

IFCS 2009
Dresden – 17/03/2009

Visualising a text with a tree cloud

Philippe Gambette, Jean Véronis



Outline

- Tag and word clouds
- Enhanced tag clouds
- Tree clouds
- Construction steps
- Quality control

Tag clouds

- Built from a set of tags
- Font size related to frequency



What is considered the first tag cloud, from D. Coupland: *Microserfs*, HarperCollins, Toronto, 1995

Tag clouds

- Built from a set of tags
- Font size related to frequency
- Gained popularity with Flickr

All time most popular tags

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

Flickr's *all time most popular tags*

Word clouds

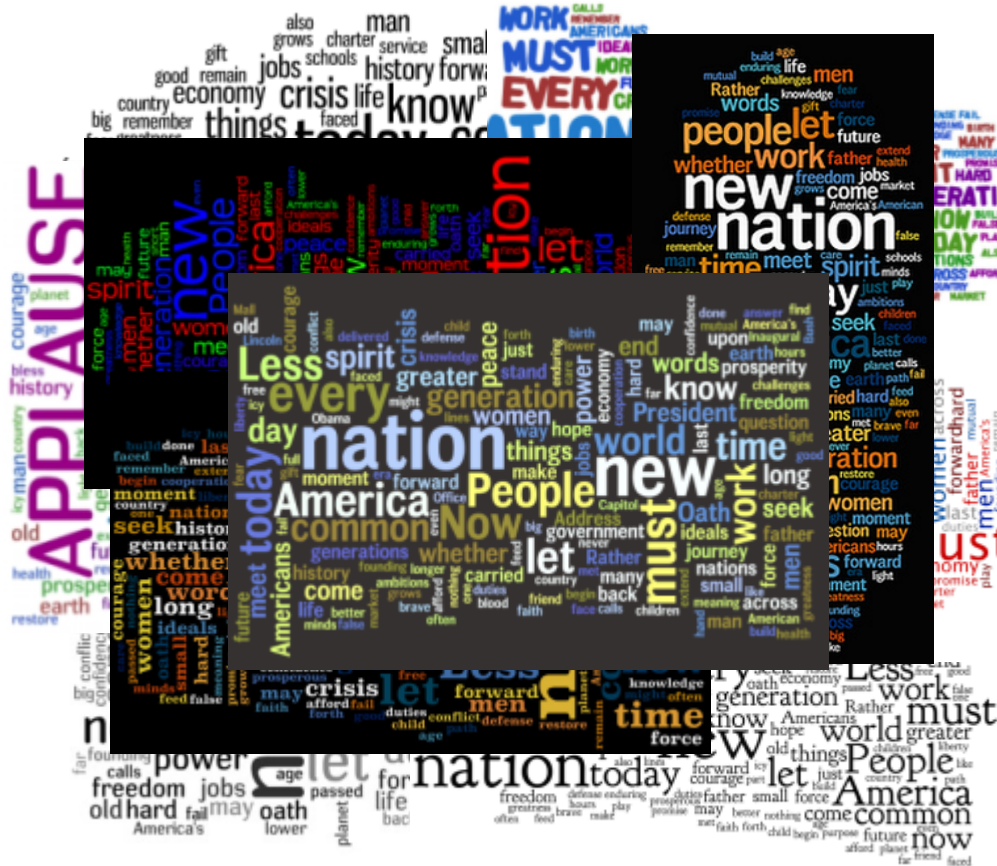
- Built from a set of **words** from a text
- Font size related to frequency
- Gained popularity with Wordle



GoogleImage(obama inaugural address wordle)

Word clouds

- Built from a set of words from a text
- Font size related to frequency
- Gained popularity with Wordle



GoogleImage(obama inaugural address wordle)

ONAFHANKELIJK DAGBLAD €1
WOENSDAG 21 JANUARI 2009 WWW.DEMORGENE

De Morgen

'The time has come'

De integrale speech van Obama >18-19
Alles over de eedaflegging >2-7

Enhanced tag / word clouds

Add information from the text:

- color intensity to express recency in Amazon
- shared tags in red on del.icio.us
- group together cooccurring tags on the same line
Hassan-Montero & Herrero-Solana, InScit'06
- optimize blank space and semantic proximity
Kaser & Lemire, WWW'07
- “topigraphy”: 2D placement according to cooccurrence
Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08

Enhanced tag / word clouds

Add information from the text:

- color intensity to express recency in Amazon
- shared tags in red on del.icio.us
- group together cooccurring tags on the same line
Hassan-Montero & Herrero-Solana, InScit'06
- optimize blank space and semantic proximity
Kaser & Lemire, WWW'07
- “topigraphy”: 2D placement according to cooccurrence
Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08

Most Popular Tags (What's this?)

Welcome to the Amazon.com tag cloud. Tags are labels customers can use to classify a product. More frequently used tags are larger and more recently used tags will appear **darker**.

1080p action adventure american history animation anime art baby best cancelled tv shows
biography blu-ray book business canon children childrens books christian christianity
christmas classic classic movie classic rock classical music comedy comics cookbook cooking
defectivebydesign digital camera disney drama dvd erotica exercise family fantasy fiction fitness fun
games gift idea graphic novel harry potter hd dvd hdtv health hip hop historical fiction historical
romance history horror humor inspirational ipod jazz kids kindle love magic manga
meditation memoir metal movie mp3 player music mystery nonfiction paranormal
romance pc game philosophy photography playstation 3 poetry politics progressive rock psychology
reference religion rock romance rpg science science fiction self-help sex soundtrack
spirituality suspense thriller toys travel tv series vampire vampire romance video
games wii women world war ii xbox 360

Enhanced tag / word clouds

Add information from the text:

- color intensity to express recency in Amazon
- shared tags in red on del.icio.us
- group together cooccurring tags on the same line
Hassan-Montero & Herrero-Solana, InScit'06
- optimize blank space and semantic proximity
Kaser & Lemire, WWW'07
- “topigraphy”: 2D placement according to cooccurrence
Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08



Enhanced tag / word clouds

Add information from the text:

- color intensity to express recency in Amazon
- shared tags in red on del.icio.us
- group together cooccurring tags on the same line
Hassan-Montero & Herrero-Solana, InScit'06
- optimize blank space and semantic proximity
Kaser & Lemire, WWW'07
- “topigraphy”: 2D placement according to cooccurrence
Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08



Enhanced tag / word clouds

Add information from the text:

- color intensity to express recency in Amazon
- shared tags in red on del.icio.us
- group together cooccurring tags on the same line
Hassan-Montero & Herrero-Solana, InScit'06
- optimize blank space and semantic proximity
Kaser & Lemire, WWW'07
- “topography”: 2D placement according to cooccurrence
Fujimura, Fujimura, Matsubayashi, Yamada & Okuda, WWW'08



Extract semantic information from a text

- literature analysis:

philological approach: only consider the text

Brody

- discourse analysis:

tree analysis or cooccurrence graph, geodesic projection

Brunet (Hyperbase), Viprey (Astartex)

- text mining:

semantic graph

Grimmer (Wordmapper)

- natural language processing:

sense desambiguation

Véronis (Hyperlex)

Extract semantic information from a text

- literature analysis:
philological approach: only consider the text

Brody

- discourse analysis:
tree analysis or cooccurrence graph, geodesic projection

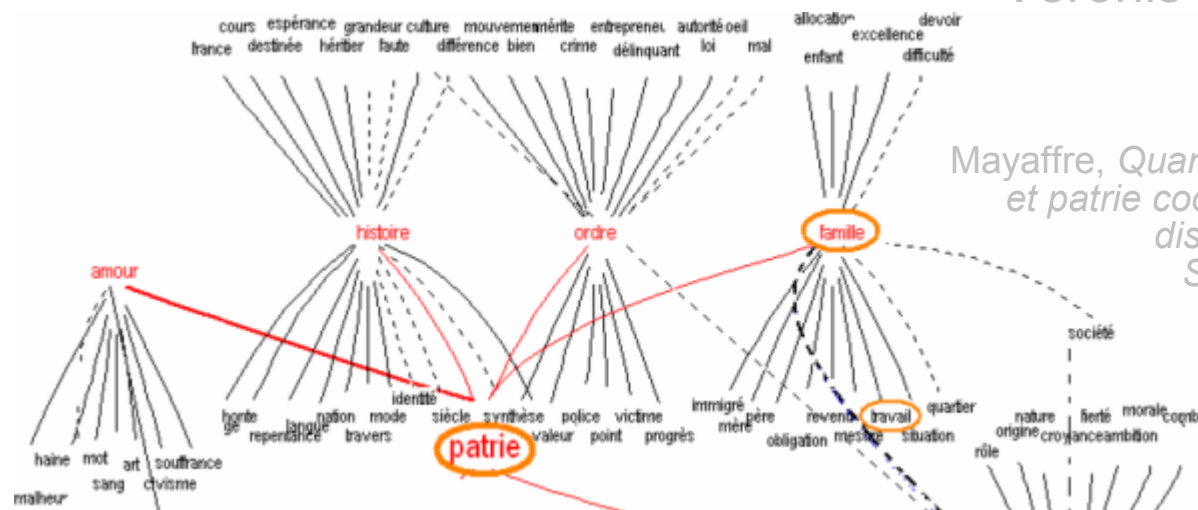
Brunet (Hyperbase), Viprey (Astartex)

- text mining:
semantic graph

Grimmer (Wordmapper)

- natural language processing:
sense desambiguation

Véronis (Hyperlex)



Mayaffre, *Quand travail, famille, et patrie cooccurrent dans le discours de Nicolas Sarkozy*, JADT'08

