# 16

# VISUALIZAÇÃO DE TEXTOS

*Profa. Raquel C. de Melo Minardi*

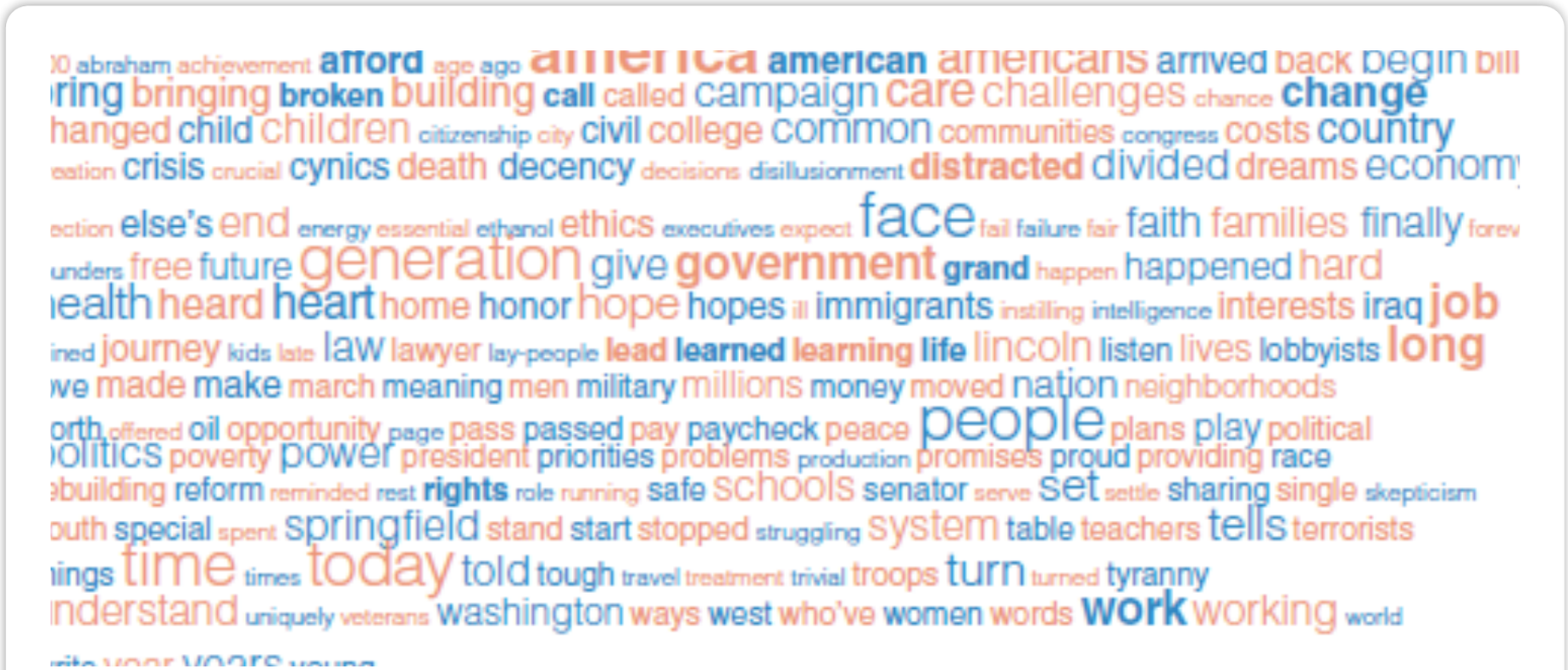
# NUVENS DE TERMOS

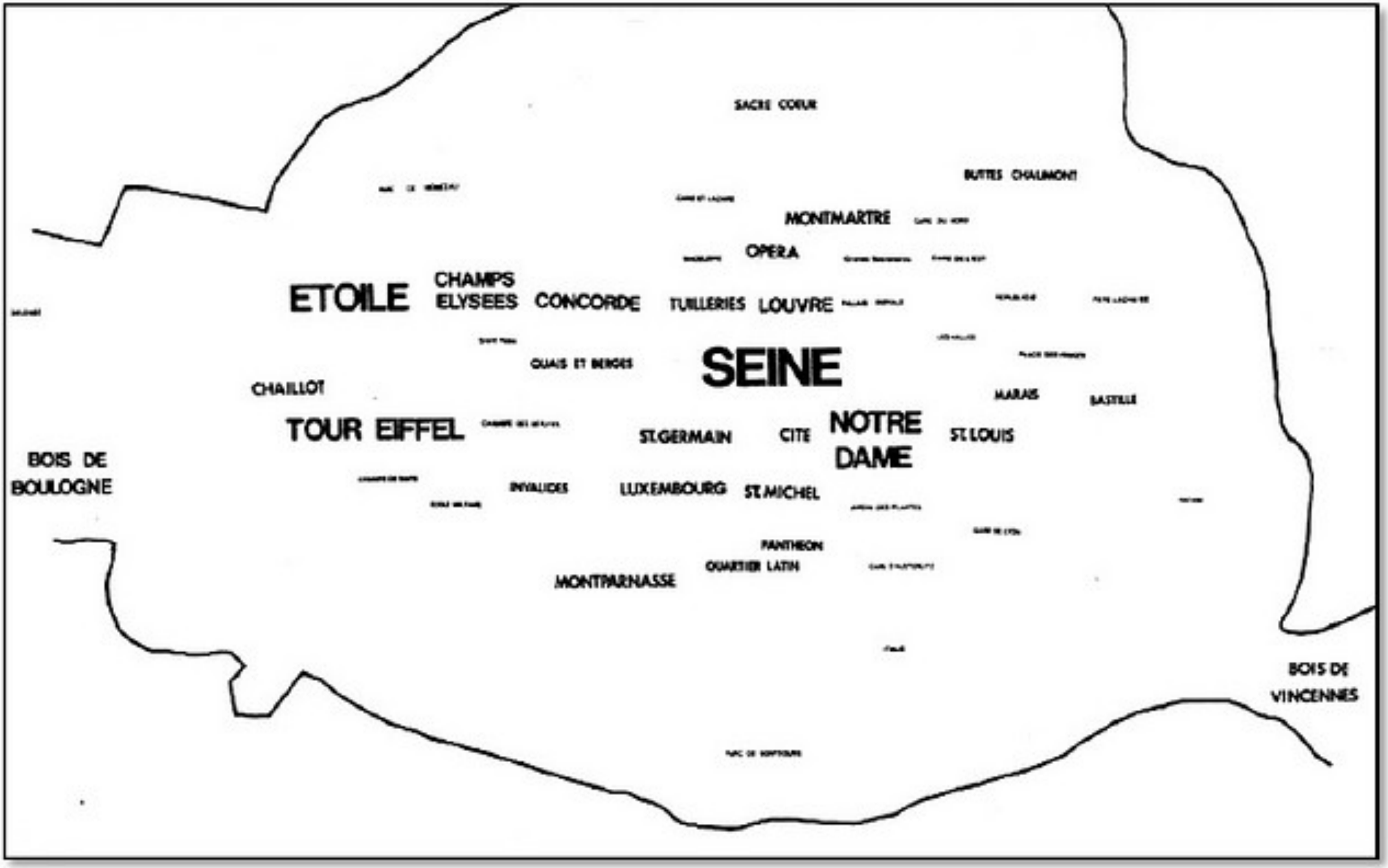
- O objetivo é apresentar um resumo visual de uma coleção de textos

# TAG CLOUDS AND THE CASE OF VERNACULAR VISUALIZATION

*F.B. Viégas e M. Wattenberg*

*Interactions*
*2009*



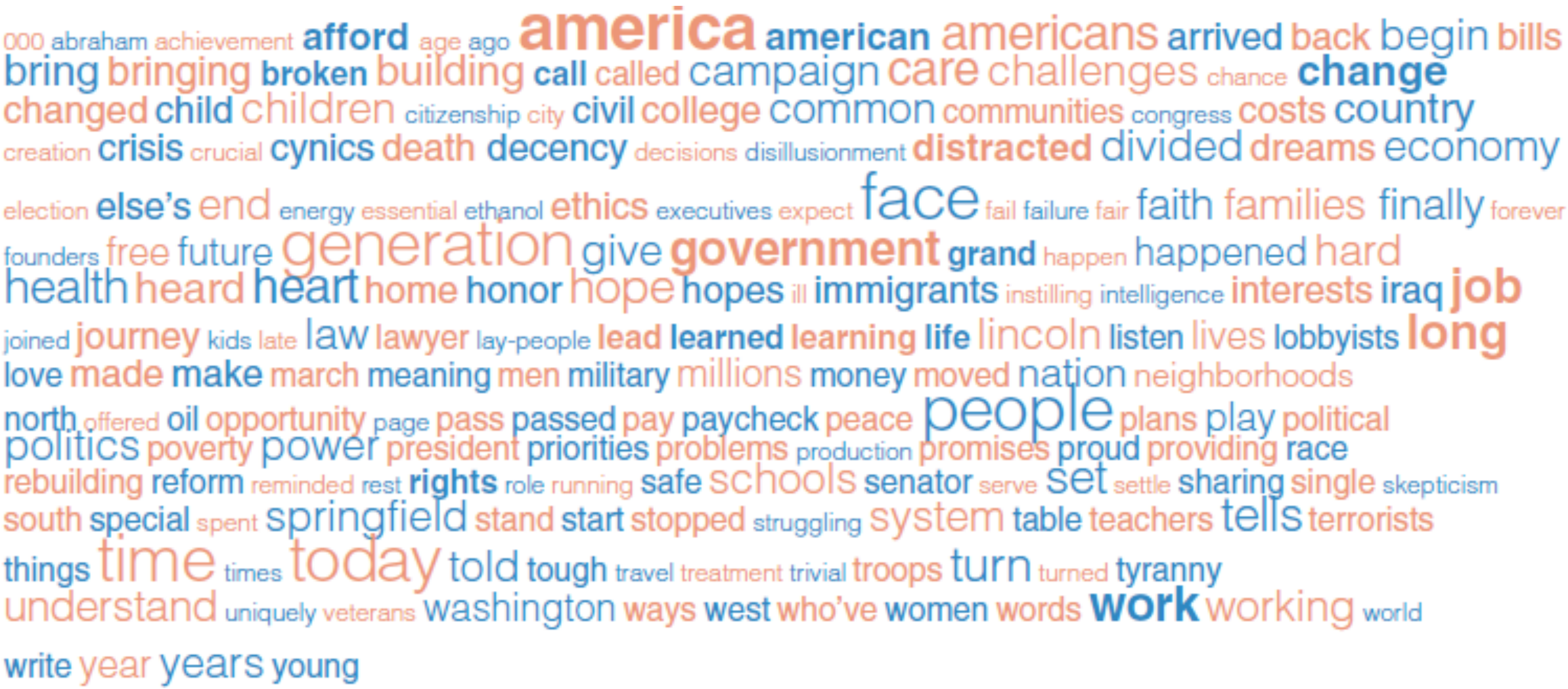



Milgram, mapa mental de Paris, 1976

07 africa amsterdam animals architecture art asia australia autumn baby band barcelona beach berlin birthday black blackandwhite blue boston bw california cameraphone camping canada canon car cat chicago china christmas church city clouds color concert cute day de dog england europe fall family festival film florida flower flowers food france friends fun garden geotagged germany girl graffiti green halloween hawaii hiking holiday home honeymoon house india ireland island italy japan july kids la lake landscape light live london macro march me mexico mountain mountains museum music nature new newyork newyorkcity newzealand night nikon nyc ocean paris park party people photo photos portrait red river rock rome san sanfrancisco scotland sea seattle show sky snow spain spring street summer sun sunset sydney taiwan texas thailand tokyo toronto tour travel tree trees trip uk urban usa vacation vancouver washington water wedding white winter yellow york zoo

Tags mais populares do Flicker, 2002

Fontes são representadas em **escala logarítmica** com relação à frequência dos termos




Discurso de Obama em 2007, Many Eyes


Versão bigrama do
discurso de Obama em 2007, Many Eyes

## VANTAGENS

- Interface amigável para visualização de informações complexas
- Diversão
- Espelho de comportamento de grupo ou individual

## DESVANTAGENS

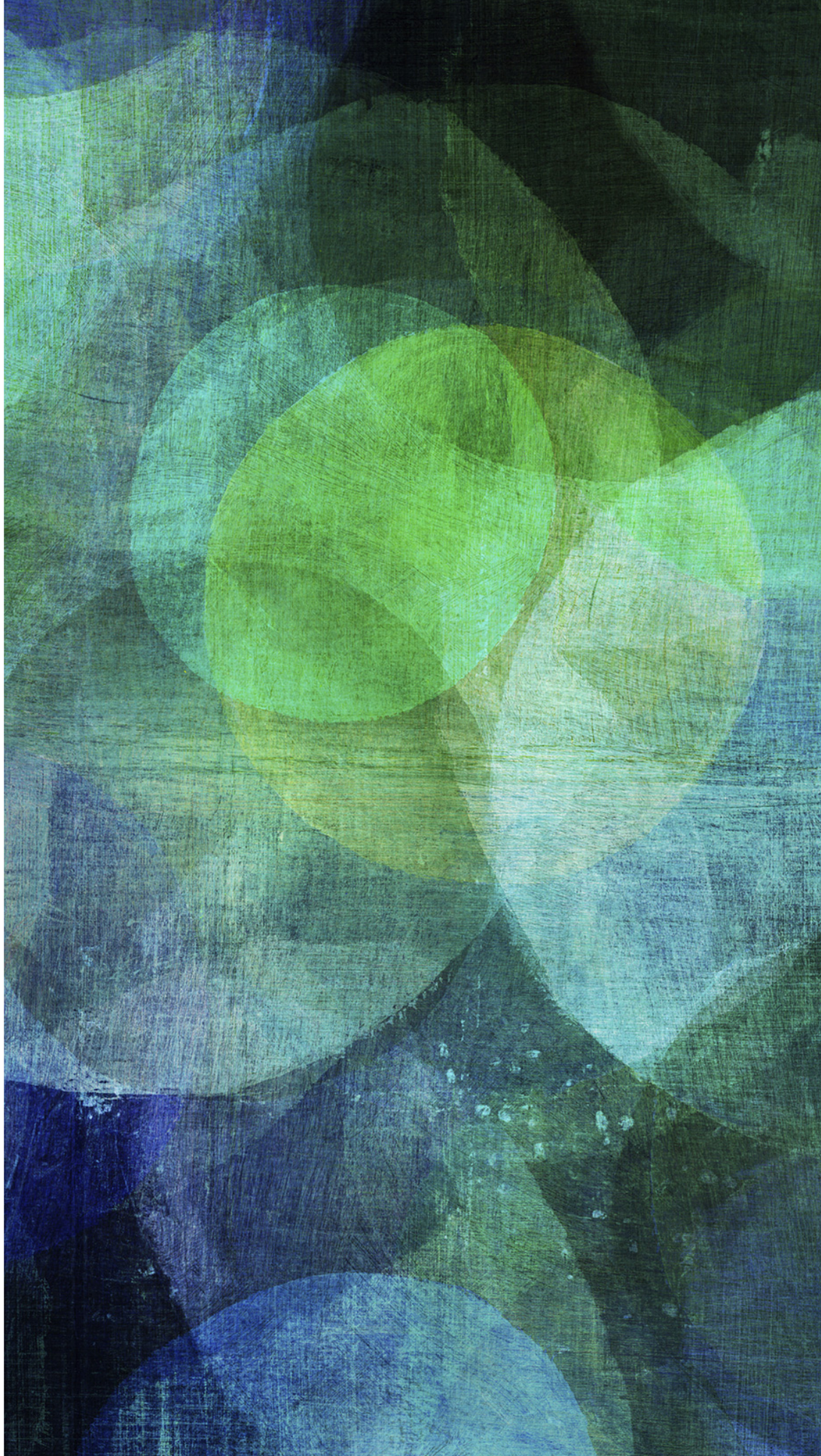
- Longas palavras têm ênfase com relação a pequenas palavras
- Dificuldade de localização de palavras
- Posicionamento não tem significado

“

Pode-se dizer que as nuvens de termos são úteis na prática mas não na teoria.

*-Fernanda Viégas e Martin Wattenberg*

# ALGORITMOS



# TAG CLOUD DRAWING: ALGORITHMS FOR CLOUD VISUALIZATIONS

*O. Kaser e D. Lemire*
*WWW*
*2007*

# DIAGRAMAÇÃO

- Sistemas de diagramação automáticos como o TEX devem ser capazes de organizar o texto em uma página de forma estética
- O resultado precisa ser visualmente atraente
- Linhas precisam ser quebradas de forma que haja um espaçamento semelhante entre as palavras

A greedy approach fits as many words per line as possible, beginning a new line whenever further words cannot be placed on the current line, with the possibility of sometimes slightly squeezing the spaces between words and letters or hyphenating a word.

- Uma **abordagem gulosa** consiste em colocar quantas palavras couberem em uma linha espaçando as igualmente e quebrando a linha assim que não couberem mais palavras
  - Abordagem utilizada pelos browsers, que não utilizam hifenação

- **Vantagem**: solução *on-line*, não é necessário carregar o parágrafo completo para escrever as palavras
- **Desvantagem**: frequentemente a heurística leva a soluções sub-ótimas

# ALGORITMO DE KNUTH-PLASS

- O algoritmo de Knuth e Plass computa uma solução ótima usando programação dinâmica
- Os autores definem uma métrica chamada *badness* para cada possível quebra de linha
  - diferença entre tamanho preferido da linha ou coluna e o tamanho obtido
- $O(n^2)$
- O algoritmo visa minimizar a soma de quadrados dos *badness* de cada linha

# ALGORITMO DE KNUTH-PLASS

- Rotula-se as palavras de um texto com identificadores entre *1* e $n$
- $B_{k,j}$ é a medida de *badness* da linha que contem as palavras entre $k$ e $j$ inclusive
- Convenciona-se que $B_{k,j} = 0$ se $k > j$
- $T_j$ é a mínima soma de quadrados de *badness* possível quando a linha termina com a palavra $j$
  - $T_j = min_{k \leq j}(T_k + B^2_{k+1,j})$ e $T_0 = 0$
  - $K_j = arg\ min_k\ (T_k + B^2_{k+1,j})$ é a última palavra da linha anterior à linha que termina com a palavra $j$

$Palavra_1$ $Palavra_2$ $Palavra_3$ $Palavra_4$ $Palavra_5$ $Palavra_6$ $Palavra_7$ ...
$Palavra_{n-3}$ $Palavra_{n-2}$ $Palavra_{n-1}$ $Palavra_n$

.....$Palavra_k$
$Palavra_{k+1}$ ... $Palavra_{j-2}$ $Palavra_{j-1}$ $Palavra_j$

Calcula-se $K_j$ para j=1,...,n em tempo $O(n^2)$ e é possível reconstruir a solução recursivamente buscando-se as quebras de linha n, $K_n$, $K_{Kn}$,...

In olden times when wish·

ing still helped one, there lived

whose daugh·ters were all

beau·ti·ful; and the youngest

beau·ti·ful that the sun it·self,

which has seen so much,

ished when·ever it shone

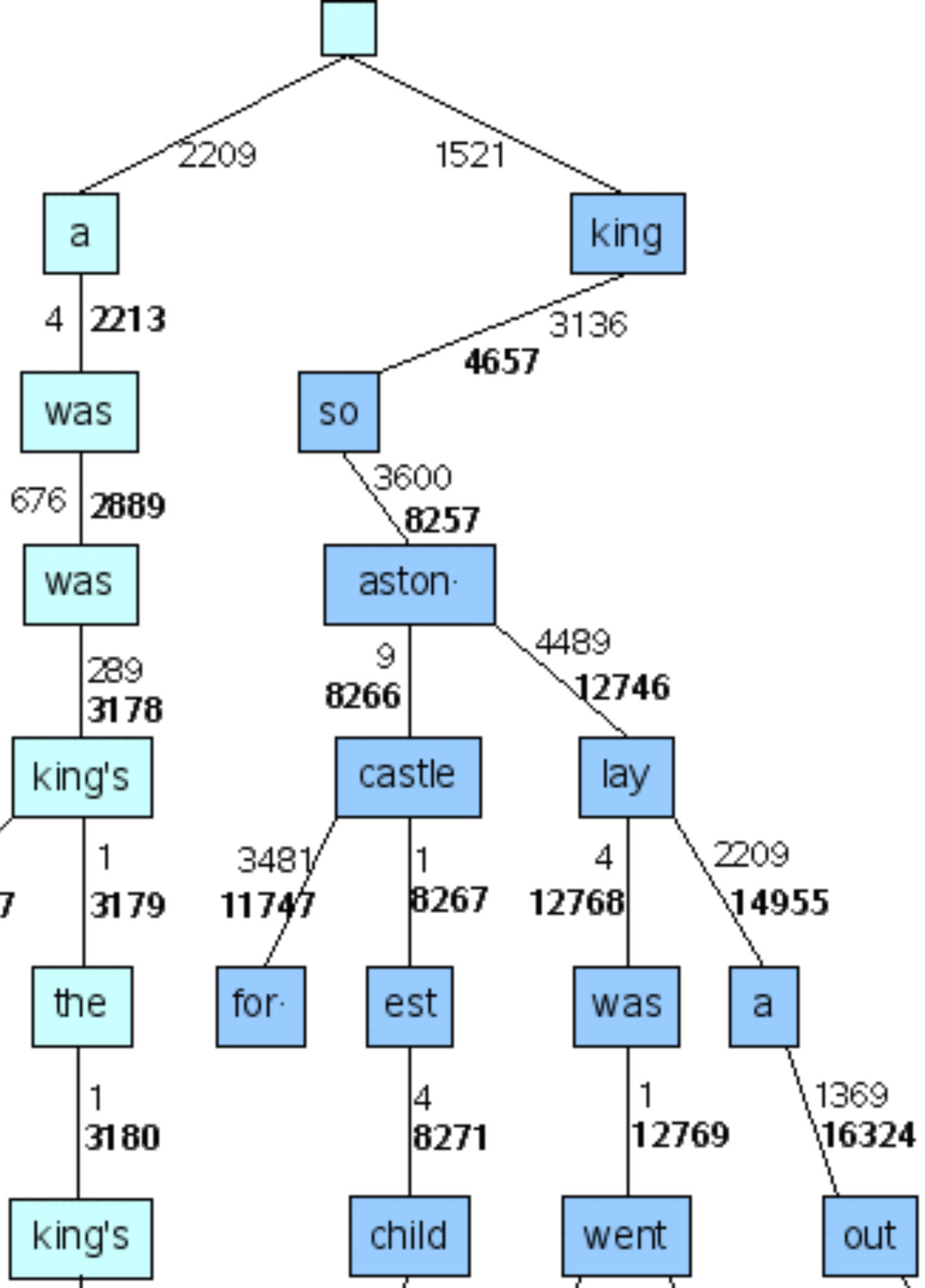
in her face. Close by the

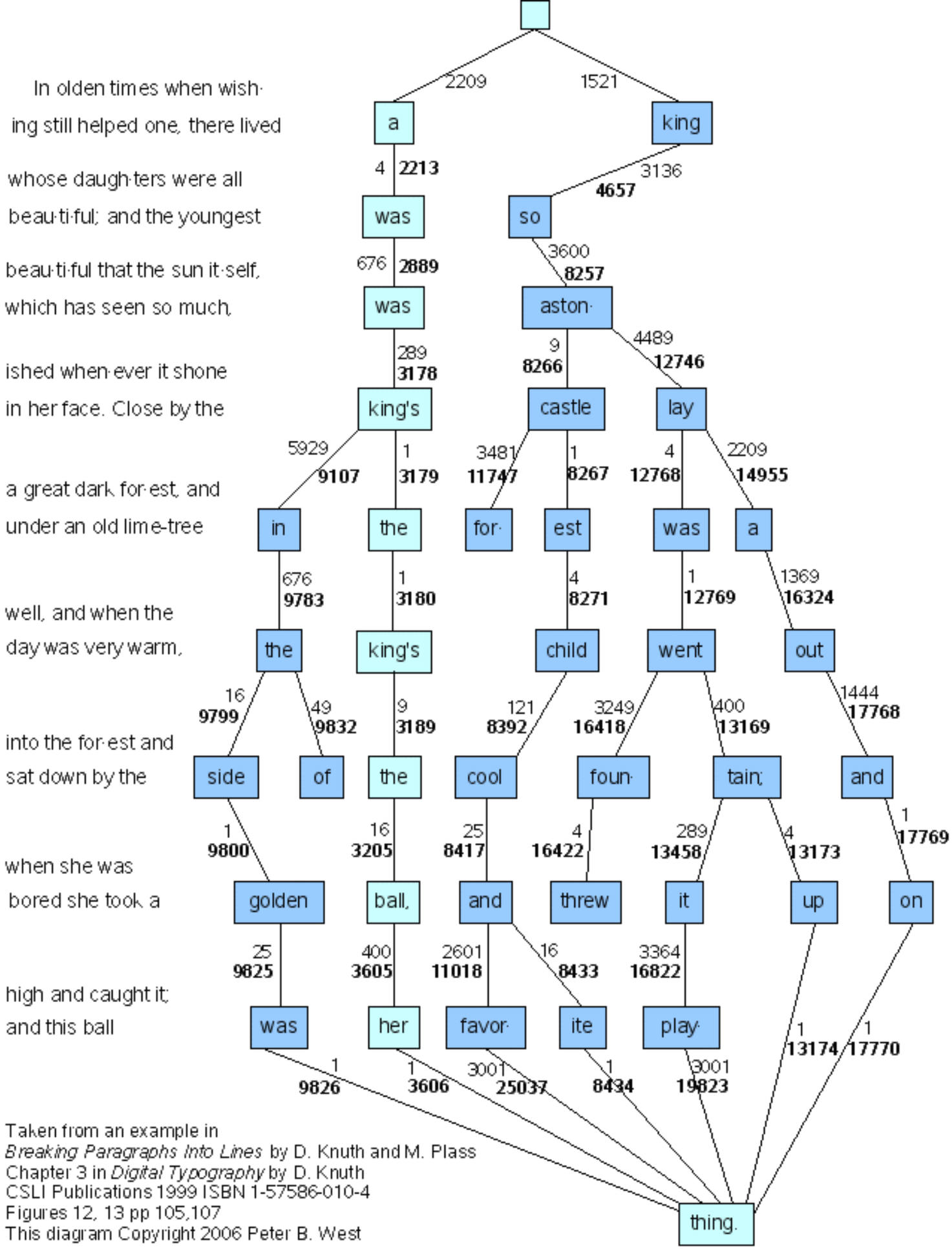
a great dark for·est, and

under an old lime-tree

well, and when the

day was very warm,






Taken from an example in
*Breaking Paragraphs Into Lines* by D. Knuth and M. Plass
Chapter 3 in *Digital Typography* by D. Knuth
CSLI Publications 1999 ISBN 1-57586-010-4
Figures 12, 13 pp 105,107

Esta abordagem pode ser usada de forma ótima quando os termos são apresentados em uma dada **ordem** e têm a **mesma altura**

Seja $w$ a largura (pré-definida) das linhas e $h$ a altura (delimitada pelo termo mais alto da linha), $w_i$ e $h_i$ a lagura e altura respectivamente do termo $i$ e seja

$$\omega = w - \Sigma\, w_i - (k-1)W$$

o espaço extra adicional requerido para separação entre os termos

O ***badness*** de uma linha é definido como

$$h \times |\omega| + \Sigma(h-h_i)w_i$$

onde $h \times |\omega|$ é uma medida do espaço em branco extra mínimo e $\Sigma(h-h_i)w_i$ é o espaço em branco devido às diferenças de alturas dos termos

