

09

Information Visualization

Daniel Gonçalves, Sandra Gama

Visualizing Text



Get Ready...

<http://etc.ch/7rGF>



01

BASICS

Text is Tough

not pre-attentive

abstract concepts

Not just words, but relationships between words!

represents similar concepts in many different ways

space ship, flying saucer, UFO, figment of imagination

very high dimensionality

every word can be a dimension...

High dimensionality means...

... most text vis do not represent the text itself

Term statistics

Clusters

Repetitions

Patterns

...

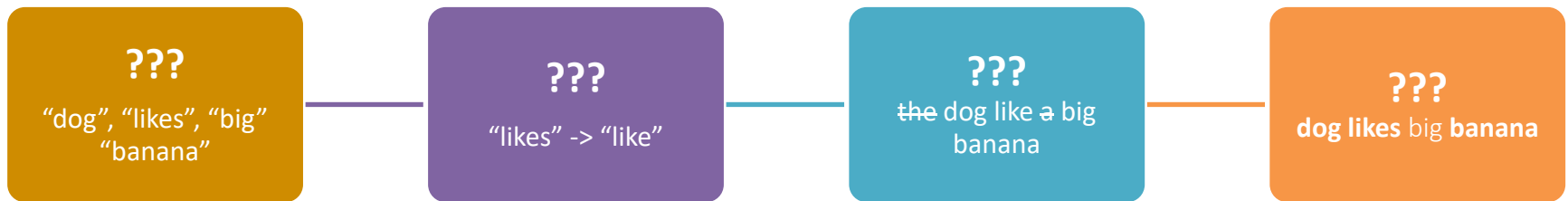
Do we still recognize the text in that?

Use the vis as an index!

Are they the right measures/models?

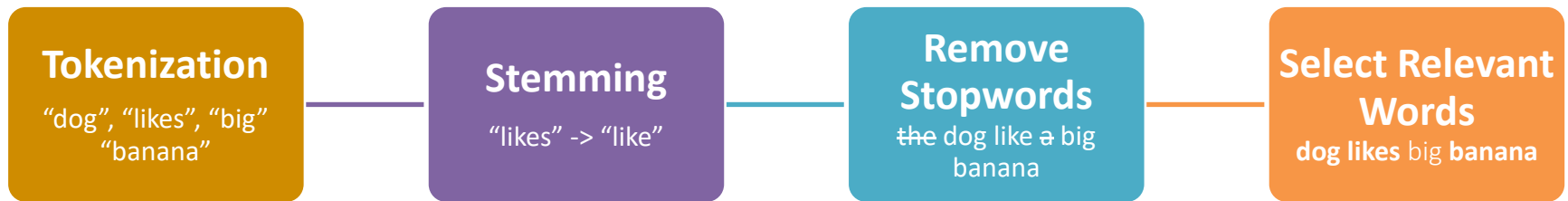
Text Preprocessing

“The dog likes a big banana”

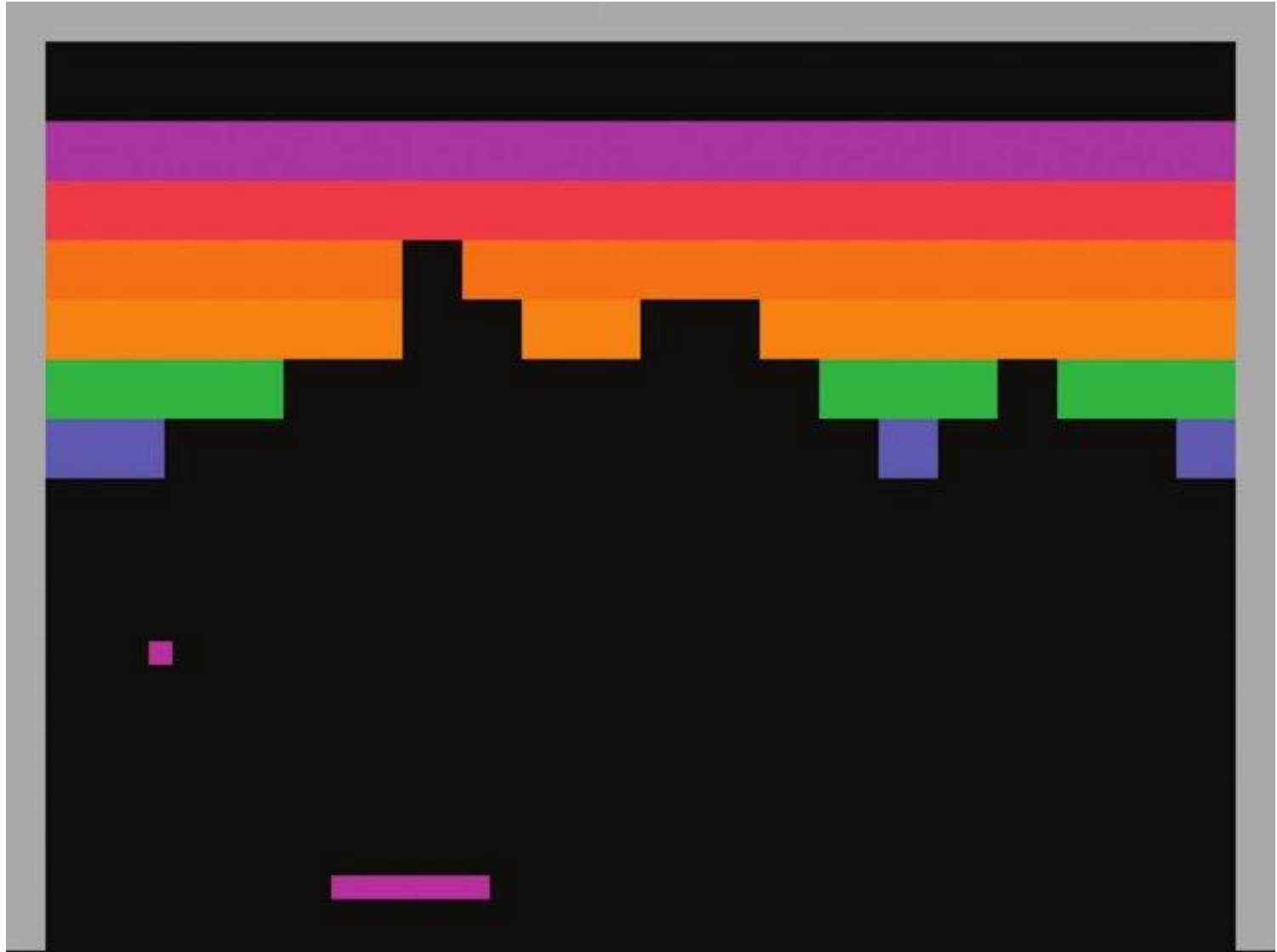


Text Preprocessing

“The dog likes a big banana”



Breakout!



Let's play some more!

Take the following text:

I must not fear.
Fear is the mind-killer.
Fear is the little-death that brings total obliteration.
I will face my fear.
I will permit it to pass over me and through me.
And when it has gone past I will turn the inner eye to see its path.
Where the fear has gone there will be nothing. Only I will remain.



<https://bit.ly/3CsK5m5>

**Run it through the four previous stages
(tokenization, stemming, removing
stopwords, term count)**

Fill the google form

Let's play some more!

Take the following text:

I must not fear.
Fear is the mind-killer.
Fear is the little-death that brings total obliteration.
I will face my fear.
I will permit it to pass over me and through me.
And when it has gone past I will turn the inner eye to see its path.
Where the fear has gone there will be nothing. Only I will remain.



<https://bit.ly/3CsK5m5>

**Run it through the four previous stages
(tokenization, stemming, removing
stopwords, term count)**

Fill the google form

05:00

Breakout Over!



Frequent == Relevant?



Finding most relevant words

Term frequency

$tf_{td} = \text{count}(t) \text{ in } d$

Normalize to show proportion: $tf_{td} / \sum_t tf_{td}$

TF.IDF

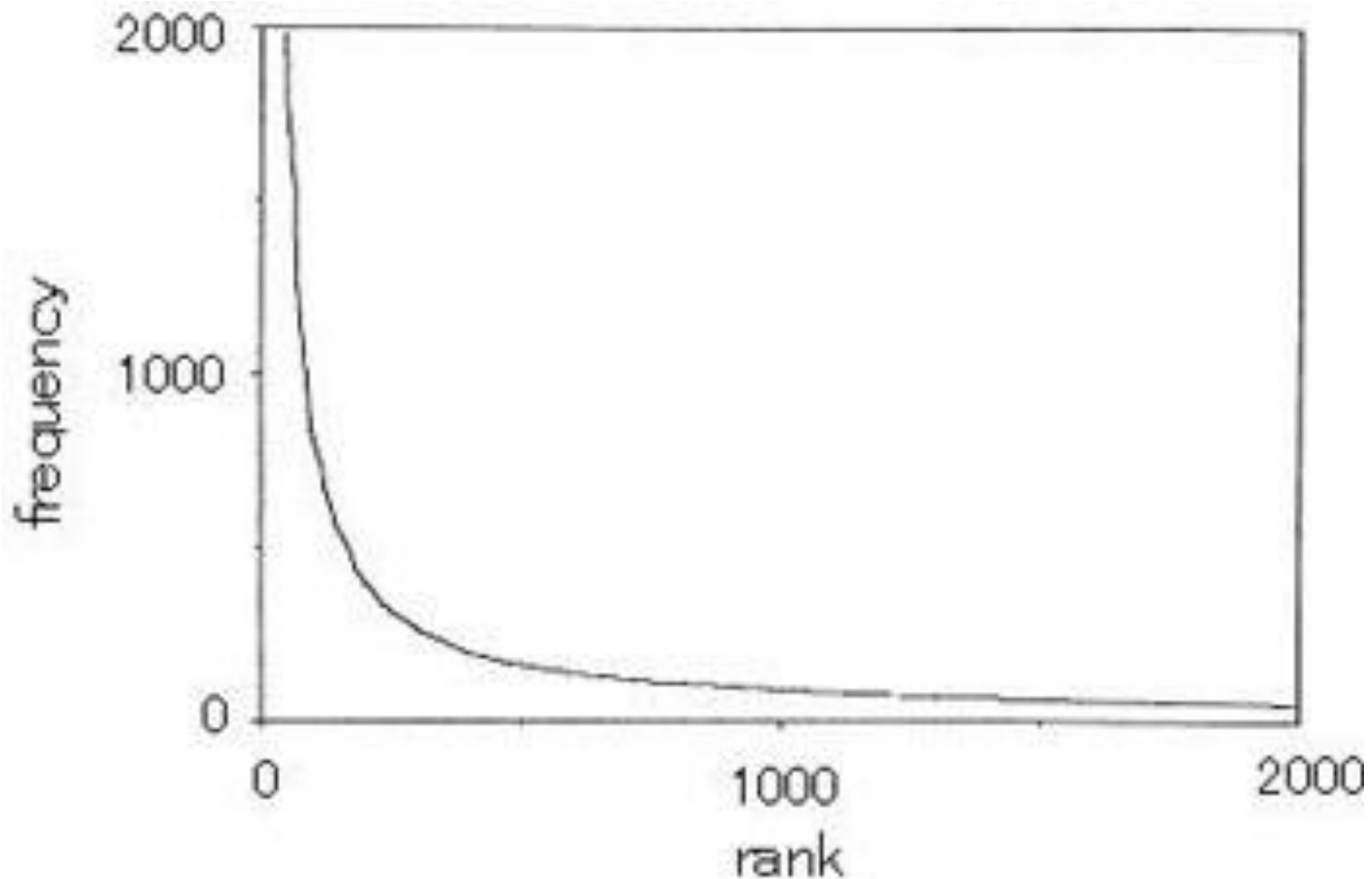
$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$