Extract semantic information from a text

- literature analysis:

philological approach: only consider the text

Brody

- discourse analysis:

tree analysis or cooccurrence graph, geodesic projection

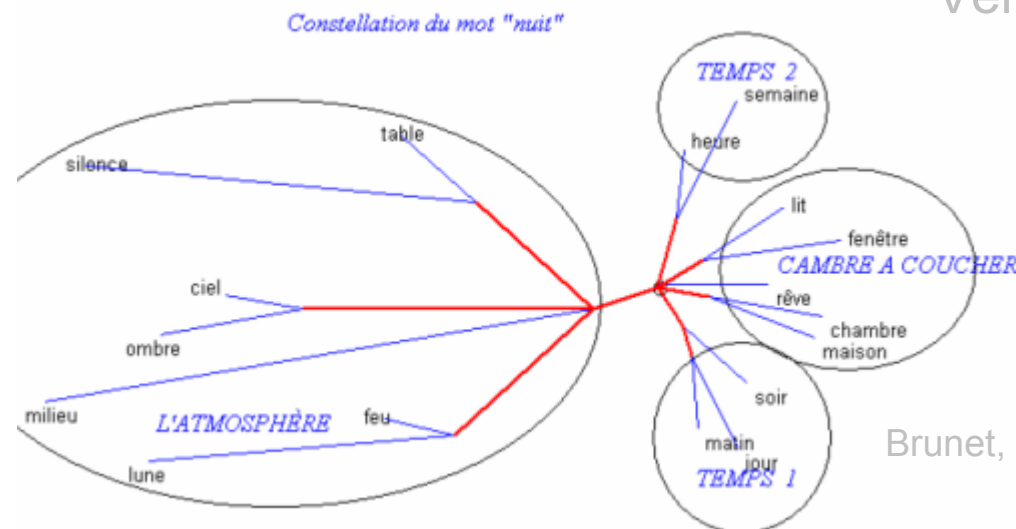
Brunet (Hyperbase), Viprey (Astartex)

- text mining:
semantic graph

Grimmer (Wordmapper)

- natural language processing:
sense desambiguation

Véronis (Hyperlex)



Brunet, *Les séquences (suite)*,
JADT'08

Extract semantic information from a text

- literature analysis:
philological approach: only consider the text

Brody

- discourse analysis:
tree analysis or cooccurrence graph, geodesic projection

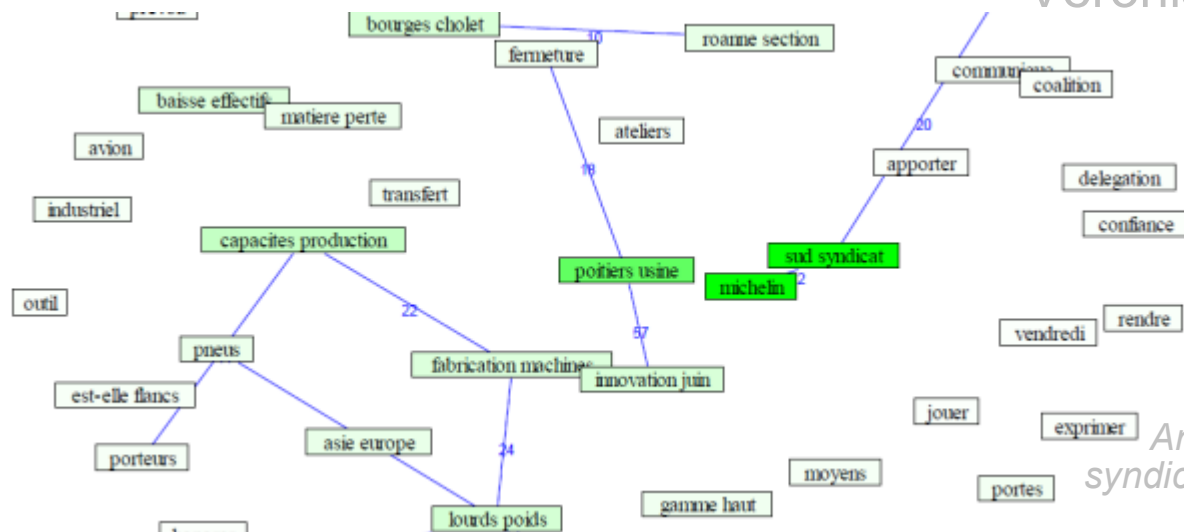
Brunet (Hyperbase), Viprey (Astartex)

- text mining:
semantic graph

Grimmer (Wordmapper)

- natural language processing:
word sense disambiguation

Véronis (Hyperlex)



Peyrat-Guillard,
*Analyse du discours
syndical sur l'entreprise*,
JADT'08

Extract semantic information from a text

- literature analysis:
philological approach: only consider the text

Brody

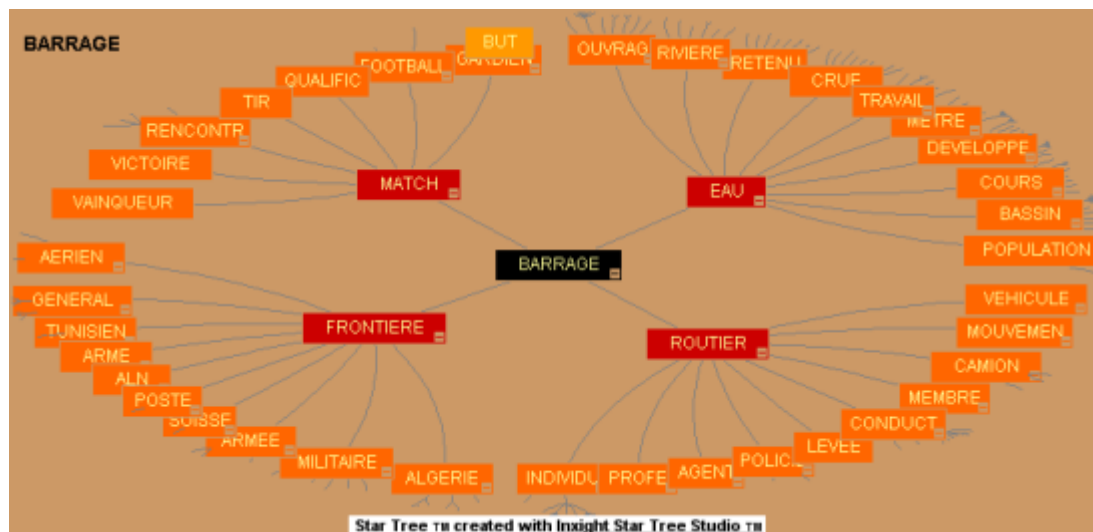
- discourse analysis:
tree analysis or cooccurrence graph, geodesic projection
Brunet (Hyperbase), Viprey (Astartex)

- text mining:
semantic graph

Grimmer (Wordmapper)

- natural language processing:
word sense disambiguation

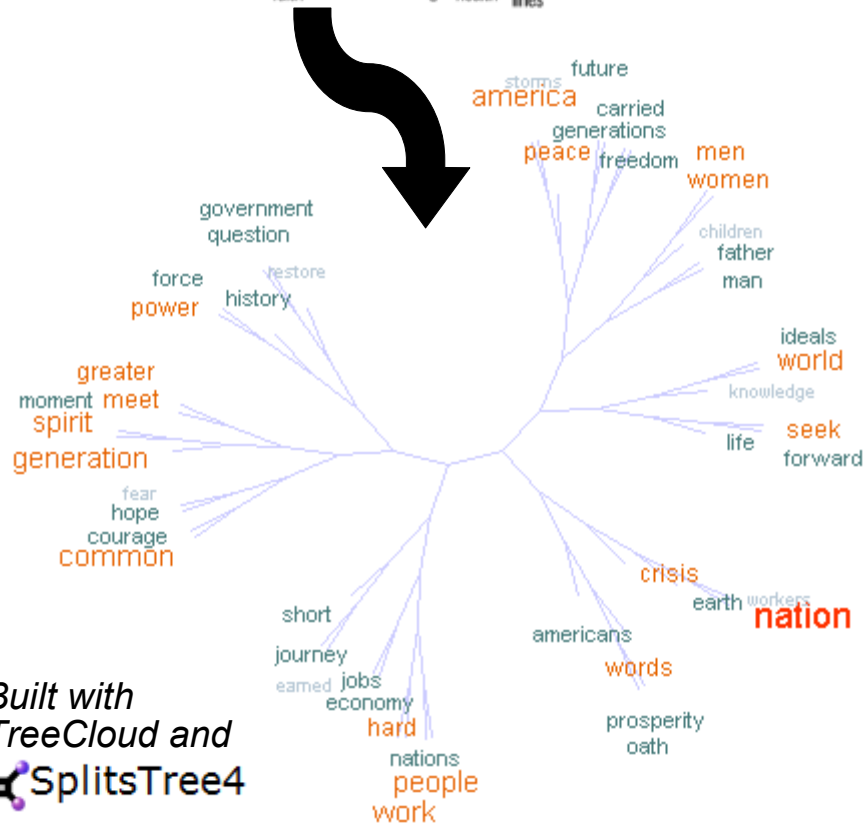
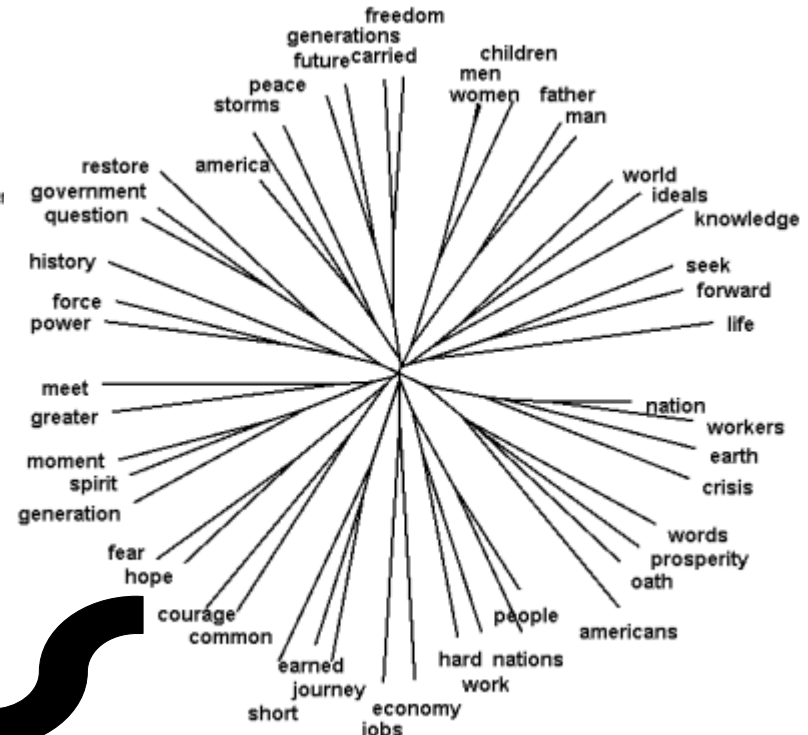
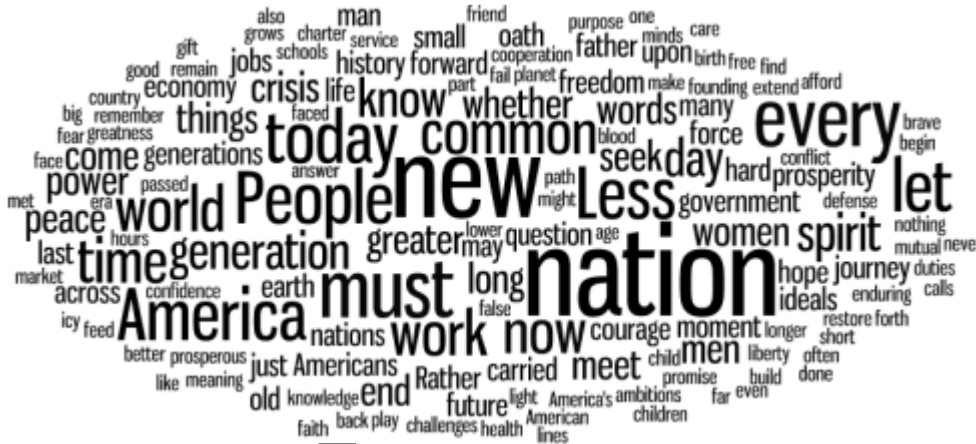
Véronis (Hyperlex)