# HÁ DUAS ABORDAGENS PARA IMPLEMENTAÇÃO DO LEIAUTE IN-LINE

A. Palavras pré-ordenadas

    1. Adição de palavras uma a uma até que não caibam mais palavras na linha ( O(n) )

    2. Programação dinâmica similar ao método de Knuth-Plass com o *badness* baseado em largura e altura ( $O(n^2)$ )

# HÁ DUAS ABORDAGENS PARA IMPLEMENTAÇÃO DO LEIAUTE IN-LINE

B. Palavras reordenadas na tentativa de melhorar o *badness*

    1. *Next fit decreasing height* (NFDH): itens ordenados de forma decrescente e empacotados em um nível até que não caibam mais itens e um novo nível é criado. Níveis nunca são revisitados

    2. *First fit decreasing height* (FFDH): itens ordenados de forma decrescente e empacotados no menor nível em que caibam. Um novo nível é criado quando o item não couber em nenhum nível


(a) Alphabetically sorted tags, greedy algorithm


(b) Alphabetically sorted tags, dynamic programming


(c) Tags sorted by weight, greedy algorithm


(d) FFDH heuristic

# TAG–CLOUD DRAWING: ALGORITHMS FOR CLOUD VISUALIZATIONS

*O. Kaser e D. Lemire*

*WWW*
*2007*

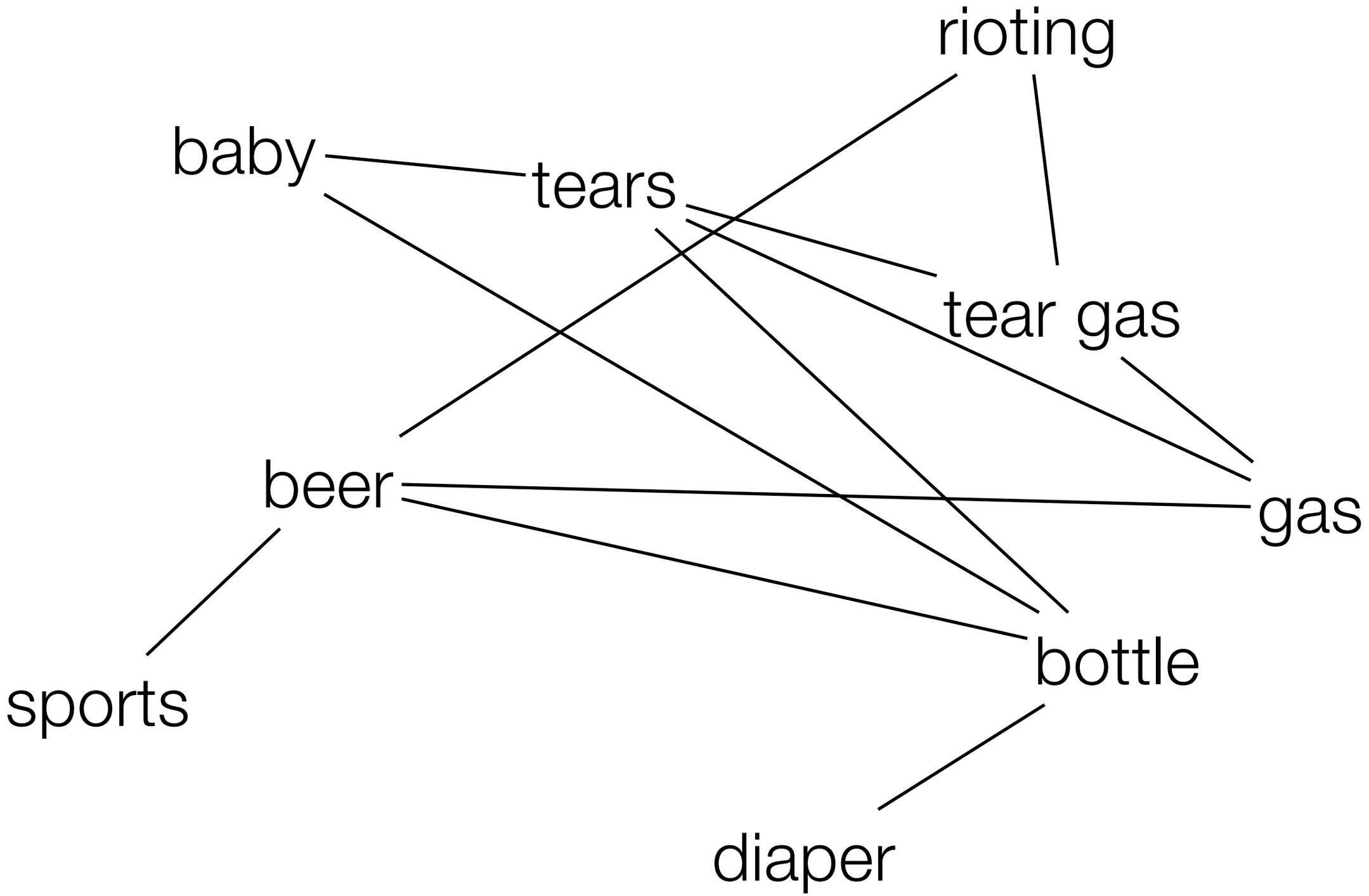

**Figure 8: Large tag cloud generated from a Project Gutenberg e-text.**

# NUVENS DE TERMOS COM POSICIONAMENTO ARBITRÁRIO

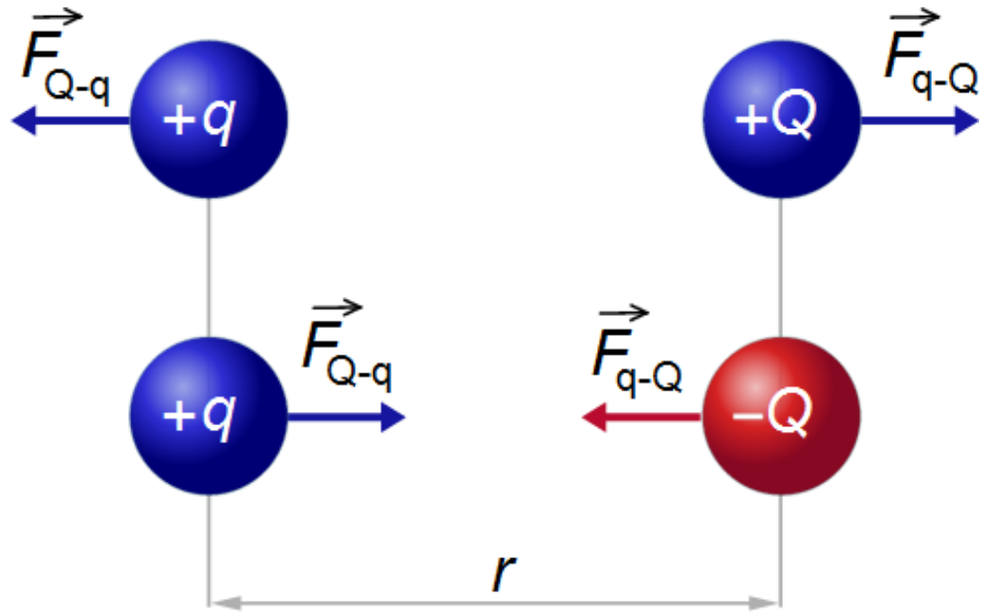
- Requisitos
  - Termos podem ser reordenados e posicionados arbitrariamente (sem sobreposição e rotação) no plano
  - Os relacionamentos entre os termos são conhecidos e termos fortemente relacionados devem se posicionados proximamente
- A largura da nuvem tem um limite superior

# RELACIONAMENTO ENTRE TERMOS

- Contagem das ocorrências dos termos na descrição de diferentes recursos
- Relacionamento entre termos é binário e pode ser modelado como um grafo


rioting
baby
tears
tear gas
beer
gas
sports
bottle
diaper

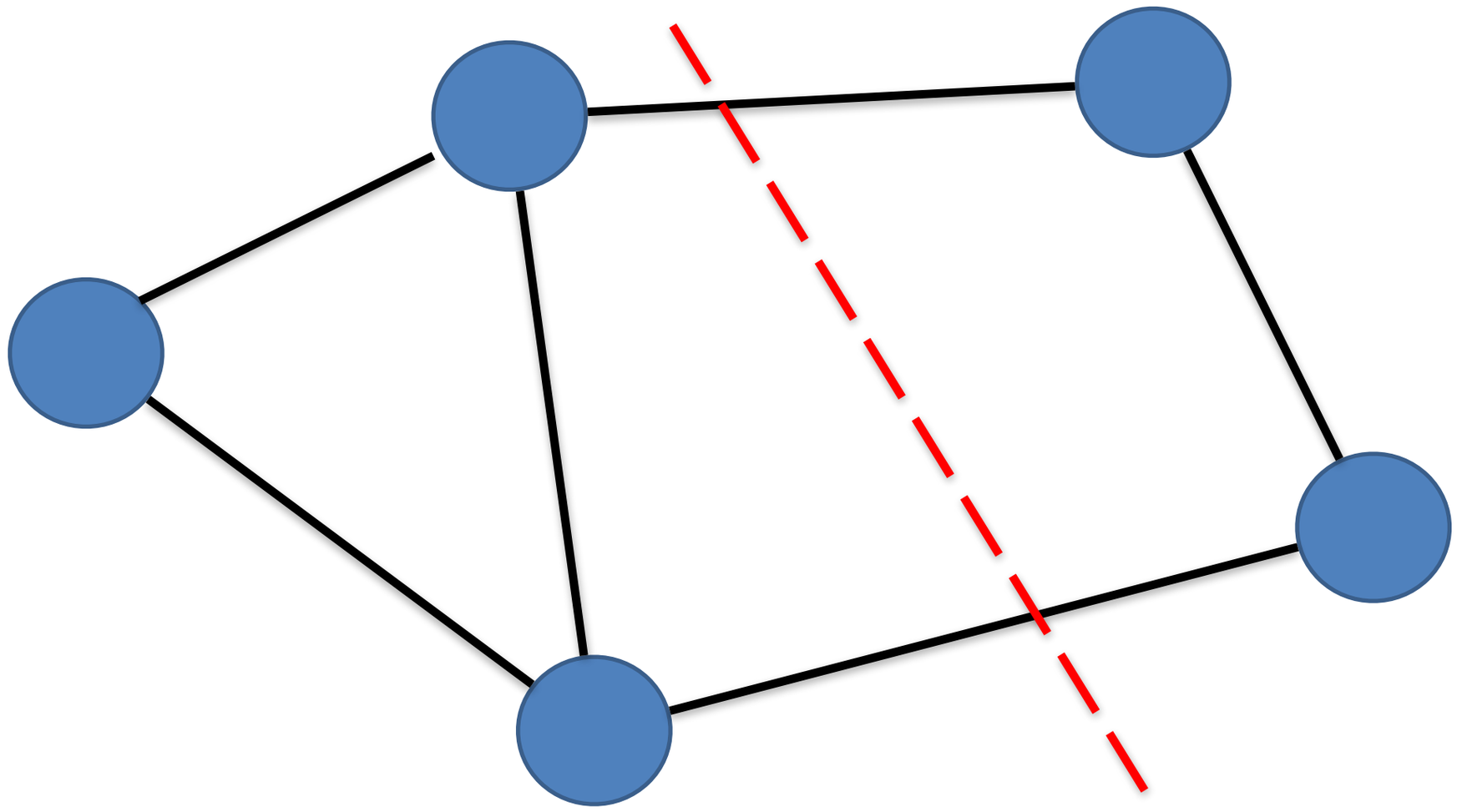
- Abordagens baseadas em força tradicionalmente usados para desenho de grafos



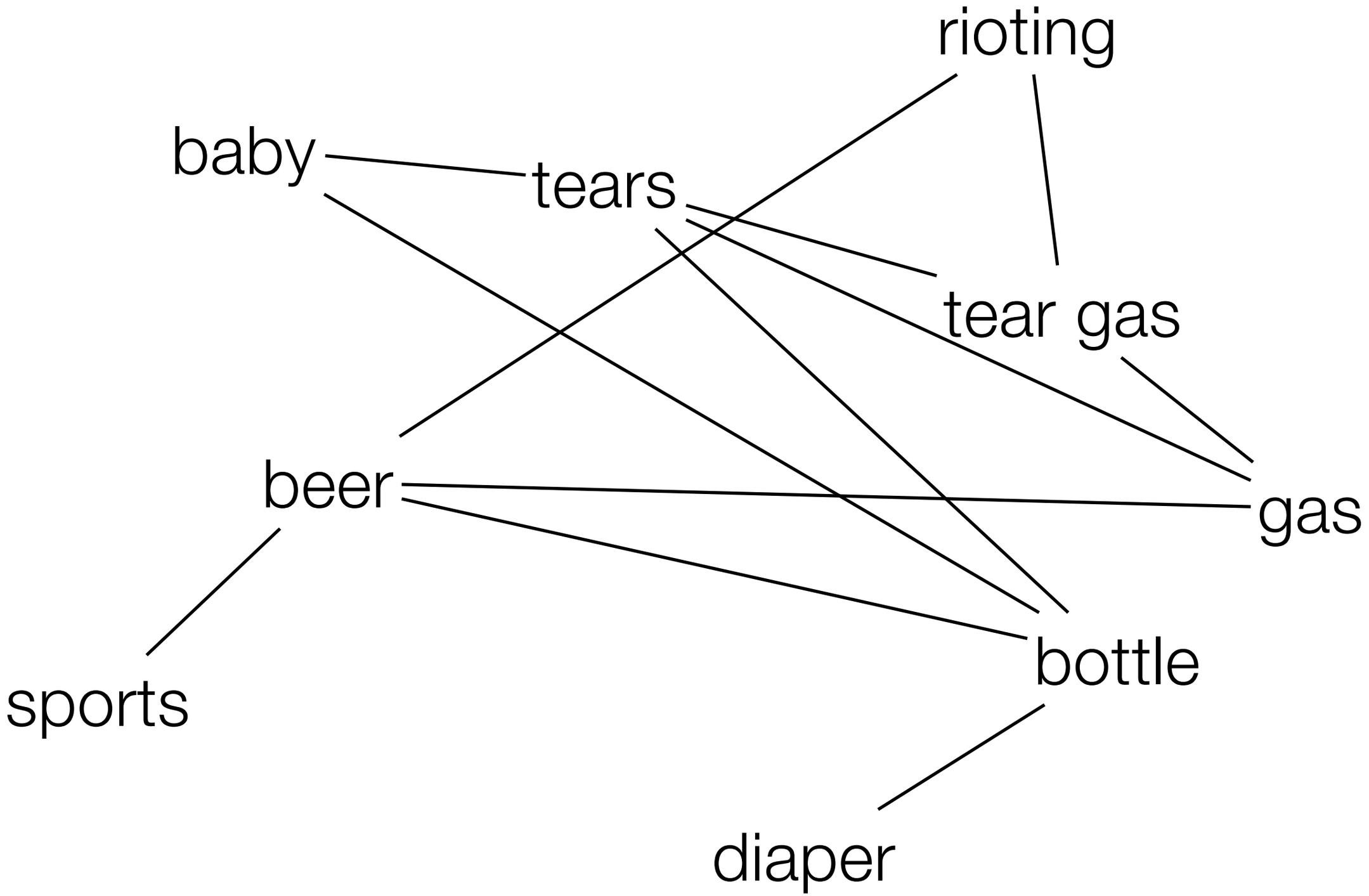

$$|\vec{F}_{Q\text{-}q}| = |\vec{F}_{q\text{-}Q}| = k\,\frac{|q \times Q|}{r^2}$$

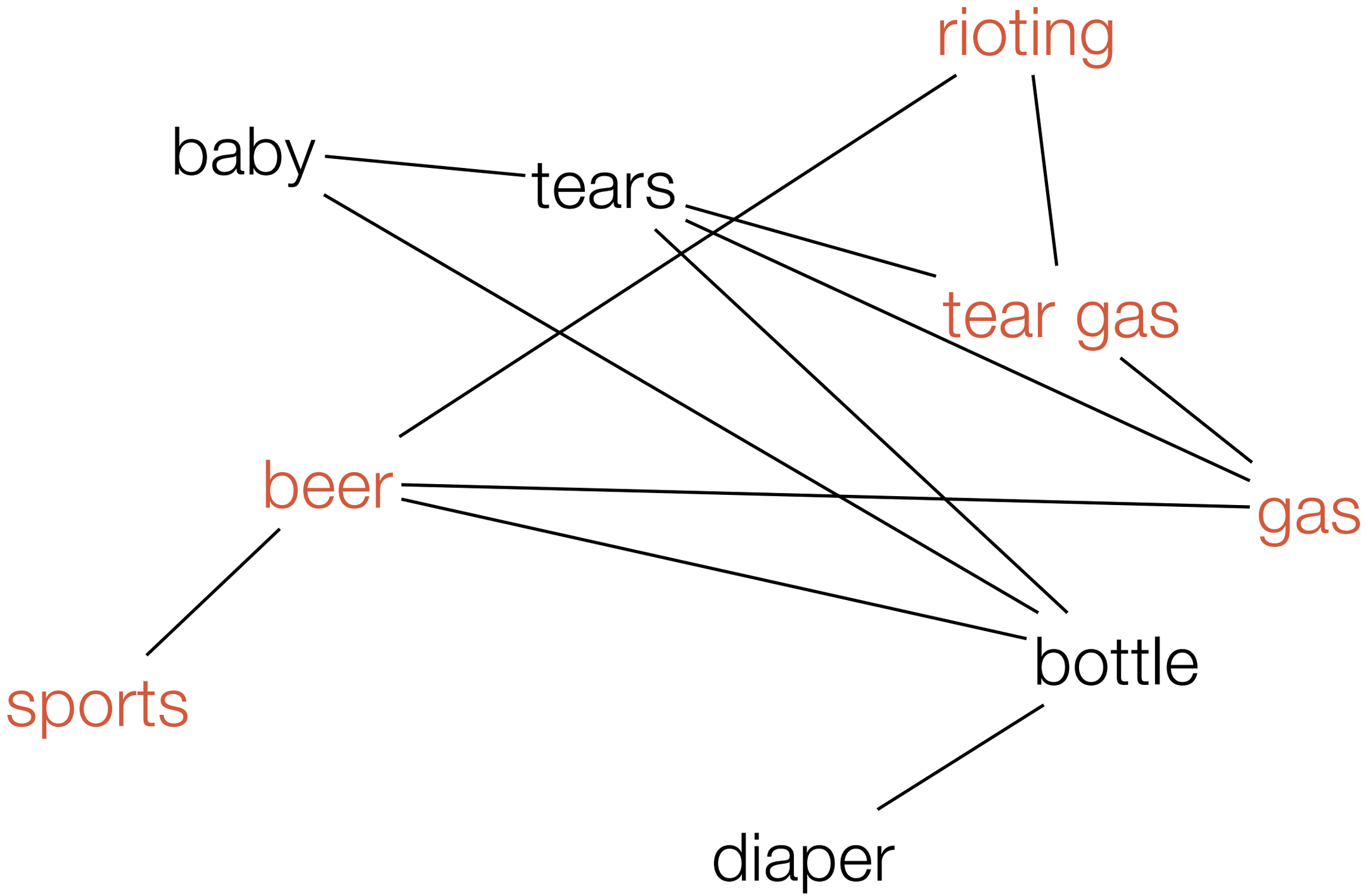
- Quando a qualidade é mais importante que o tempo de execução meta-heurísticas como *simulated annealing* podem ser usadas

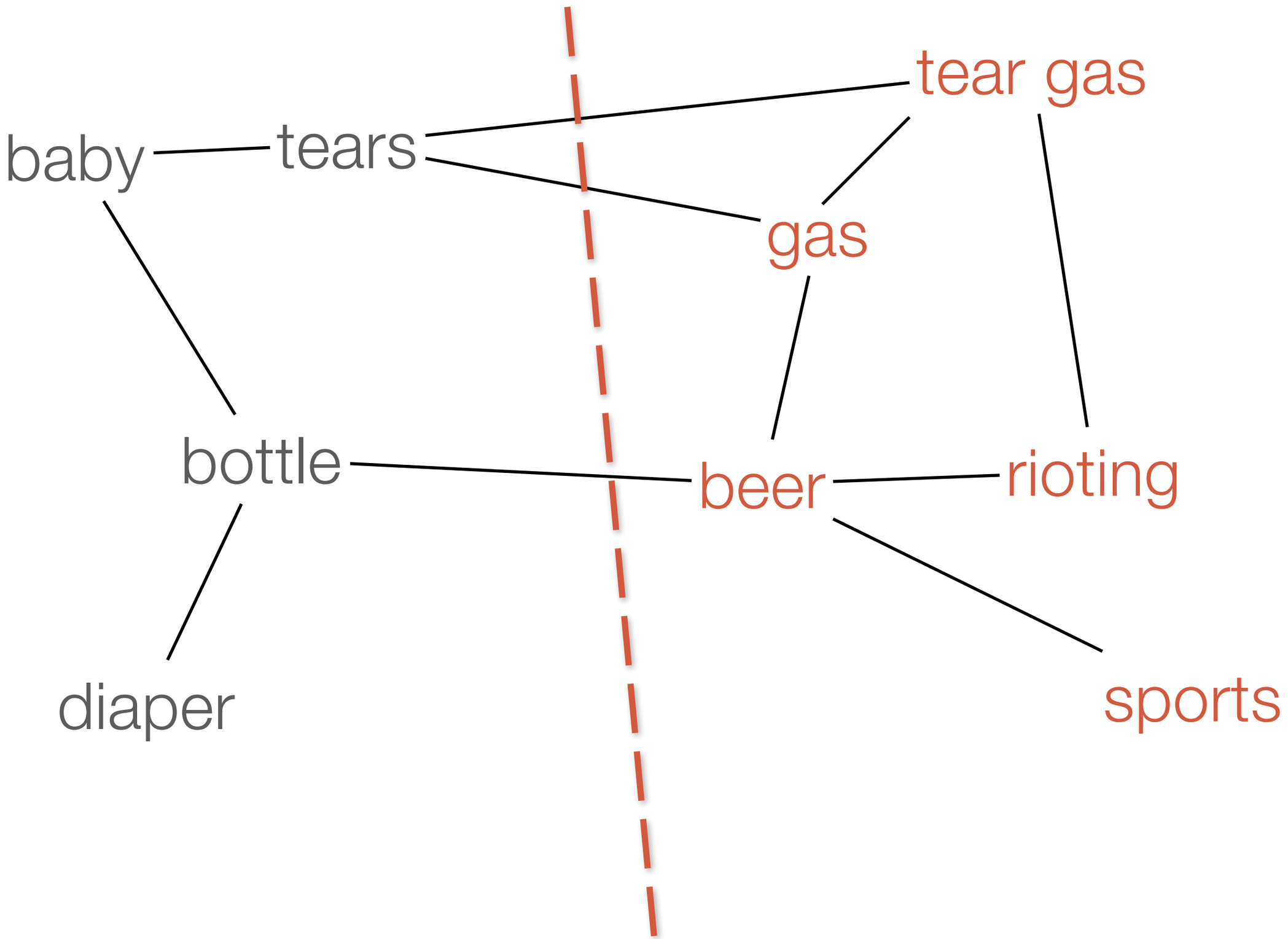
- Algoritmos baseados em estratégias de projeto físico de circuitos (EDA): *placement* e *floorplanning*
- Considerando a necessidade de eficiência, utiliza-se abordagem baseada no min-cut

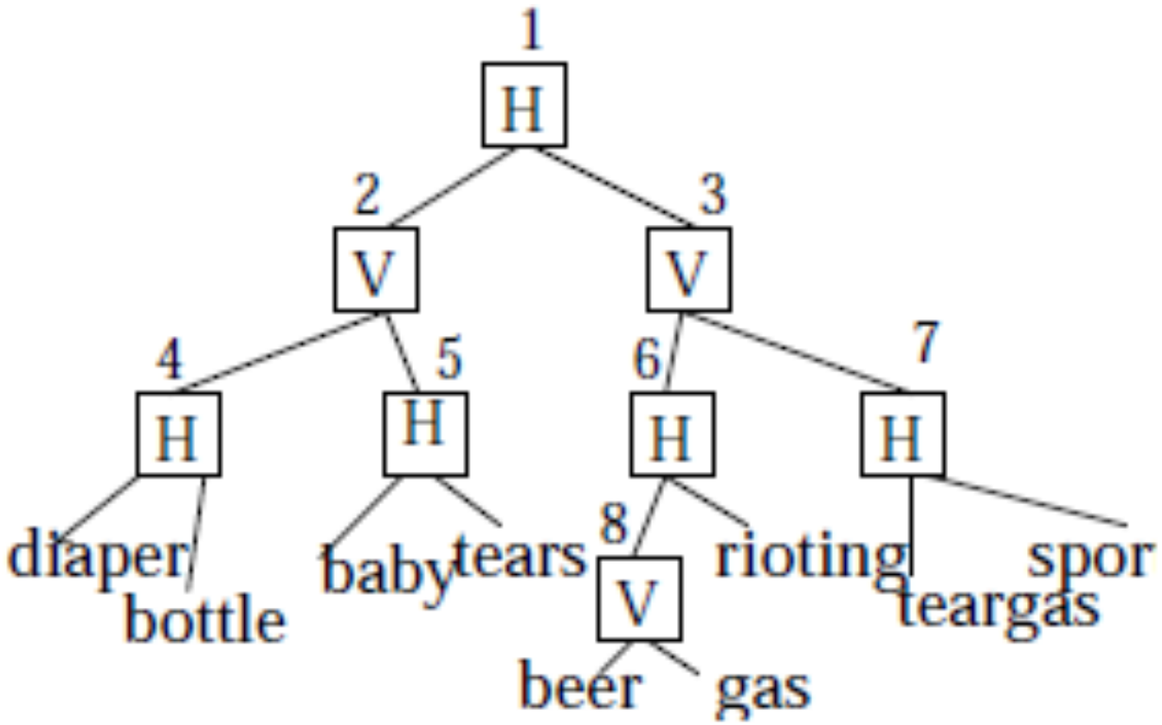


- Decompor uma coleção de termos pelo biparticionamento
- Tentar manter o balanceamento entre o número de termos nos dois lados

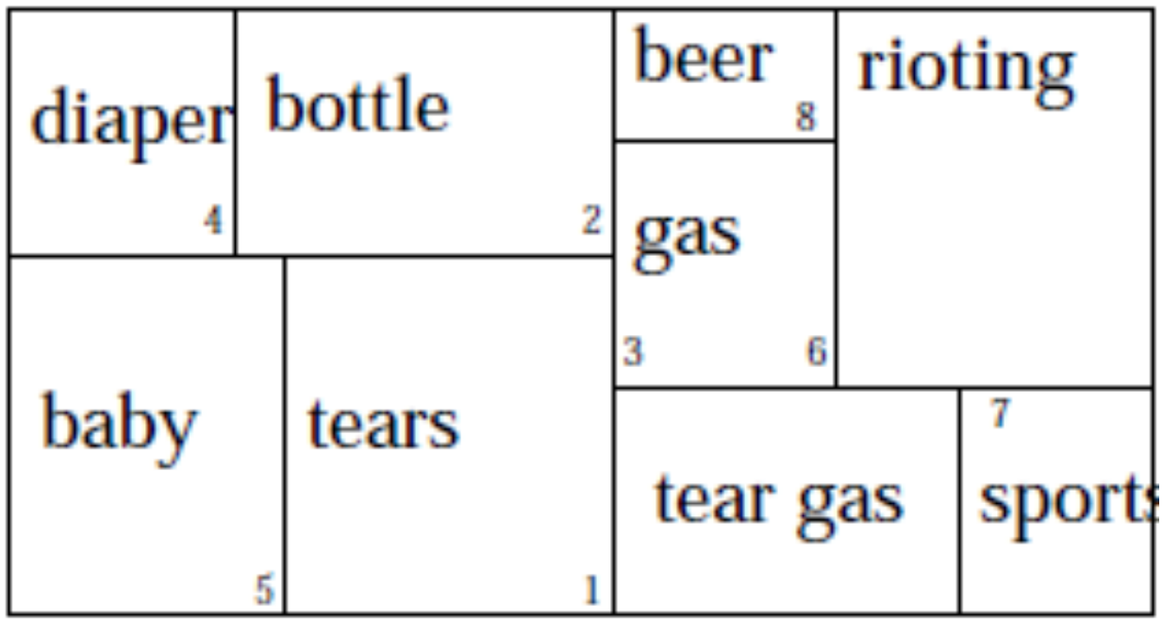

rioting
baby
tears
tear gas
beer
gas
sports
bottle
diaper


rioting
baby
tears
tear gas
beer
gas
sports
bottle
diaper


tear gas
baby
tears
gas
bottle
beer
rioting
diaper
sports




(a) Slicing tree. Numbers next to nodes relate to areas in the slicing floorplan.



