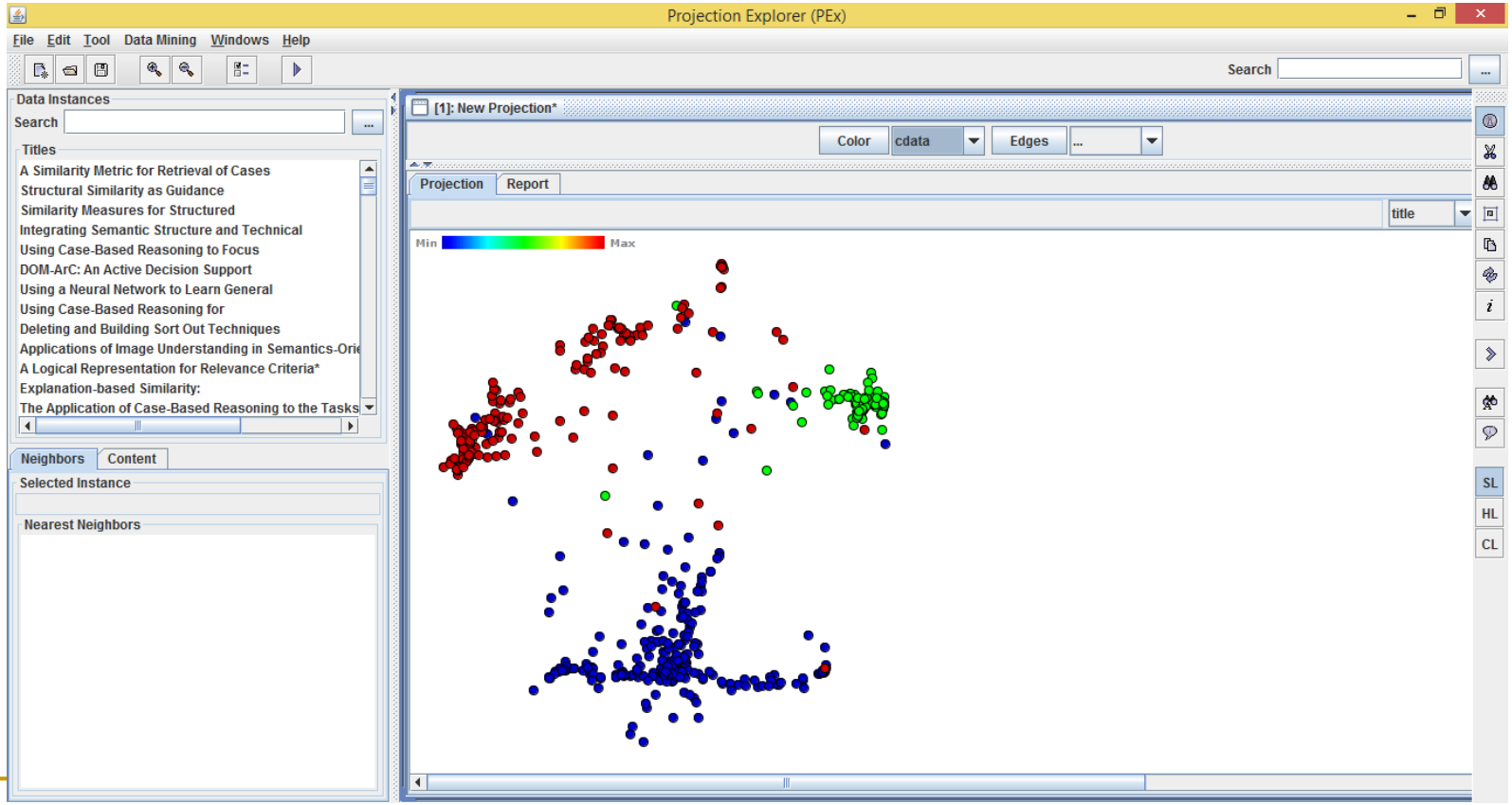
574 artigos da área de Inteligência Artificial

- Case based reasoning
- Inductive logic programming
- Information retrieval



Projection Explorer (PEx)

- A PEx é uma ferramenta desenvolvida para explorar conjuntos de dados por meio de projeções multidimensionais