Disambiguation of word
“barrage”: dam, play-off,
roadblock, police cordon.

Véronis, *HyperLex:
Lexical Cartography for
Information Retrieval*, 2004

Tag cloud + tree = tree cloud

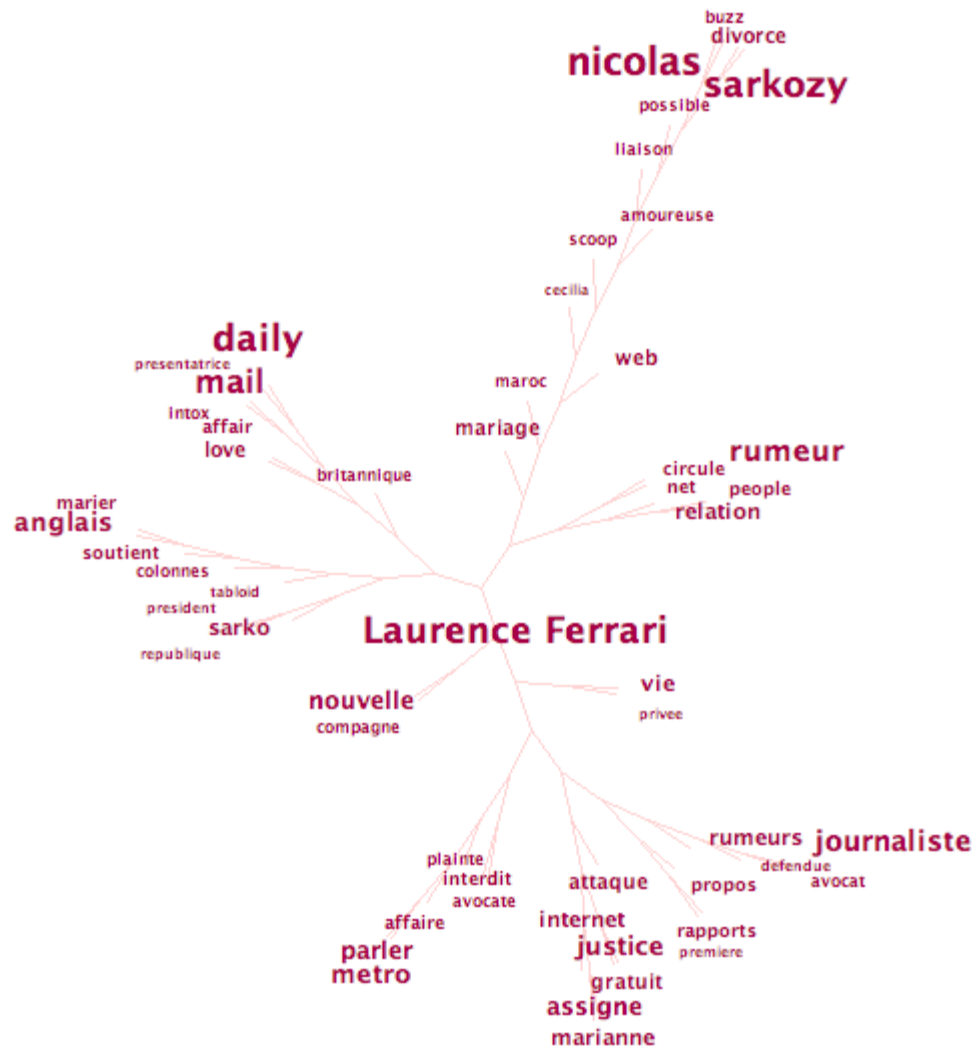


SplitsTree: Huson 1998,
Huson & Bryant 2006

GPL-licensed Treecloud in Python,
available at <http://www.treecloud.org>

Built with
TreeCloud and
SplitsTree4

The first tree cloud



Tree cloud of the blog posts containing "Laurence Ferrari"
from 25/11/2007 to 10/12/2007, by Jean Véronis

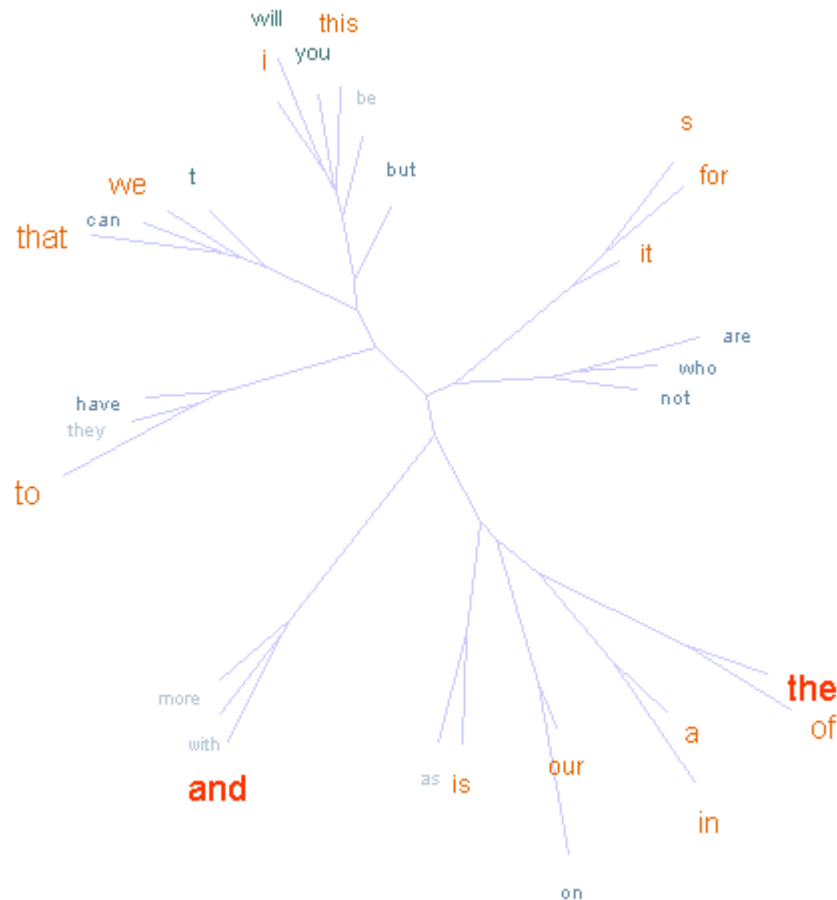
<http://aixtal.blogspot.com/2007/12/actu-une-ferrari-dans-un-arbre.html>