$df_t = \text{num docs containing } t$

$N = \text{num of docs}$

Zipf's Law

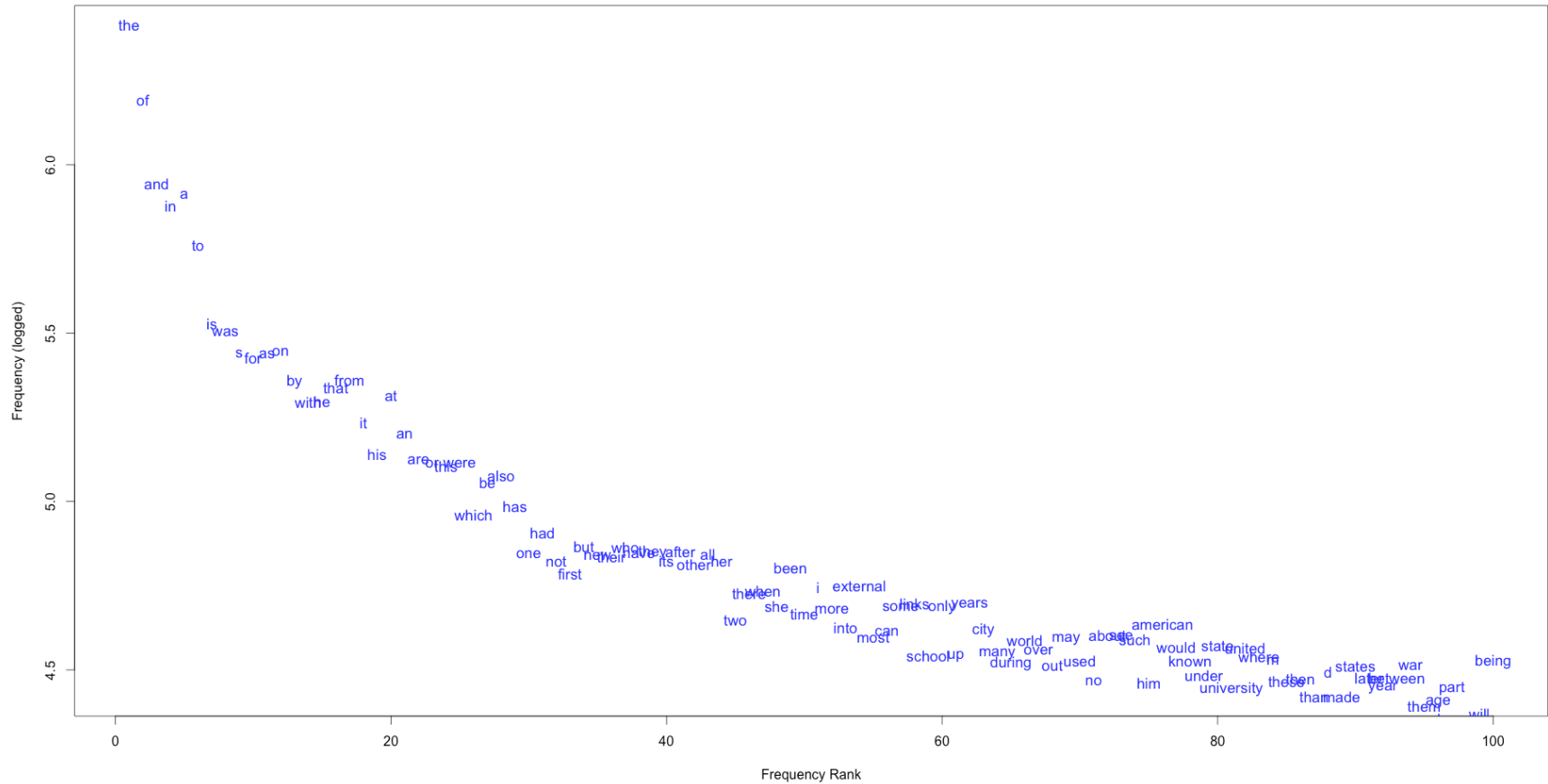
20% of words appear 80% of the time



Zipf's Law

<http://wugology.com/zipfs-law/>

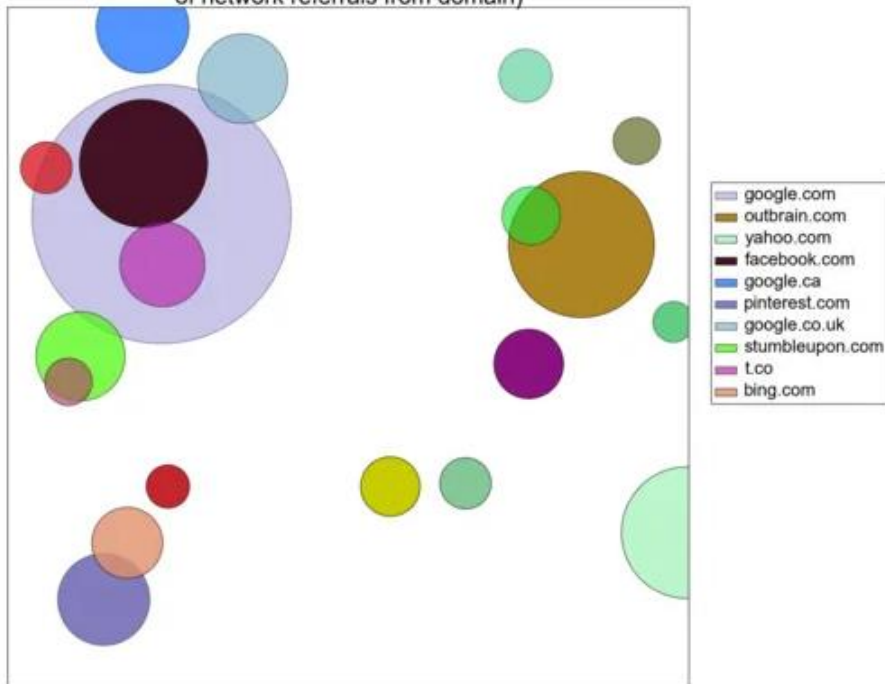
100 Most Frequent Words in Wikipedia



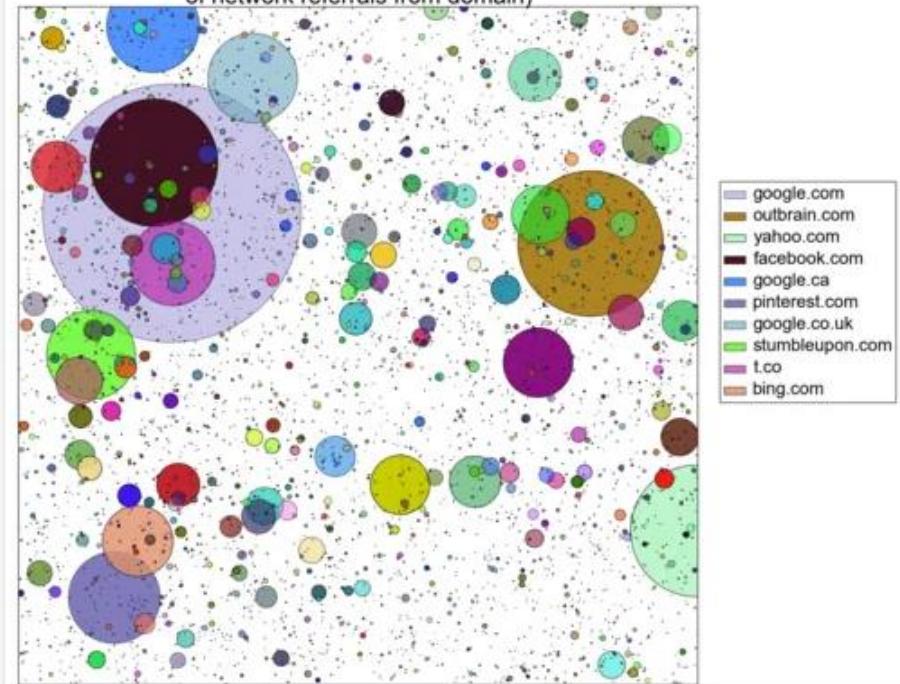
Impact in InfoVis!

<https://blog.parse.ly/post/10/zipfs-law-of-the-internet-explaining-online-behavior/>

Domains referring top 75% of network traffic
(area of circle proportional to number
of network referrals from domain)

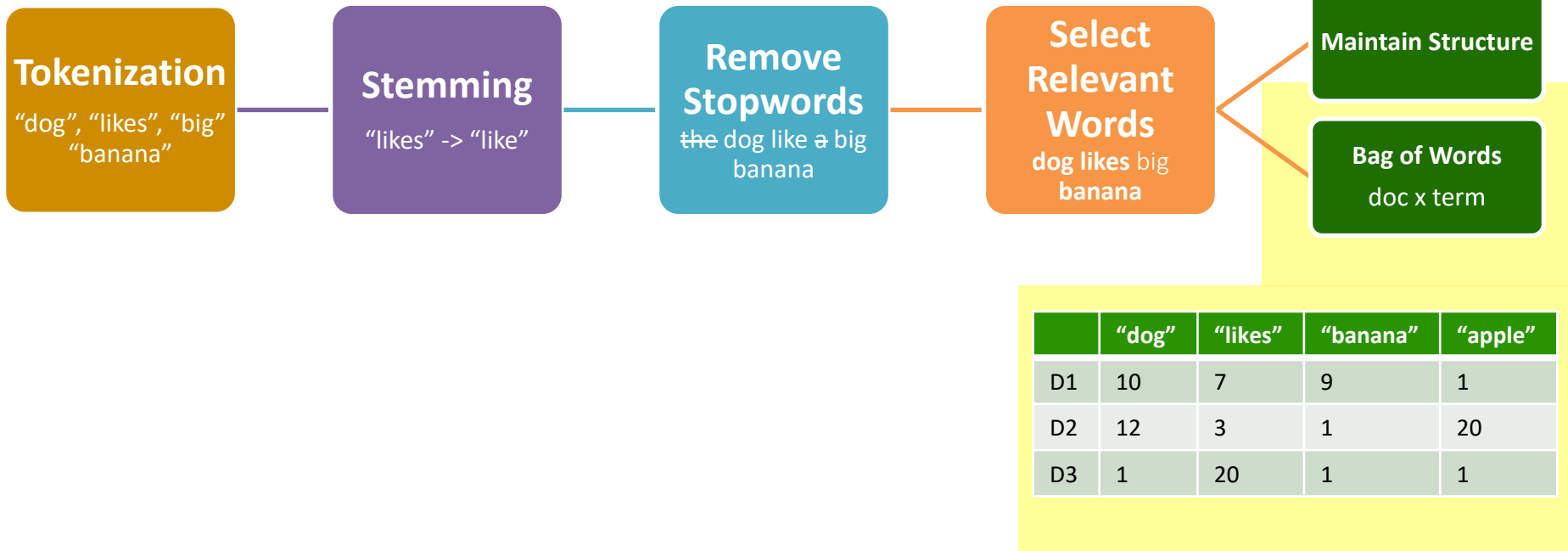


Domains referring top 99% of network traffic
(area of circle proportional to number
of network referrals from domain)



Two Approaches

“The dog likes a big banana”





KEEP
CALM

TELL ME
WHAT TO DO

Keeper of the Seven Keys



<https://youtu.be/iDpAahKUNB4?t=186>

Lyrics!