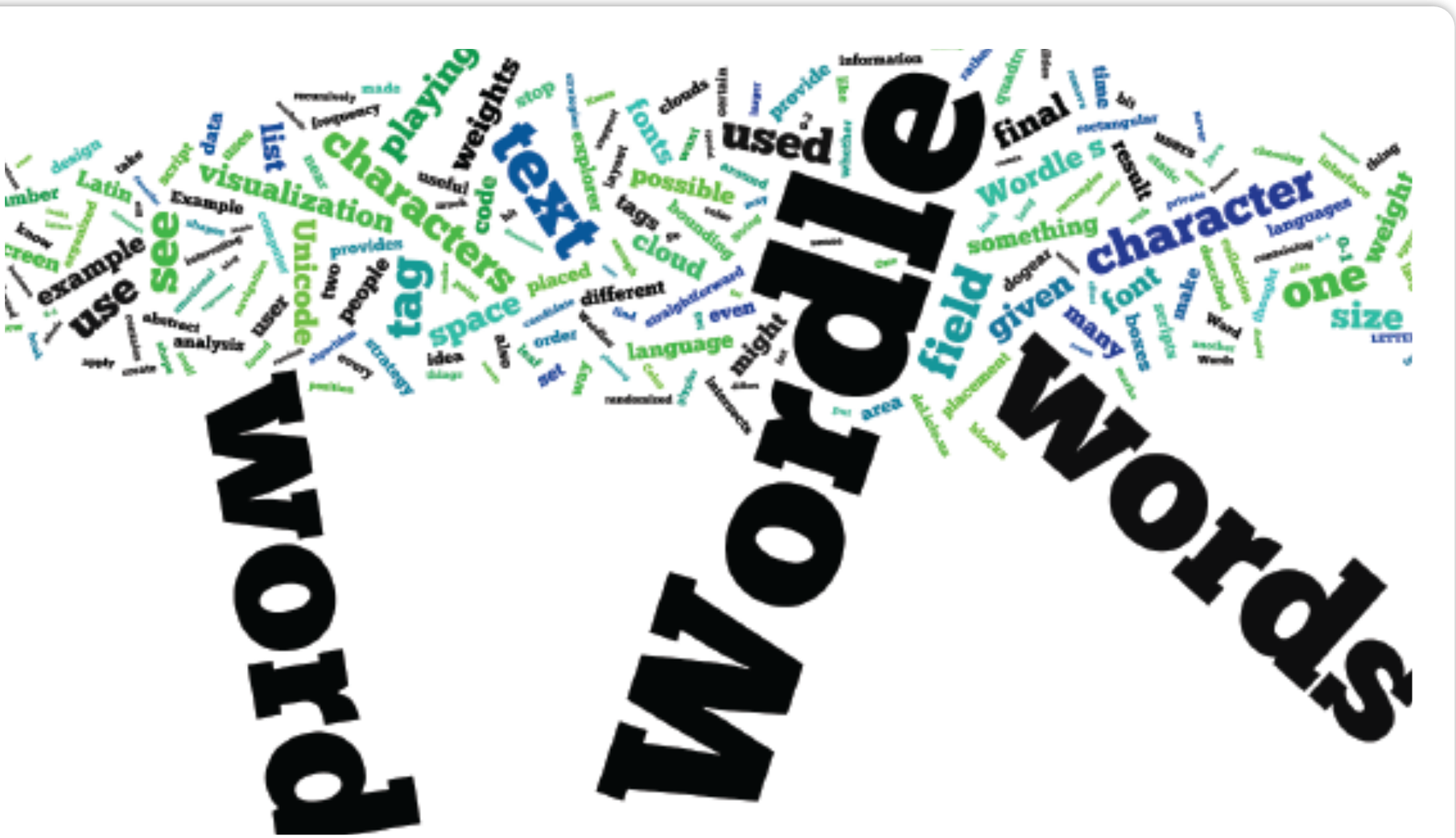

(b) Slicing floorplan


Figure 8: Large tag cloud generated from a Project Gutenberg e-text.

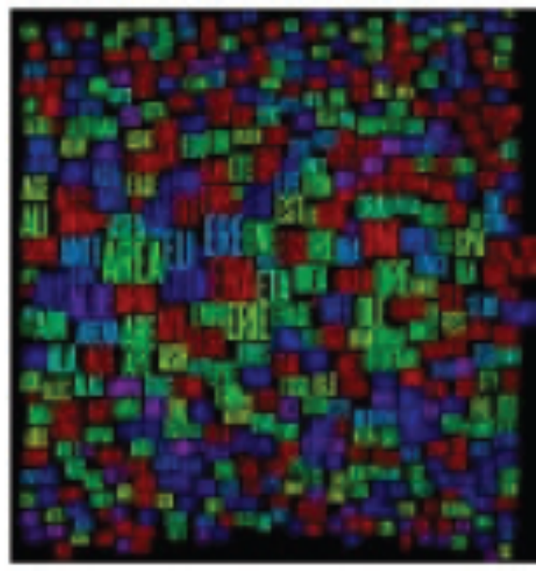
# BEAUTIFUL VISUALIZATION, CAPÍTULO 3

*J. Feinberg*

*2010*





By now, even people who have never heard of “information visualization” are familiar with colorful word collage known as wordle, “the gateway drug to textual analysis”.

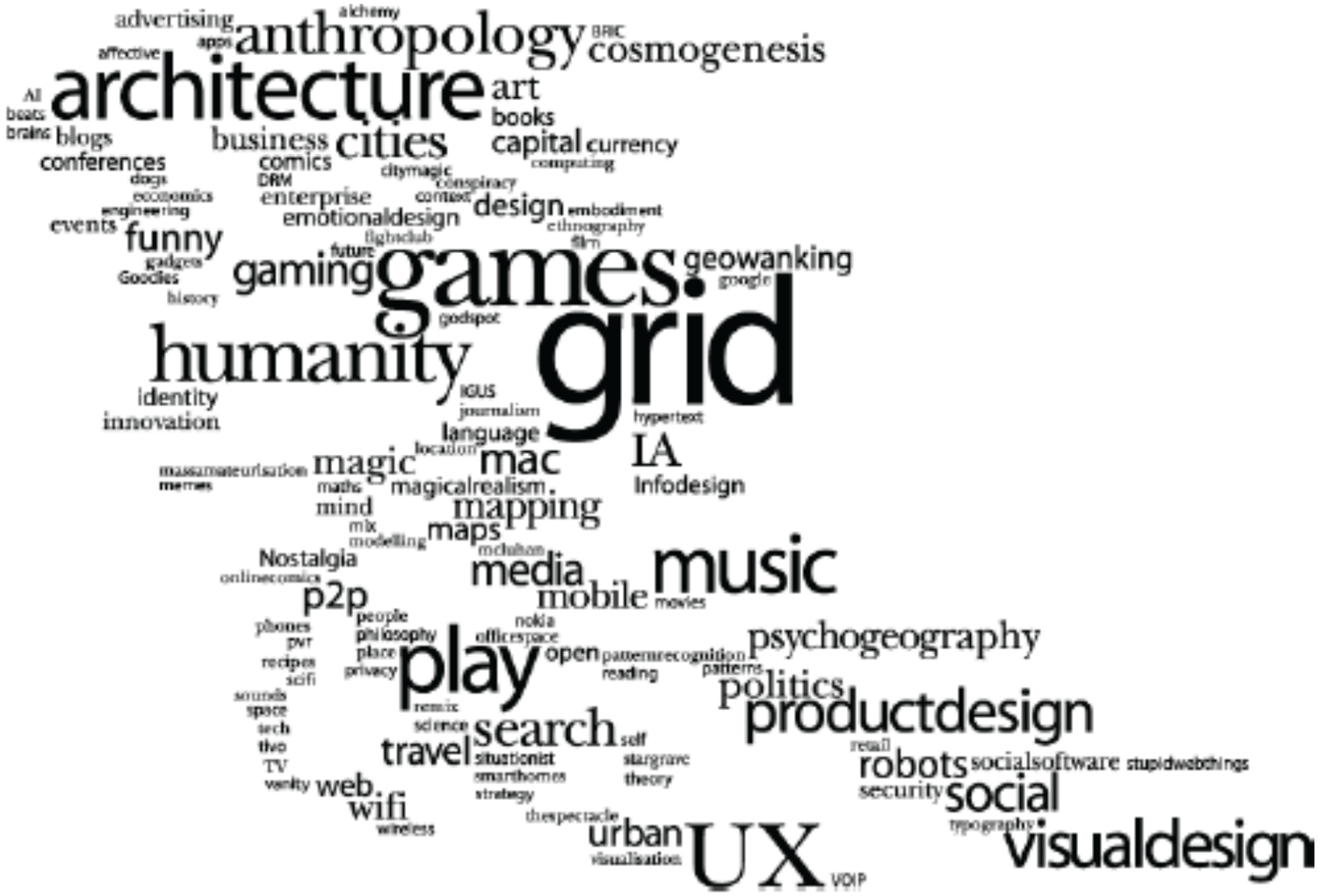



Figure 3-3. *Matt Jones's typographically aware tag cloud*


Figure 3-4. *The dogear tag explorer*[*]


Jonathan Feinberg
Senior Software Engineer
Collaborative User Experience
IBM Research
Cambridge, MA

Figure 3-5. *The author's 2006 work email signature*


Figure 3-13. *Word treemap of an Obama speech*

Estratégia **gulosa aleatória** para preencher uma região (não necessariamente retangular)

Conjunto de palavras de tamanhos variados podendo ser usadas um número variado de vezes

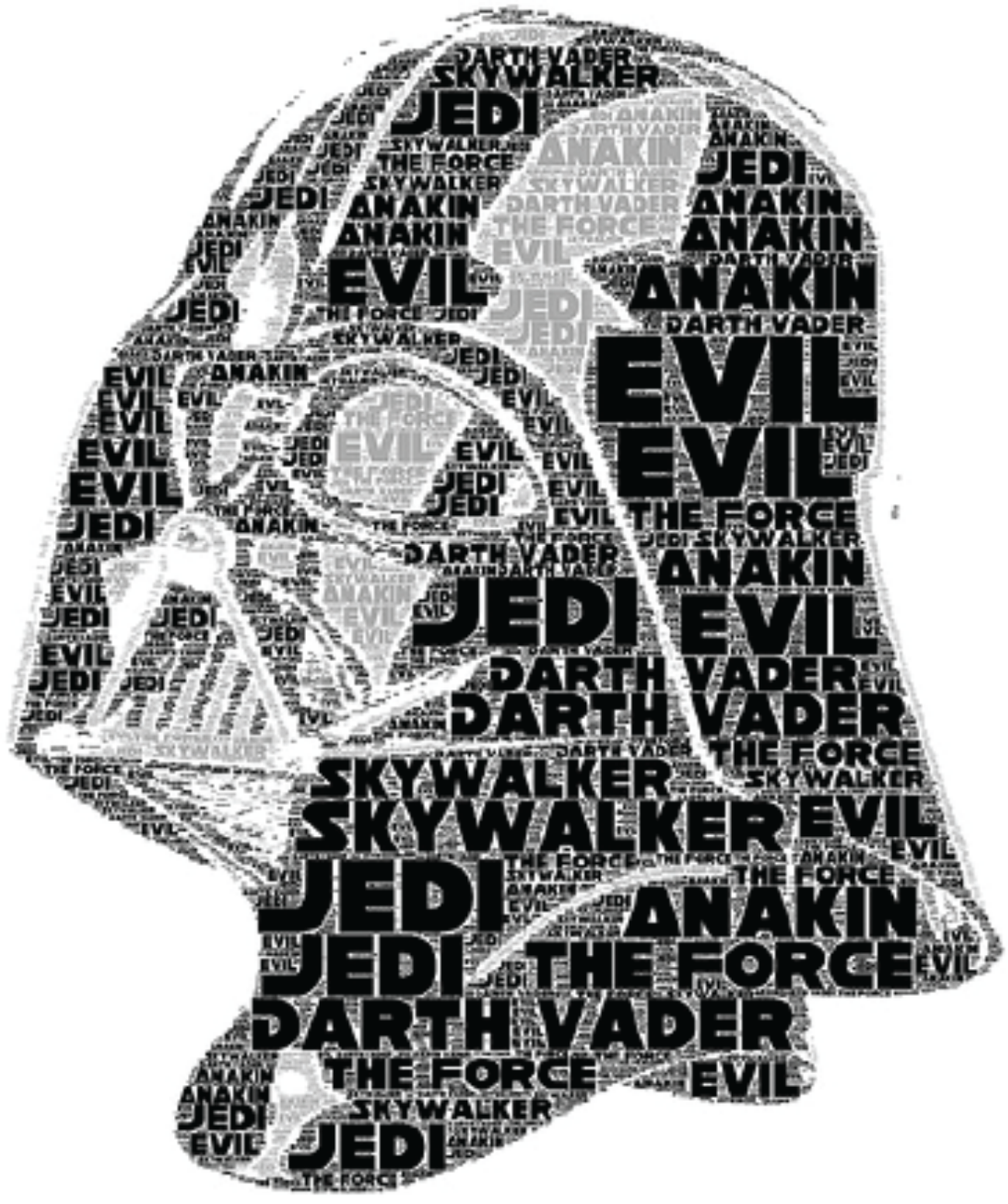



Figure 3-11. *Do not underestimate the power of the randomized greedy algorithm*

# WORDLE

- Implementado como uma applet Java
- Análise simples do texto
  - Determinação do *script*: alfabeto ou o conjunto de símbolos que podem ser usados em linguagens
    - O Wordle suporta Latin, Hebreu, Arabe e grego entre outros. Não suporta o CJKV
  - Remoção de *stop words*
- Cálculo do peso das palavras
- Cálculo do leiaute

# STOP WORDS E IDIOMA DO TEXTO

- Coleções de textos em vários idiomas foram processados buscando as palavras mais frequentes: *stop words*
- Muitas *stop words* foram acrescentadas por usuários
- Descobre as 50 palavras mais frequentes no texto e verifica sua ocorrência em listas de *stop words* cadastradas
- O idioma que tiver mais *hits* é o idioma hipotético do texto

# PESOS DAS PALAVRAS

$$\text{peso} = \text{frequência das palavras}$$

- Segundo Feinberg, o uso de logaritmo torna os wordles desinteressantes

# LEIAUTE

- Uso da função Shape do Java 2D
- Estimativa da **caixa delimitadora** para cada palavra e, através da soma de todas as caixas, estimar a área total do wordle
- Posicionamento
  - Estratégia gulosa aleatória
  - As palavras são posicionadas uma a uma
  - Uma palavra não muda de lugar após posicionada

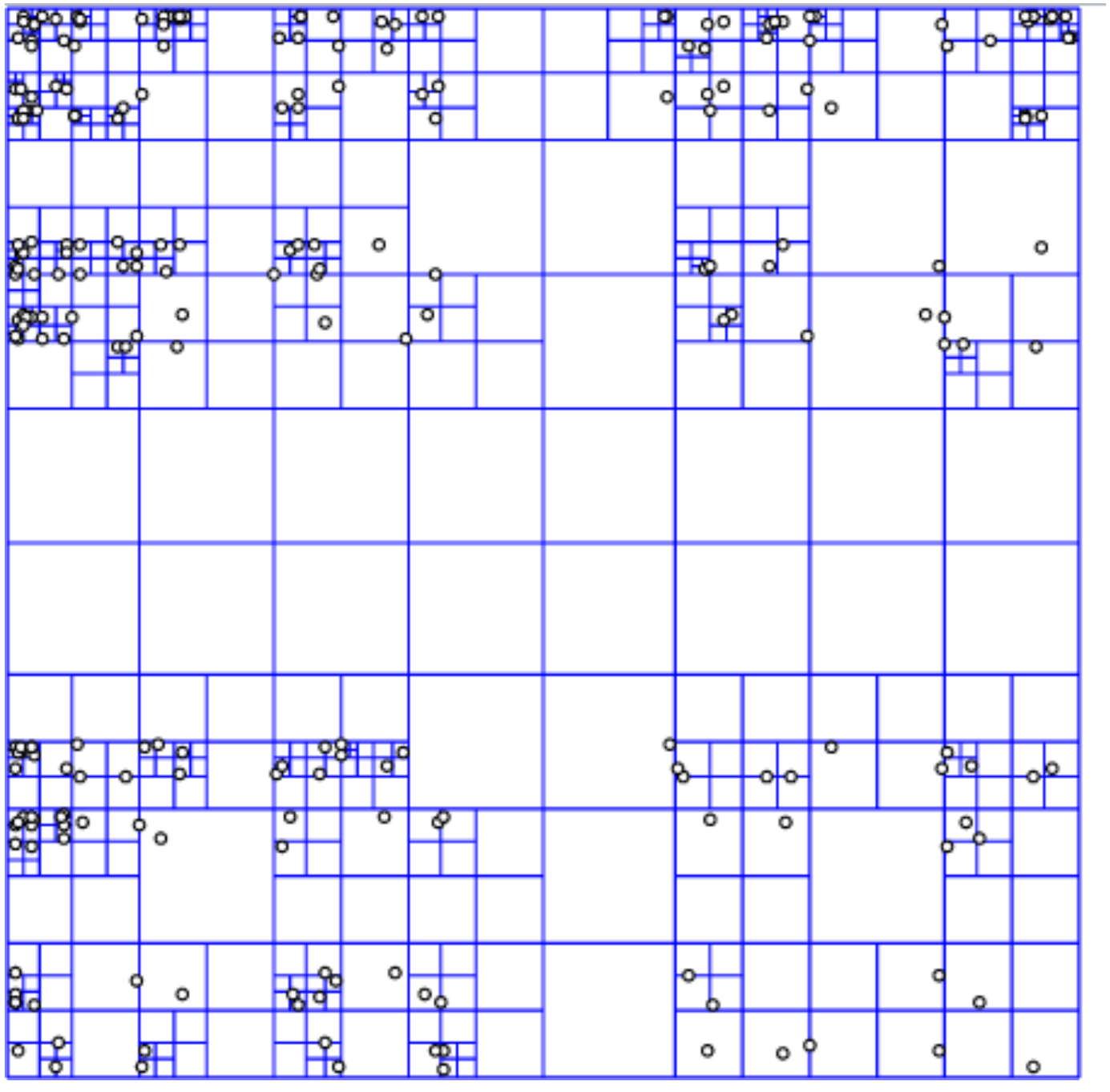
Example 3-3. *The secret Wordle algorithm revealed at last!*

```
For each word w in sorted words:
    placementStrategy.place(w)
    while w intersects any previously placed words:
        move w a little bit along a spiral path
```


Figure 3-17. *The path taken by the word "Denmark"*

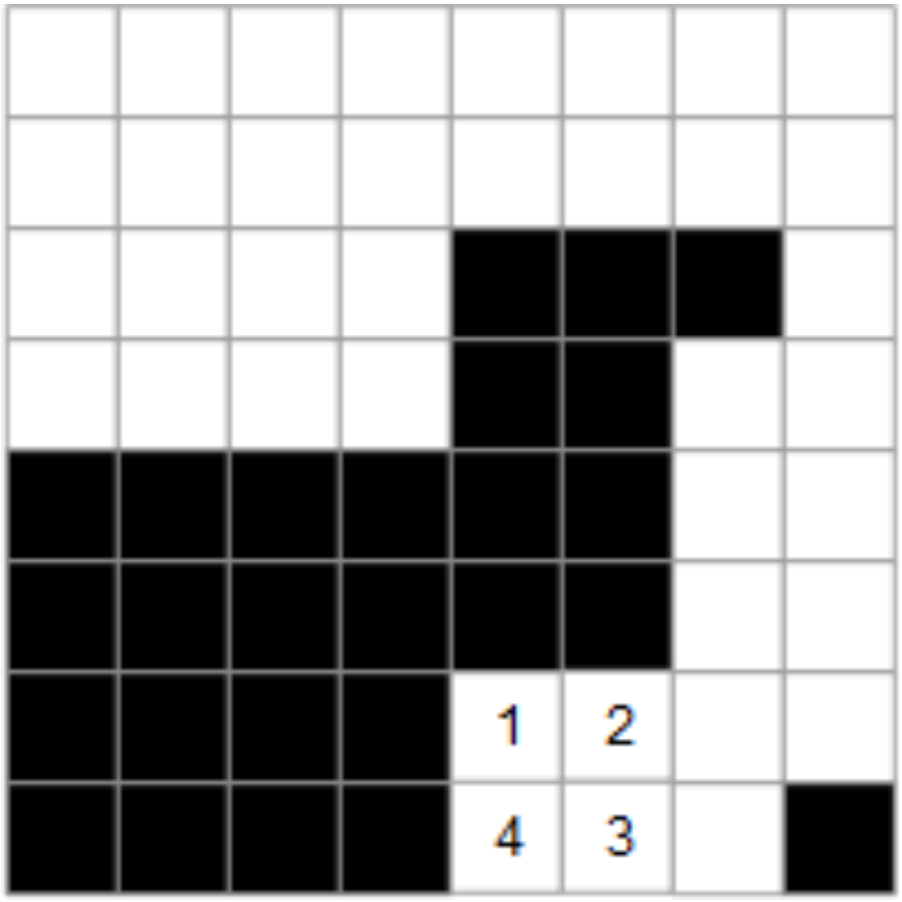
# TESTE DE INTERSEÇÃO

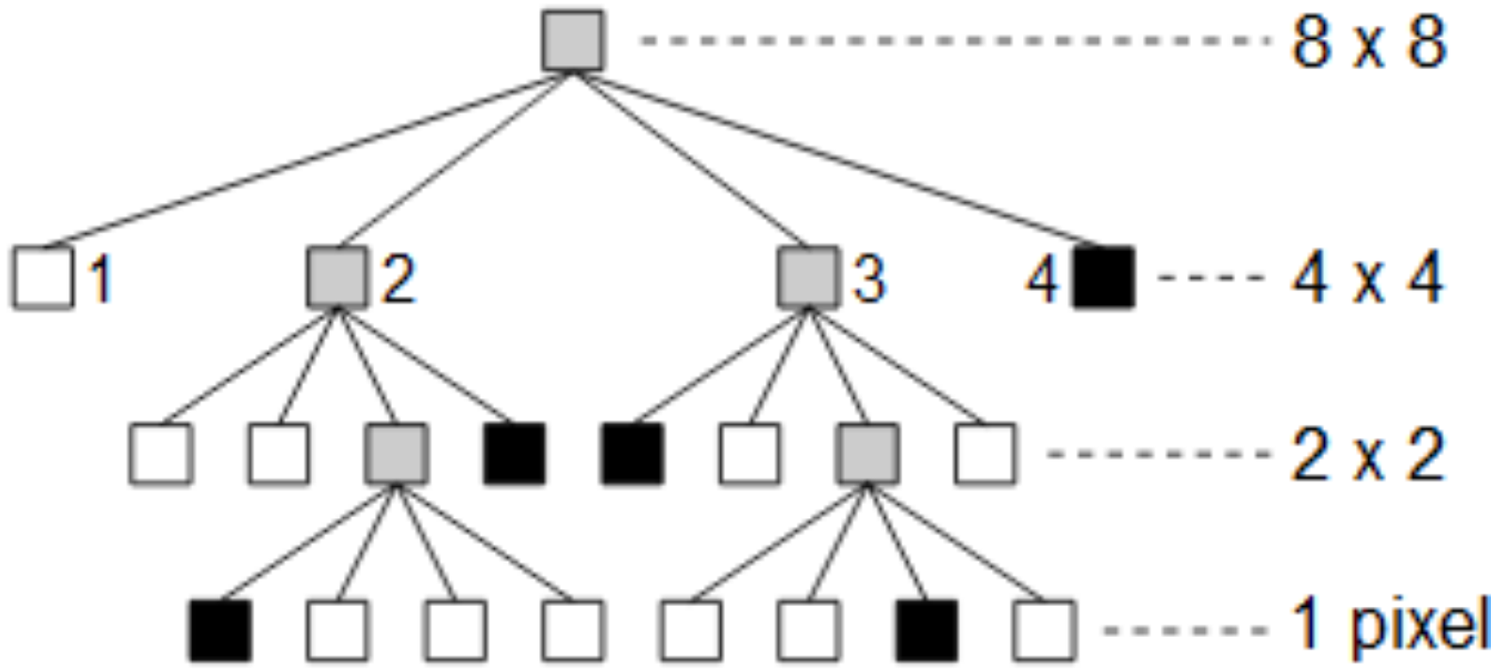
- Indexação espacial usando quad-trees



# TESTE DE INTERSEÇÃO

- Indexação espacial usando quad-trees



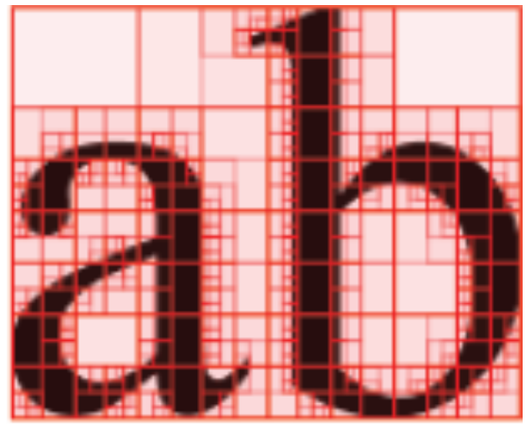



Figure 3-18. *Hierarchical bounding boxes*

# TESTE DE INTERSEÇÃO

- *Caching*
  - Pela exploração em espiral, há alta probabilidade de uma palavra A que colidiu com B colidir novamente nas próximas tentativas de posicionamento de A
  - Manter o *last-hit* em *cache* para testar por colisão primeiro