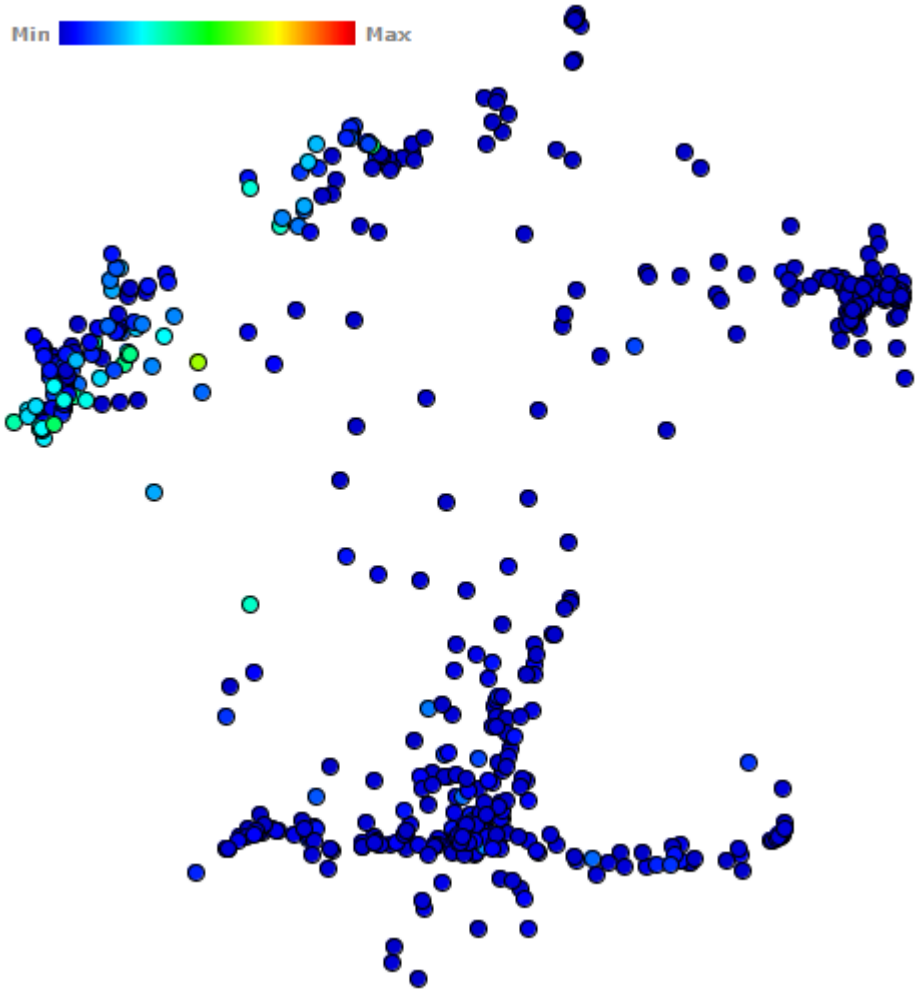


Projection Explorer (PEx)

- Nem sempre os conjuntos de dados são rotulados
  - Como no exemplo ao lado
- Por isso, algumas ferramentas de interação auxiliam a exploração do conjunto