Building a tree cloud – extracting the words

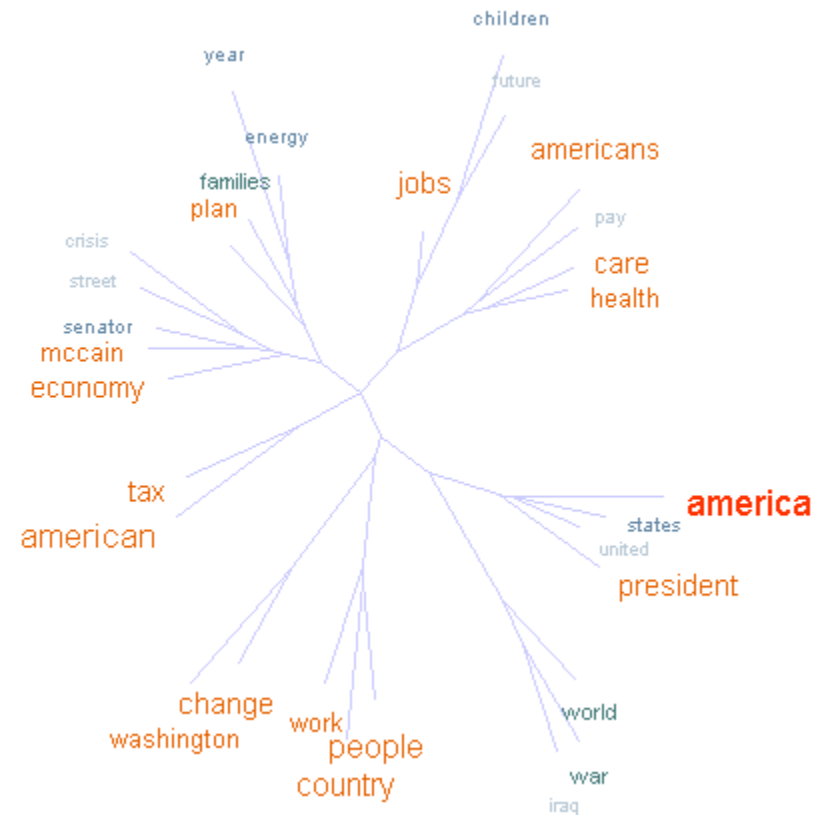
Extract words with frequency:

- stoplist?

without stoplist



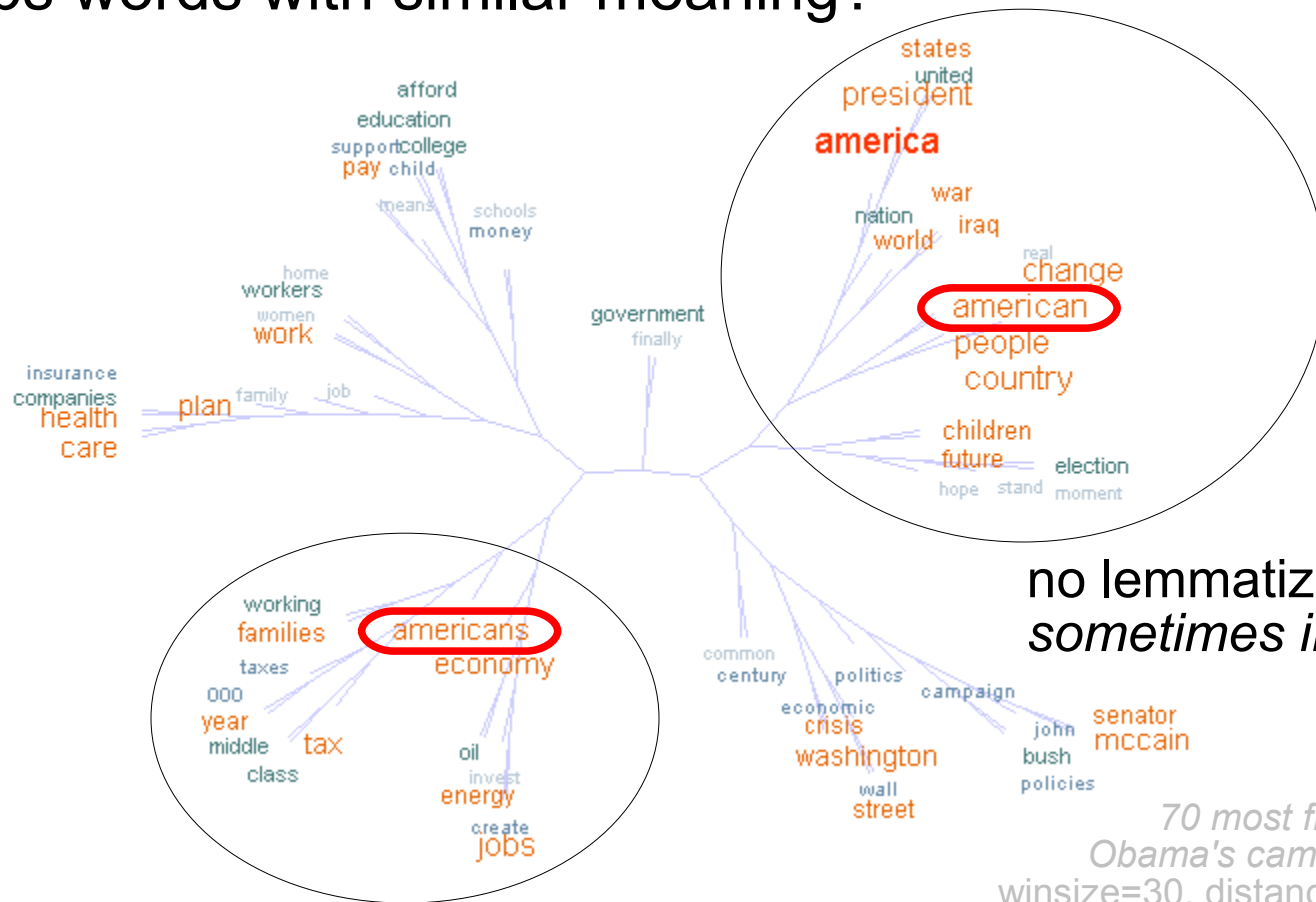
with stoplist



Building a tree cloud – extracting the words

Extract words with frequency:

- lemmatization?
- groups words with similar meaning?

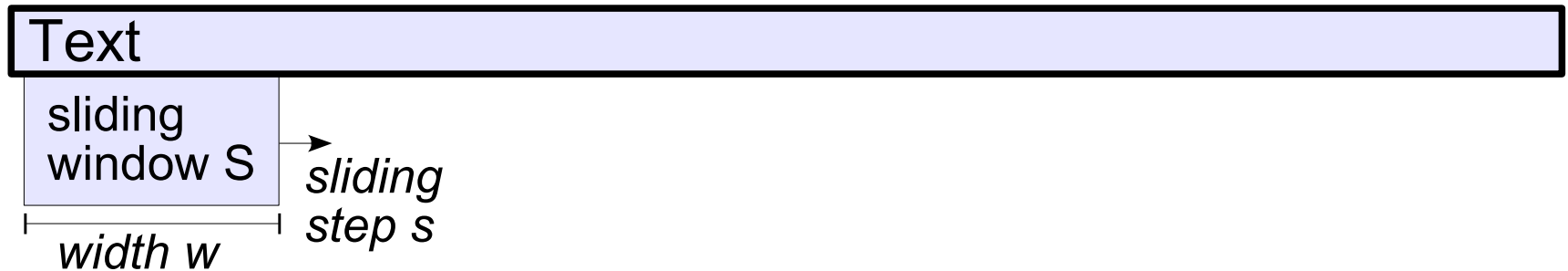


Building a tree cloud – dissimilarity matrix

Many semantic distance formulas based on cooccurrence

Building a tree cloud – dissimilarity matrix

Many semantic distance formulas based on cooccurrence



cooccurrence matrices

O_{11} , O_{12} , O_{21} , O_{22}

| | $v \in S$ | $v \notin S$ |
|--------------|-----------|--------------|
| $u \in S$ | O_{11} | O_{12} |
| $u \notin S$ | O_{21} | O_{22} |



semantic dissimilarity matrix

chi squared, mutual information, liddel, dice, jaccard, gmean, hyperlex, minimum sensitivity, odds ratio, zscore, log likelihood, poisson-stirling...

Building a tree cloud – dissimilarity matrix

Transformations needed on the dissimilarity:

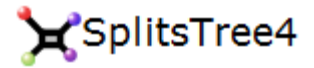
- transform similarity into dissimilarity
- linear normalization for positive matrices to get distances in $[0, 1]$
- affine normalization for matrices with positive or negative numbers, to get distances in $[\alpha, 1]$ (for example $\alpha=0.1$)

Building a tree cloud – tree reconstruction

Many existing methods:

- Neighbor-Joining

Saitou & Nei, 1987



- Addtree variants

Barthelemy & Luong, 1987

- Quartet heuristic

Cilibrasi & Vitanyi, 2007

Building a tree cloud – tree decoration

Choice of word sizes:

- computed directly from **frequency** (apply a log!)