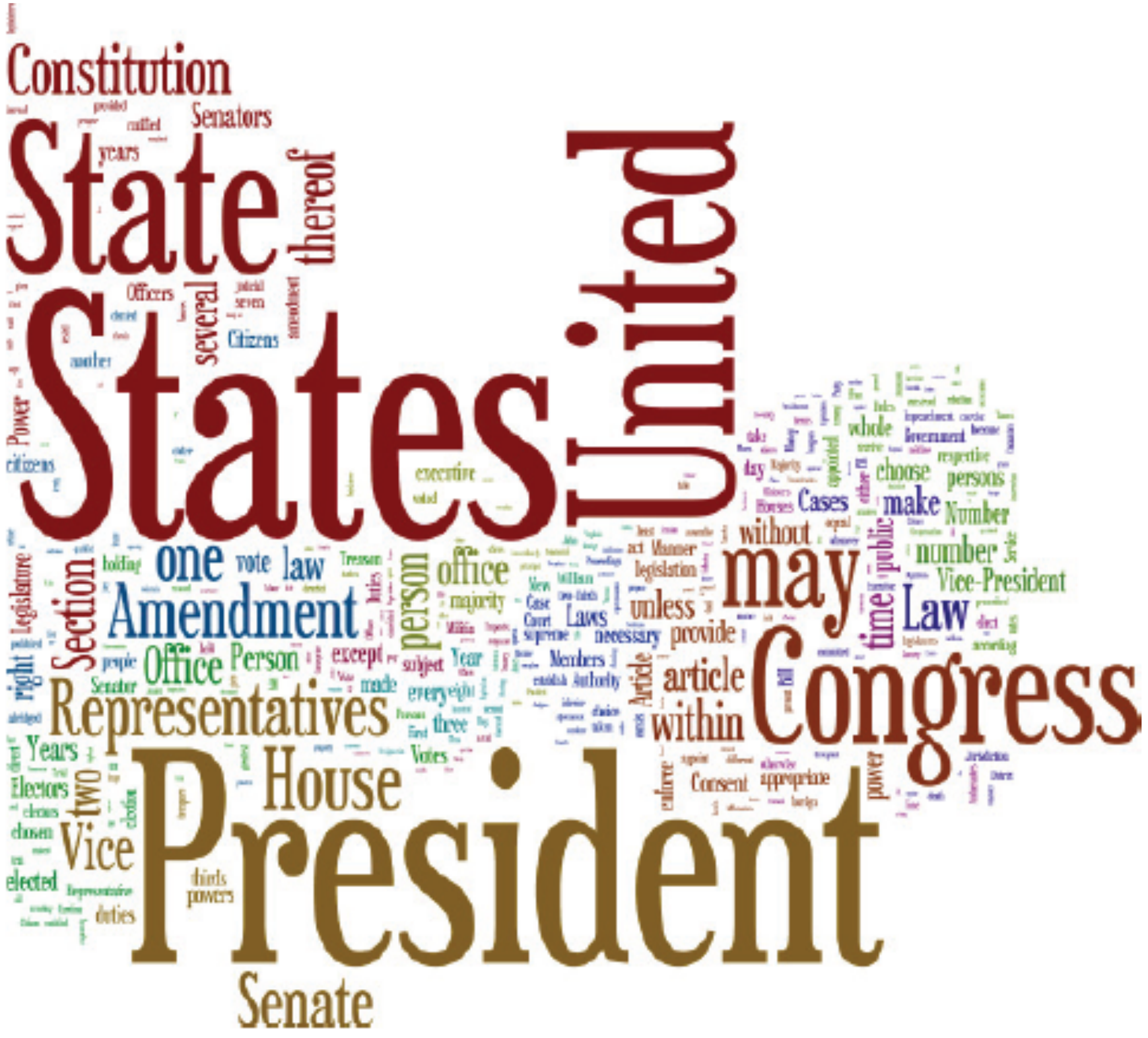
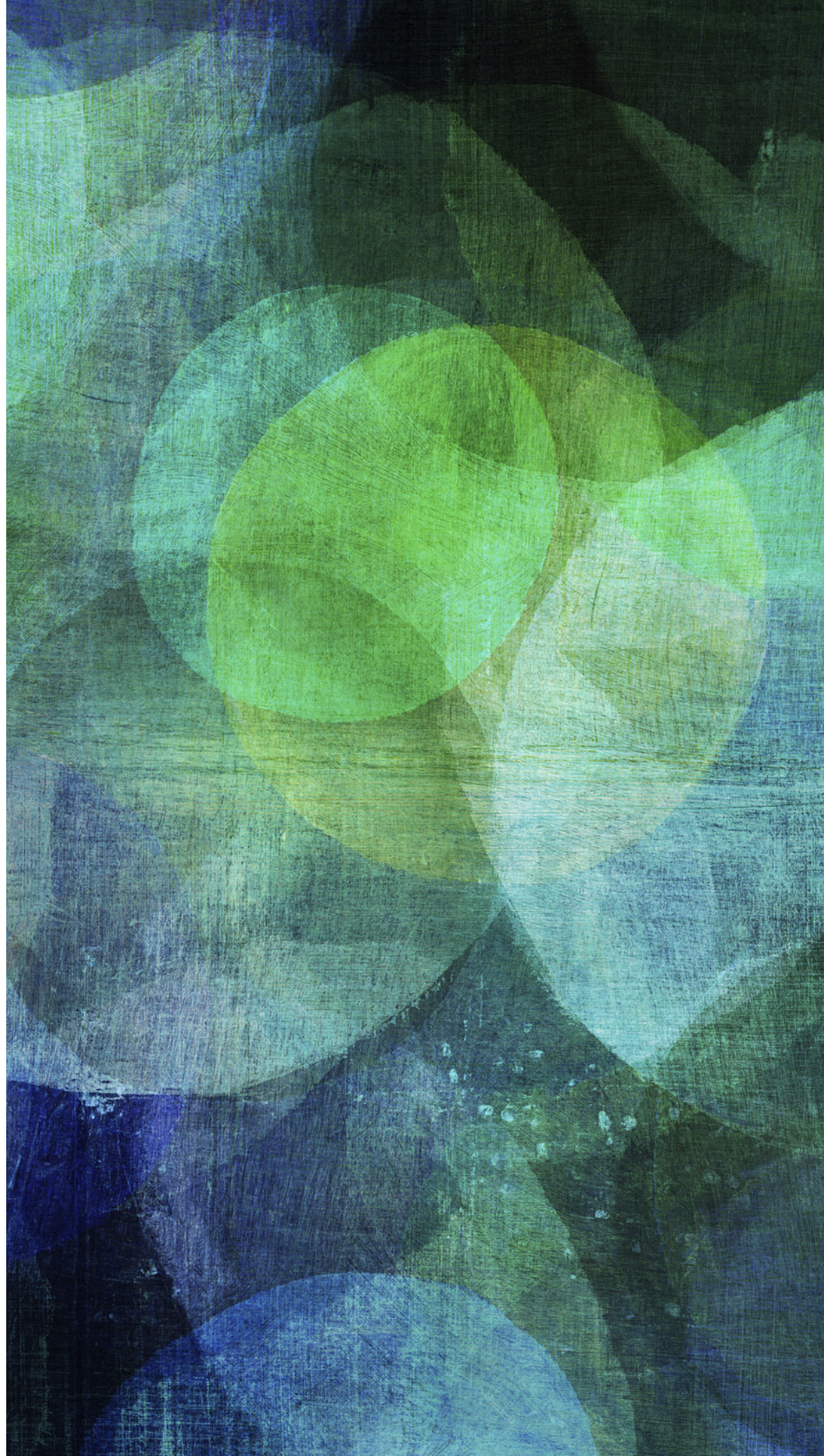



Figure 3-16. *The result of a clustering placement strategy*

# AVALIAÇÕES

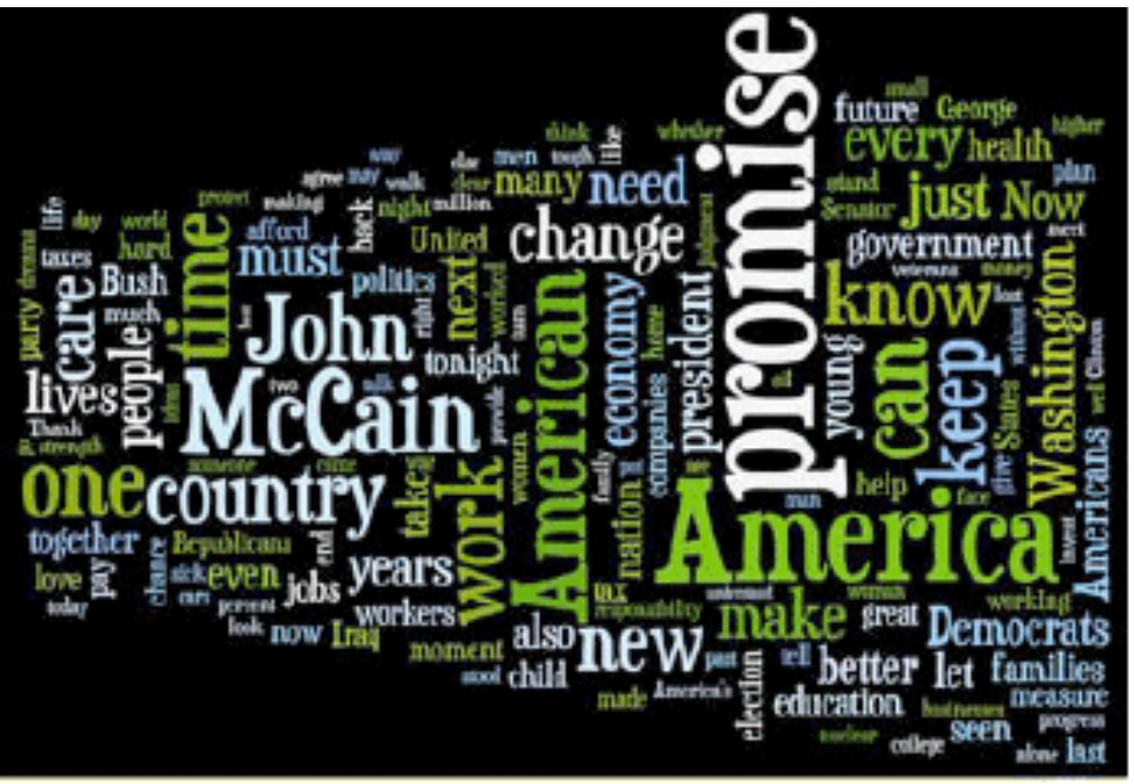


# PARTICIPATORY VISUALIZATION WITH WORDLE

*F.B. Viégas, M. Wattenberg e J. Feinberg*
*IEEE Transactions on Visualization and Computer Graphics*
*2009*

# OBJETIVO DO WORDLE

- Versão mais agradável da nuvem de termos tradicional
- Reduzir os espaços em branco não permitindo que as fontes se interceptem



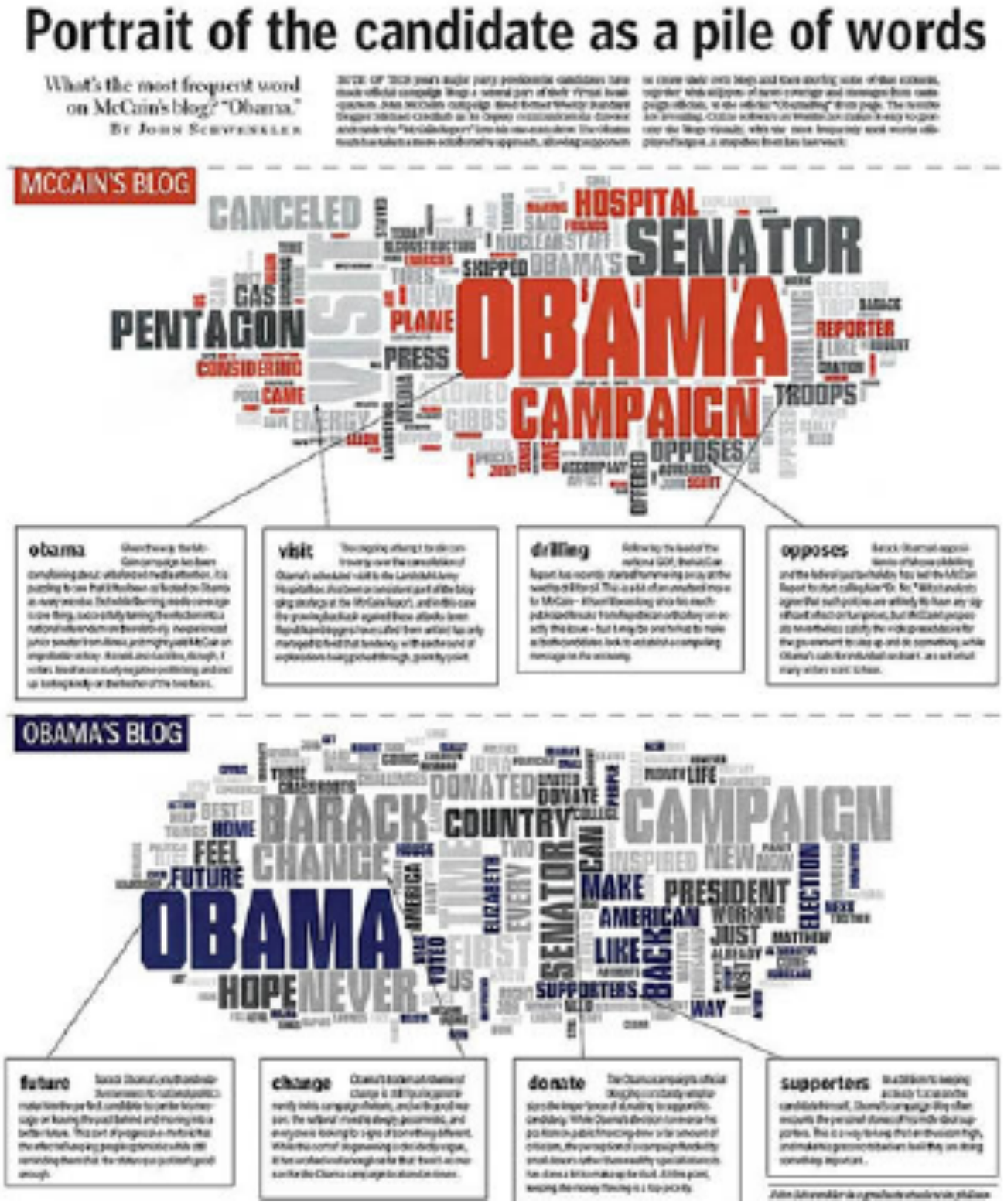


Wordle e nuvem de termos do discurso de Obama na convenção democrata de 2008

# REPRESENTAÇÃO VISUAL

- **Fonte**: o tamanho está relacionado linearmente à frequência das palavras
- **Cor**: sem significado
- **Conteúdo / linguagem**: remoção de stop words (26 linguagens)
- **Leiaute**: sem significado

- Uso pela mídia
- Uso pessoal
- Uso educacional
- Por que?




Wordle do Boston Globe sobre os blogs de campanha dos democratas e republicanos

- 720 questionários respondidos na web

  - 67% mulheres
  - 33% de homens

  - 22% abaixo dos 20 anos
  - 20% entre 21 e 30
  - 19% entre 31 e 40
  - 39% acima de 40

  - 29% estudantes
  - 17% professores
  - Nenhuma outra profissão alcançou mais de 6%

# WORLDE X NUVENS DE TERMOS TRADICIONAIS

- 70% dos participantes achou o wordle mais efetivo que as tradicionais nuvens de termos
- 11% preferem as nuvens de termos tradicionais
- 19% acham que o desempenho é igual

- Os participantes que preferem o wordle indicam três motivos:
  - Impacto emocional
  - Visual que chama atenção
  - Não linear

Table 1: Familiarity with text in Wordle

| **How familiar were you with the text before making the Wordle?** | **Percentage** |
|---|---|
| **I wrote it** | **57** |
| I read it many times | 19 |
| I read it once | 9 |
| I skimmed it | 6 |
| I had never looked at it before | 7 |

Table 2: Users' experience with Wordle

| | **Agree %** | **Neutral %** | **Disagree %** |
|---|---|---|---|
| **I felt creative** | **88** | **9** | **4** |
| I felt an emotional reaction | 66 | 22 | 12 |
| I learned something new about the text | 63 | 24 | 13 |
| It confirmed my understanding of the text | 57 | 33 | 10 |
| It jogged my memory | 50 | 35 | 15 |
| The Wordle confused me | 5 | 9 | 86 |

Table 3: How respondents used Wordles

| | Agree % | Neutral % | Disagree % |
|---|---|---|---|
| **I was just trying it out for fun** | **81** | **9** | **9** |
| To decorate websites, presentation, print, etc. | 51 | 18 | 29 |
| To illustrate a point I was making | 50 | 22 | 27 |
| I used it as a memento or souvenir | 47 | 19 | 32 |
| I used it as an analytic tool | 46 | 22 | 31 |
| I used it as a teaching tool | 37 | 23 | 39 |
| I used it as a gift | 30 | 23 | 46 |
| I didn't use it for anything | 18 | 17 | 64 |

# PROBLEMA

- Perguntas colocadas
    1. O que o tamanho de cada palavra significa?
    2. O que a direção de cada palavra significa?
    3. O que a cor de cada palavra significa?

- Respostas propostas para cada pergunta acima
    A. Número de vezes que a palavra é usada
    B. Importância emocional
    C. Significado da palavra
    D. Nada

Table 4: Percentage of respondents who did not know what font size means in a Wordle

| | Male | Female |
|---|---|---|
| **Under 20** | **35** | **49** |
| 20-30 | 12 | 18 |
| Above 30 | 19 | 31 |

# IMPROVING TAG-CLOUDS AS VISUAL INFORMATION RETRIEVAL INTERFACES

*Y. Hassan-Montero e V. Herrero-Solana*
*International Conference on Multidisciplinary Sciences and Technologies*
*2006*

ajax apple art article audio blog blogging blogs books business code comics community computer cool
css culture daily del.icio.us delicious design development diy firefox flash flickr free freeware fun
funny games geek google graphics gtd hacks hardware history howto html humor images internet
java javascript language lifehacks linux mac maps media movies mp3 music news opensource
osx photo photography photos php politics productivity programming python rails reference
research rss ruby science search security shopping social software tech technology tips tool tools
toread travel tutorial tutorials usability video web web2.0 webdesign webdev wiki windows writing xml

Figure 1: Traditional Tag-Cloud. Tags have been selected and visually weighted according to its frequency of use.

lisp perl python ruby rails

database wordpress fonts wiki gtd

books writing language math science philosophy religion history politics

media news blog blogs internet technology business web2.0 rss search google

firefox accessibility usability php xml ajax javascript html css webdesign

design web reference howto tutorial java programming development tools software opensource free

windows linux unix security networking hardware apple mac osx

game games fun funny humor art photography flash animation comics

cinema film movies movie video tv

audio music mp3 ipod radio podcast podcasting

mobile treo psp xbox fashion shopping

travel food health marketing advertising

Figure 2: Improved Tag-Cloud. Tags have been selected and visually weighted according to function 1.

# OBJETIVOS

- Avaliar a capacidade de localização de informação em nuvens de termo
- O método tradicional é baseado unicamente em frequências
  - Nem sempre os termos mais frequentes são os mais discriminantes
  - Normalmente o conjunto de termos menos frequentes é mais discriminante
- Organização alfabética não facilita a inferência de relacionamentos semânticos entre termos

# SIMILARIDADE ENTRE TERMOS

- Número de co-ocorrências, ou seja, número de vezes que dois termos são associadas ao mesmo recurso ou documento
  - Outra opção: co-ocorrência relativa ou coeficiente de Jaccard

$$RC(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

- Seja $D_i = (d_{i0},..., d_{in})$ um documento e $T_j = (t_{j0},..., t_{jm})$ um termo, então $d_{ij}$ representa a frequência de uso do termo $T_j$ para descrever o documento $D_i$

- A utilidade do termo $T_j$ será dada então por:

$$F(T_j) = \sum_{i=1}^{i=n} \left( \frac{\log(d_{ij})}{m^2} \right) \quad (2)$$

$log(d_{ij})$ atenua o efeito de termos muito utilizados em um único documento e $m^2$ diminui o efeito de termos pouco discriminantes

Table 1. Selection method comparison over 95 highest weighted tags by each function. Coverage is the number of resources that have been described by at least one of the 95 selected tags. Overlapping is the relative co-occurrence between tags.

| | Coverage | Overlapping average | Overlapping standard deviation |
|---|---|---|---|
| a) $F(T_j) = n$ | 188,761 (86.56%) | 0.0503 | 0.0414 |
| b) $F(T_j) = \sum_{i=1}^{i=n} (d_{ij})$ | 187,907 (86.17%) | 0.0399 | 0.0425 |
| c) $F(T_j) = \sum_{i=1}^{i=n} \left( \frac{\log(d_{ij})}{m} \right)$ | 191,567 (87.85%) | 0.0329 | 0.0406 |
| d) $F(T_j) = \sum_{i=1}^{i=n} \left( \frac{\log(d_{ij})}{m^2} \right)$ | 190,405 (87.32%) | 0.0242 | 0.0372 |

ajax apple art article audio blog blogging blogs books business code comics community computer cool
css culture daily del.icio.us delicious design development diy firefox flash flickr free freeware fun
funny games geek google graphics gtd hacks hardware history howto html humor images internet
java javascript language lifehacks linux mac maps media movies mp3 music news opensource
osx photo photography photos php politics productivity programming python rails reference
research rss ruby science search security shopping social software tech technology tips tool tools
toread travel tutorial tutorials usability video web web2.0 webdesign webdev wiki windows writing xml

Figure 1: Traditional Tag-Cloud. Tags have been selected and visually weighted according to its frequency of use.

lisp perl python ruby rails

database wordpress fonts wiki gtd

books writing language math science philosophy religion history politics

media news blog blogs internet technology business web2.0 rss search google

firefox accessibility usability php xml ajax javascript html css webdesign

design web reference howto tutorial java programming development tools software opensource free

windows linux unix security networking hardware apple mac osx

game games fun funny humor art photography flash animation comics

cinema film movies movie video tv

audio music mp3 ipod radio podcast podcasting

mobile treo psp xbox fashion shopping

travel food health marketing advertising

Figure 2: Improved Tag-Cloud. Tags have been selected and visually weighted according to function 1.

# SEEING THINGS IN THE CLOUDS: THE EFFECT OF VISUAL FEATURES ON TAG CLOUD SELECTIONS

*S. Bateman, C. Gitwin e M. Nacenta*
*ACM Conference on Hypertext and Hypermidia*
*2008*

# ATRIBUTOS PRÉ-ATENTIVOS

- O que são atributos pré-atentivos?
  - São atributos (forma, cor, posição, movimentação) percebidos em um momento anterior ao da atenção consciente

- Forma, cor e posição são usados em nuvens de termos de forma combinada normalmente

- Como estes atributos afetam os usuários?

1. Tamanho da fonte dos termos
2. Espessura da fonte
3. Cor
4. Intensidade da cor
5. Número de pixels
6. Comprimento do termo
7. Número de caracteres do termo
8. Área do termo
9. Posição

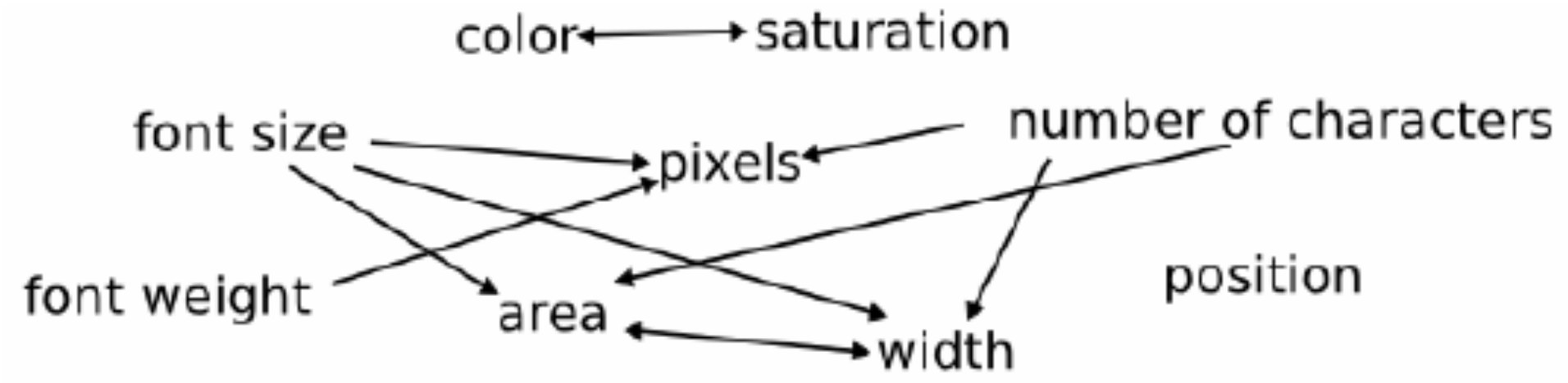



**Figure 2. The interdependencies of the selected visual features.**

| | Tag Cloud Set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Visual Property** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Font Size | | | | | | | X | X | X | X |
| Tag Area | | | | | | | X | X | X | X |
| Num. of Characters | | | X | | | | | X | X | X |
| Tag Width | X | | X | | | | X | X | X | X |
| Font Weight | | | | X | X | | | | | X |
| Colour | | | | | | X | | | | X |
| Intensity | | | | | X | X | | | X | X |
| Num. of Pixels | | X | | X | | | X | X | X | |

**Table 1. Visual properties manipulated for each cloud set.**

| Visual Property | Value Range |
|---|---|
| Font Size | 26-36 pt.; *31 pt.* |
| Tag Area | Dependent on font size and number of characters, and the use of variable or fixed fonts.; *2625 px.* |
| Num. of Characters | 3-7 characters; *5 characters* |
| Tag Width | Dependent on font size, spacing, # of characters, and the use of variable or fixed fonts.; *75px.* |
| Font Weight | bold or normal; *normal* |
| Colour | blue or red; *blue* |
| Intensity | 100%, 87.5%, 75%,62.5%,50%; *100%* |
| Num. of Pixels | tags varied over a 300 pixel range; *varied over 10 pixel range* |

**Table 2. Value ranges for independent variables. Italicized values show the variable when not varied.**

# AVALIAÇÃO COM USUÁRIOS

- Nuvens de termos foram apresentadas aos usuários com algumas perguntas para que eles selecionassem os termos correspondentes

- Após as seleções, um *pop-up* perguntou qual atributo foi usado para a seleção

- O valor esperado foi calculado com base no valor de seleções ao acaso, por exemplo, havendo 150 termos, sendo 30 com uma certa propriedade visual, o valor esperado ao acaso seria 0,2

| | Strength of Effect | | | Reliability (Wilcoxon) | |
|---|---|---|---|---|---|
| **Tag Visual Properties** | **Mean Selected** | **Mean Expected** | **Difference** | **Z** | **p** |
| Font Size | 8.6 | 2.0 | 6.6 | -5.88 | < .001 |
| Tag Area | 2.0 | 0.8 | 1.2 | -5.11 | <.001 |
| Num. of Characters | 2.7 | 3.6 | - 0.9 | -5.12 | <.001 |
| Tag Width | 2.5 | 2.4 | 0.1 | -1.97 | <.05 |
| Font Weight | 8.2 | 3.3 | 4.9 | -5.51 | <.001 |
| Colour (blue before red) | 5.2 | 4.9 | 0.3 | -0.37 | >.05 |
| Colour (red before blue) | 4.8 | 5.1 | - 0.3 | -0.37 | >.05 |
| Intensity | 6.0 | 4.0 | 2.0 | -6.07 | <.001 |
| Num. of Pixels | 1.3 | 0.7 | 0.6 | -6.29 | <.001 |