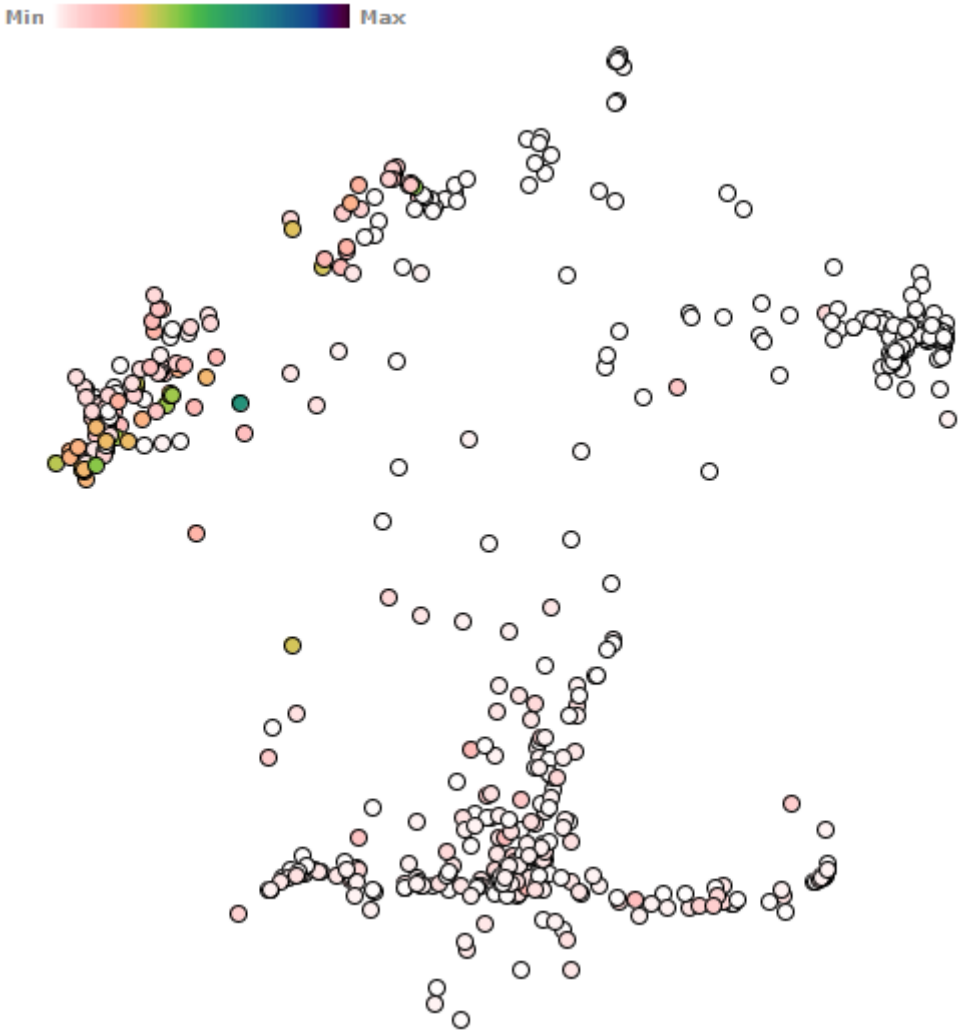
Projection Explorer (PEx)

- Consulta por palavras chave
  - No caso, a chave de busca foi a palavra “retrieval”



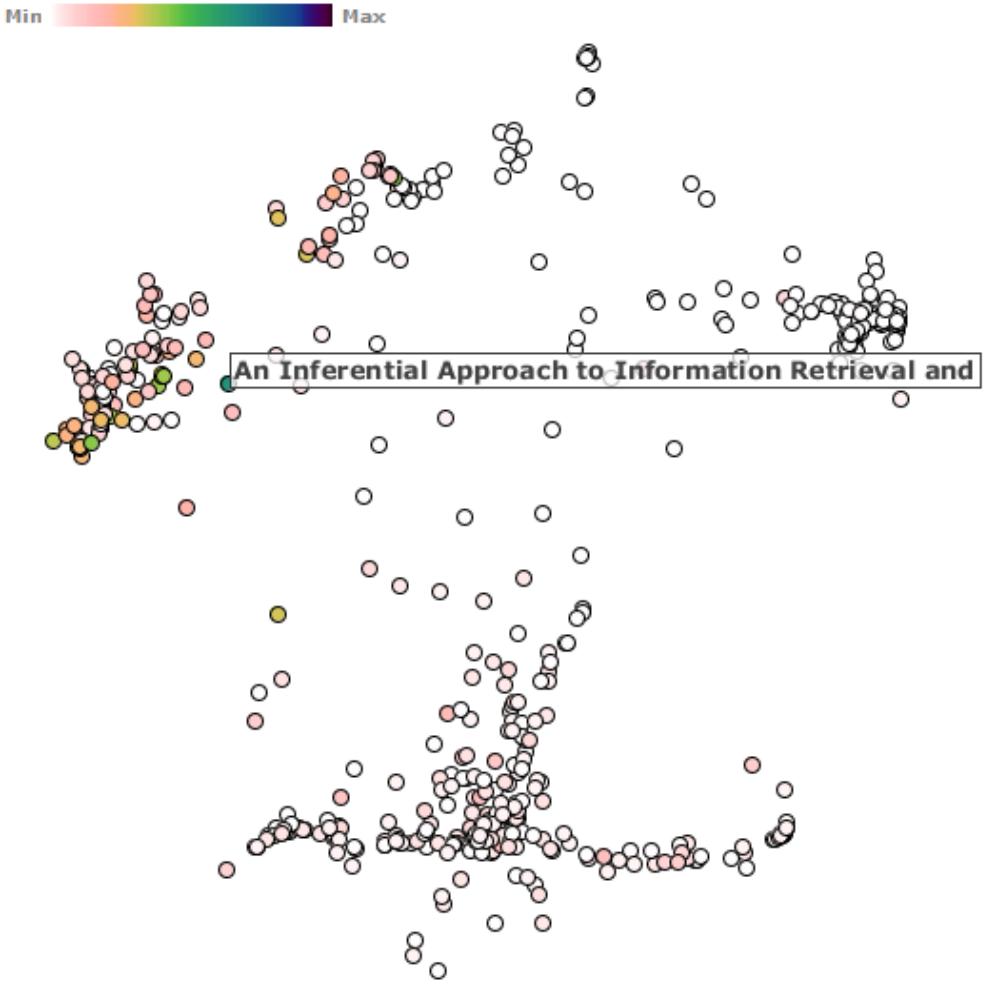
Projection Explorer (PEx)