or

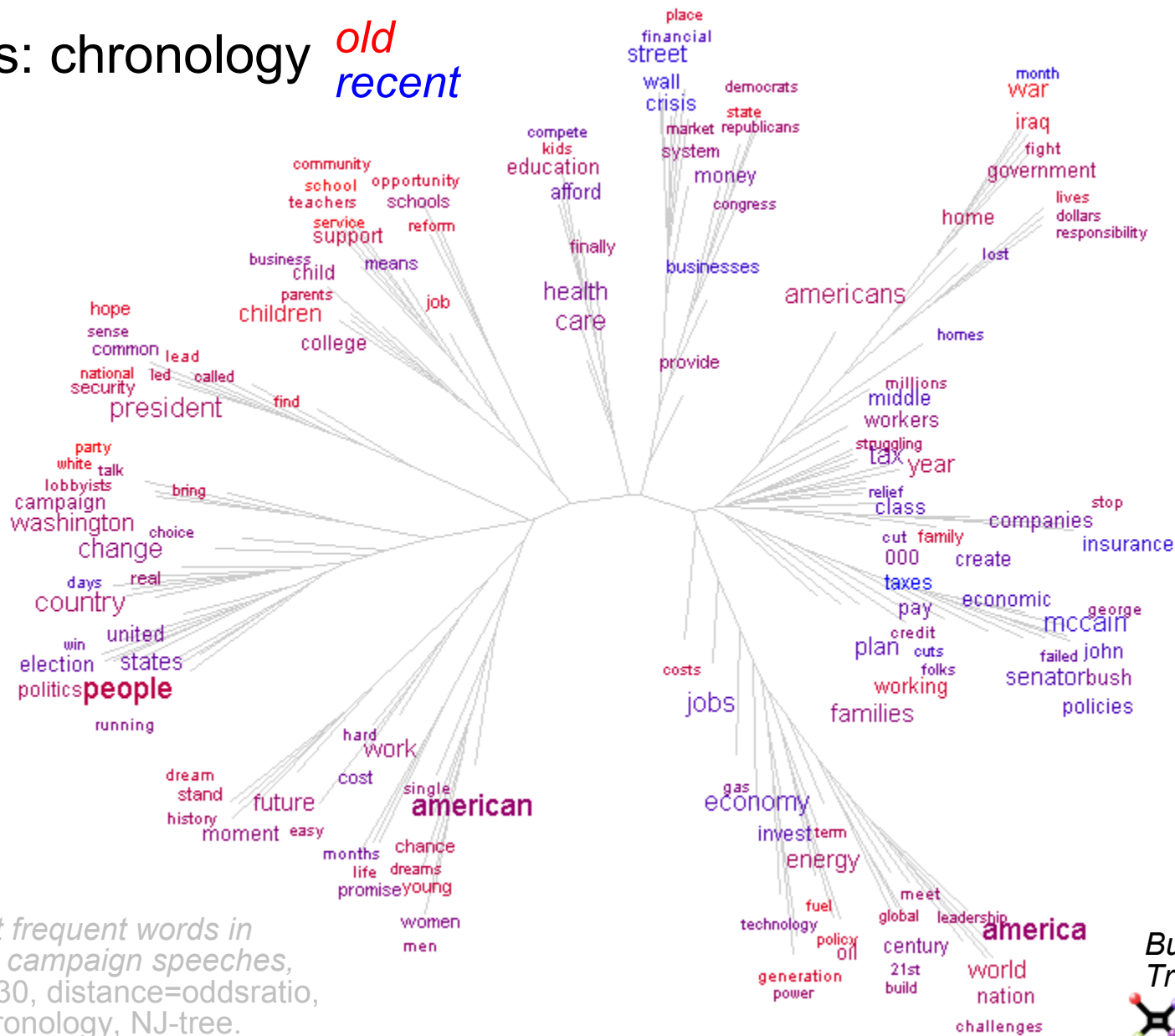
- computed from **frequency ranking** (exponential distribution)

or

- **statistical significance** with respect to a reference corpus

Building a tree cloud – tree decoration

Colors: chronology *old*
recent

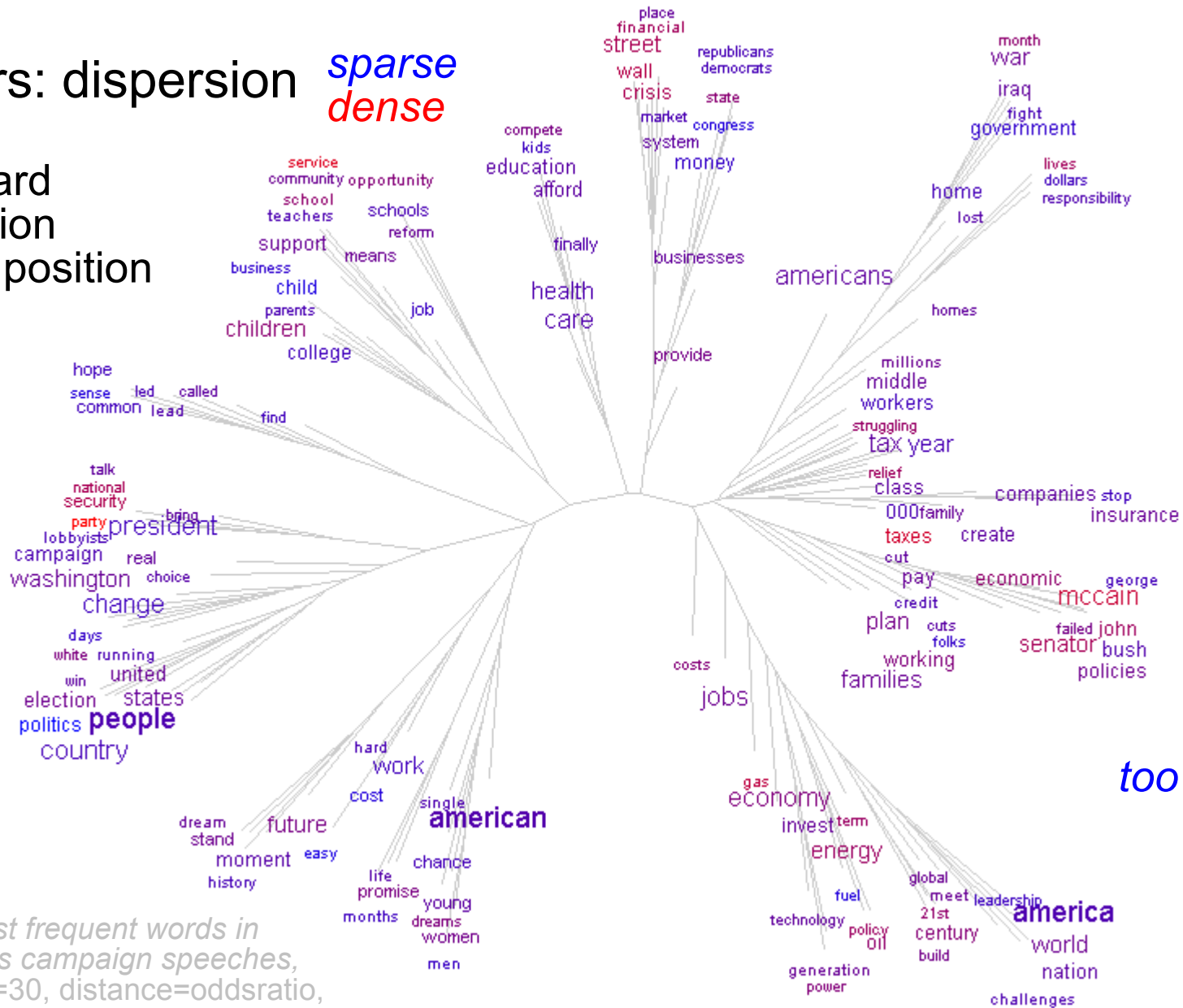


150 most frequent words in Obama's campaign speeches, winsize=30, distance=oddsratio, color=chronology, NJ-tree.

Building a tree cloud – tree decoration

Colors: dispersion *sparse*
dense

standard
deviation
of the position



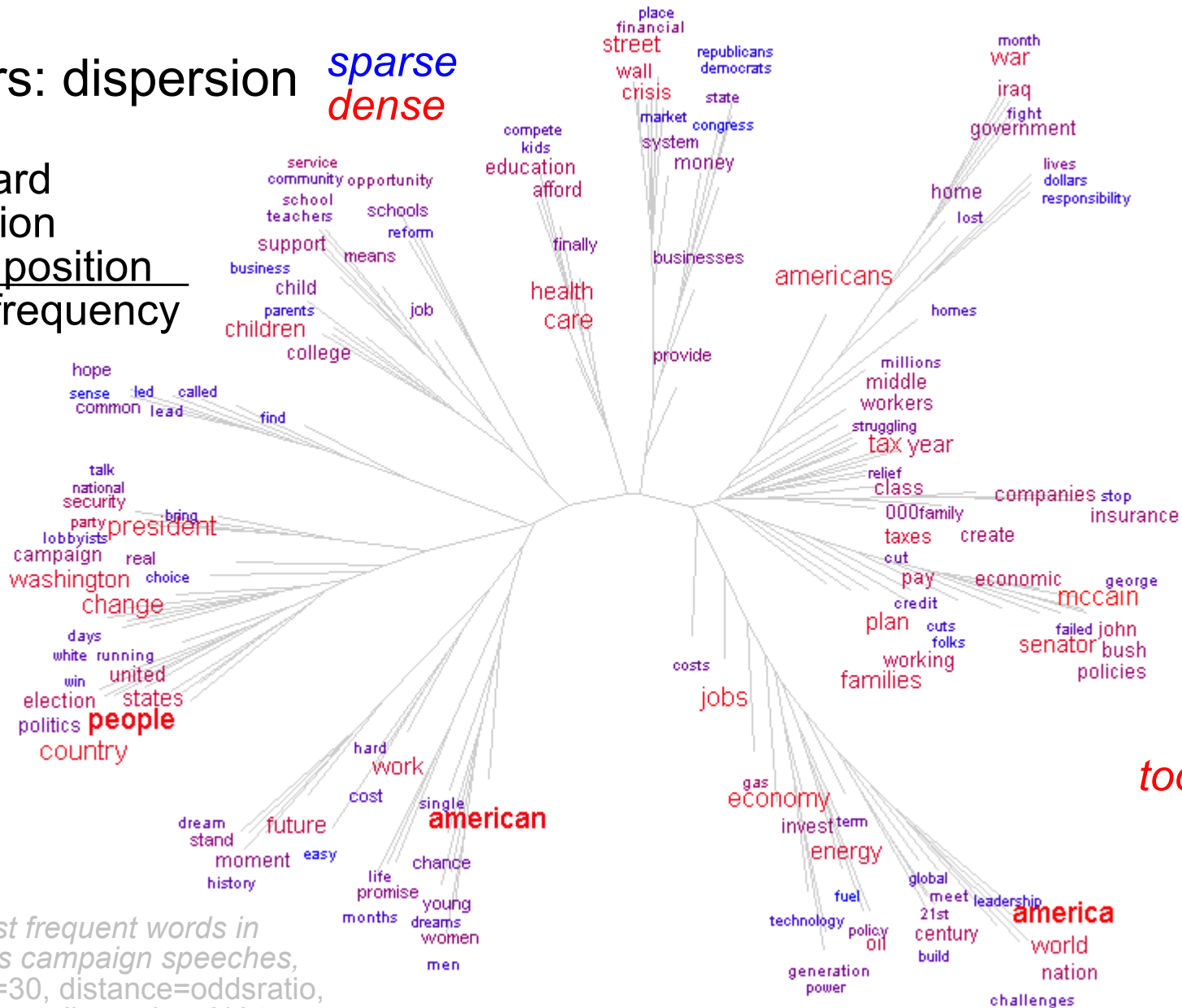
150 most frequent words in
Obama's campaign speeches,
winsize=30, distance=oddsratio,
color=dispersion, NJ-tree.