**Table 3. Effects of visual properties on selection rates. (All relevant tag cloud sets are grouped for each visual property)**

| | Tag Cloud Sets | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Tag Visual Properties** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Font Size | | | | | | | 25 | 24 | 24 | 10 |
| Tag Area | | | | | | | 0 | 0 | 0 | 0 |
| Num. of Characters | | | 3 | | | | | 0 | 0 | 0 |
| Tag Width | 1 | | 3 | | | | 0 | 0 | 0 | 1 |
| Font Weight | | | | 27 | 24 | | | | | 15 |
| Colour | | | | | | 22 | | | | 16 |
| Intensity | | | | | 8 | 14 | | | 10 | 6 |
| Number of Pixels | | 0 | | 0 | | | 0 | 0 | 0 | |
| Position | 7 | 1 | 5 | 4 | 2 | 2 | 2 | 6 | 5 | 2 |
| Other | 25 | 28 | 25 | 6 | 7 | 5 | 7 | 7 | 1 | 1 |

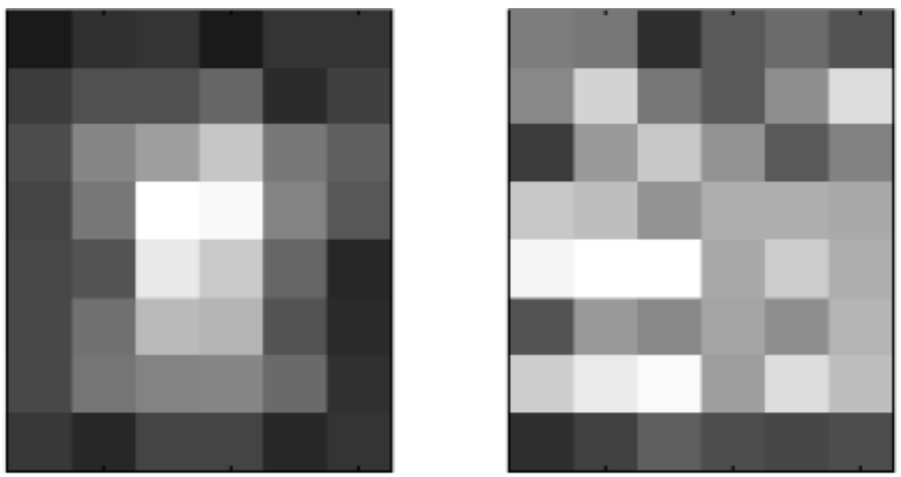
**Table 4. Visual properties and number of participants stating that they used that property in making their selection.**

| | Tag Cloud Set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Visual Property** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Font Size | | | | | | | X | X | X | X |
| Tag Area | | | | | | | X | X | X | X |
| Num. of Characters | | | X | | | | | X | X | X |
| Tag Width | X | | X | | | | X | X | X | X |
| Font Weight | | | | X | X | | | | | X |
| Colour | | | | | | X | | | | X |
| Intensity | | | | | X | X | | | X | X |
| Num. of Pixels | | X | | X | | | X | X | X | |

**Table 1. Visual properties manipulated for each cloud set.**

| | Tag Cloud Set | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Visual Property** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Font Size | | | | | | | X | X | X | X |
| Tag Area | | | | | | | X | X | X | X |
| Num. of Characters | | | X | | | | | X | X | X |
| Tag Width | X | | X | | | | X | X | X | X |
| Font Weight | | | | X | X | | | | | X |
| Colour | | | | | | X | | | | X |
| Intensity | | | | | X | X | | | X | X |
| Num. of Pixels | | X | | X | | | X | X | X | |

**Table 1. Visual properties manipulated for each cloud set.**



**Cloud Set 2** **Cloud Set 10**

**Figure 4. Clickmaps displaying the concentration of clicks for the clouds in two tag cloud sets. Lighter areas indicate more clicks than darker areas.**

Tendência de visualizar o centro da nuvem quando poucos atributos são usados

Visualização tende a se espalhar quando se usa mais atributos visuais

Tendência a usar pouco as regiões superiores e inferiores

# CONCLUSÕES

- Atributos **mais** importantes
  - **Tamanho da fonte**: influência consistente, usuários detectam até mesmo pequenas variações no tamanho
  - **Espessura da fonte**: muito influente, usar em casos binários e com cuidado pois tendem a atenuar o efeito de outros atributos
- **Intensidade da cor**: relativamente bom na captura da atenção
  - Não se sabe ao certo como são percebidas as variações mas, aparentemente, 10% de variação é perceptível

# CONCLUSÕES

- Atributos **menos** importantes
  - **Número de pixels**
  - **Comprimento** do termo
  - **Área** do termo: apesar da correlação com outros atributos visuais, foram raramente mencionados

- Os autores sugerem que esses atributos podem ser ignorados para uso em nuvens de termos

# CONCLUSÕES

- Atributos a serem usados **com precauções**
  - **Cor**: facilmente diferenciáveis, mas não ficou claro em que grau cada cor pode chamar a atenção do usuário
  - **Posição**: não ficou claro, exceto pela ênfase na visualização de termos centrais
  - O uso de variações em vários atributos favorece a visualização de todas as partes da nuvem mas outros resultados mostraram que isto também reduz a eficácia da nuvem

# COMPARISON OF TAG CLOUD LAYOUTS: TASK-RELATED PERFORMANCE AND VISUAL EXPLORATION

*S. Lohmann, J. Ziegler e L. Tetzlaff*

*Interact*
*2009*

# OBJETIVO

- Avaliar tipos de leiaute de nuvens de termos com relação ao **posicionamento dos termos**

  - Geração de nuvens de mesmo tamanho e razão de aspecto 3:2 com 100 termos cada uma
  - Fontes de 6 tamanhos (entre 30 e 15pts)
  - Mesmo número de termos em cada quadrante


**Fig. 2.** Tag cloud layouts for the corpus 'France': (a) sequential (alphabetical sorting), (b) circular (decreasing popularity), (c) clustered (thematic clusters), (d) reference (sequential, alphabetical sorting, no weighting of tags)

# AVALIAÇÃO COM USUÁRIOS

- Medida do tempo gasto para desempenhar tarefas e contagem dos votos

- Tarefas

1. Encontrar um termo **específico**
2. Encontrar o termo **mais popular**
3. Encontrar termos que pertençam a um determinado **tópico**

**Table 2.** Performance values of the tag cloud layouts for the 12 participants (N) of each task: Kruskal-Wallis (KW) mean rank, user votes, mean, median, minimum, and maximum (in sec)

| Task | Layout | N | KW Mean Rank | User Votes | Mean | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|
| 1<br>p=.131<br>3df<br>$\chi^2$=5,6 | sequential | 12 | 22.5 | 8 | 8.6 | 6.5 | 2.5 | 18.8 |
| | circular | 12 | 27.8 | 1 | 13.6 | 12.8 | 3.3 | 42.3 |
| | clustered | 12 | 30.0 | 0 | 14.3 | 12.2 | 2.8 | 33.5 |
| | reference | 12 | 17.8 | 3 | 6.9 | 5.3 | 2.3 | 18.7 |
| 2<br>p=.036<br>3df<br>$\chi^2$=8,5 | sequential | 12 | 20.8 | 1 | 2.6 | 1.6 | 1.0 | 6.6 |
| | circular | 12 | 18.7 | 8 | 2.2 | 1.8 | 0.8 | 4.7 |
| | clustered | 12 | 24.4 | 3 | 3.1 | 2.3 | 1.0 | 7.8 |
| | reference | 12 | 34.1 | 0 | 4.5 | 3.4 | 1.2 | 9.8 |
| 3<br>p=.239<br>3df<br>$\chi^2$=4,2 | sequential | 12 | 26.5 | 4 | 7.3 | 5.0 | 1.7 | 23.0 |
| | circular | 12 | 22.0 | 3 | 5.4 | 4.7 | 1.9 | 12.2 |
| | clustered | 12 | 19.4 | 5 | 4.9 | 3.7 | 1.5 | 13.2 |
| | reference | 12 | 30.2 | 0 | 6.9 | 6.5 | 3.0 | 11.5 |

Nos primeiros 6 segundos (fase de busca), toda fixação de olhar com mais de 100ms de duração foram analisados

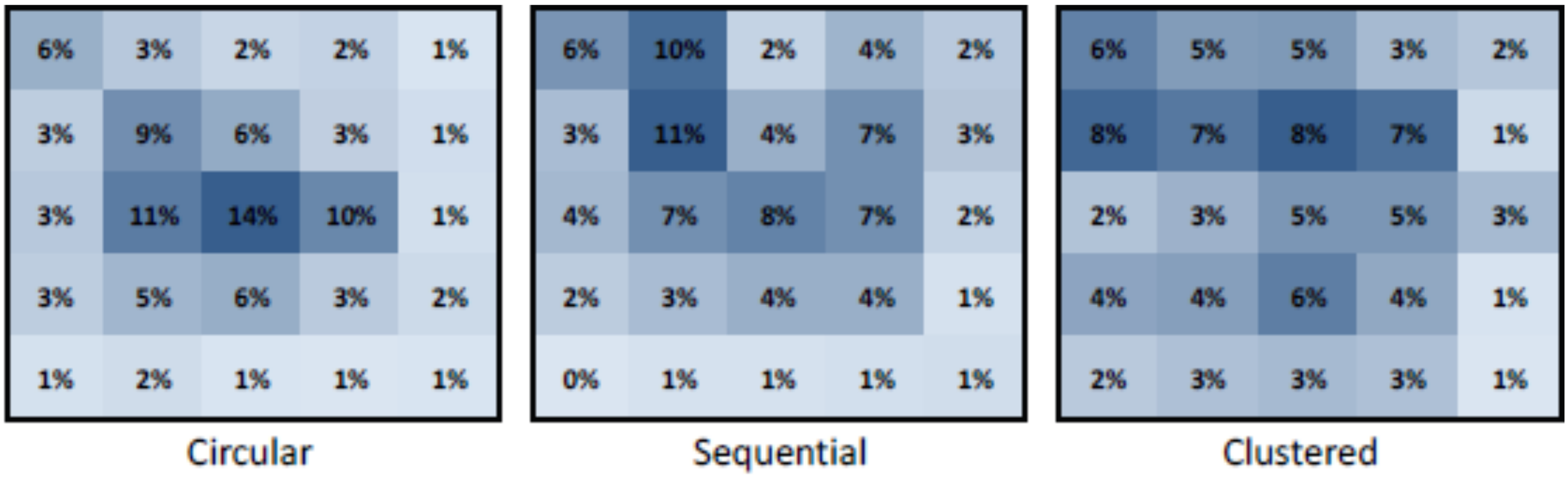



**Fig. 4.** Distribution of fixations in percent over the 5 x 5 subareas for the three tag cloud layouts. The five-level coloring illustrates the pattern of the distribution.

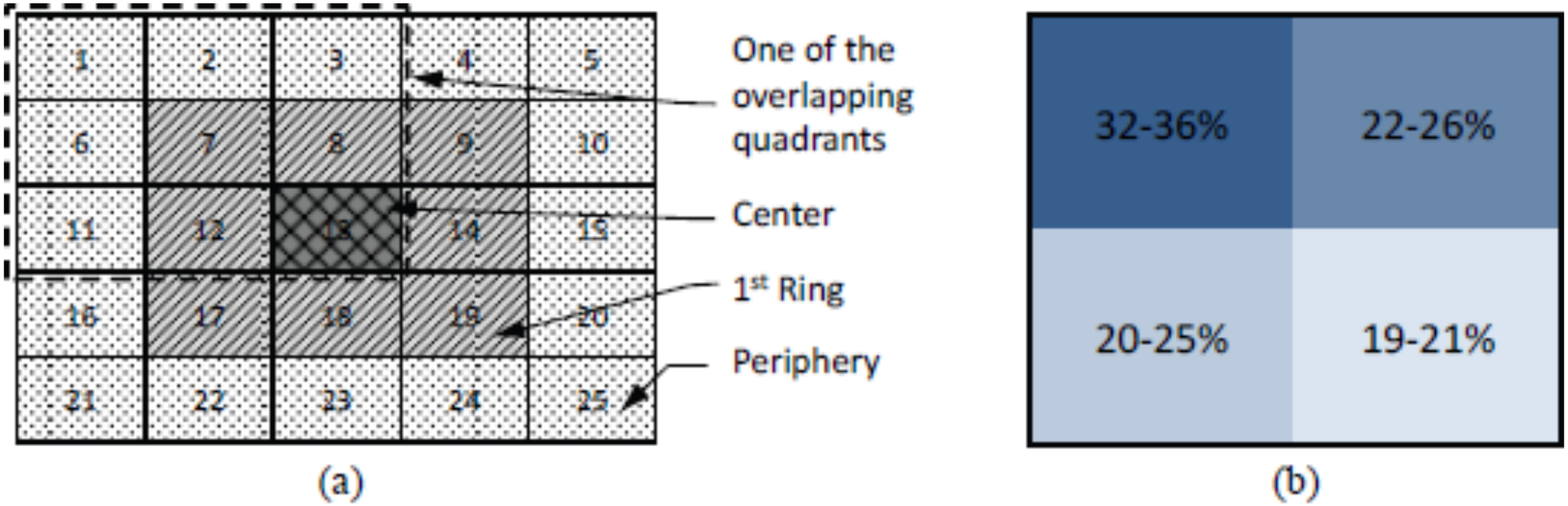



**Fig. 5.** a) Definition of areas of interest, b) Distribution of fixations over the quadrants

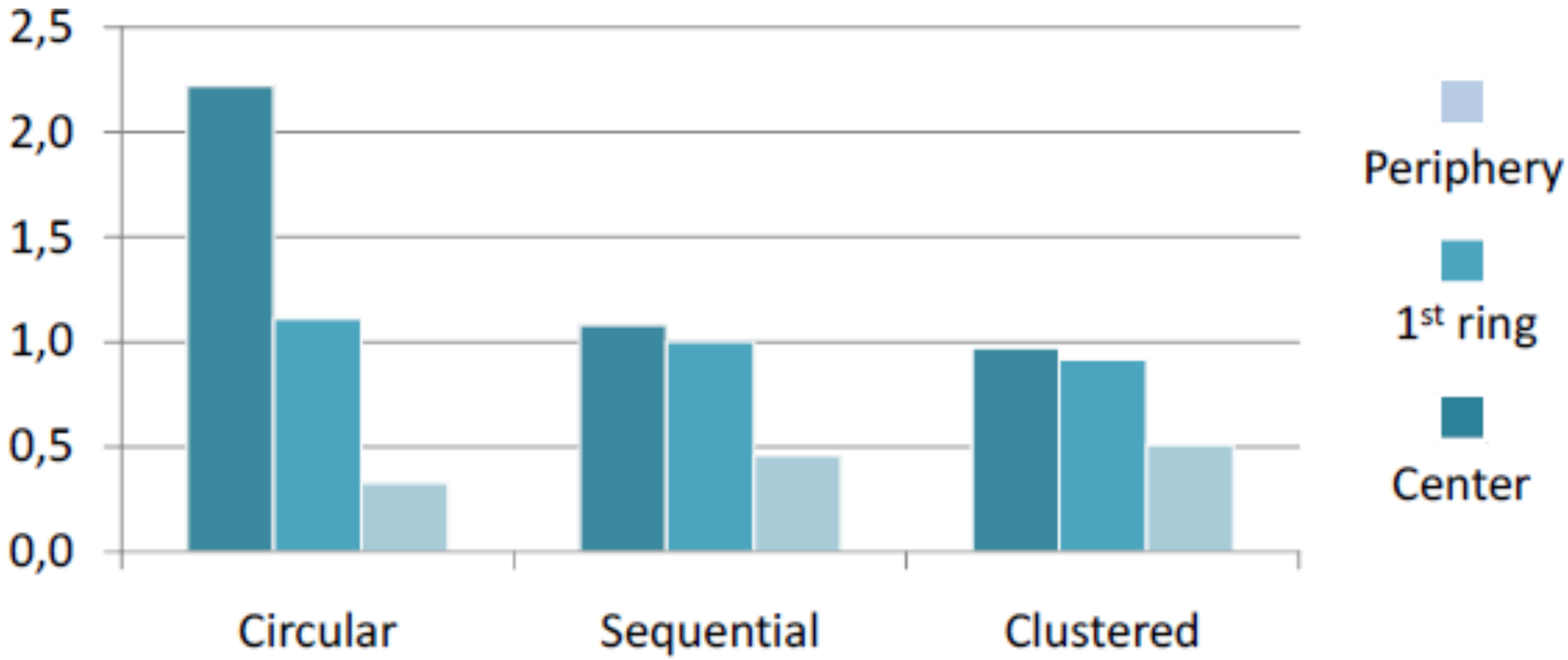



**Fig. 6.** Fixation density for the central-to-peripheral zones

P1 = 0-1s
P2 = 1-3s
P3 = 3-6s

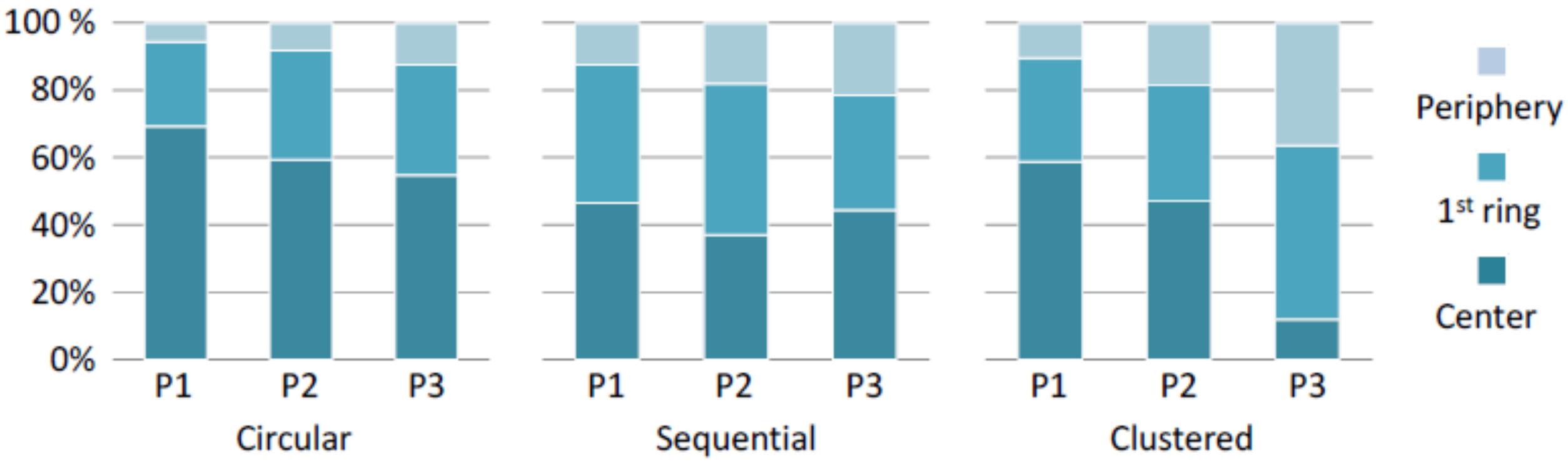



**Fig. 7**: Distribution of fixation density in the center, 1st ring and periphery zones for the three time periods

| Análise | Suposição | Avaliação do autor |
|---|---|---|
| Tamanho do termo | Grandes atraem mais atenção e são encontradas mais rapidamente | **Confirmam** mas outros fatores são importantes como: posição na nuvem e termos vizinhos |
| Forma de leitura | Usuários leem a nuvem da esquerda para a direita e de cima para baixo | **Confirmam** |
| Centralização | Termos centrais atraem mais atenção | **Confirmam** |
| Posição | Termos no quadrante da esquerda e superior são encontrados mais rapidamente | **Confirmam** para todos os leiaoutes |
| Leiaute | Influencia na percepção | **Confirmam** |
| Exploração | Provêm suporte sub-ótimo na busca por termos | **Confirmam** porém termos maiores são facilmente encontrados |

# CONCLUSÕES

- **Encontrar um termo específico:**
  - sequencial com ordenação alfabética

- **Encontrar o termo mais popular:**
  - circular com popularidade decrescente

- **Encontrar termos relacionados a um tópico:**
  - agrupamento por tópicos

(a) Alphabetically sorted tags, greedy algorithm

(b) Alphabetically sorted tags, dynamic programming

(c) Tags sorted by weight, greedy algorithm

(d) FFDH heuristic

Figure 8: Large tag cloud generated from a Project Gutenberg e-text.

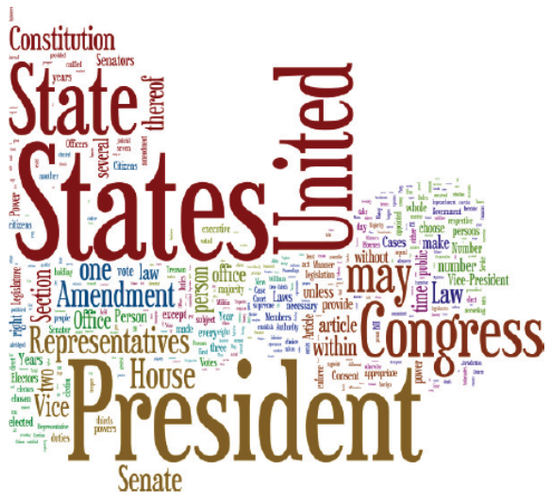



Figure 3-16. *The result of a clustering placement strategy*

Quais as limitações dessa abordagem?

# PARTICIPATORY VISUALIZATION WITH WORDLE

*C. Collins, F.B. Viégas e M. Wattenberg*
*IEEE Transactions on Visual Analytics Science and Technology*
*2009*

# OBJETIVOS

- Compreensão de textos multi-facetados
  - **Facetas**: dimensões ortogonais nas quais podem-se dividir uma fonte de informação

- Como as comédias e as tragédias de Shakespeare se comparam em termos de linguagem?

- Como diferentes cortes tratam os mesmos tipos de casos? Em diferentes períodos?

**Figure 2:** US Court *Circuits* are multi-state regions.

Inspirado em **coordenadas paralelas** e **nuvens de termos**

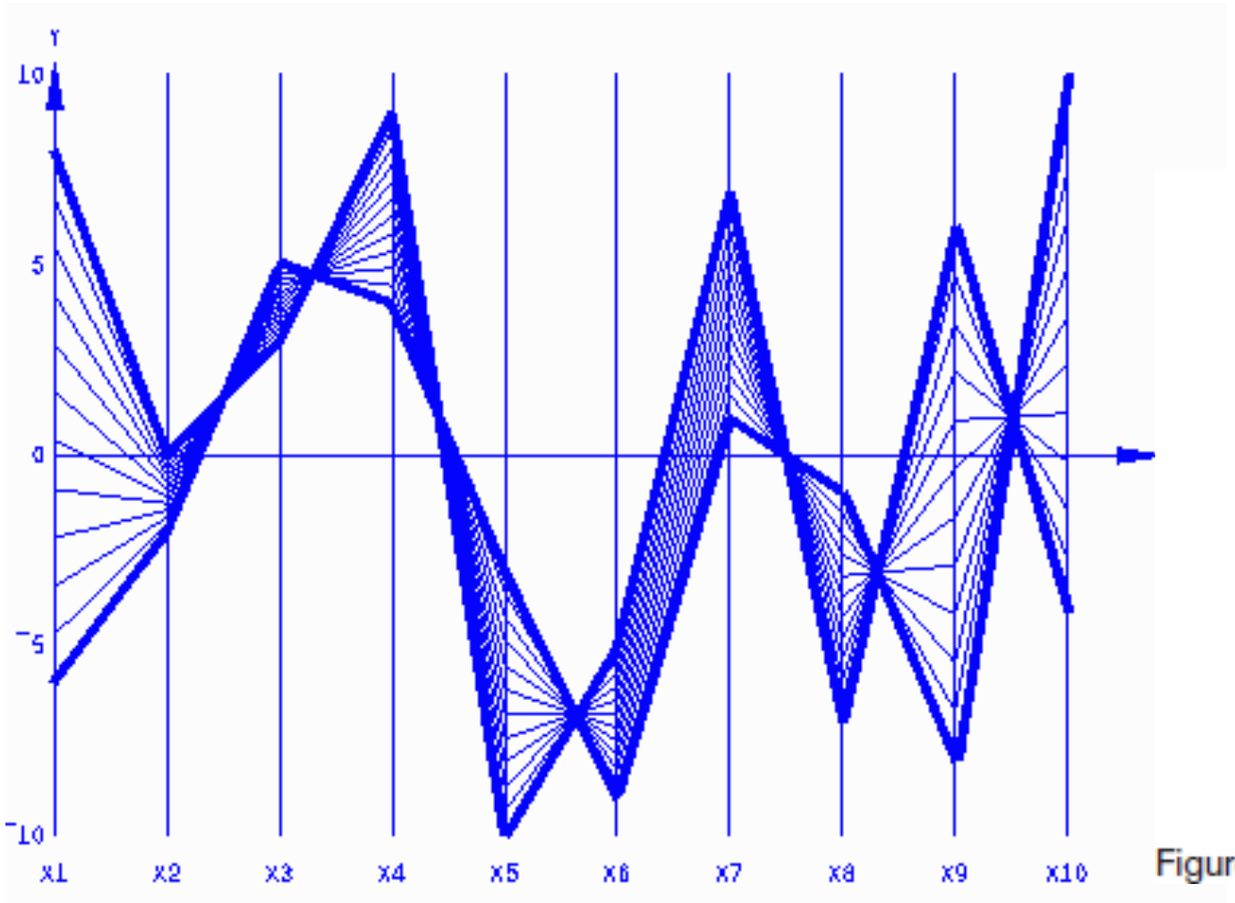




Figure 3-4. *The dogear tag explorer*

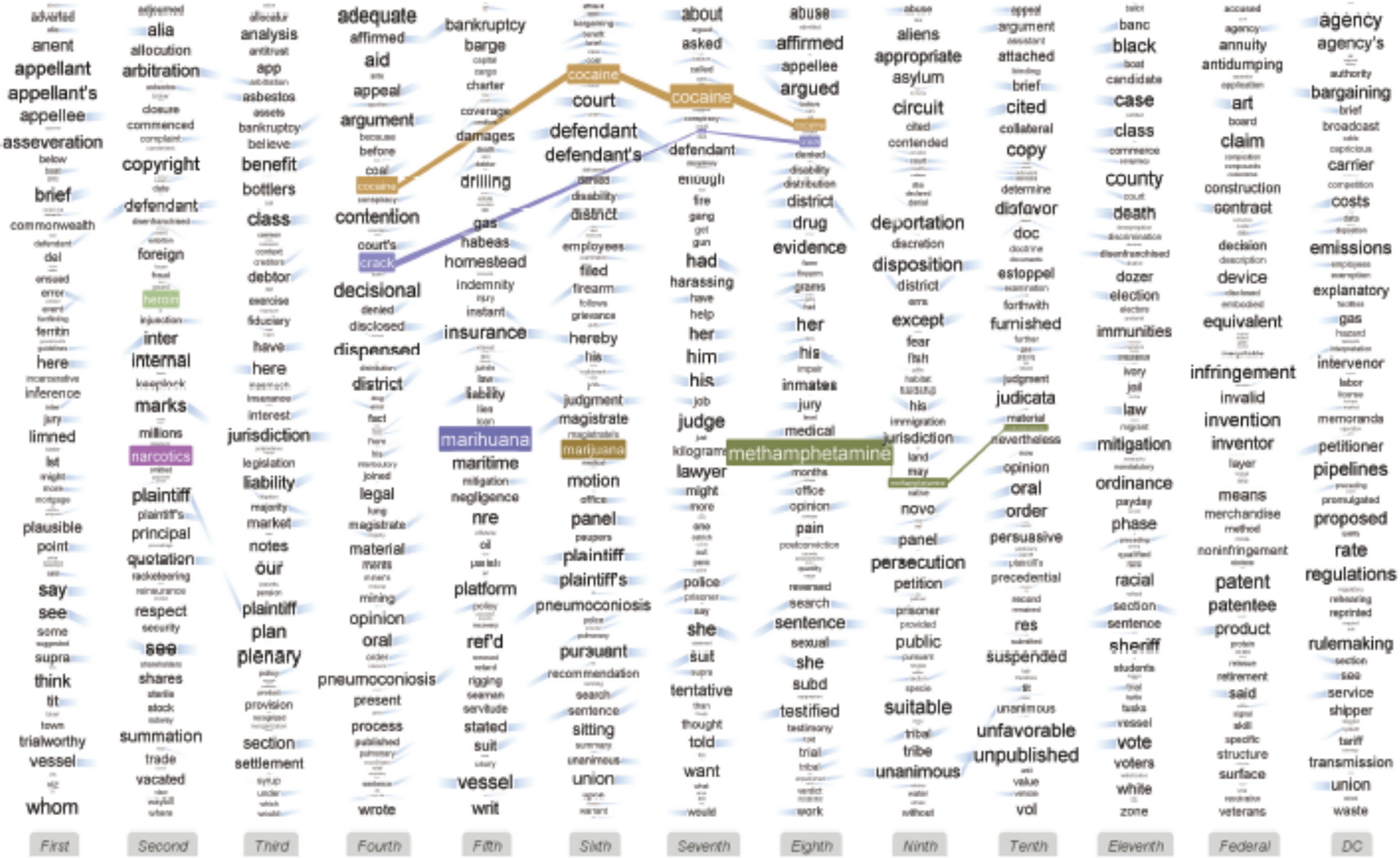



**Figure 1:** A PTC revealing the differences in drug prevalence amongst the circuits.

# REPRESENTAÇÃO VISUAL

- **Tamanho da fonte**: frequência do termo
- **Arestas** entre as colunas: indicam termos em comum entre as facetas
  - **Espessura da aresta**: frequência do termo, enfatizando esta propriedade
- Termos em **ordem alfabética**