(...)

You can feel cold sweat
Running down your neck
And the dwarfs of falseness
Throw mud at your back

Guided by spells
From the old Seer's hand
You're suffering pain
Only steel can stand

Stay well on your way and follow the sign
Fulfill your own promise and do what's divined
The seven seas are far away
Placed in the valley of dust heat and sway

You're the Keeper of the Seven Keys
That lock up the seven seas
And the Seer of Visions said before he went blind
Hide them from demons and rescue mankind
Or the world we're all in will soon be sold
To the throne of the evil paid with Lucifer's gold

Throw the first key into the sea of hate

Throw the second key into the sea of fear
Throw the third key into the sea of senselessness
And make the people hold each other's hands
The fourth key belongs into the sea of greed

And the fifth into the sea of ignorance
Disease, disease, disease my friend
For this whole world's in devil's hand
Disease, disease, disease my friend
Throw the key or you may die

On a mound at the shore of the last sea
He is sitting, fixing your sight
With his high iron voice causing sickness
He is playing you out with delight

"Man who do you just think you are?
A silly bum with seven stars
Don't throw the key or you will see
Dimensions cruel as they can be"

Don't let him suck off your power
Throw the key...!

(...)

Word frequency

28 the

12 of

8 and

7 throw

7 key

6 your

6 you

6 sea

6 disease

5 into

4 with

4 seven

3 or

3 in

3 he

3 can

2 you're

2 will

2 seas

2 on

2 my

2 is

2 hand

2 friend

2 don't

2 do

2 be

2 at

2 are

2 a

1 world's

1 world

1 whole

1 who

1 what's

(...)

Challenge One: Visualize Word Frequency

Wordcloud



Tag Clouds

Explore Flickr Through Tags

art australia baby beach birthday blue bw california cameraphone canon
canon cat chicago china christmas city dog england europe family flower
flowers food france friends fun germany halloween holiday india italy japan
london me music nature new newyork night nikon nyc paris park
party people portrait sanfrancisco sky snow spain summer sunset taiwan tokyo

Direct
Poll

Concordance [\(learn more\)](#)

These are the 100 most frequently used words in this book.

accuracy algorithm another approach authority between called case chapter
class classification classifier clustering clusters collection content crawler
data different distribution documents end engines estimate even
example features fetch few figure find first form given graph hits hub
hyperlinks hypertext index information ir labels large learning let link may
measure model need network new nodes number pages
parameters point possible precision probability problem query random ranking
relevant representation response root scores search section see server set
shown shows similar similarity simple sites space system techniques terms
test text time topic training two urls use used user values vector web
words work

delicious Home Bookmarks People Tags

cscie12's Tags
Bookmarks | Network | Tags | Subscriptions

Also try exploring Popular tags

cscie12 Type a tag Tags 200

Show all tags Sort: Alphabetically | By size

 <i> 3GP accessibility Adi ajax anti-phishing apache apache2 api application art
asp.net attribute AVI Backup browser browsers by CGI Chaobai characters clipart code codes
color colors colour columns compliant content conversion cool cost create cscie12 css
customized customized-error-page-creation debug design development developer development
dhtml digital DivX DOM domain download dynamic echo editor element emacs email embed
error ETags Example examples favicon file firebug firefox fixes flash float format forms
framework free games google graphics guide hex home host htaccess html http ico image
introduction iPod java javascript joomla jquery js lamp layout lists lovers lynx management
map Maps menu microsoft MOV MP3 MP4 MPEG4 mysql navigation Neilson Newegg object
open page palette palettes Patchong Patricia perl photoshop php pico png Print programming
properties PSD PSP public Rain rap rapid REALAUDIO reference reset rfc RSS RTF script
scripts SecondLife server shows source SSH ssi ssi standard standards Submitted support
survey SWF SWF tables tags templates Terminal testing text The timeline timelines to tool
tools tutorial tutorials tweaks TXT typography uie unix upload url usability useragent Utility
validator vi video VOB w3 w3c WAI web web2.0 web20 webdesign webdev webservices
website websites, Wiki windows WK1 WMA WMV xhtml xml yahoo yui

How to make a word easier to find?

the

of

and

throw

key

your

you

sea

disease

into

with

seven

How to make a word easier to find?

and

disease

key

into

of

sea

seven

the

throw

with

you

your

Less impacting

the _____
of _____
and _____
throw _____
key _____
your _____
you _____
sea _____
disease _____
into _____
with _____
seven _____

Why bother? INTEGRATION!

and	43, 65, 123
disease	12, 43
key	23, 34, 165
into	1, 56
of	54, 67, 76
sea	12, 45, 87, 154, 243
seven	76, 43
the	54, 67, 68, 124, 246, 276
throw	43, 65, 187
with	45, 32
you	76, 34
your	87, 98



KEEP
CALM

TELL ME
WHAT TO DO



KEEP
CALM

TELL ME
WHAT TO DO

Challenge 2: Where is 'key' in the text?

You can feel cold sweat
Running down your neck
And the dwarfs of falseness
Throw mud at your back

Guided by spells
From the old Seer's hand
You're suffering pain
Only steel can stand

Stay well on your way and follow the sign
Fulfill your own promise and do what's divined
The seven seas are far away
Placed in the valley of dust heat and sway

You're the Keeper of the Seven Keys
That lock up the seven seas
And the Seer of Visions said before he went blind
Hide them from demons and rescue mankind
Or the world we're all in will soon be sold
To the throne of the evil paid with Lucifer's gold

Throw the first key into the sea of hate

Throw the second key into the sea of fear
Throw the third key into the sea of senselessness
And make the people hold each other's hands
The fourth key belongs into the sea of greed

And the fifth into the sea of ignorance
Disease, disease, disease my friend
For this whole world's in devil's hand
Disease, disease, disease my friend
Throw the key or you may die

On a mound at the shore of the last sea
He is sitting, fixing your sight
With his high iron voice causing sickness
He is playing you out with delight

"Man who do you just think you are?
A silly bum with seven stars
Don't throw the key or you will see
Dimensions cruel as they can be"

Don't let him suck off your power
Throw the key...!

Obvious?

You can feel cold sweat
Running down your neck
And the dwarfs of falseness
Throw mud at your back

Guided by spells
From the old Seer's hand
You're suffering pain
Only steel can stand

Stay well on your way and follow the sign
Fulfill your own promise and do what's divined
The seven seas are far away
Placed in the valley of dust heat and sway

You're the Keeper of the Seven **Keys**
That lock up the seven seas
And the Seer of Visions said before he went blind
Hide them from demons and rescue mankind
Or the world we're all in will soon be sold
To the throne of the evil paid with Lucifer's gold

Throw the first **key** into the sea of hate

Throw the second **key** into the sea of fear
Throw the third **key** into the sea of senselessness
And make the people hold each other's hands
The fourth **key** belongs into the sea of greed

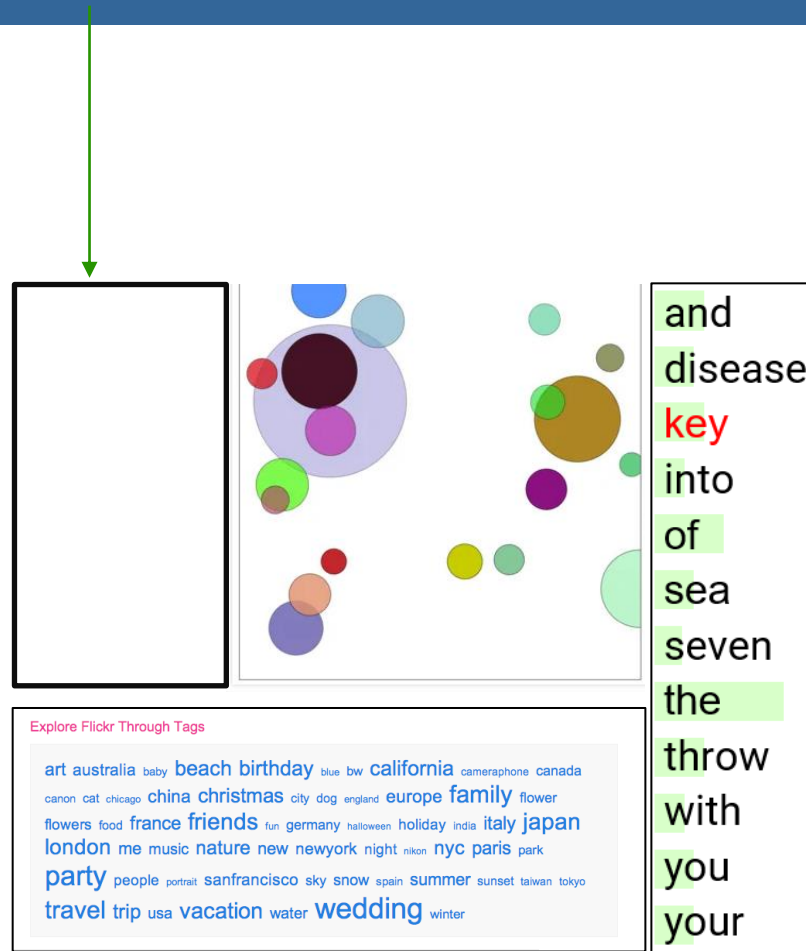
And the fifth into the sea of ignorance
Disease, disease, disease my friend
For this whole world's in devil's hand
Disease, disease, disease my friend
Throw the **key** or you may die

On a mound at the shore of the last sea
He is sitting, fixing your sight
With his high iron voice causing sickness
He is playing you out with delight

"Man who do you just think you are?
A silly bum with seven stars
Don't throw the key or you will see
Dimensions cruel as they can be"

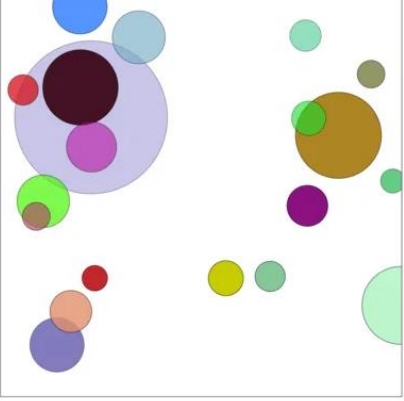
Don't let him suck off your power
Throw the **key**...!