- Consulta por palavras chave
  - No caso, a chave de busca foi a palavra “retrieval”
- Outra escala de cor pode facilitar a identificação da frequência da palavra



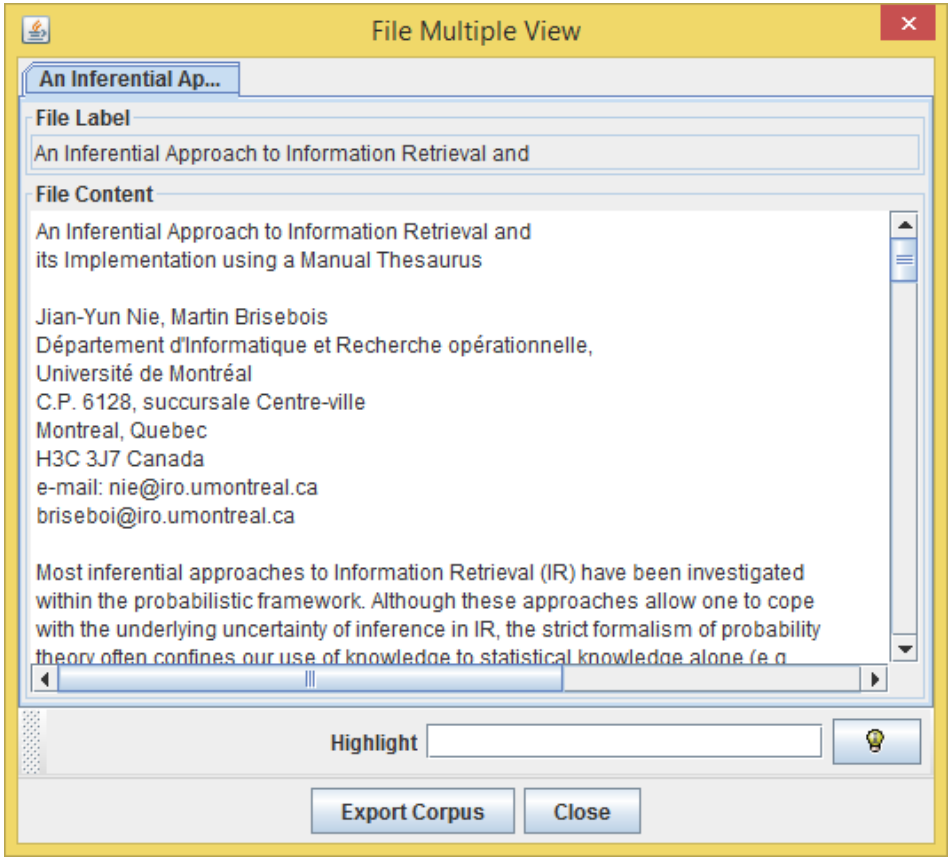
Projection Explorer (PEx)