# TAMANHO PELO RANK OU PELA FREQUÊNCIA?

- Como todas as colunas possuem o mesmo número de palavras, a escolha do tamanho da fonte pela **posição do termo no *rank*** otimiza o aproveitamento do espaço
  - Porém, impossibilita a comparação entre as frequências das palavras em diferentes facetas
- A escolha do tamanho **frequência global** é mais apropriada nestes casos
- O usuário pode selecionar entre estas duas opções além de remover palavras conforme sua necessidade

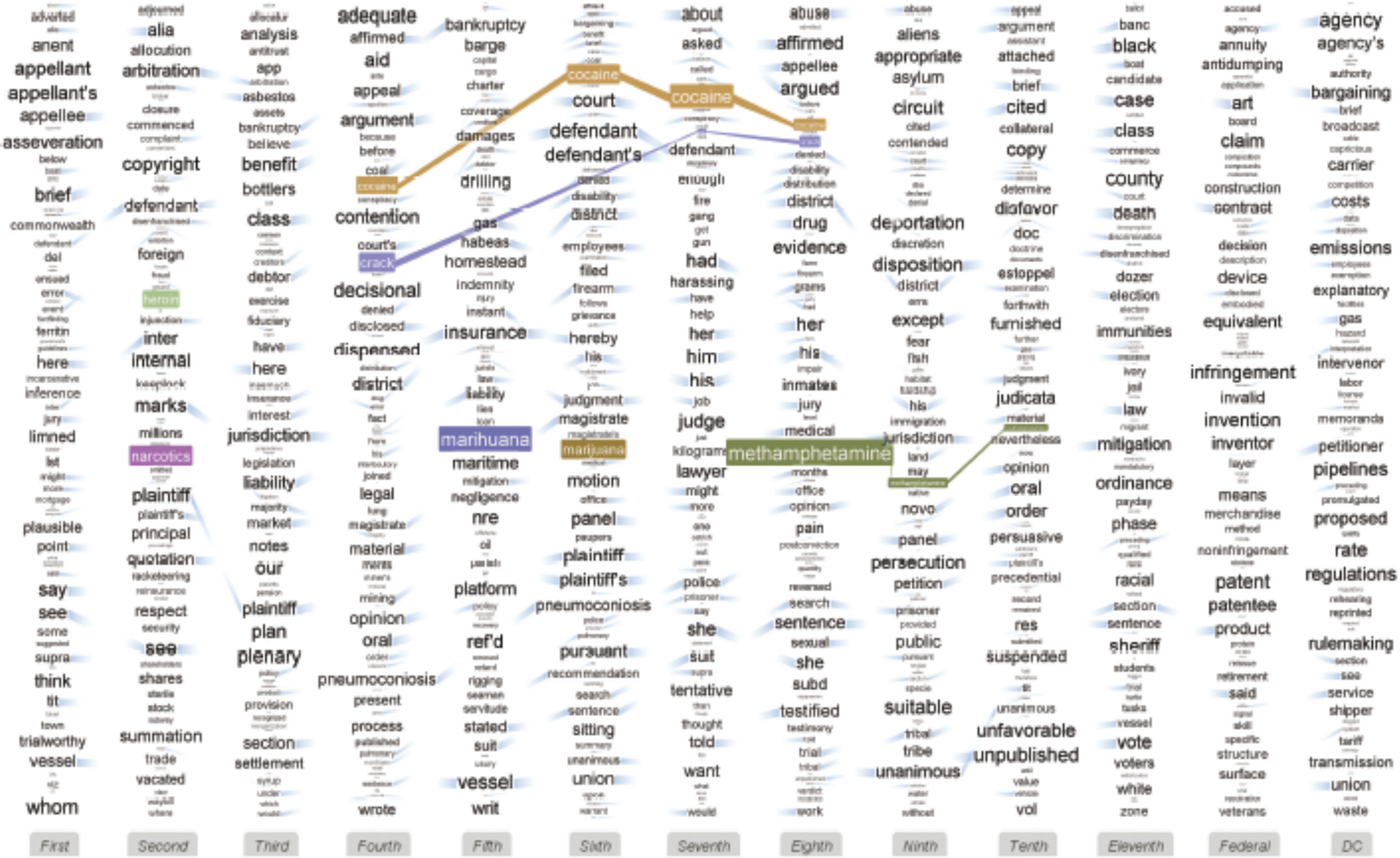



**Figure 1:** A PTC revealing the differences in drug prevalence amongst the circuits.


**Figure 3:** Sizing by score reveals that the Federal Circuit (far right) is the most different from other courts, and that the word 'patent' is overall the most differentiating word in the selected time period of the corpus.

# FILTROS

- ***Stop words***: as 0.5% palavras mais frequentes: juiz, corte, circuito, etc

- Remoção dos 40% menos frequentes: palavras que ocorrem uma ou duas vezes (cauda pesada)

- Remoção de palavras com iniciais maiúsculas: nomes

# TRATAMENTO DE SUFIXOS

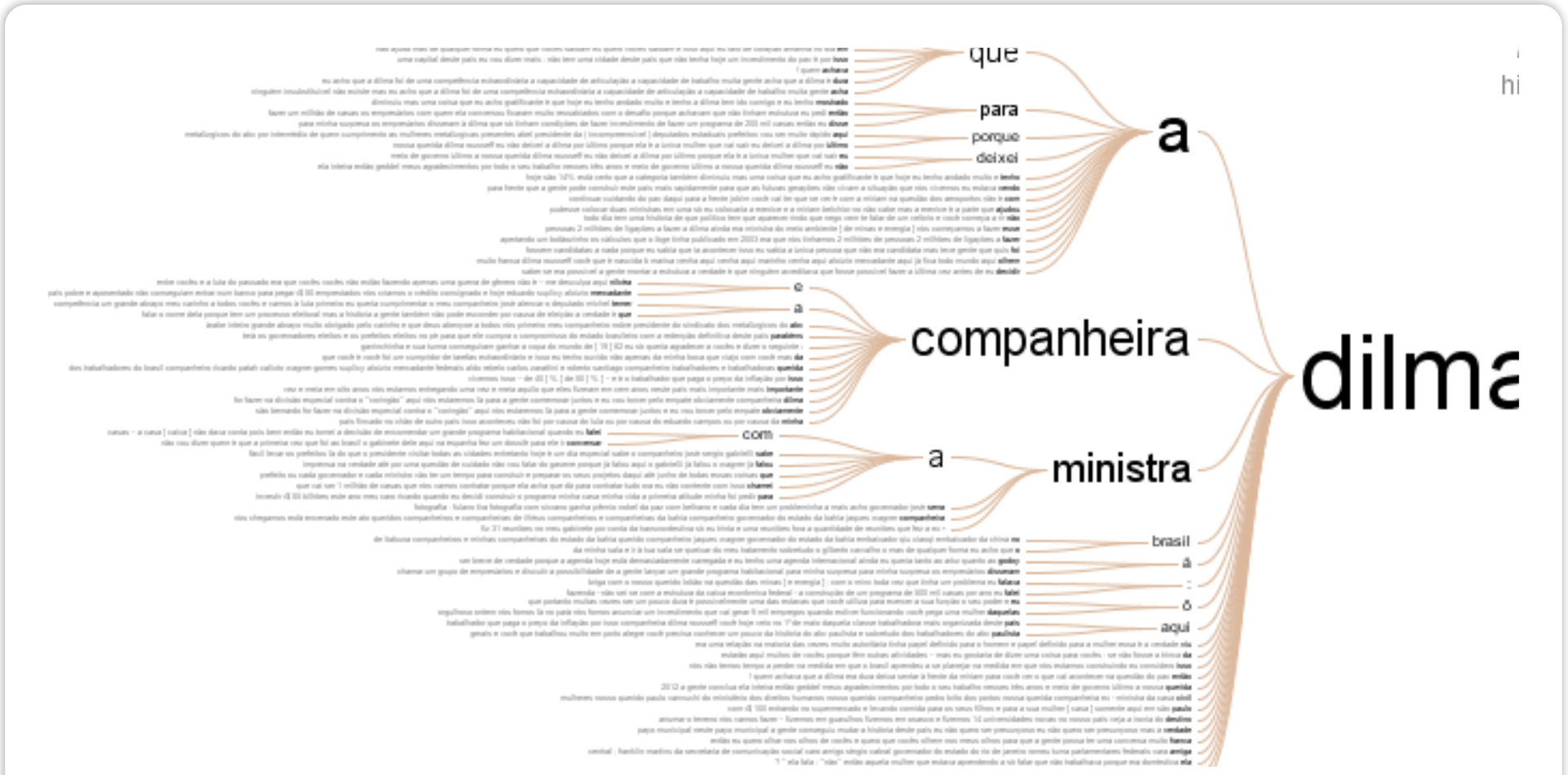
- Palavras como *jurisdictions* e *jurisdicion* devem ser consideradas a mesma ou ser relacionadas ao mesmo radical
  - Algoritmo de *stemming* de Porter

- Teste com 10.000 palavras
  - 6.370 palavras distintas (aproximadamente 1/3 das palavras é reduzida)

- União de todas as palavras de mesmo radical na contagem
- Exibição da mais frequente

# THE WORD TREE, AN INTERACTIVE VISUAL CONCORDANCE

*M. Wattenberg e F.B. Viégas*
*IEEE Transactions on Visualization and Computer Graphics*
*2008*



- Inspirado nas concordâncias: índices que provêm informação sobre o contexto de uso das palavras

- Atualmente, chamadas de *KeyWord In Context* (KWIC)

if love be rough with you , be rough with love .

if love be blind , love cannot hit the mark .

if love be blind , it best agrees with night .

Fig 1. All instances of “if love” in *Romeo and Juliet*.

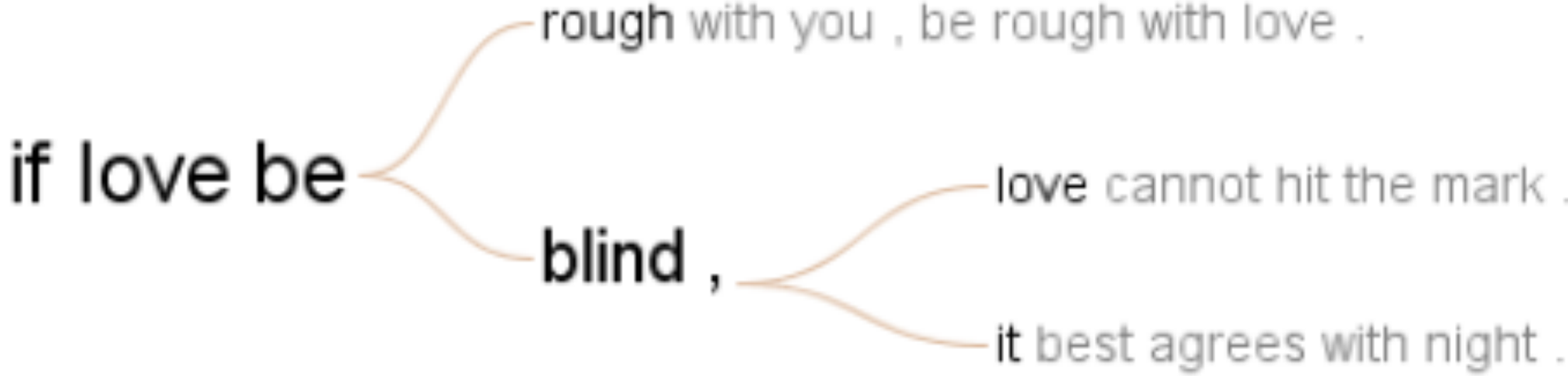



Fig 2. Word tree showing all instances of “if love” in *Romeo and Juliet*.

- Preserva a linearidade do texto

- O tamanho das fontes é proporcional à raiz quadrada da frequência do termo

- Sem descarte de *stop words* e pontuação

- Não desenham sub-árvores com menos de 3 pixels de altura

- Apenas ramos com mais de 1% do total das folhas são exibidos

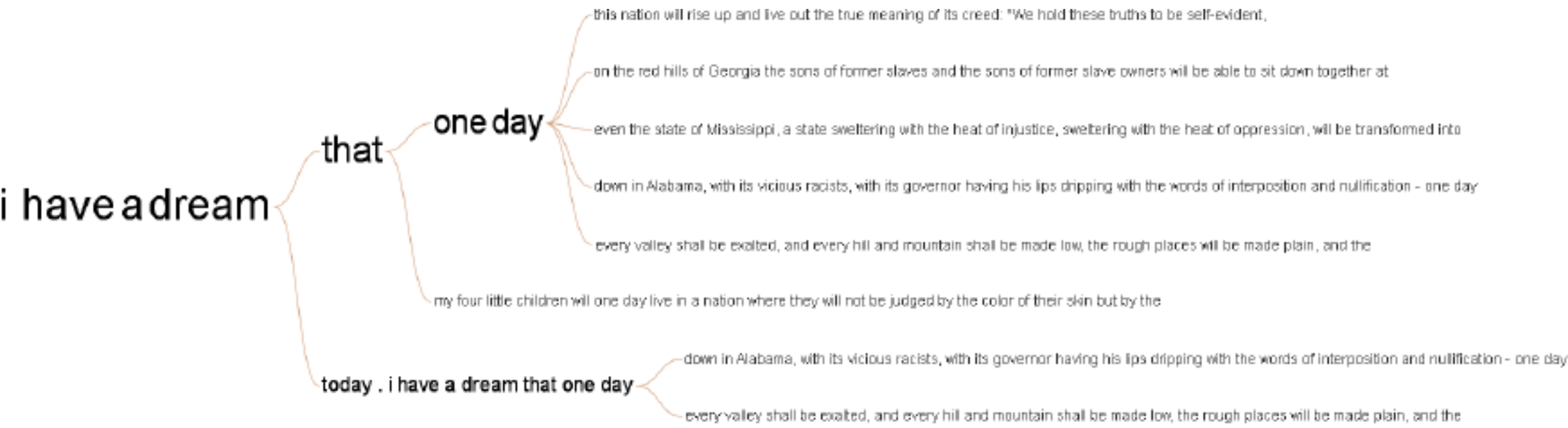



Fig 10: Word Tree showing all occurrences of “I have a dream” in Martin Luther King’s historical speech.


Fig 9. Word tree of the King James Bible showing all occurrences of “love the.”

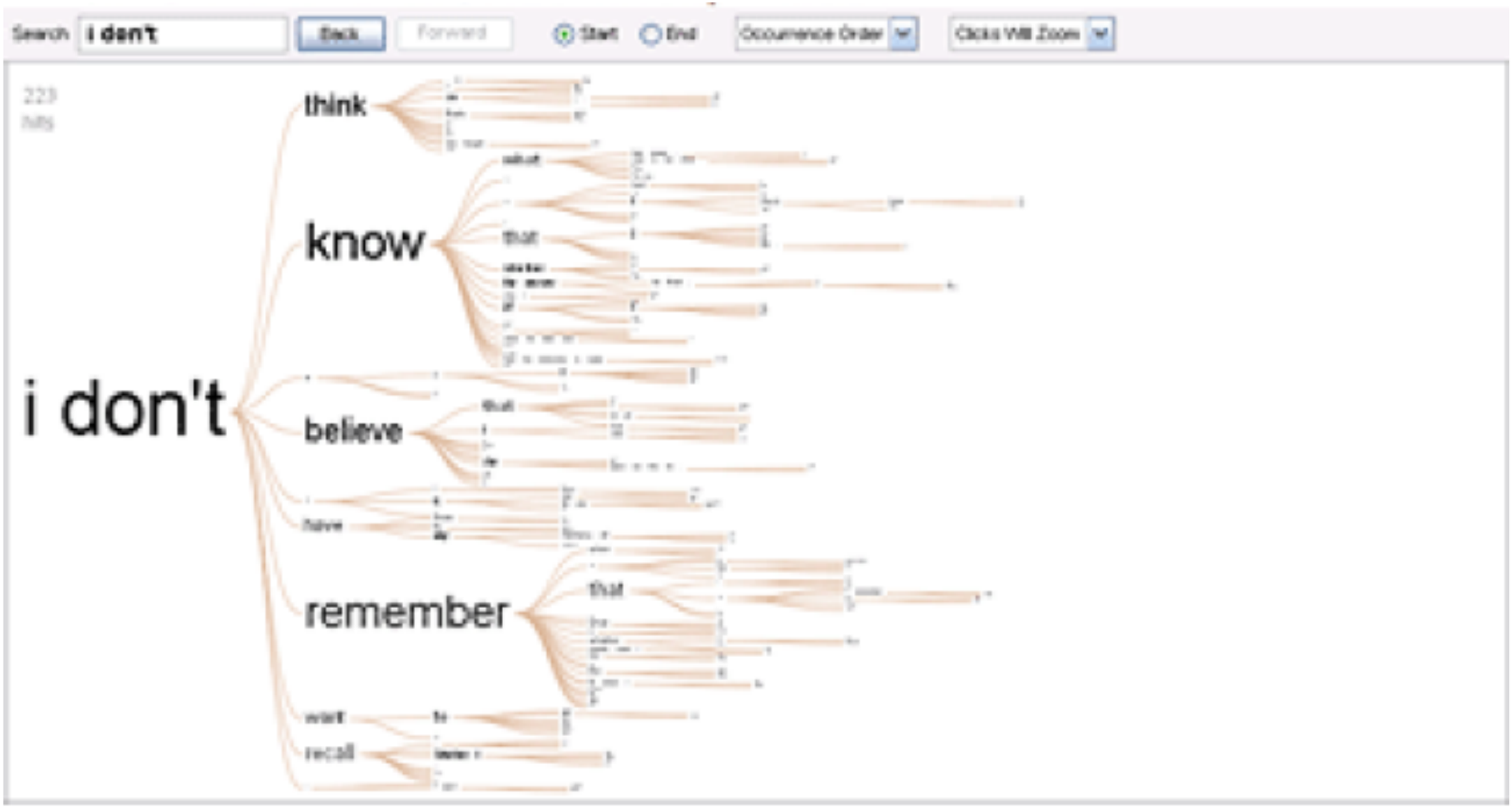



Fig 6: Bill Clinton's testimony in 1998.

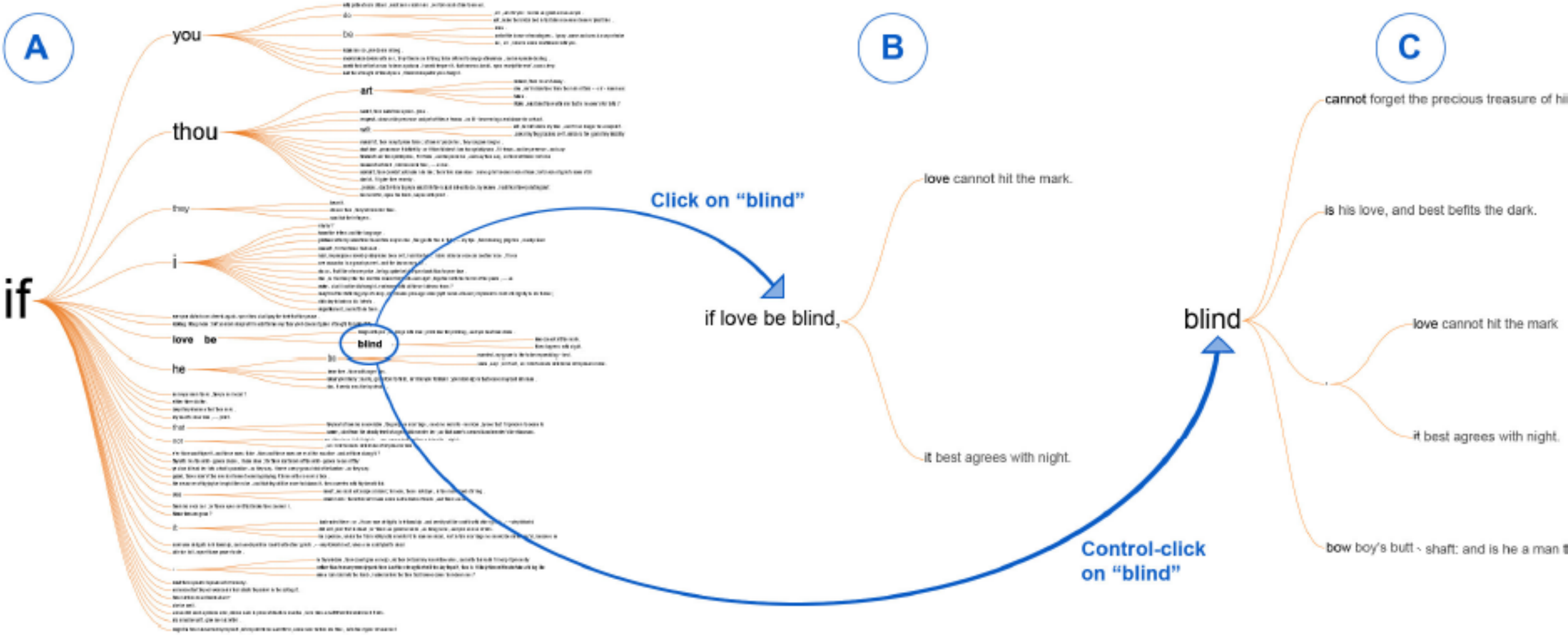



Fig 3. Sequence showing some of the interaction options in the word tree. In figure A, the user has typed the word “if” in *Romeo and Juliet*. In B, the user has clicked on “blind,” which appears in one of the branches under “if.” This causes the visualization to recenter to the longer phrase “if love be blind.” In C, the user Control-clicks on “blind,” which causes the visualization to recenter to blind by itself, revealing that there are

## Data set Structure

```
President  Governor  George_W_  Bush
President  Governor  Governor  Attorney_General  Clinton
President  Vice_President  Ambassador  CIA  Liason  Representative  George_HW_Bush
President  Governor  Reagan
President  Governor  State_Senator  Carter
President  Vice_President  Representative  Ford
President  Vice_President  Senator  Representative  Nixon
President  Vice_President  Senator  Representative  Johnson
President  Senator  Representative  Kennedy
President  General  Eisenhower
President  Vice_President  Senator  Truman
President  Governor  Secretary_of_the_Navy  State_Senator  Franklin_Roosevelt
President  Secretary_of_Commmerce  Humanitarian  Hoover
President  Vice_President  Governor  Coolidge
President  Senator  Lt_Governor  State_Senator  Harding
President  Governor  University_President  Wilson
President  Secretary_of_War  Governor_General_of_Phillipines  Federal_Judge
```

## Pathways to the Presidency

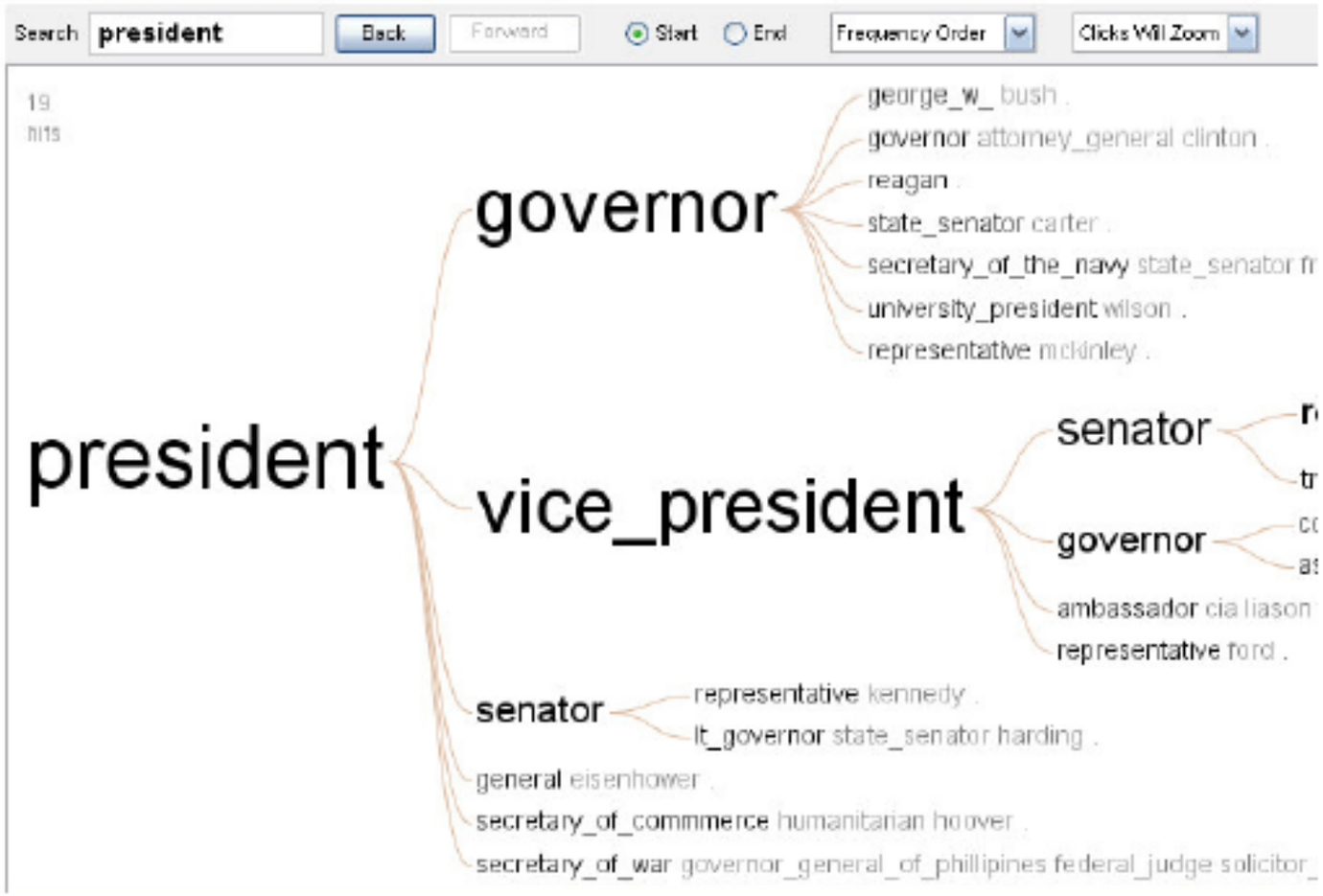



Fig 8. Data set and word tree of pathways to the U.S. presidency.