Now fit it here!



There you go! Minimap for the win!

You can feel cold sweat
Running down your neck
And the dwarf's falconess
Throw mud at your back
Guided by spells
From the old Seer's hand
You're suffering pain
Only steel can stand
Stay well on your way and follow the sign
Fulfill your own promise and do what's divined
The seven seas are far away
Placed in the valley of dust heat and sway
You're the keeper of the Seven **key**
That lock up the seven seas
And the Seer of Visions said before he went blind
Hide them from demons and rescue mankind
Or the world we're all in will soon be sold
To the throne of the evil paid with Lucifer's gold
Throw the first **key** into the sea of hope
Throw the second **key** into the sea of fear
Throw the third **key** into the sea of consciousness
And make the people hold each other's hands
The fourth **key** belongs into the sea of greed
And the fifth into the sea of ignorance
Disease, disease, disease my friend
For this whole world's in death's hand
Disease, disease, disease my friend
Throw the **key** or you may die
On a mound at the shore of the last sea
He is sitting, fixing your sight
With his high iron voice causing sickness
He is playing you out with delight
"Nan who do you just think you are?
A silly bum with seven stars
Don't throw the **key** or you will see
Dimensions cruel as they can be"
Don't let him suck off your power
Throw the **key**...



and
disease
key
into
of
sea
seven
the
throw
with
you
your

Explore Flickr Through Tags

art australia baby beach birthday blue bw california cameraphone canada
canon cat chicago china christmas city dog england europe family flower
flowers food france friends fun germany halloween holiday india italy japan
london me music nature new newyork night nikon nyc paris park
party people portrait sanfrancisco sky snow spain summer sunset taiwan tokyo
travel trip usa vacation water wedding winter

Overview & Timeline

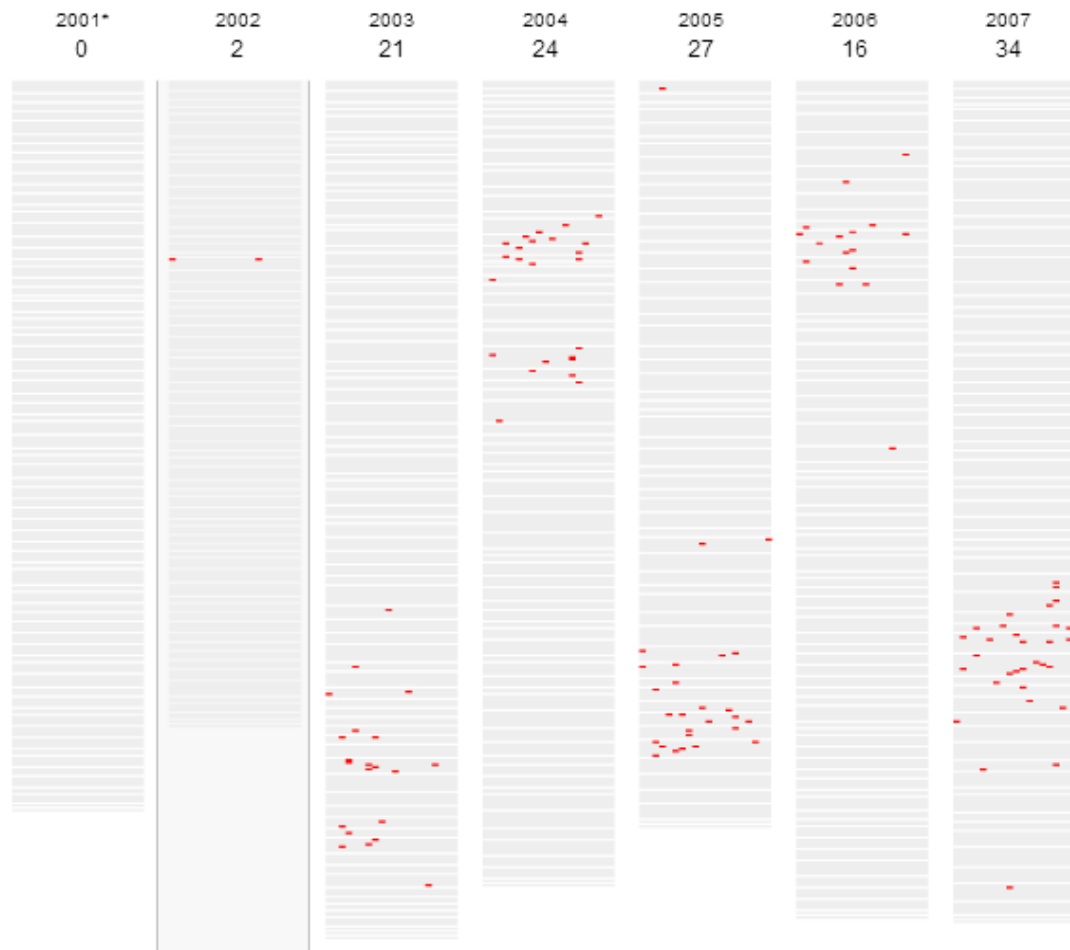
http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?initialWord=iraq

The 2007 State of the Union Address

Over the years, President Bush's State of the Union address has averaged almost 5,000 words each, meaning the the President has delivered over 34,000 words. Some words appear frequently while others appear only sporadically. Use the tools below to analyze what Mr. Bush has said.

 or choose a word here. 

Use of the phrase "Iraq" in past State of the Union Addresses



Compared with other words



How about now?

the
of
and
throw
key
your
you
sea
disease
into
with
seven

Use sparklines!

the
of
and
throw
key
your
you
sea
disease
into
with
seven

Sparkclouds

http://research.microsoft.com/en-us/um/people/nath/docs/sparkclouds_infovvis2010.pdf

boy called day did door dudley eyes gone
good harry head lemon little mrs privet professor reason
say saying sir sister think turned wall yes

blankets boy called cat cloak come corner day did
door drills drive dudley dumbledore dursley eyes fall going
gone good got hagrid harry head himself inside just

know lamp left lemon letter like little looked man moonspell



KEEP
CALM

TELL ME
WHAT TO DO



KEEP
CALM

TELL ME
WHAT TO DO

Let's see how words relate...

What usually comes after “key”?

Bigrams!

you can
can feel
feel cold
cold sweat
sweat running
running down
down your
your neck
neck and
and the
the dwarfs
dwarfs of
of falseness
falseness throw
throw mud
mud at
at your
your back
back guided
guided by
by spells

Bigrams starting with “key”!

keys that

key into (x3)

key belongs

key or (x2)

Google nGram Viewer

Google books Ngram Viewer

Graph these comma-separated phrases: ☐ case-insensitive

between and from the corpus with smoothing of

[Search lots of books](#)



What about larger sequences?

you can feel
can feel cold
feel cold sweat
cold sweat running
sweat running down
running down your
down your neck
your neck and
neck and the
and the dwarfs
the dwarfs of
dwarfs of falseness
of falseness throw
falseness throw mud
throw mud at
mud at your
at your back
your back guided
back guided by
guided by spells
by spells of

“Key” trigrams...

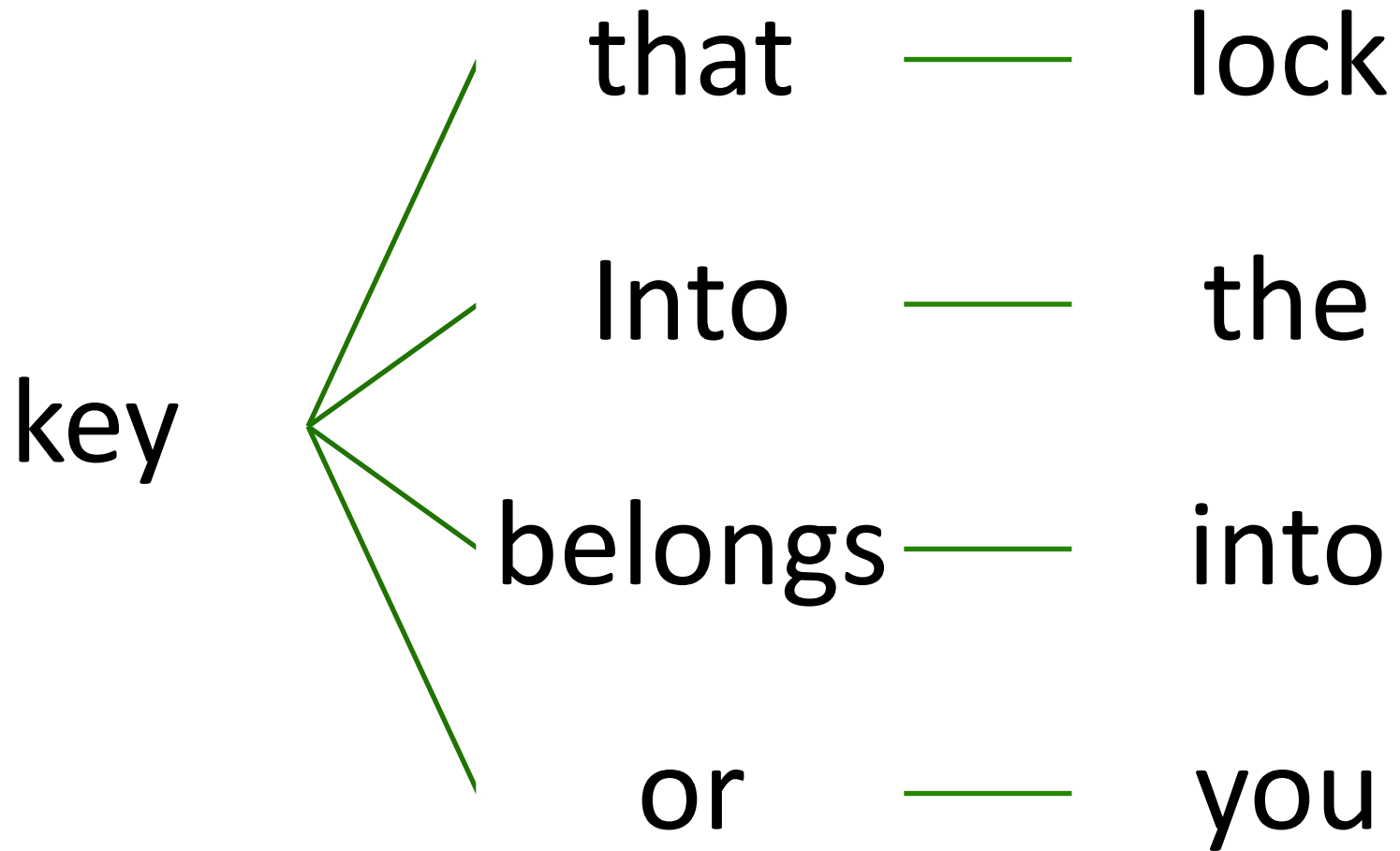
keys that lock

key into the (x3)

key belongs into

key or you (x2)

Tree!