Building a tree cloud – tree decoration

Colors: dispersion

sparse
dense

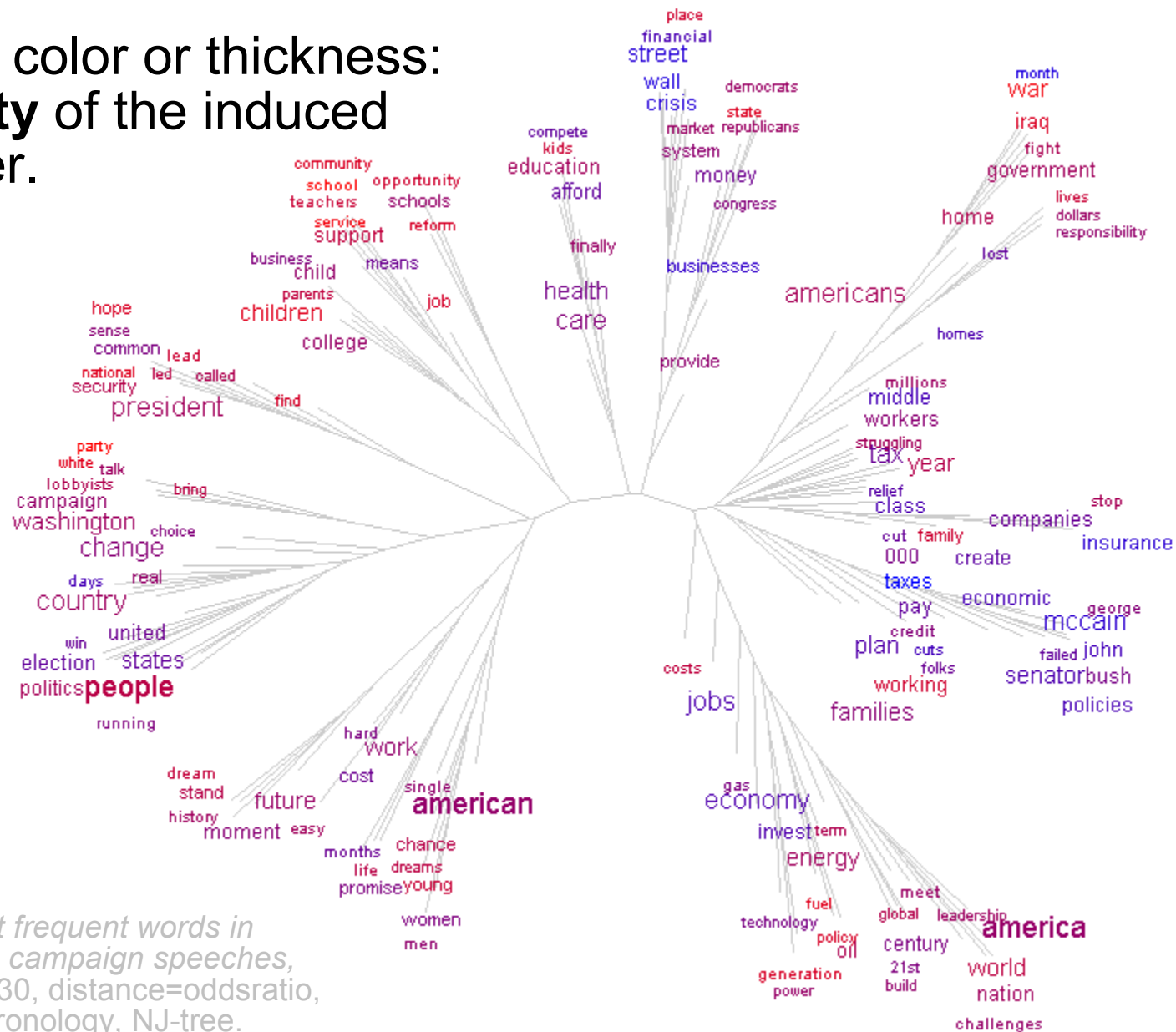
standard
deviation
of the position
word frequency



150 most frequent words in
Obama's campaign speeches,
winsize=30, distance=oddsratio,
color=norm-dispersion, NJ-tree.

Building a tree cloud – tree decoration

Edge color or thickness:
quality of the induced
cluster.



Quality control

Is there an objective quality measure of tree clouds?

Quality control

Is there an objective quality measure of tree clouds?

What is the best method to build a tree cloud from my data?

Quality control

Is there an objective quality measure of tree clouds?

What is the best method to build a tree cloud from my data?

Tree cloud variations if small changes?

➔ **bootstrap** to evaluate:

- **stability of the result**
- **robustness of the method**

Quality control

Is there an objective quality measure of tree clouds?

What is the best method to build a tree cloud from my data?

Tree cloud variations if small changes?

➡ **bootstrap** to evaluate:

- **stability of the result**
- **robustness of the method**

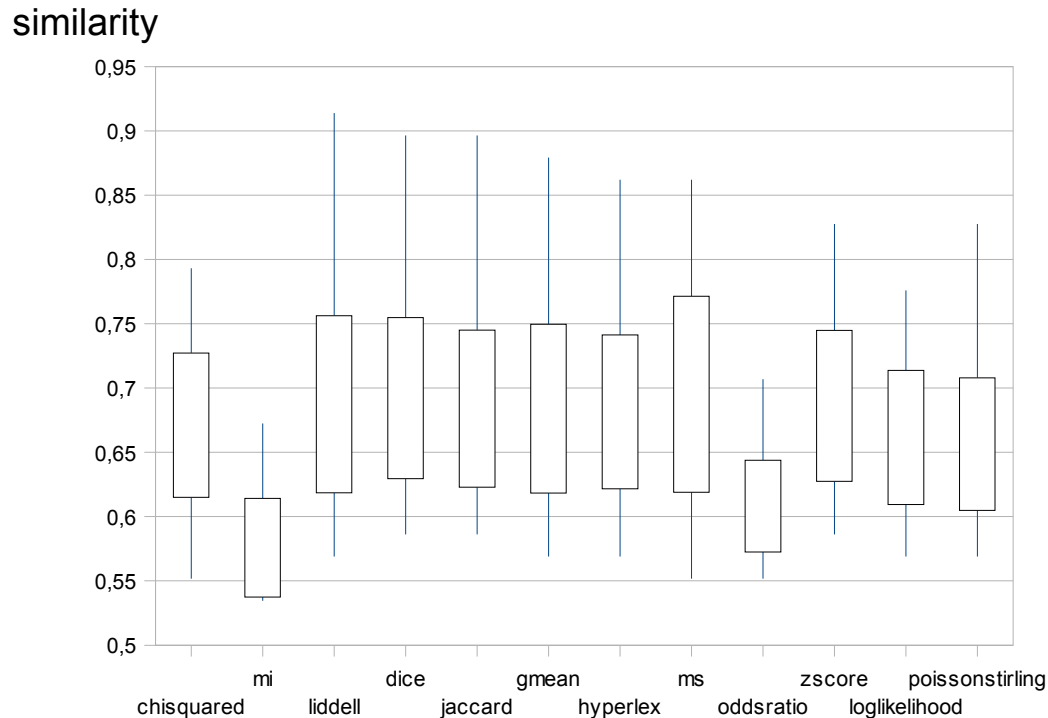
Still, is there a more direct method?

➡ **arboricity** to show whether the distance matrix **fits with a tree**, which should **imply stability**?

Guénoche & Garreta, 2001
Guénoche & Darlu, 2009

Quality control – bootstrap

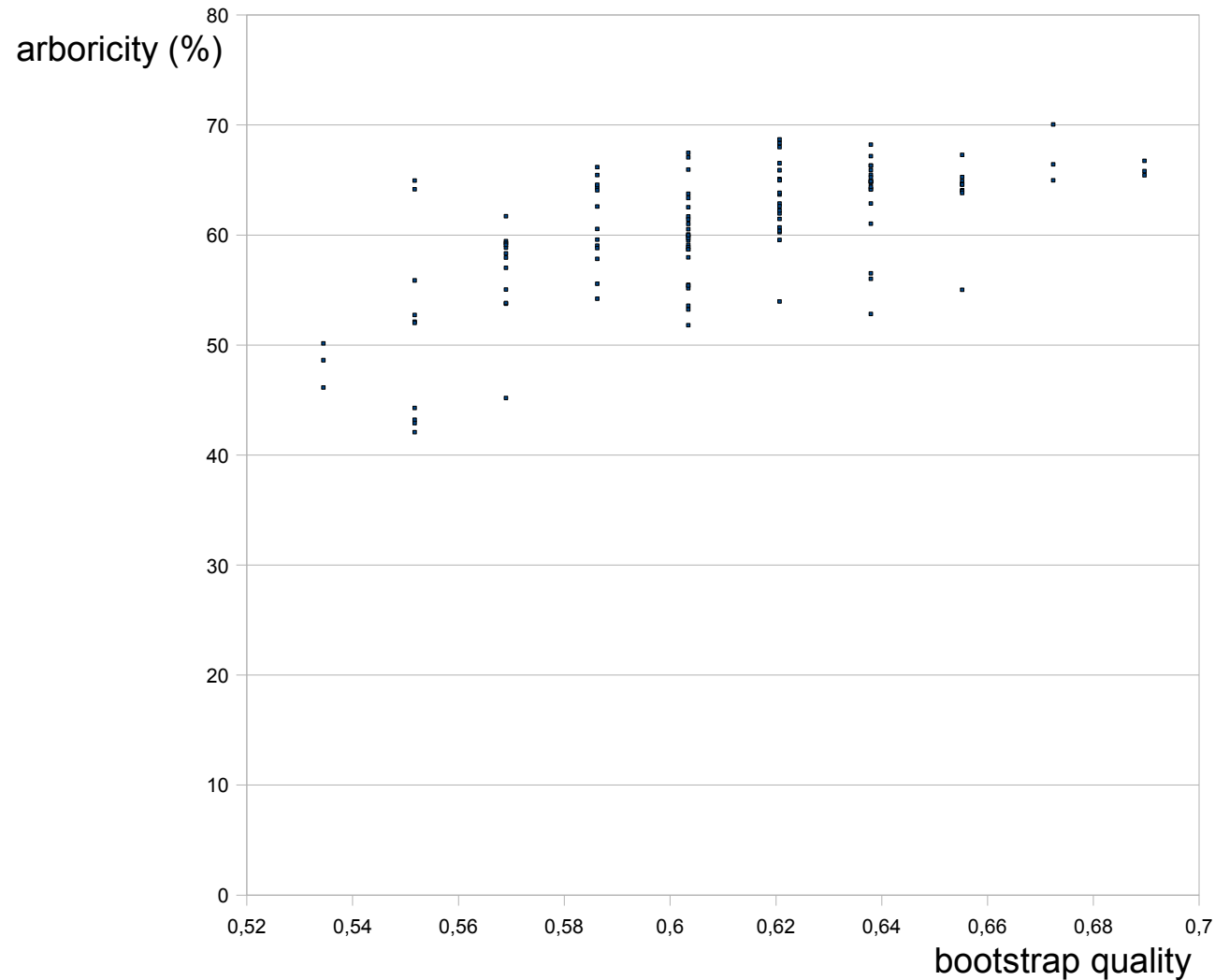
- Randomly delete words with probability 50%.
- Built tree cloud of original text, and altered text.
- Compute similarity of both trees (1-normalized RobinsonFoulds)