## Data set structure

A I G U P T I O I NPM
G A L I L A I O I NPM
A U Q A I R E T O I NPM
A U Q A D E I S NPM
A U T O X E I R E S NPM
B A R E I S NPM
B D E L U K T O I NPM
B L A S F H M O I NPM
B A R B A R O I NPM
G N W S T O I NPM
G U M N O I NPM
D E I L O I NPM
D E K A O K T W NPM
D E U T E R A I O I NPM
O M O I A I NPF
P L E I O U S NPF
P L E I S T A I NPF
A I G U P T I O I NPM

## New Testament Greek Nominal Suffixes

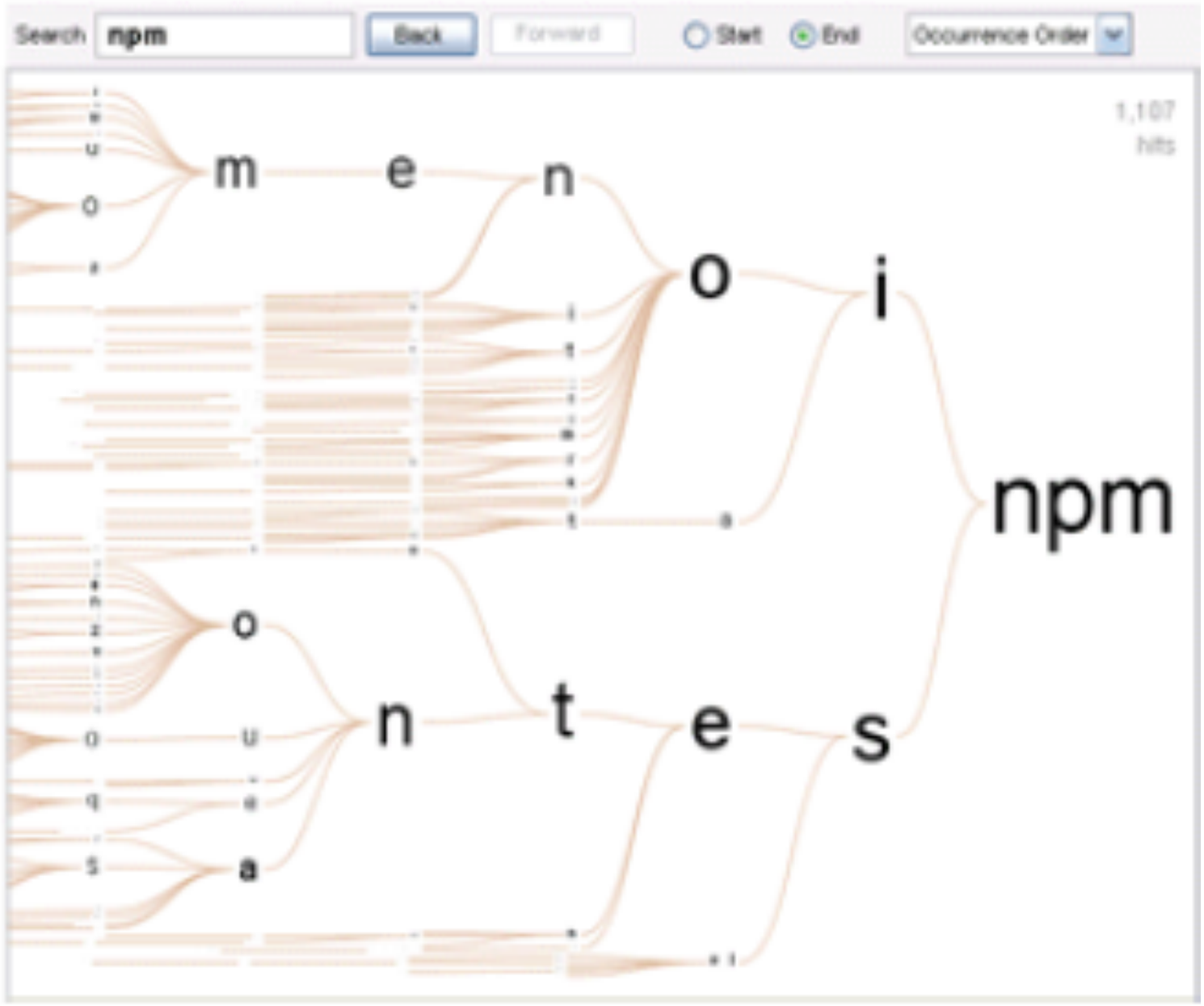
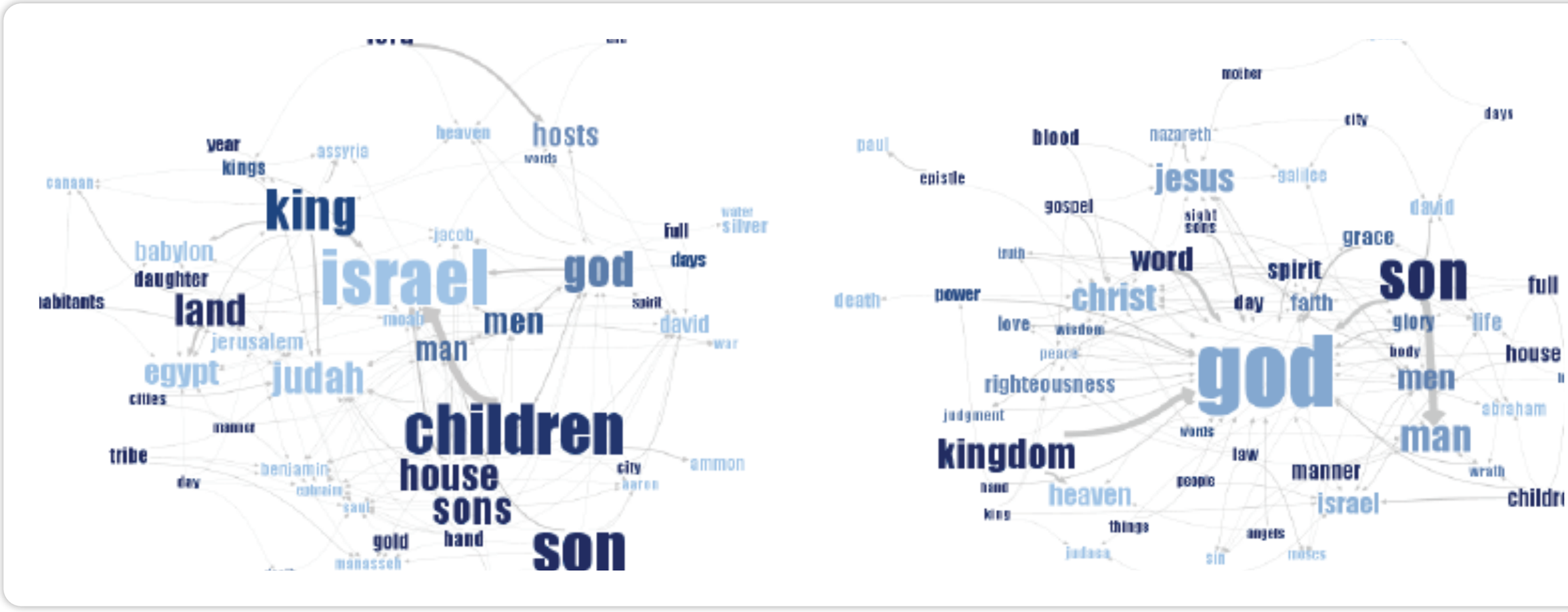



Fig 7. Data set and word tree of Greek nominal suffixes in the Bible. Here, “npm” refers to nominative, plural, masculine nouns.

# MAPPING TEXT WITH PHRASENETS

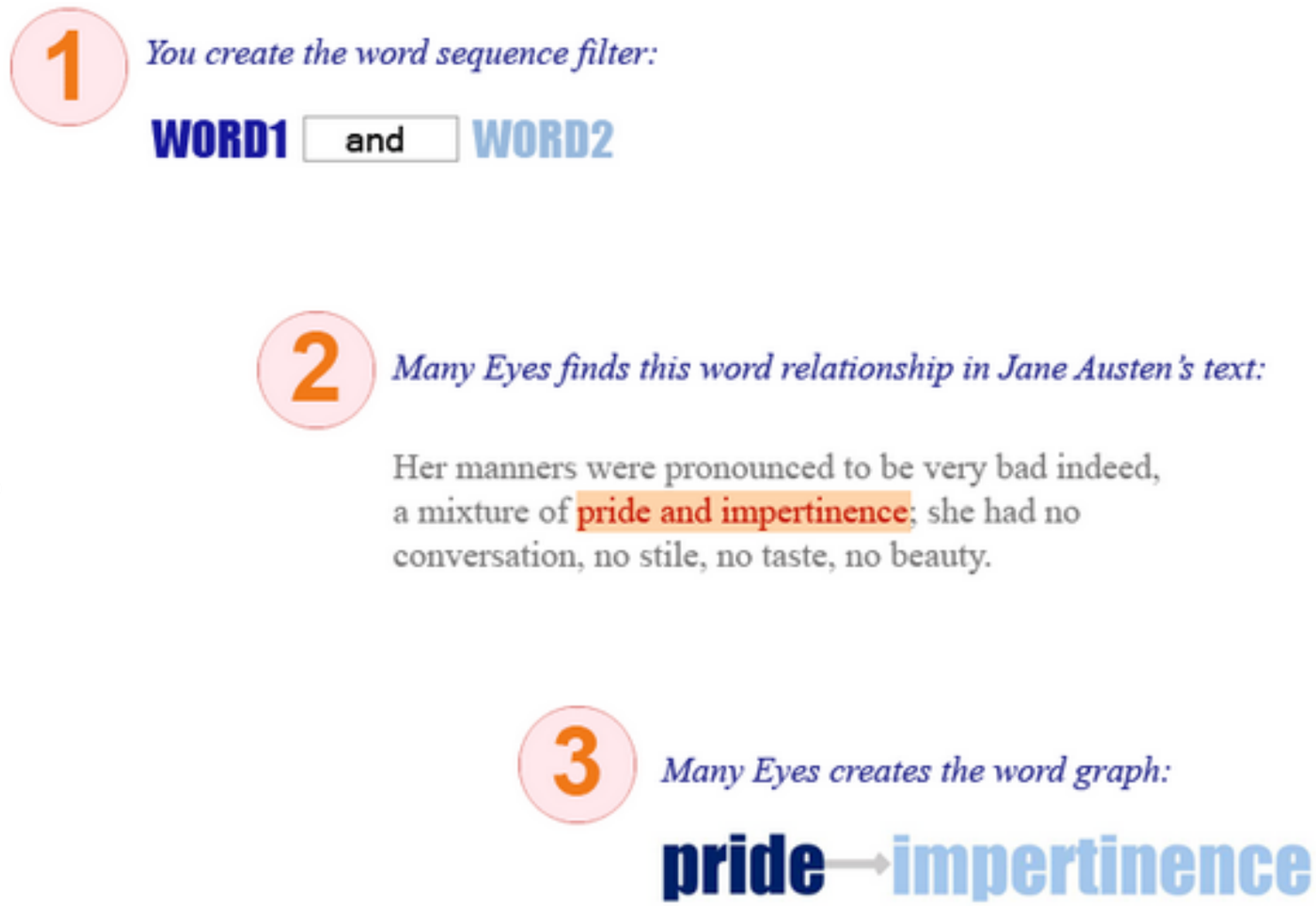
*F. van Ham, M. Wattenberg e F.B. Viégas*
*IEEE Transactions on Visualization and Computer Graphics*
*2009*



# ANÁLISE DO TEXTO

- Busca de ligações semânticas ou padrões com base na estrutura sintática de sentenças
- Casamento de padrão baseado no texto simples
  - Ao invés de usar técnicas de processamento de linguagem natural, para inferência de relacionamentos de posse, usar
    - "... X's Y ..."
    - "... X at Y ..."
  - "... because X (is|are|was|were) Y ..."

- Grafos (redes semânticas)
  - **nós** são palavras
  - **arestas** indicam palavras ligadas por relações definidas pelo usuário



# REPRESENTAÇÃO VISUAL

- Após a etapa de *parser*, tem-se um grafo cujos
  - **nós** são palavras: **tamanhos das fontes** indicando suas **frequências**
  - **arestas** são direcionadas: **espessura** indicando a **frequência** de ocorrência do padrão

- Grandes redes: problema de legibilidade


Fig 2. Comparing phrase nets of Pride and Prejudice. The left network was generated by an orthographical parser matching ‘X at Y’, while the right network was generated by a syntactical parser looking for the preposition ‘at’. Both manage to identify major locations in the novel., yet the orthographical parser finished in under a second, whereas the syntactical parser took over 24 hours.

# FILTRAGEM

- Remoção de *stop words*

- Ordenação de palavras pela relevância
  - relevância = frequência no texto

# COMPRESSÃO DAS ARESTAS

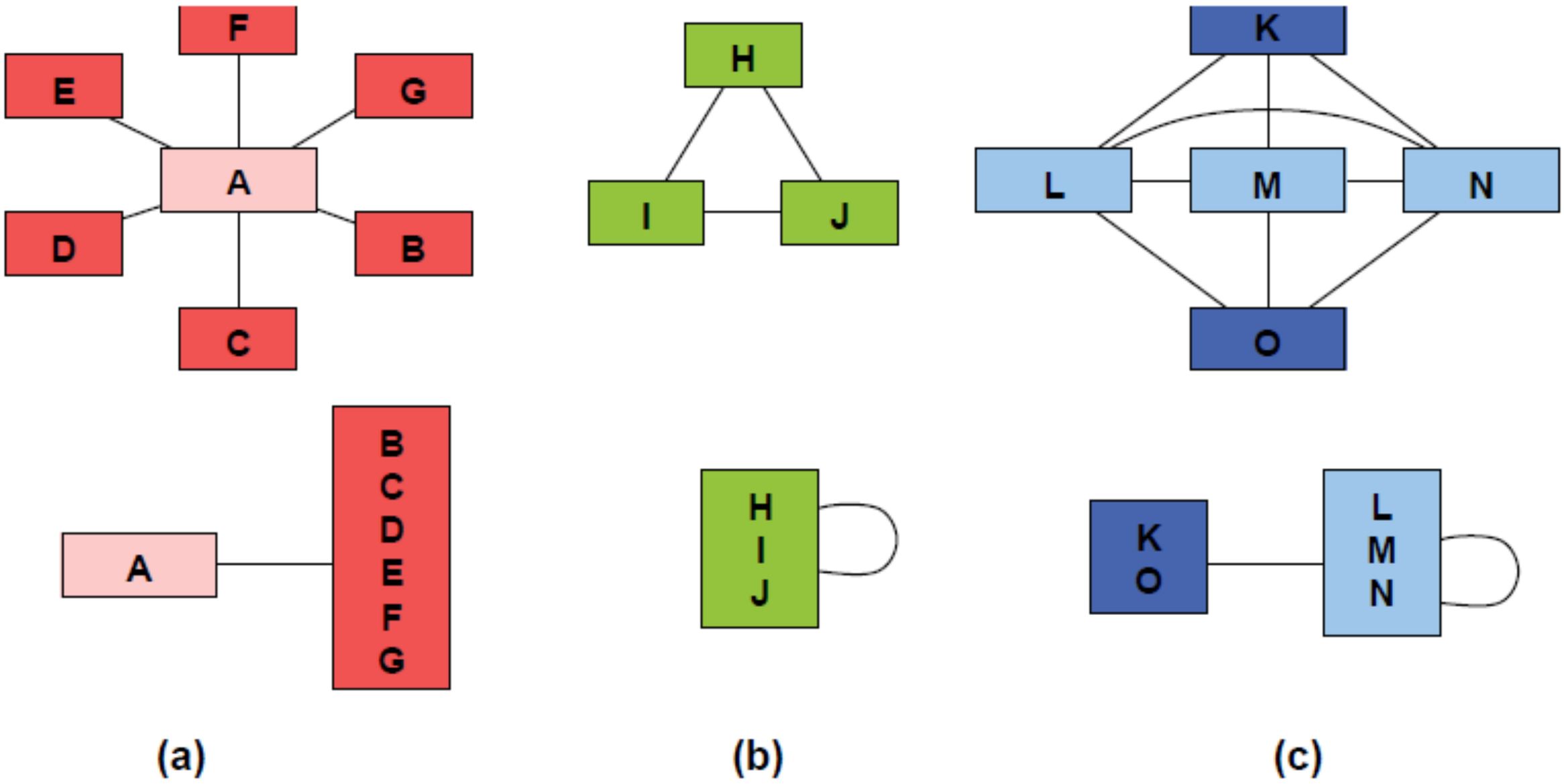



Fig 1. **Edge compression**: Collapsing networks based on identical network neighborhoods. The darker nodes in **(a)** all have identical neighbor sets {a} and can be collapsed into a singe clustered node. Although the nodes in **(b)** are all structurally interchangeable they have different neighbor sets {i,j}, {h,j} and {h,i} respectively; we can still merge them if we consider $N_{self}$ set {h,i,j} instead. The graph in **(c)** has neighbor sets {l,m,n} for both K and O (dark blue) and $N_{self}$ set {k,l,m,n,o} (light blue) for L, M and N.

# LEIAUTE

- Algoritmo *stress majorization*

- Pós-processamento com algoritmo baseado em transferência de força para reduzir as distâncias entre as palavras

# STRESS MAJORIZATION

- Uma variação de uma técnica chamada *Multidimensional Scaling* (MDS)

- MDS denomina uma família de técnicas para análise de proximidade entre itens de um conjunto de dados

- As técnicas consistem em calcular estímulos dados a cada item em um espaço *n*-dimensional de forma a posicionar os itens em distâncias proporcionais às similaridades entre eles

| | Boston | NY | DC | Miami | Chicago | Seattle | SF | LA | Denver |
|---|---|---|---|---|---|---|---|---|---|
| Boston | 0 | 206 | 429 | 1504 | 963 | 2976 | 3095 | 2979 | 1949 |
| NY | 206 | 0 | 233 | 1308 | 802 | 2815 | 2934 | 2786 | 1771 |
| DC | 429 | 233 | 0 | 1075 | 671 | 2684 | 2799 | 2631 | 1616 |
| Miami | 1504 | 1308 | 1075 | 0 | 1329 | 3273 | 3053 | 2687 | 2037 |
| Chicago | 963 | 802 | 671 | 1329 | 0 | 2013 | 2142 | 2054 | 996 |
| Seattle | 2976 | 2815 | 2684 | 3273 | 2013 | 0 | 808 | 1131 | 1307 |
| SF | 3095 | 2934 | 2799 | 3053 | 2132 | 808 | 0 | 379 | 1235 |
| LA | 2979 | 2786 | 2631 | 2687 | 2054 | 1131 | 379 | 0 | 1059 |
| Denver | 1949 | 1771 | 1616 | 2037 | 996 | 1307 | 1235 | 1059 | 0 |

Miami

LA

SF

Denver

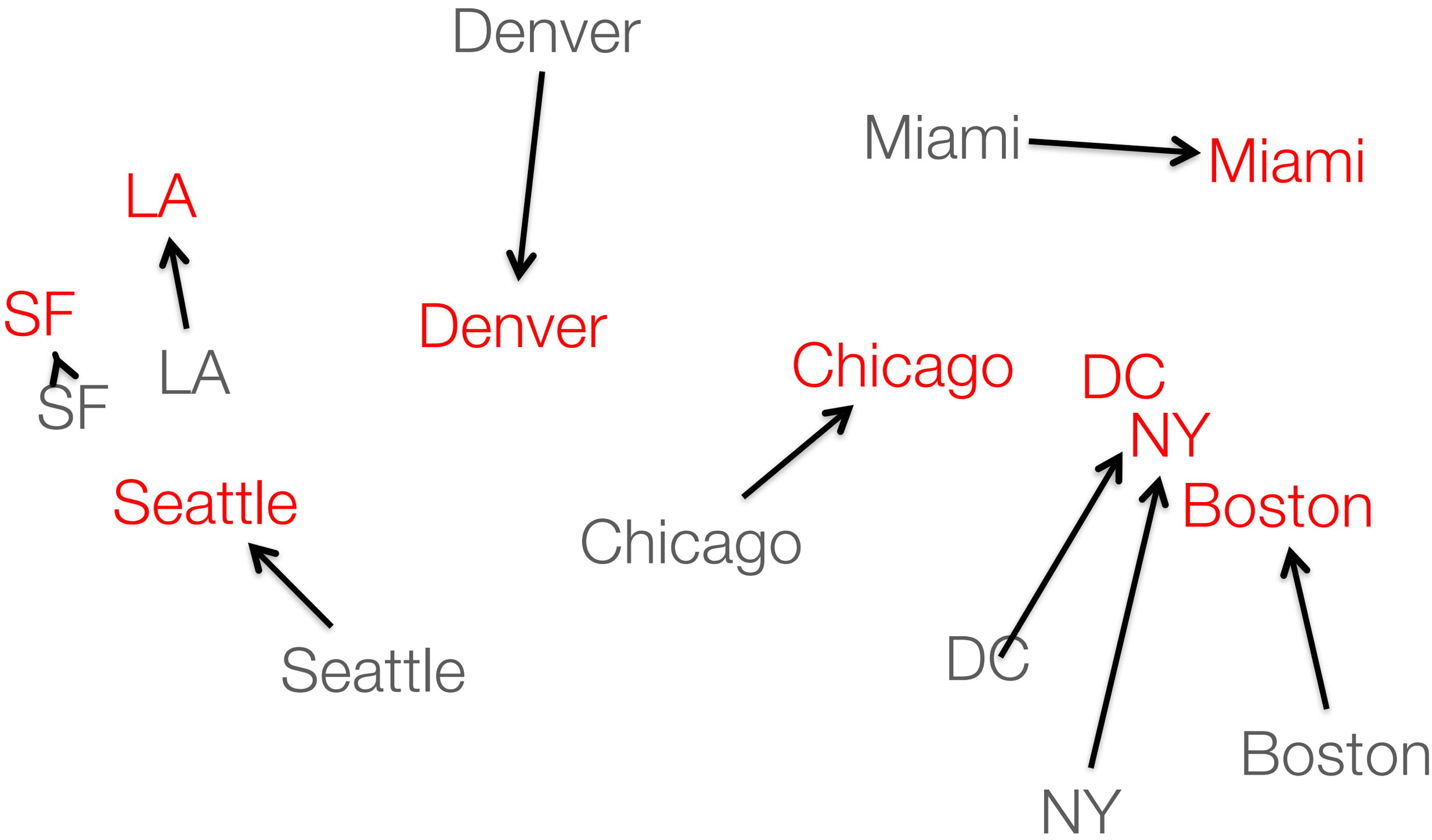
Chicago

DC

NY

Seattle

Boston




Solução através do método de gradiente conjugado iterativo

# STRESS MAJORIZATION

1. Coloque os pontos em posições arbitrárias no espaço $n$-dimensional

2. Compute as distâncias euclidianas entre os todos os possíveis pares de pontos ($d_{ij}$)

3. Compare a matriz de distâncias ($f(x_{ij})$) à matriz de similaridades para avaliação da função de stress

$$\sqrt{\frac{\sum\sum\left(f(x_{ij})-d_{ij}\right)^2}{\sum\sum d_{ij}^2}}$$

4. Ajuste as coordenadas dos pontos na direção que minimiza o stress

5. Repita os passos 2 a 4 até a convergência

# REPRESENTAÇÃO VISUAL

- **Tamanho da fonte**: frequência da palavra
- **Espessura da aresta**: frequência de ocorrência do relacionamento
- **Intensidade da cor**: razão *out-degree/in-degree*
- Usuários podem usar zoom e detalhes sob demanda mostrar os trechos de texto onde os padrões foram encontrados