- Consulta por palavras chave
  - No caso, a chave de busca foi a palavra “retrieval”
- Outra escala de cor pode facilitar a identificação da frequência da palavra



Projection Explorer (PEx)

- Consulta por palavras chave
  - No caso, a chave de busca foi a palavra “retrieval”
- O conteúdo do documento selecionado pode ser visualizado

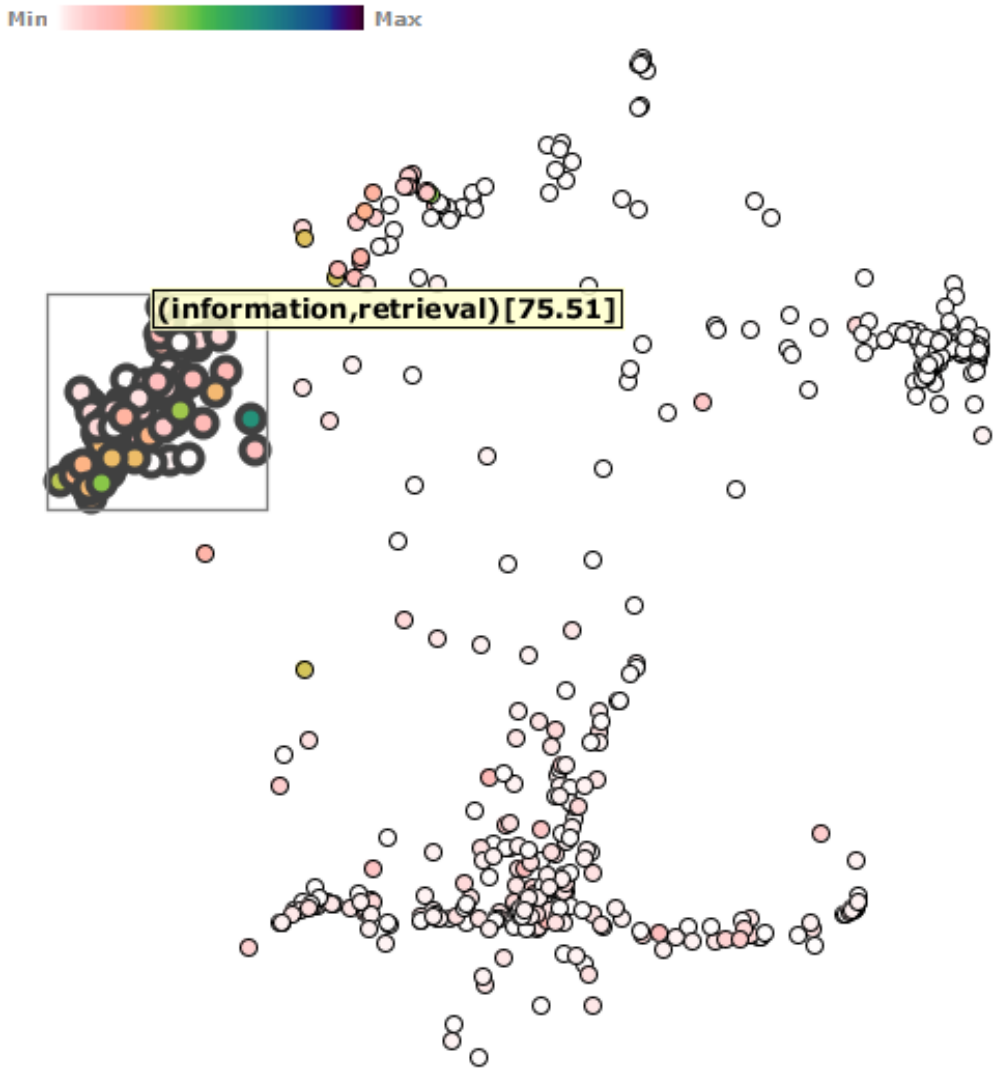


Detecção de Tópicos

- Uma técnica importante na exploração de coleção de documentos é a detecção de tópicos