4 altered versions of 10
Obama's speeches,
3000 words in average,
width=30, NJ-tree.

Quality control – arboricity

Relationship between “bootstrap quality” and arboricity:

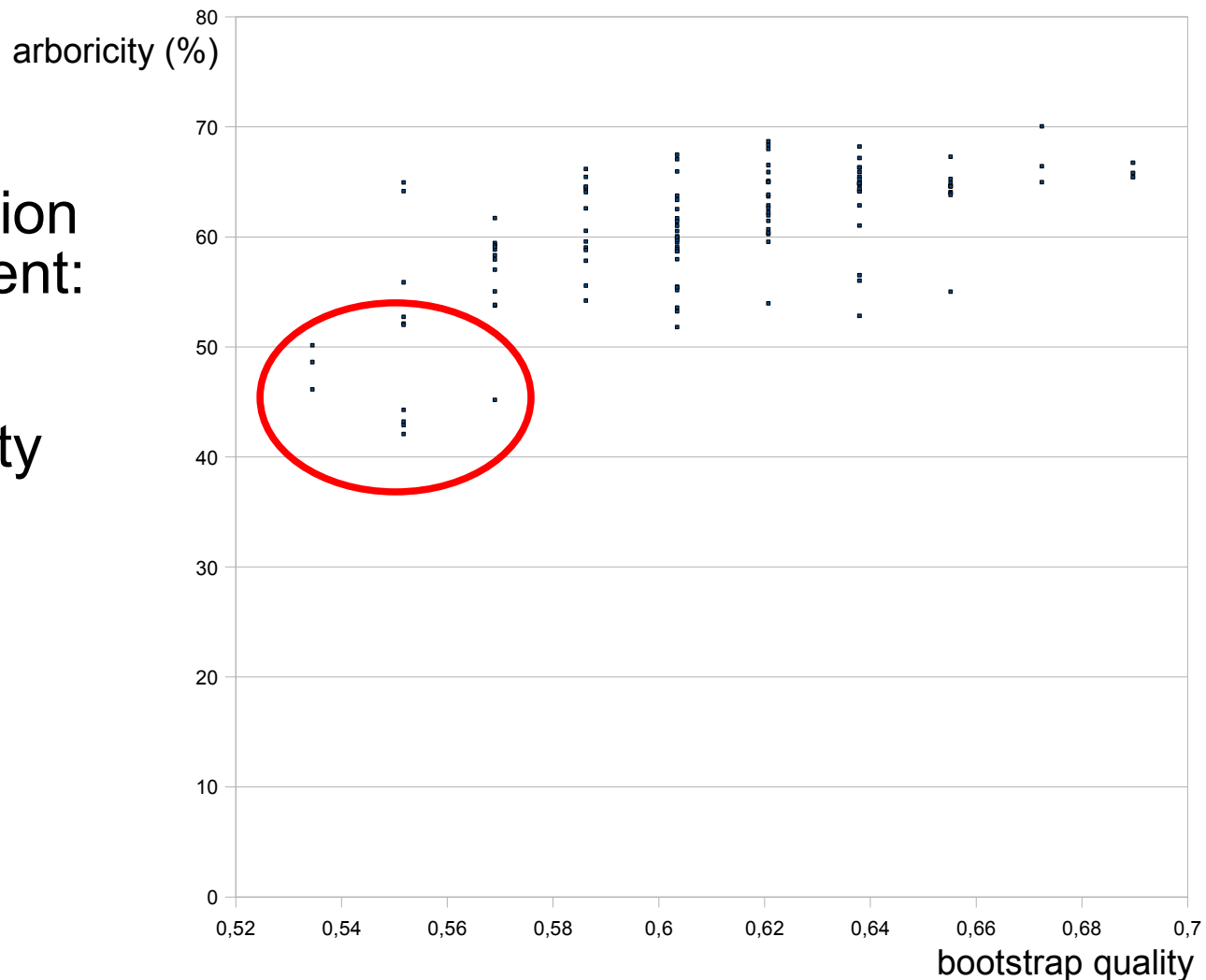


Quality control – arboricity

Relationship between “bootstrap quality” and arboricity:

correlation
coefficient:
0.64

arboricity
below
50%:
danger!



Perspectives

- Make the tool available on a web interface <http://www.treecloud.org>
- Evaluate tree clouds for discourse analysis
- Build the daily tree cloud of people popular on blogs,

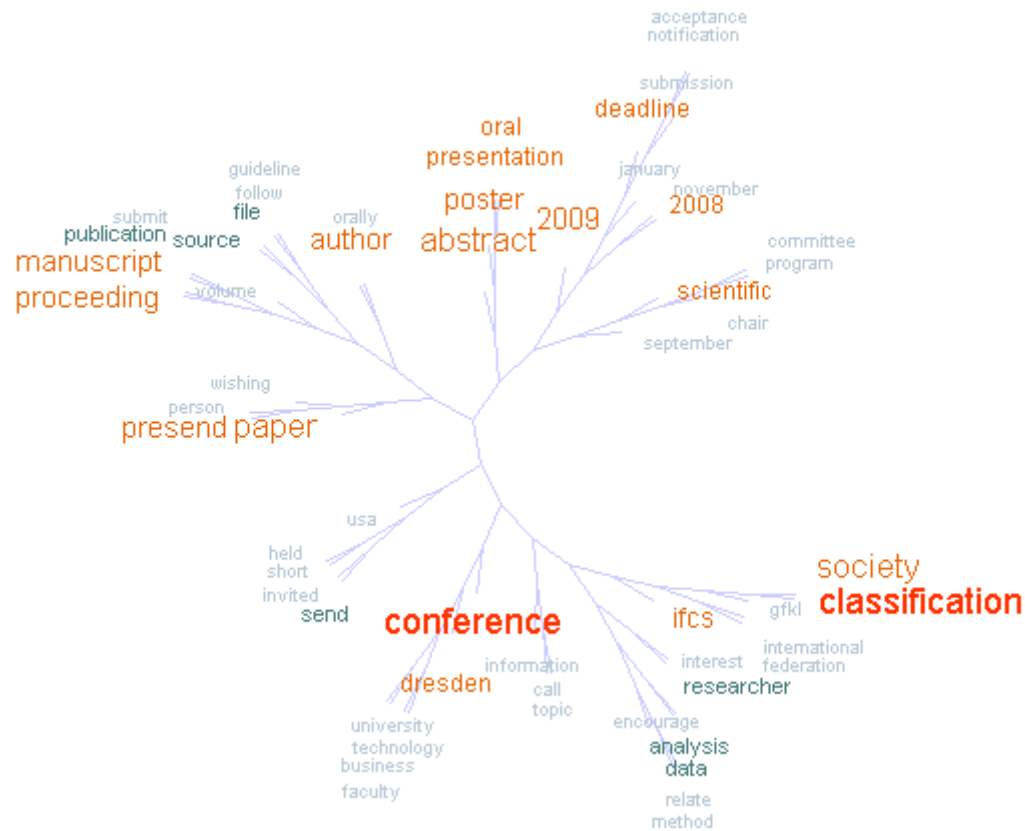
with  WIKIO Labs

 **Allemagne**

Albert Streit Arnd Peiffer Ben Bernanke Benjamin
Netanjahu Christopher Flowers Claudio Pizarro Filippo
Inzaghi Gerhard Tremmel Henning Grieneisen Israel Beitenu
Jenny Wolf Joachim Löw **Josef Fritzl** Markus
Rosenberg Martin Jol Mauricio Funes Michael Essien
Mladen Petric Piotr Trochowski Sebastian
Boenisch Sebastien Grainger Simon Schempp Susanne Klatten
Thomas Cichon Thomas Reichenberger Tim Wiese Torsten
Albig Tyra Banks Uli Hoeness Wolfgang Reitzle