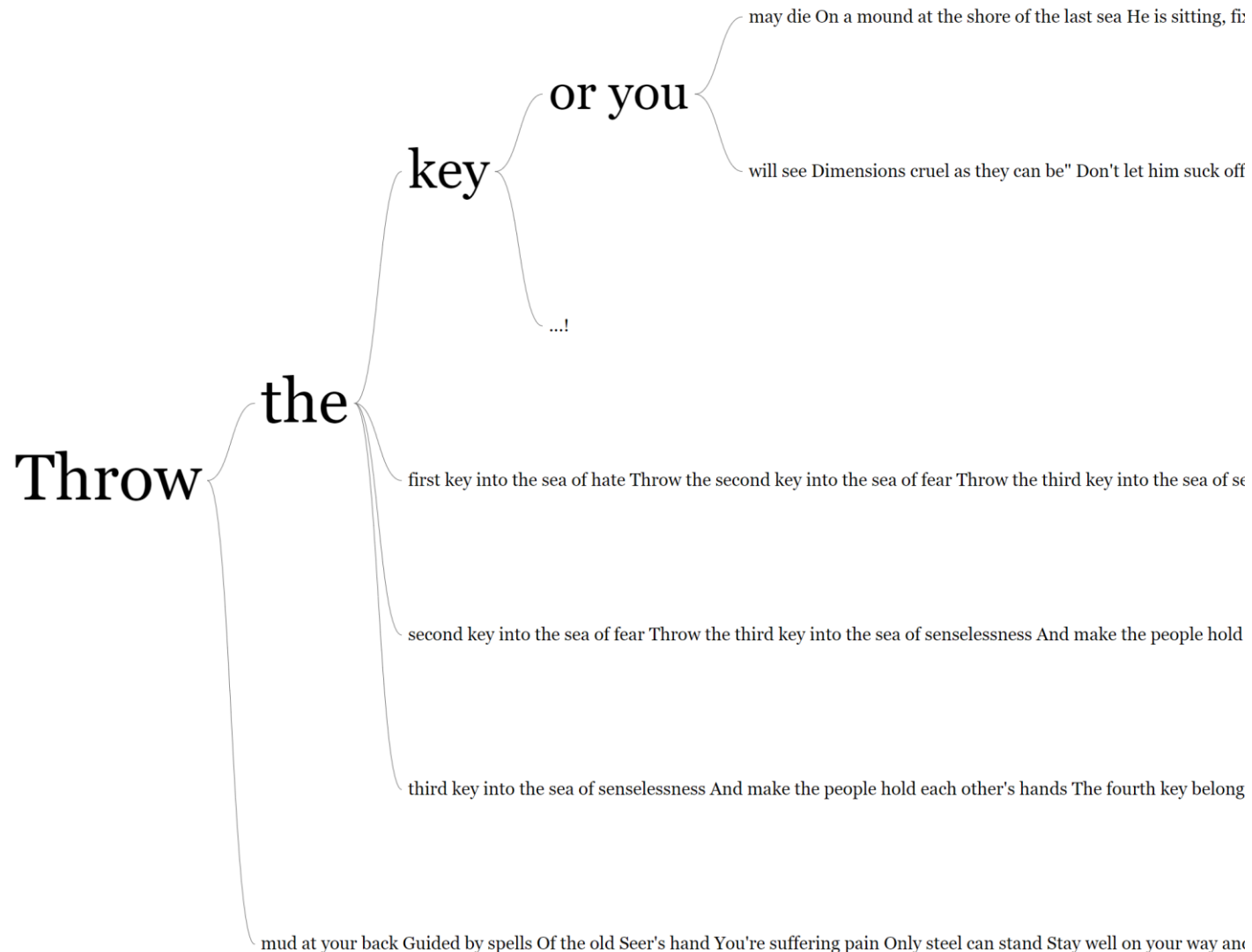


More than just a tree?

Let's do it for “throw”, instead...

Wordtree for “throw”

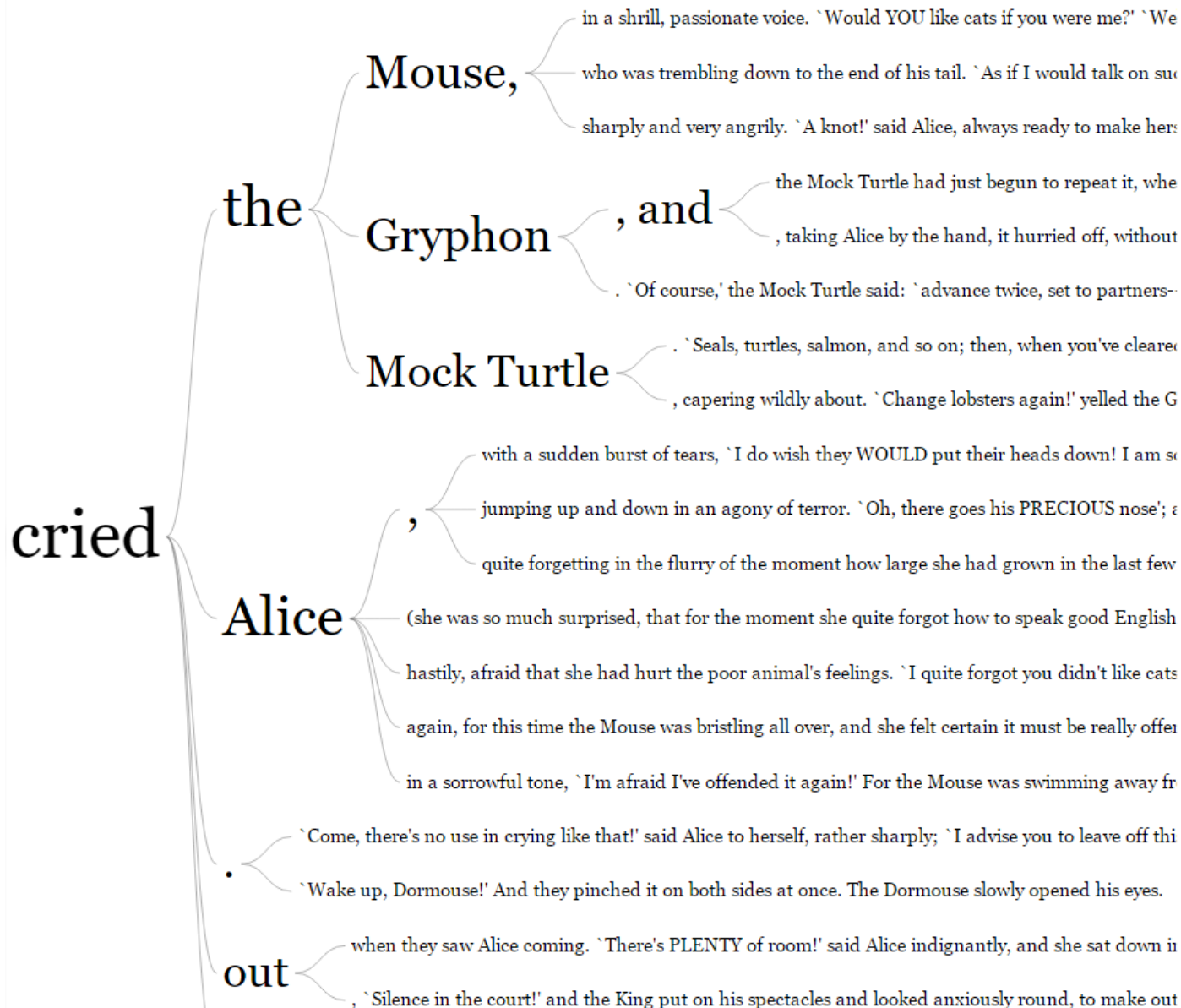
<https://www.jasondavies.com/wordtree/?source=b16fa834cb7301ec7c106783523fb6dd&prefix=throw>



WordTree

<https://www.jasondavies.com/wordtree/?source=alice-in-wonderland.txt&prefix=cried>

Shift-click to make that word the root.



you didn't like cats.'

'Not like cats!' cried the Mouse, in a shrill, passionate voice. 'Would YOU like cats if you were me?'

'Well, perhaps not,' said Alice in a soothing tone: 'don't be angry about it. And yet I wish I could show you our cat Dinah: I think you'd take a fancy to cats if you could only see her. She is such a dear quiet thing,' Alice went on, half to herself, as she swam lazily about in the pool, 'and she sits purring so nicely by the fire, licking her paws and washing her face--and she is such a nice soft thing to nurse--and she 's such a capital one for catching mice--oh, I beg your pardon!' cried Alice again, for this time the Mouse was bristling all over, and she felt certain it must be really offended. 'We won't talk about her any more if you'd rather not.'

'We indeed!' cried the Mouse, who was trembling down to the end of his tail. 'As if I would talk on such a subject! Our family always HATED cats: nasty, low, vulgar things! Don't let me hear the name again!'