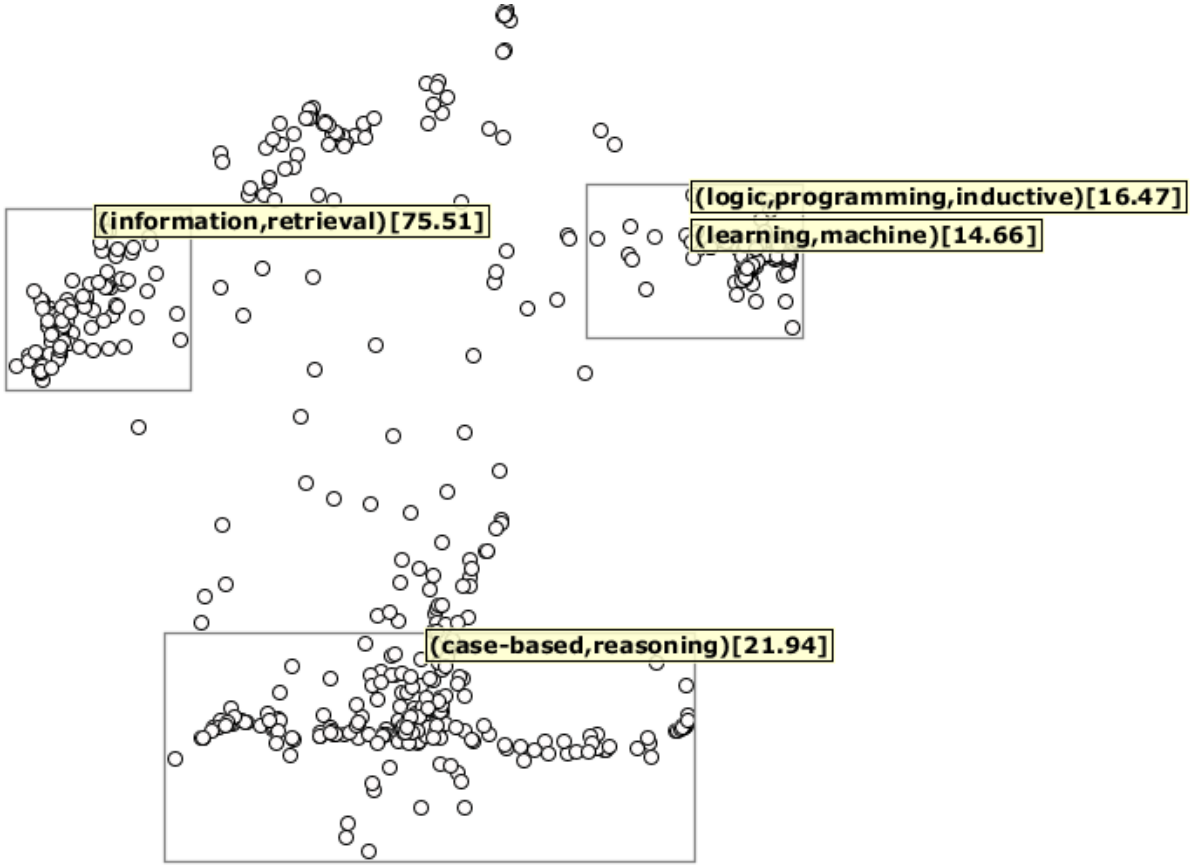
Detecção de Tópicos

- Uma técnica importante na exploração de coleção de documentos é a detecção de tópicos