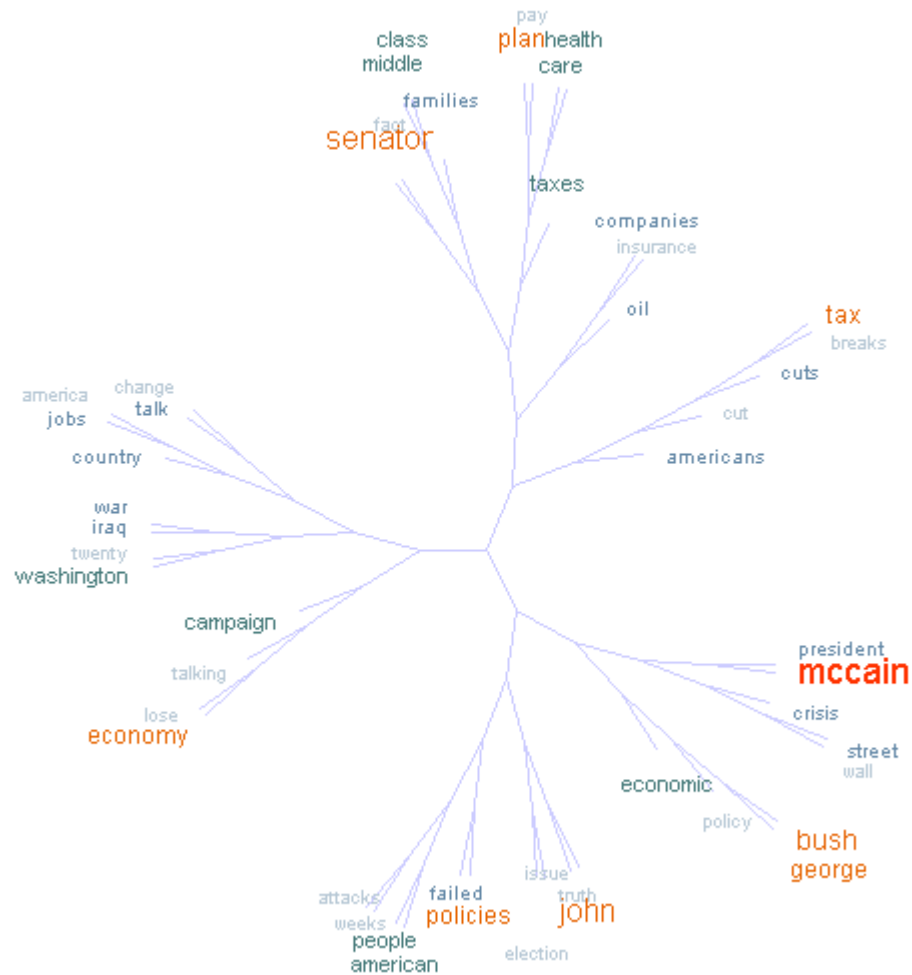
<http://labs.wikio.net>

Thank you for your attention!



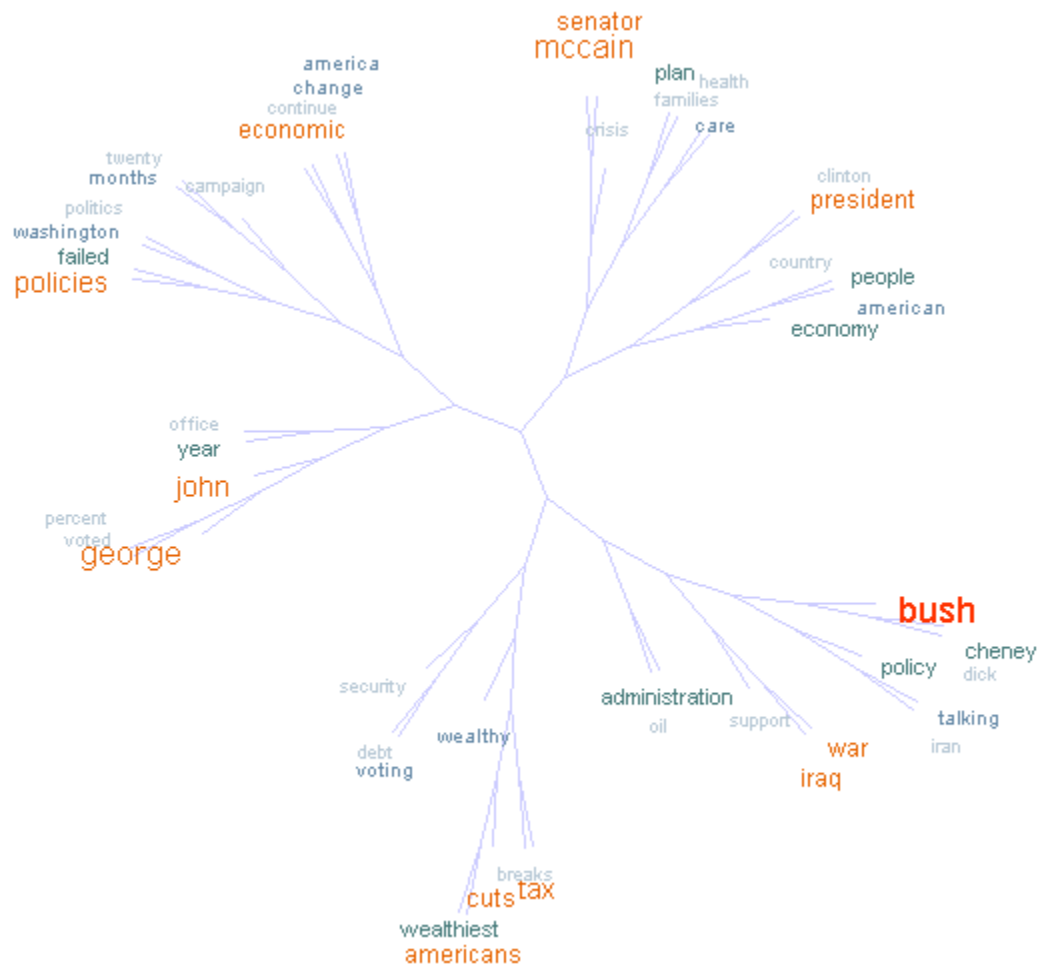
Tree cloud of the words appearing twice or more in the IFCS 2009 call for paper lemmatization, width=20, distance=dice, NJ-tree.

Tree clouds focused on one word



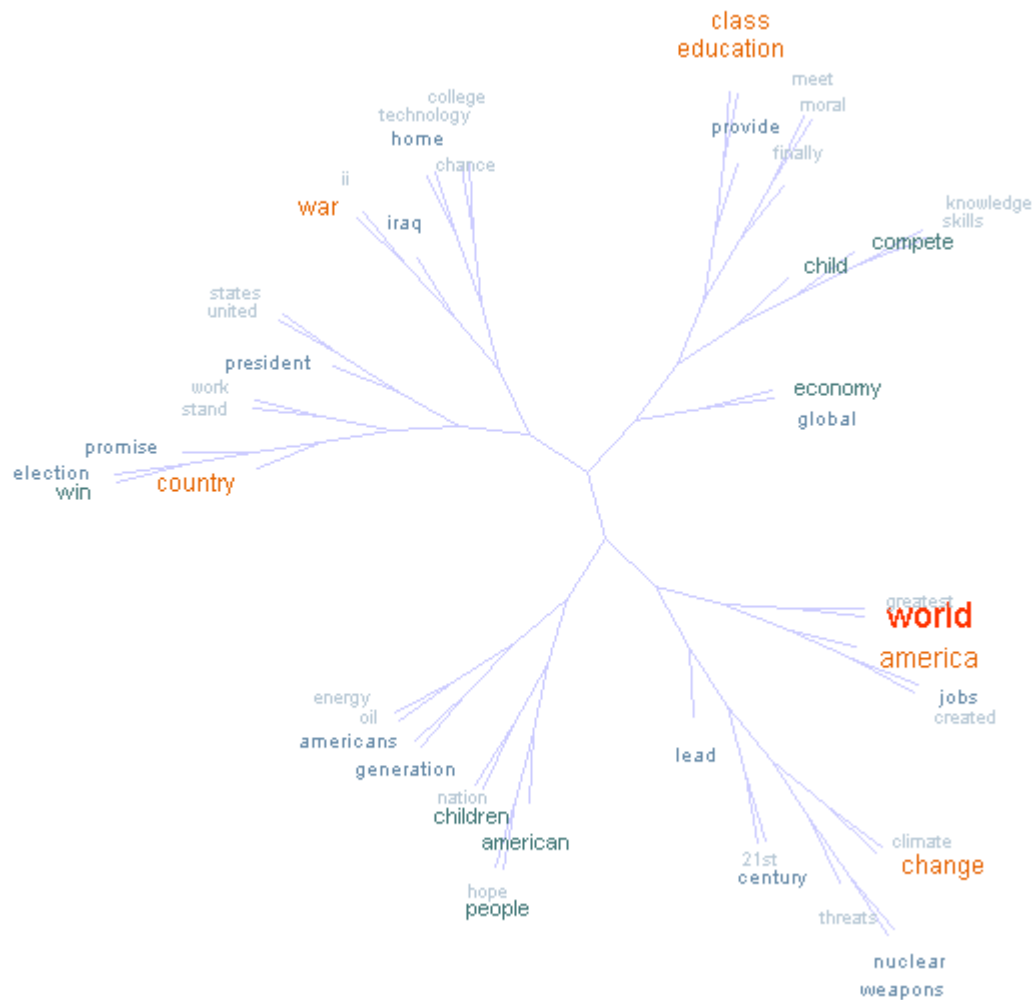
Tree cloud of the neighborhood of "McCain" in Obama's campaign speeches

Tree clouds focused on one word



Tree cloud of the neighborhood of "Bush" in Obama's campaign speeches

Tree clouds focused on one word



Tree cloud of the neighborhood of "world" in Obama's campaign speeches

<http://www.treecloud.org> - <http://www.splitstree.org>