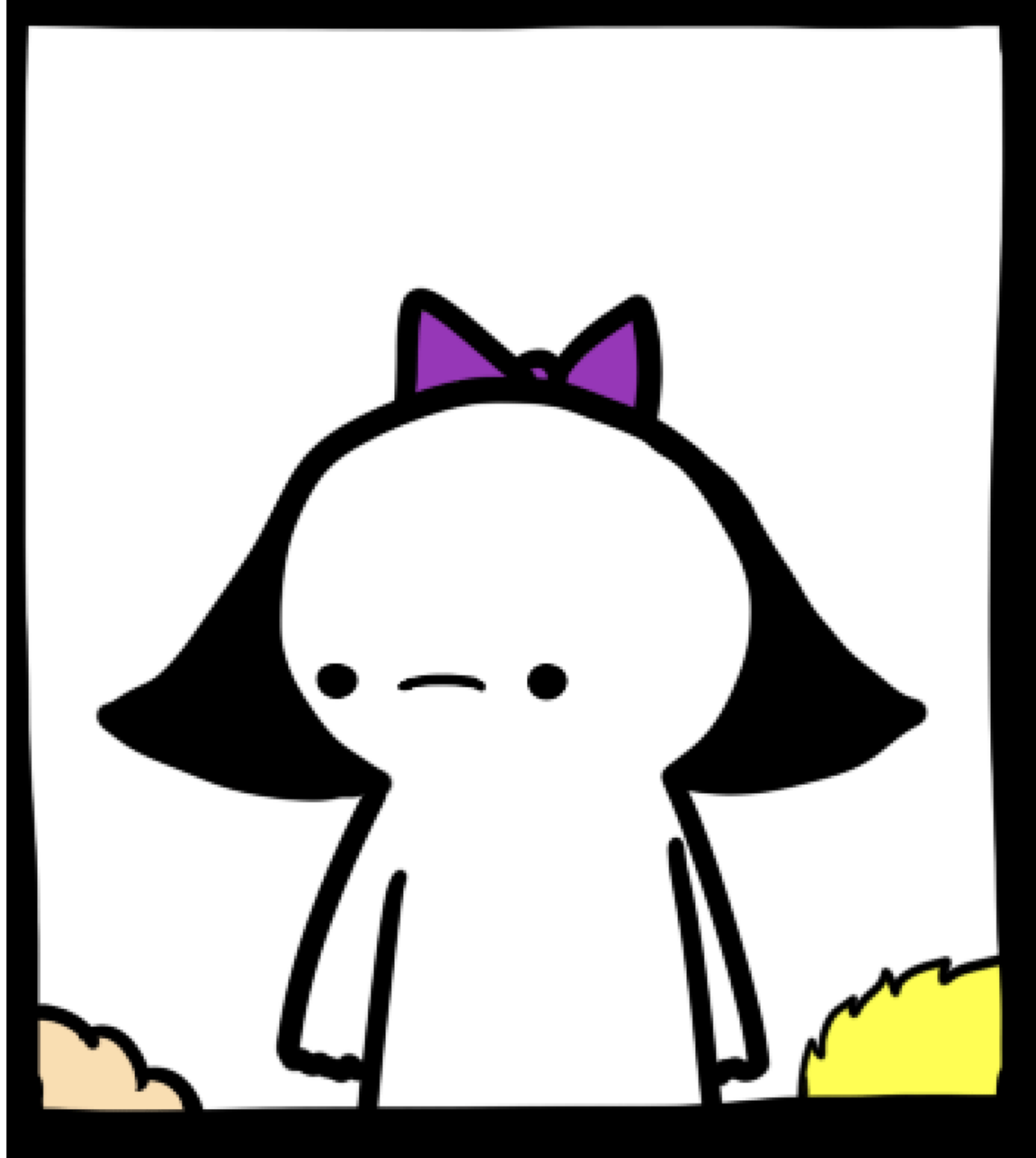


KEEP
CALM

TELL ME
WHAT TO DO



THERE WILL
NOW BE A
BRIEF
INTERMISSION





THERE WILL
NOW BE A
BRIEF
INTERMISSION

GOOD or EVIL?



EXHIBIT A

<https://github.com/adodell/AcademicVsCurrentUse-ProgrammingLanguages>

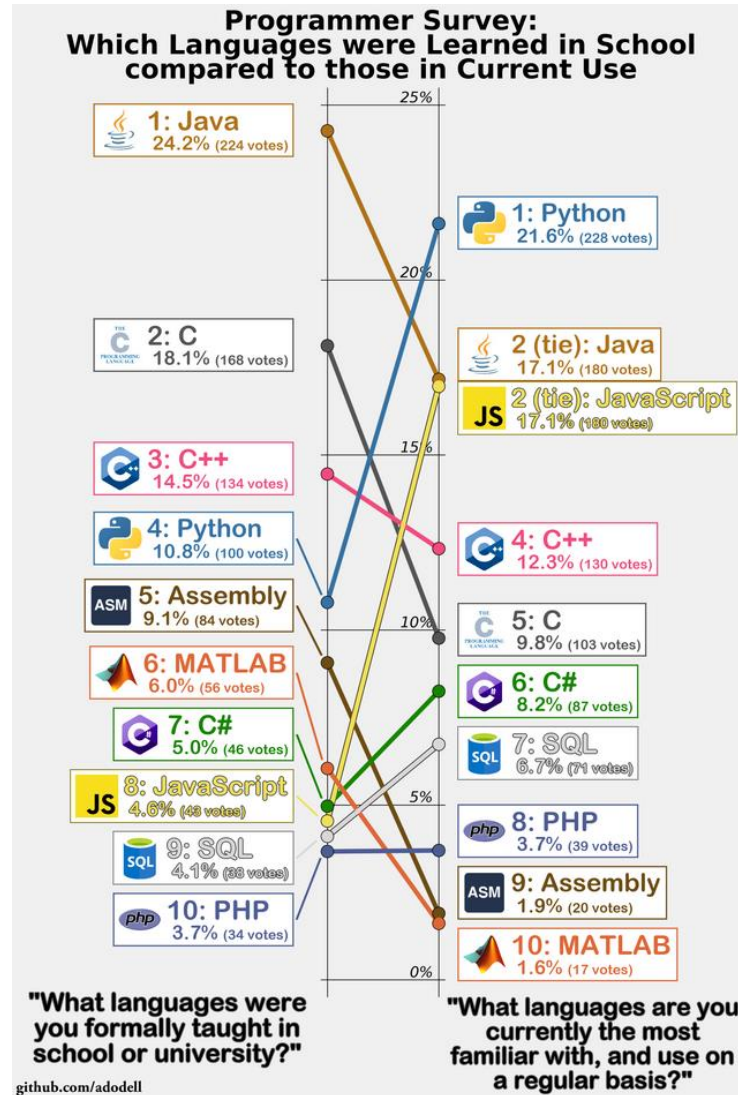


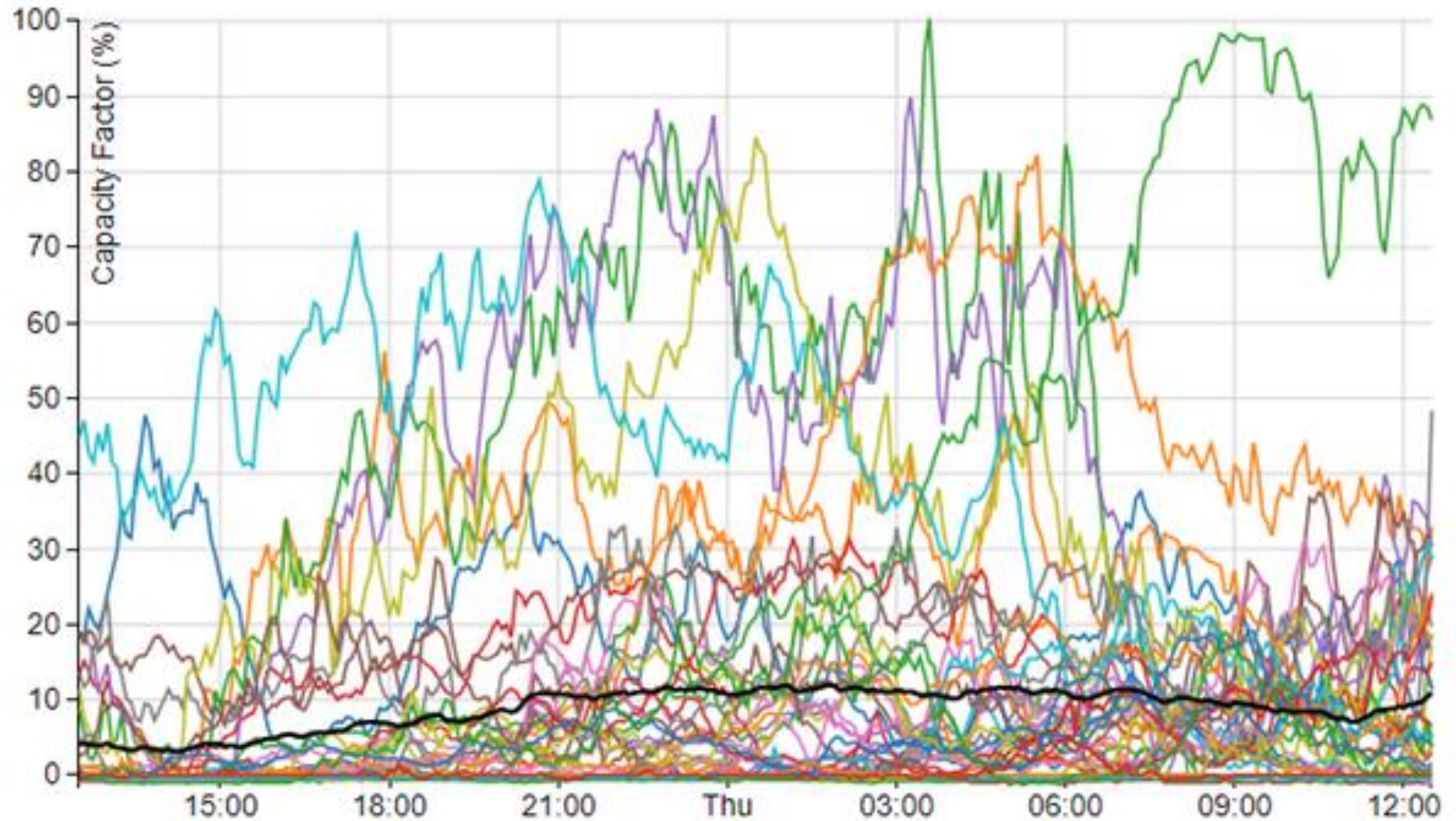
EXHIBIT B

https://www.reddit.com/r/dataisbeautiful/comments/9n1044/bee_movie_plot_using_sentiment_analysis_oc/