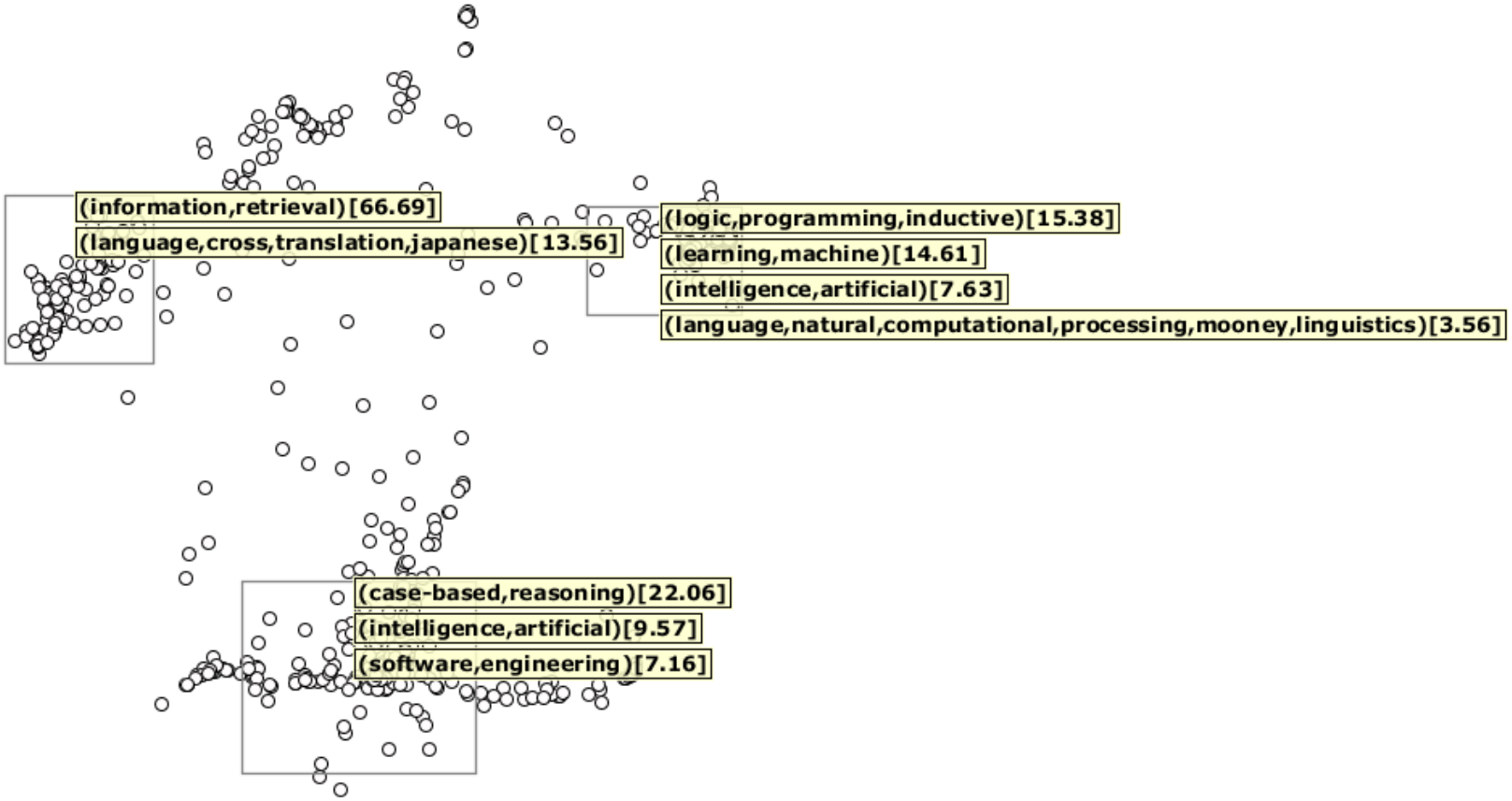
Detecção de Tópicos

- Seleção dos três grupos bem definidos



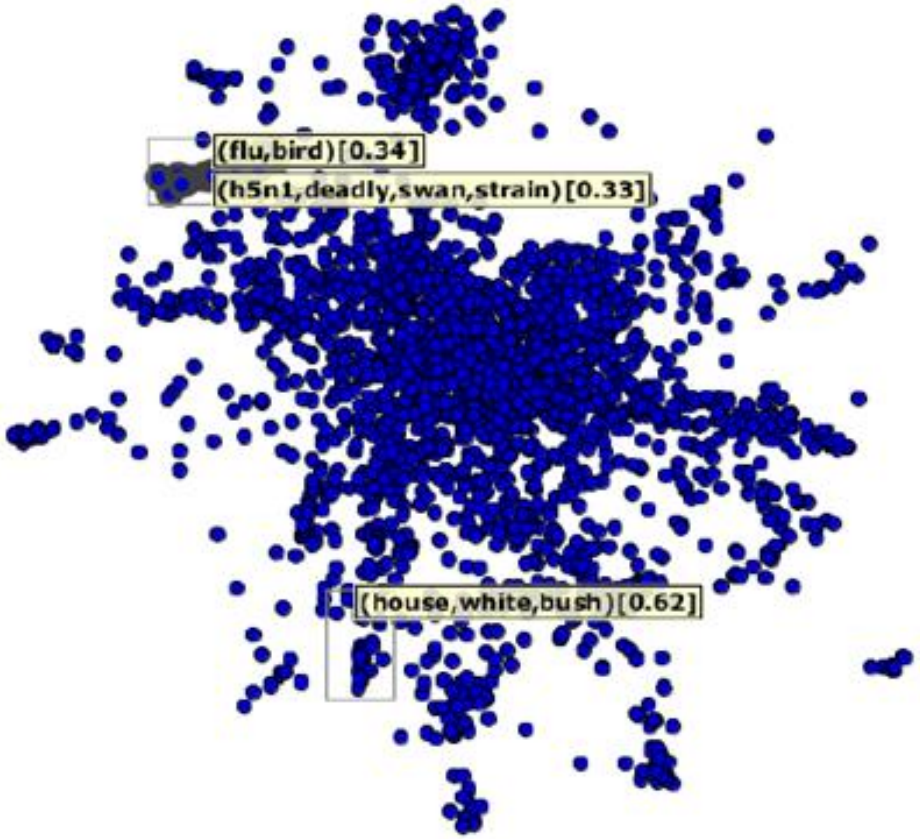
Detecção de Tópicos

- Seleção dos três grupos bem definidos
  - Alteração das configurações do método de detecção



Detecção de Tópicos

- Seleção dos dois grupos bem definidos em uma coleção de notícias



Referências

- Ward, M., Grinstein, G. G., Keim, D.
  - Interactive data visualization foundations, techniques, and applications. Natick, Mass., A K Peters, 2a Edição, 2010.
    - Capítulo 10 (Text and Document Visualization)
- G. Salton, A. Wong, and C. S. Yang.
  - “A Vector Space Model for Automatic Indexing.” Commun. ACM 18:11 (1975), 613–620
- M. Wattenberg and F. B. Viégas
  - “The Word Tree, an Interactive Visual Concordance.” IEEE Transactions on Visualization and Computer Graphics 14:6 (2008), 1221–1228.
- Jonathan Feinberg.
  - “Wordle Home Page.” http://www.wordle.net/, accessed August 31, 2009.
- WordTree
  - IBM. “Many Eyes Home Page.” http://manyeyes.alphaworks.ibm.com/, accessed August 31, 2009.

Referências

- W. B. Paley.
  - “TextArc: ShowingWord Frequency and Distribution in Text.” Poster presented at IEEE Symposium on Information Visualization, Boston, MA, October 27–November 1, 2002.
  - http://www.textarc.org/
- D. A. Keim and D. Oelke.