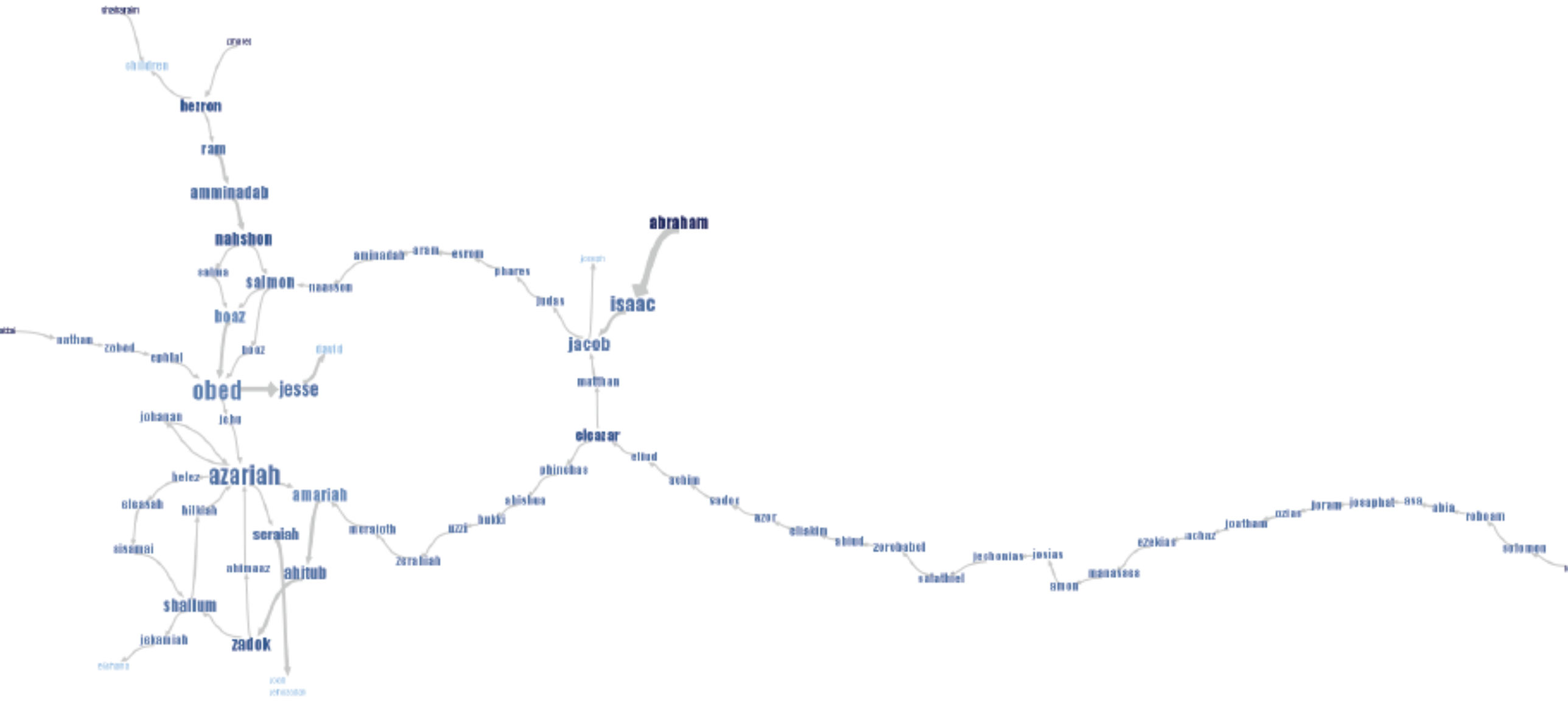
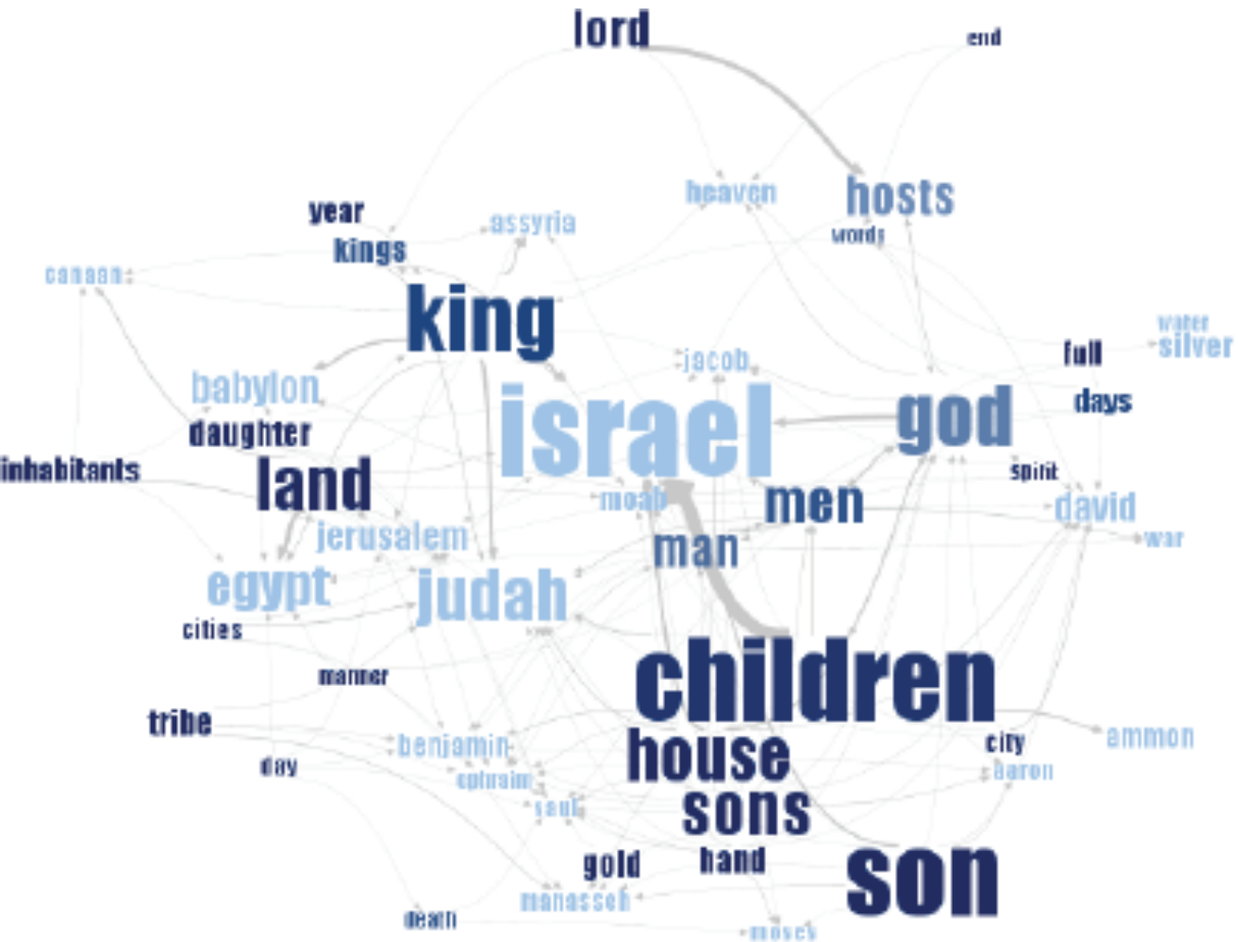



Fig. 1. Scanning the bible for textual matches to the pattern '*X begat Y*' reveals a network of family relations.



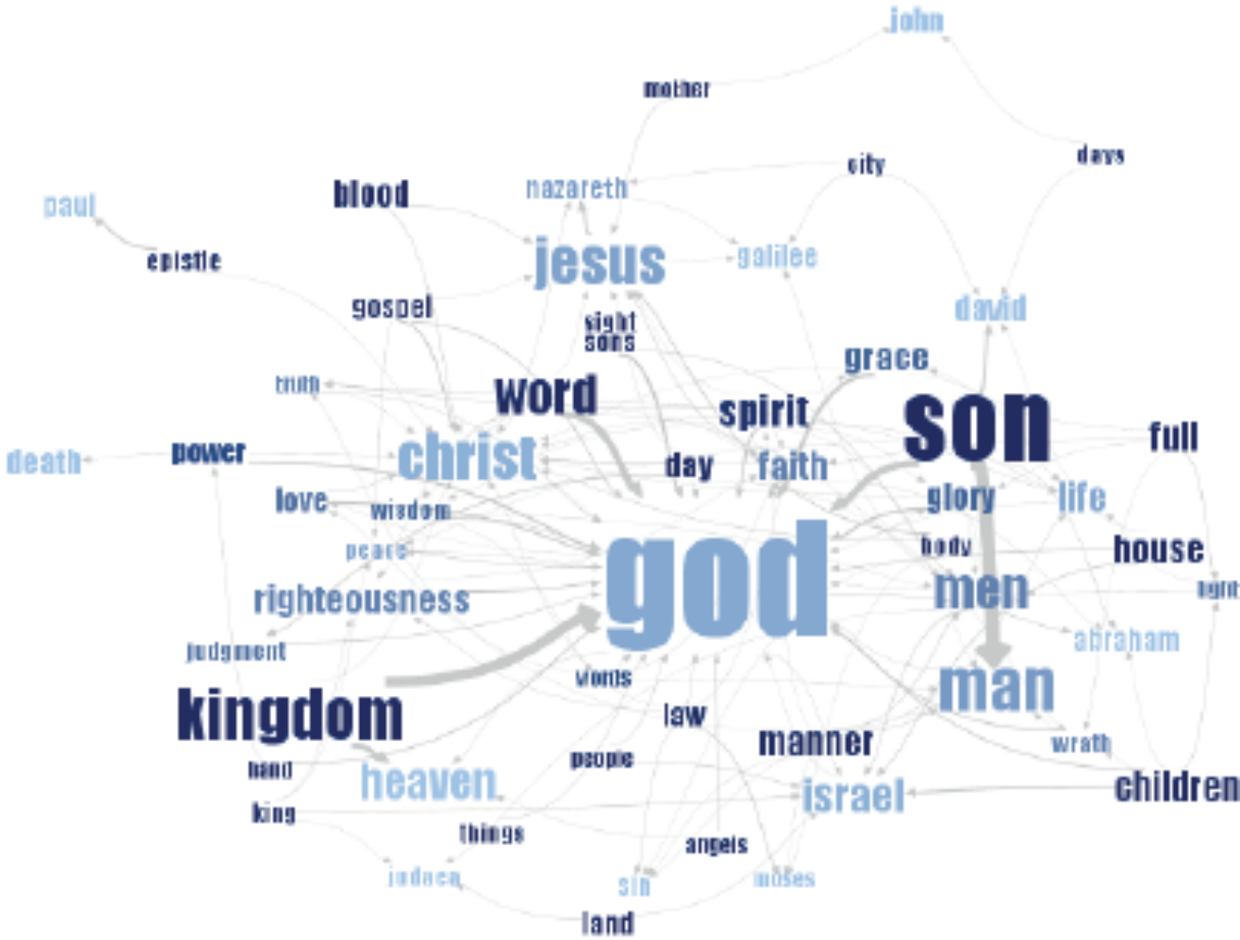




Fig 4. Matching the same pattern on different texts. Here we used the pattern “X of Y” to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

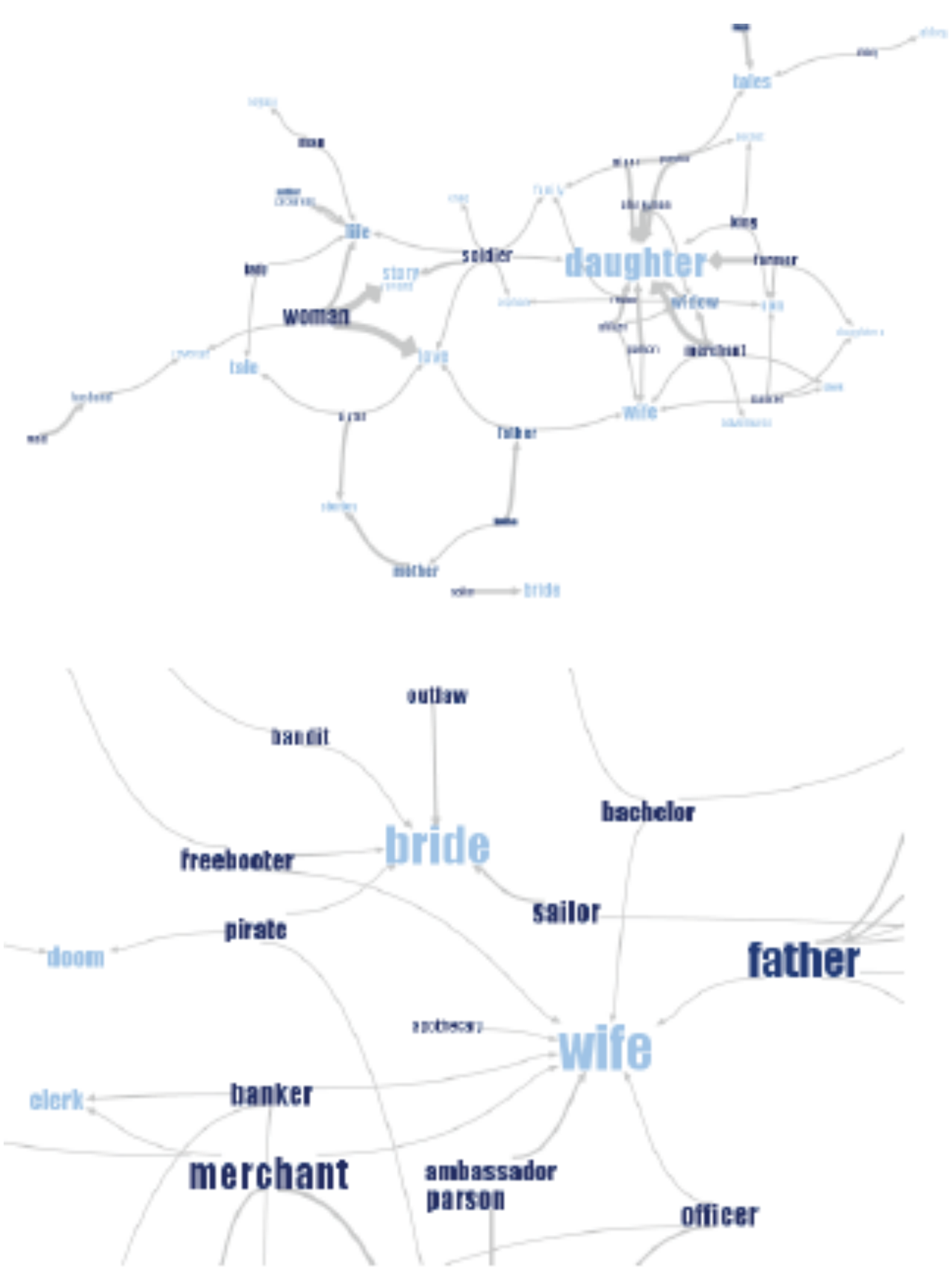



Fig 6. Phrase nets from a collection of titles of 18th and 19th century novels. The text was analyzed with the "X's Y" pattern. The top phrase net shows the top 50 terms. The bottom detail image

# Beautiful

is the most common word translated to English with Google

*when looking only at nouns and their adjectives*

A collaboration between Google News Lab and Visual Cinnamon

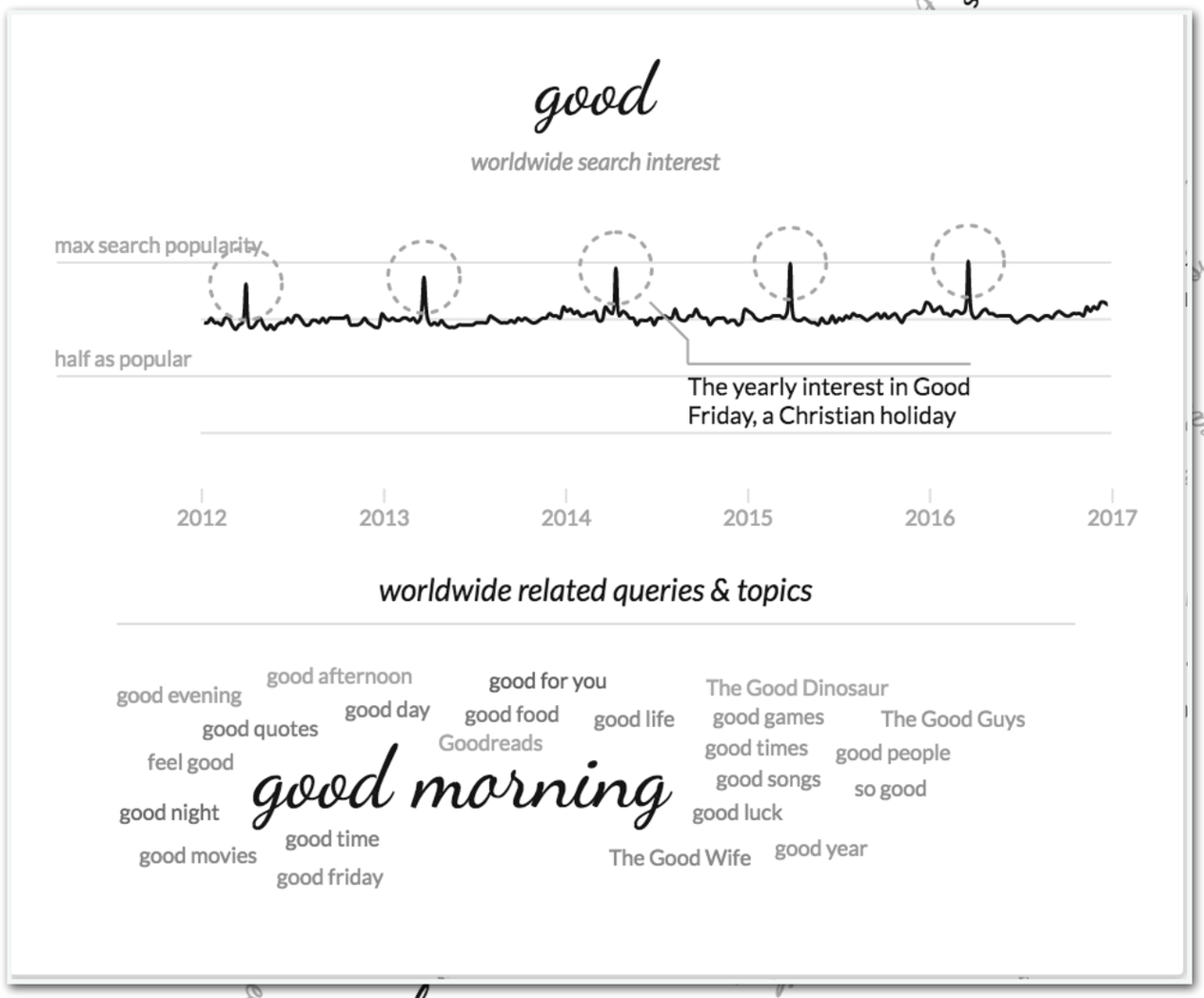
• THE MOST •
translated
• NOUN or ADJECTIVE •
word
in the language of
Google

colleagu
colleague

holiday regards reply 60 decision important girl life bad right

the most translated words across all languages

French
bon
good




summer target power colleague height price small dividend 80 can add

Italian
bene
good

boa
good

inquiry expensive error 100 high

1 beautiful thursday tuesday wednesday good

Russian
мама
mama

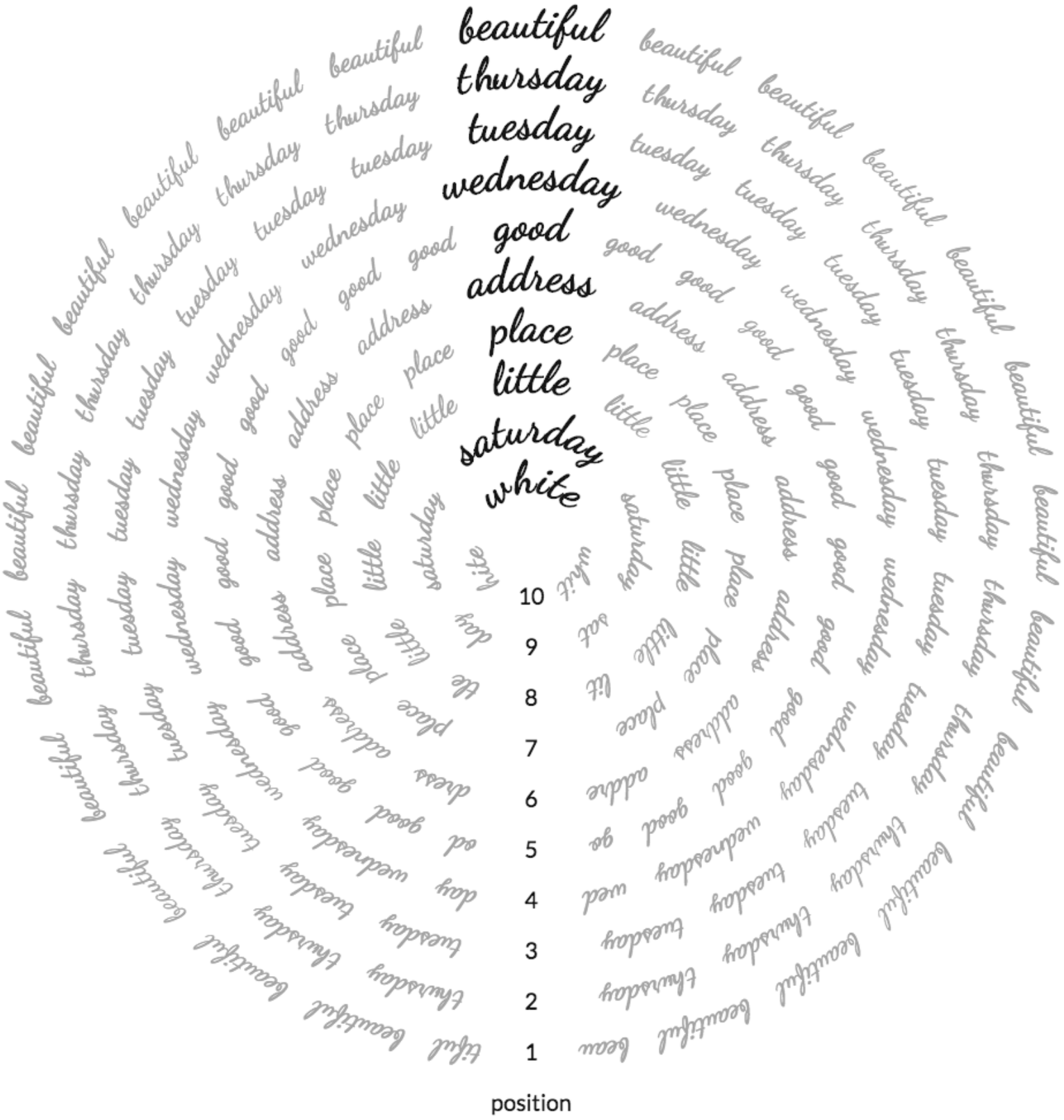
Spanish
hermoso
beautiful

10 white love request

:DIVING INTO

Top 10 Words Per Language

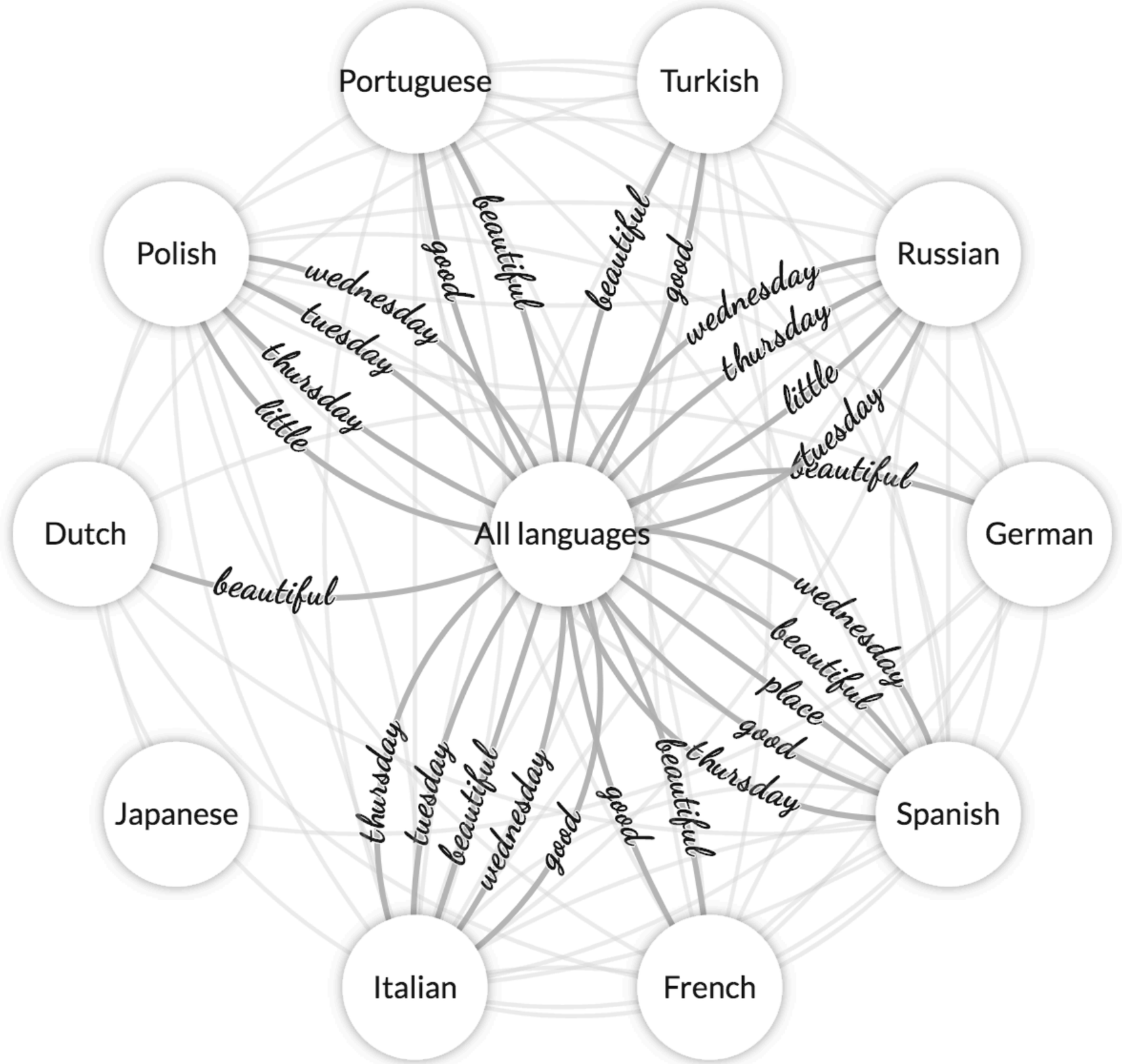
The top 10 words that are translated from **all 10 languages combined** into English※



Similarities BETWEEN languages

The similarities between the top 10 words* per language

**Click on an outside circle to move it to the center**



# TEXT VISUALIZATION BROWSER: A VISUAL SURVEY OF TEXT VISUALIZATION TECHNIQUES

*K. Kucher,, and A. Kerren,*
*Poster Abstracts of IEEE VIS*
2014

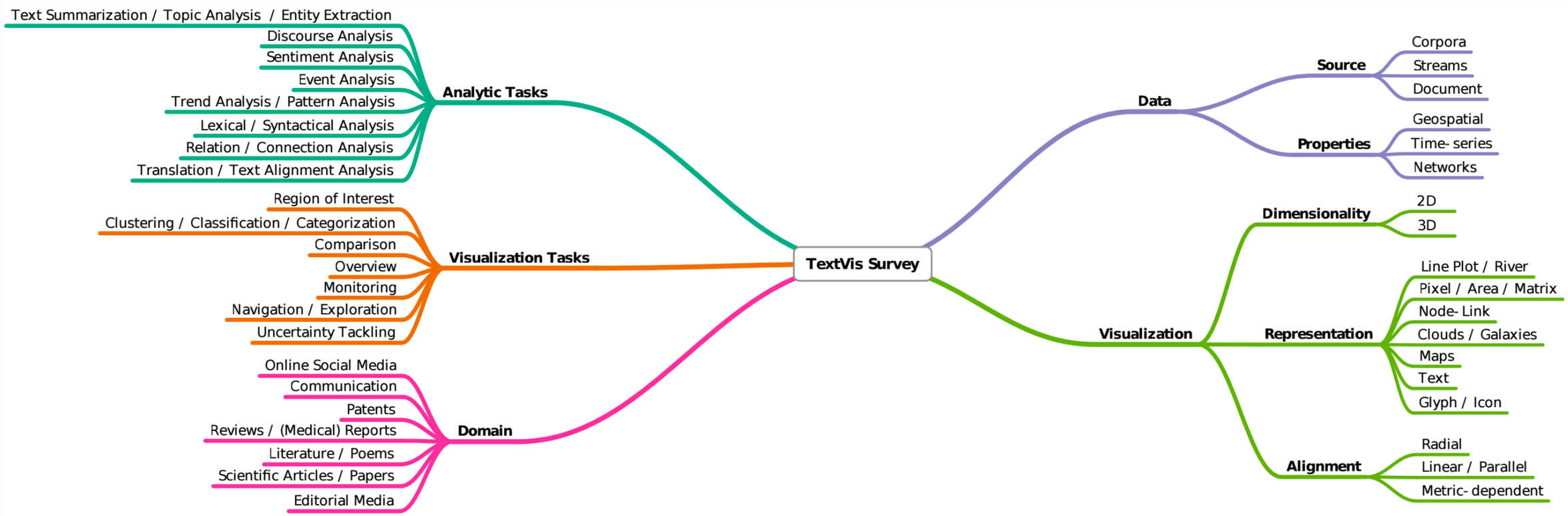



Figure 2: The taxonomy of text visualization techniques used in our visual survey.

# Text Visualization Browser

**A Visual Survey of Text Visualization Techniques (IEEE PacificVis 2015 short paper)**

Provided by ISOVIS group



**Techniques displayed:**

**380**

**Search:**

**Time filter:**

1976 2017

...

## Analytic Tasks

Σ text ...

## Visualization Tasks