Today's Wind Energy Production

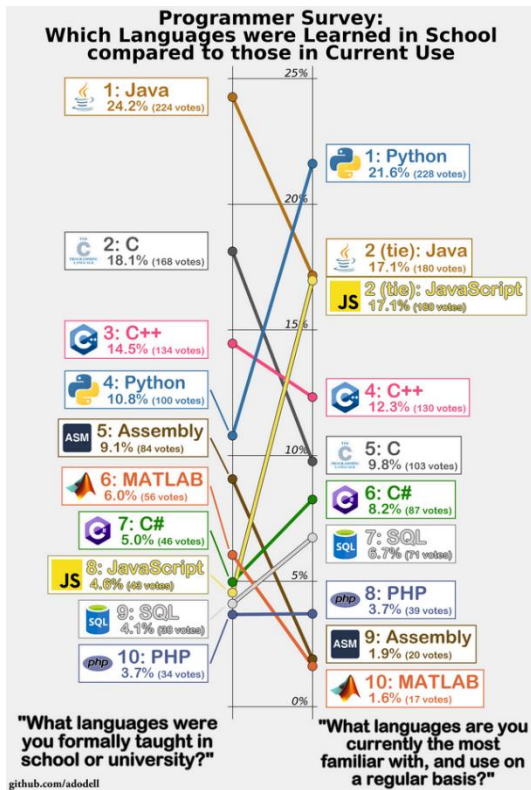
%

MW



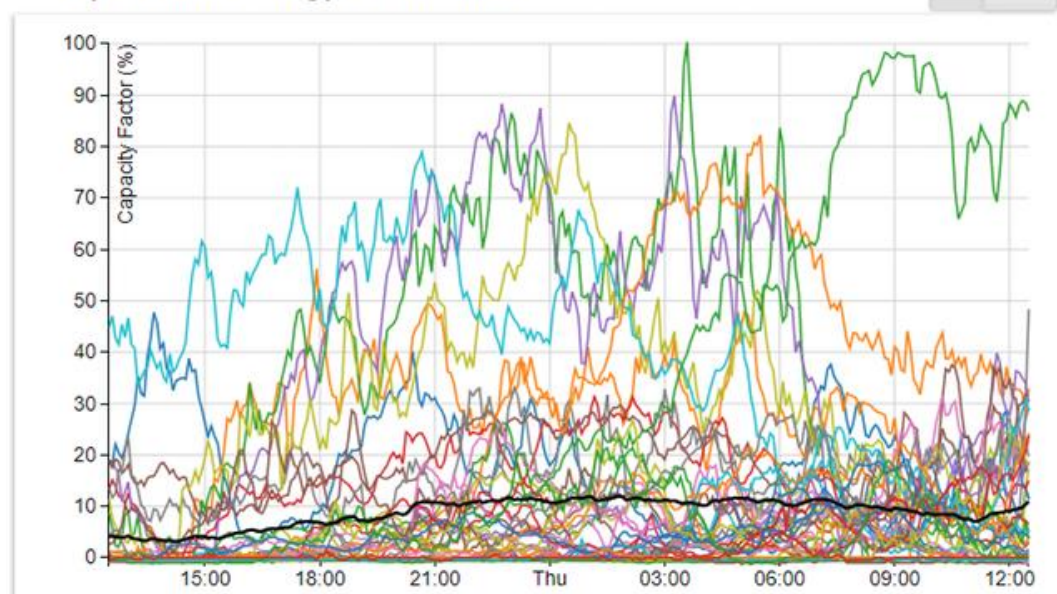
Where does the Evil hide?

A



B

Today's Wind Energy Production



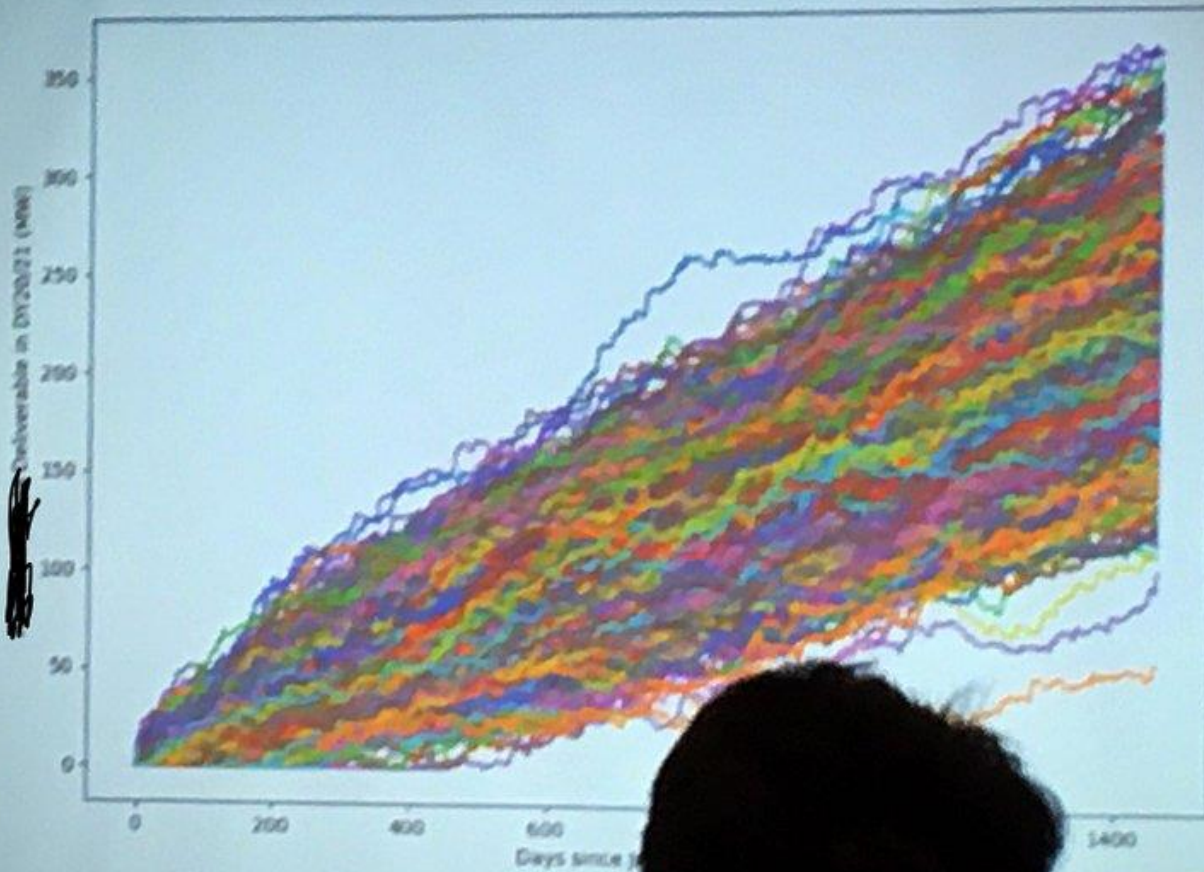
Direct
Poll

GOOD or EVIL?

Wait for it...

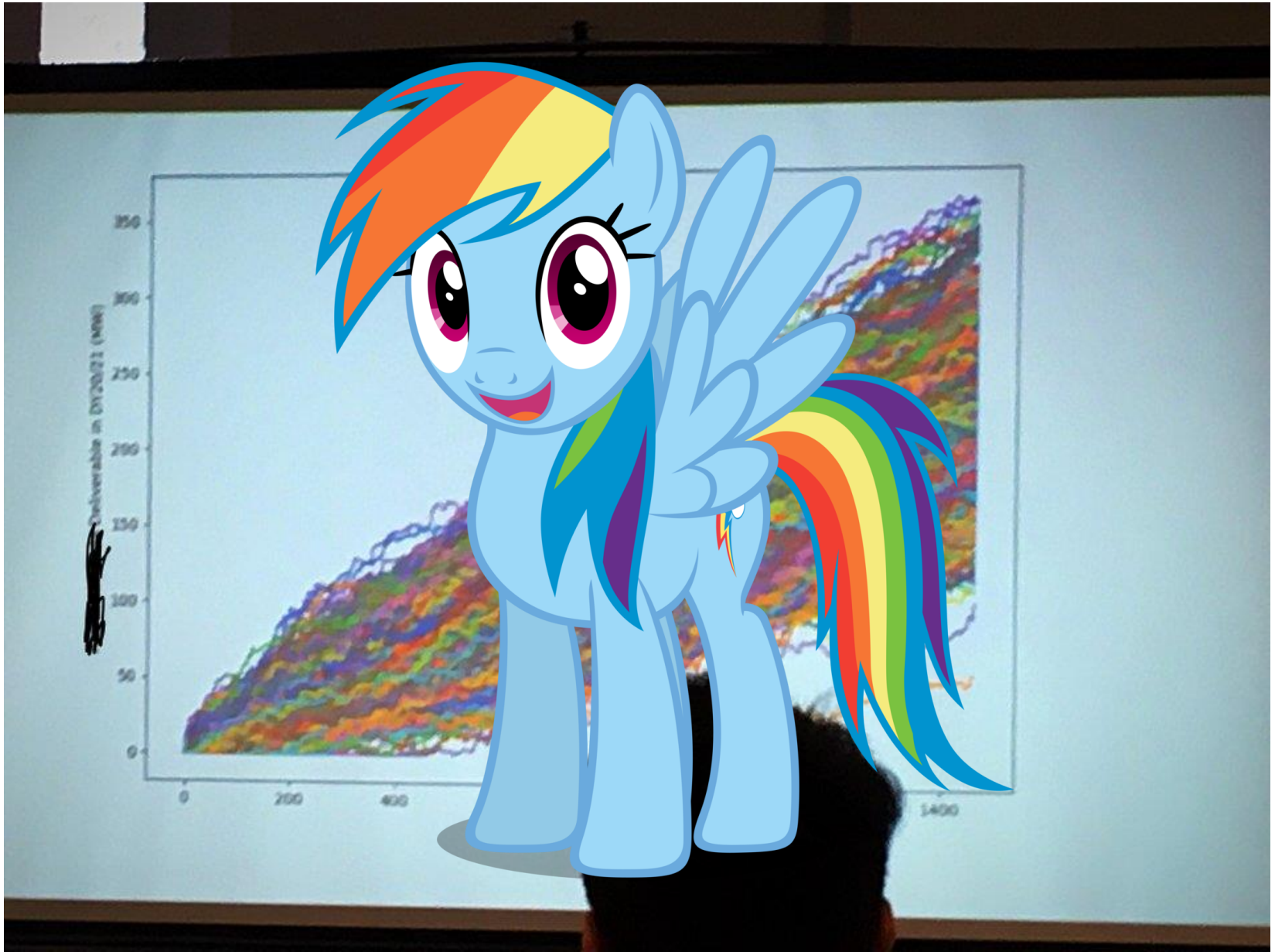
GOOD or EVIL?

https://www.reddit.com/r/dataisugly/comments/9jfn6/chart_regarding_statistical_modeling_techniques/



GOOD or EVIL?

https://www.reddit.com/r/dataisugly/comments/9jfn6/chart_regarding_statistical_modeling_techniques/



02

**MAKE IT
QUANTITATIVE!
(METRICS!)**

Sentiment Analysis

What was the attitude towards something?

Positive?

Negative?