  - “Literature Fingerprinting: A New Method for Visual Literary Analysis.” In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST 2007), pp. 115–122. Los Alamitos, CA: IEEE Computer Society Press, 2007.
- T. Kohonen.
  - Self-Organizing Maps, Springer Series in Information Sciences, 30, Third edition. Berlin: Springer, 2001.
- Steffen Lohmann ; Florian Heimerl ; Fabian Bopp ; Michael Burch ; Thomas Ertl
  - Concentri Cloud: Word Cloud Visualization for Multiple Text Documents. In 19th International Conference on Information Visualisation, 2015
  - https://ieeexplore.ieee.org/abstract/document/7272588

Referências

- Projection Explorer (PEx)
  - F. V. Paulovich, M. C. F. Oliveira, and R. Minghim
    - “The projection explorer: A flexible tool for projection-based multidimensional visualization”, in XX Brazilian Symposium on Computer Graphics and Image Processing. Washington, DC, USA: IEEE Computer Society, 2007, pp. 27–36.
      - http://vis.icmc.usp.br/vicg/tool/1/projection-explorer-pex
- Projection Explorer for Images (PEx-Image)
  - D. M. Eler, M. Nakazaki, F. Paulovich, D. Santos, G. Andery, M. Oliveira, J. E. S. Batista, and R. Minghim
    - “Visual analysis of image collections”, The Visual Computer, vol. 25, no. 10, pp. 923–937, 2009.
      - https://github.com/daniloeler/PEx-Image