(Humans get it right 79% of times)

Simplest form: keyword spotting

Love

hate

Enjoy

detest

Like

ugly

:)

:(

Sentiment Analysis of the Bible

<https://technowiki.wordpress.com/2011/11/22/applying-sentiment-analysis-to-the-bible-2/>



Tweet Sentiment Viz

http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/



sentiment viz

Tweet Sentiment Visualization

Sentiment

Topics

Heatmap

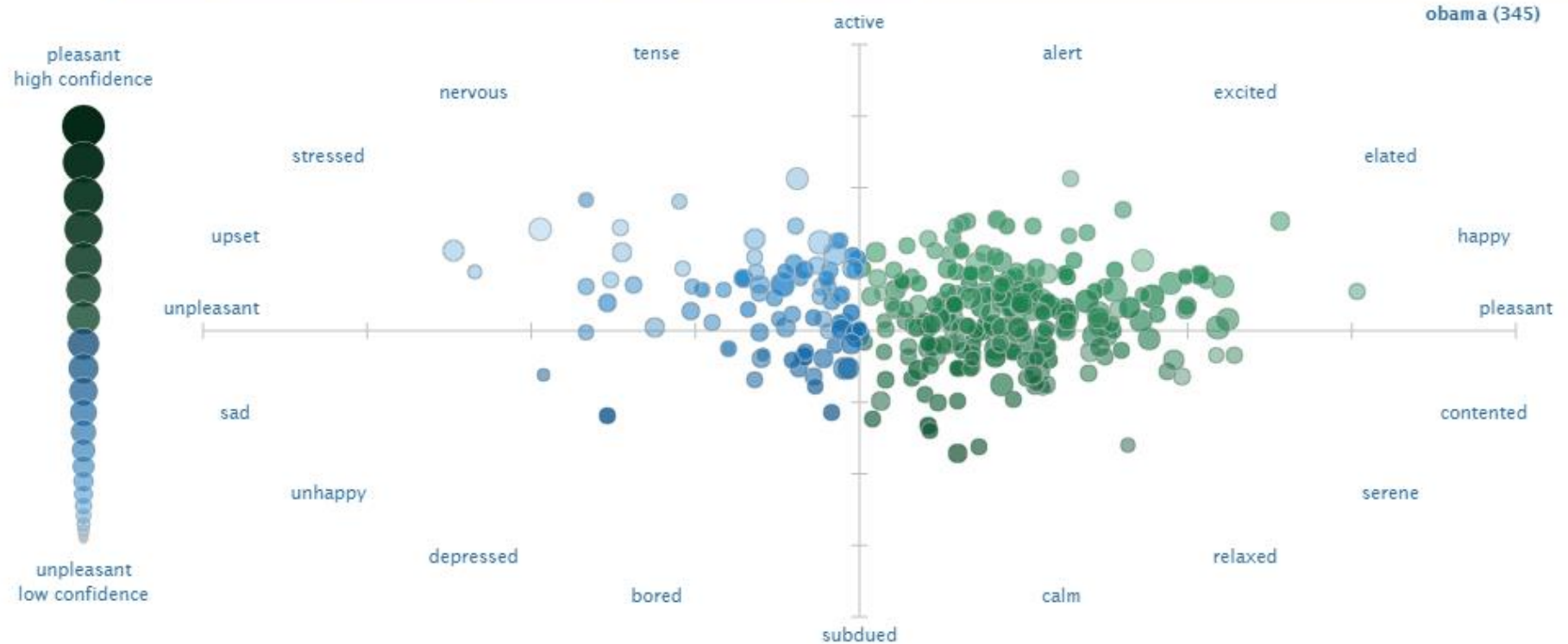
Tag Cloud

Timeline

Map

Affinity

Tweets



Keywords:



Readability Measures – How easy to understand?

<https://raven.zendesk.com/hc/en-us/articles/202308564-What-are-Readability-Metrics->

Flesch-Kincaid Reading Ease

Incorporates average number of syllables per word and the average number of words per sentence. Results measured on 1-100 (100 = easy).

Flesch-Kincaid Grade Level

As previous, results are given as an academic grade level, from 0 to 12.

Gunning-Fog Score

Takes into account "complex" words (three or more syllables)
Result is a grade-level score beginning at 1 with no upper bound.

Coleman-Liau Index

Relies on the number of characters instead of syllables per words for its calculation.

Returns a U.S. grade-level score from 1-12

SMOG Index

Take 30 sentences (10 from the beginning, middle and end of your text), calculate sqrt of count of words with +3 syllables in each group, round to nearest 10 and add 3. Returns U.S. grade level.

Readability Metrics

<http://read-able.com/check.php>

"In the beginning the Universe was created. This has made a lot of people very angry and been widely regarded as a bad move. Many races believe that it was created by some sort of God, though the Jatravartid people of Viltvodle VI believe that the entire Universe was in fact sneezed out of the nose of a being called the Great Green Arkleseizure. The Jatravartids, who live in perpetual fear of the time they call The Coming of The Great White Handkerchief, are small blue creatures with more than fifty arms each, who are therefore unique in being the only race in history to have invented the aerosol deodorant before the wheel."

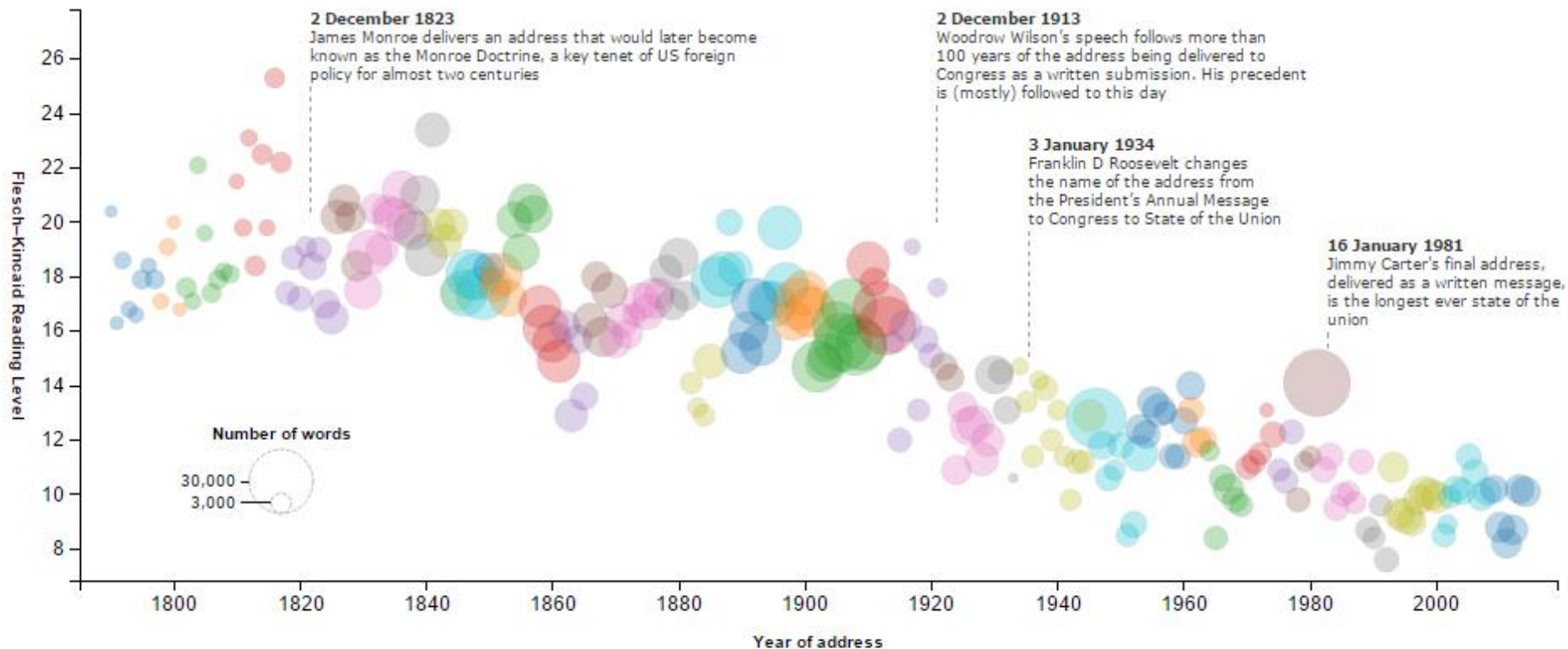
Flesch Kincaid Reading Ease	53.9
Flesch Kincaid Grade Level	12.8
Gunning Fog Score	14.1
SMOG Index	11.2
Coleman Liau Index	10.8

Dumb Presidents?

<http://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level>

The state of our union is ... dumber: How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union



Presidents in order of reading level

1. George H.W. Bush



Avg. reading level

2. Barack Obama



Avg. reading level

3. William J. Clinton



Avg. reading level

4. Lyndon B. Johnson



Avg. reading level

Literature Fingerprinting

Different variables for literary analysis

Average word length

Syllables per word

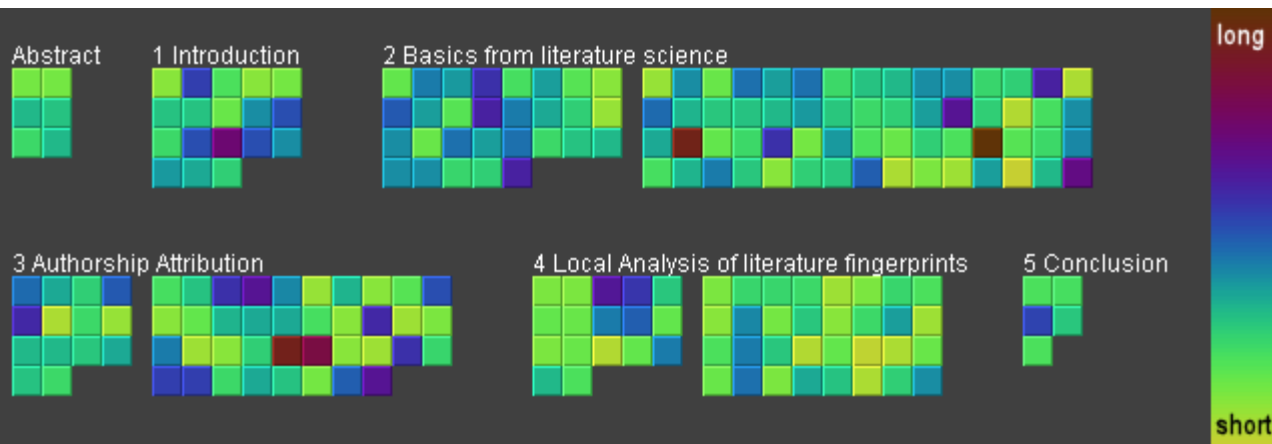
Average sentence length

Percentage of nouns, verbs, adjectives

Frequencies of specific words

Hapax Legomena –number of words that occur once

http://infovis.uni-konstanz.de/papers/2007/2007_vast.pdf



long



03

COLLECTIONS

PaperLens



Bohemian Bookshelf

The Bohemian Bookshelf

Supporting Serendipitous Book Discoveries through Visualization

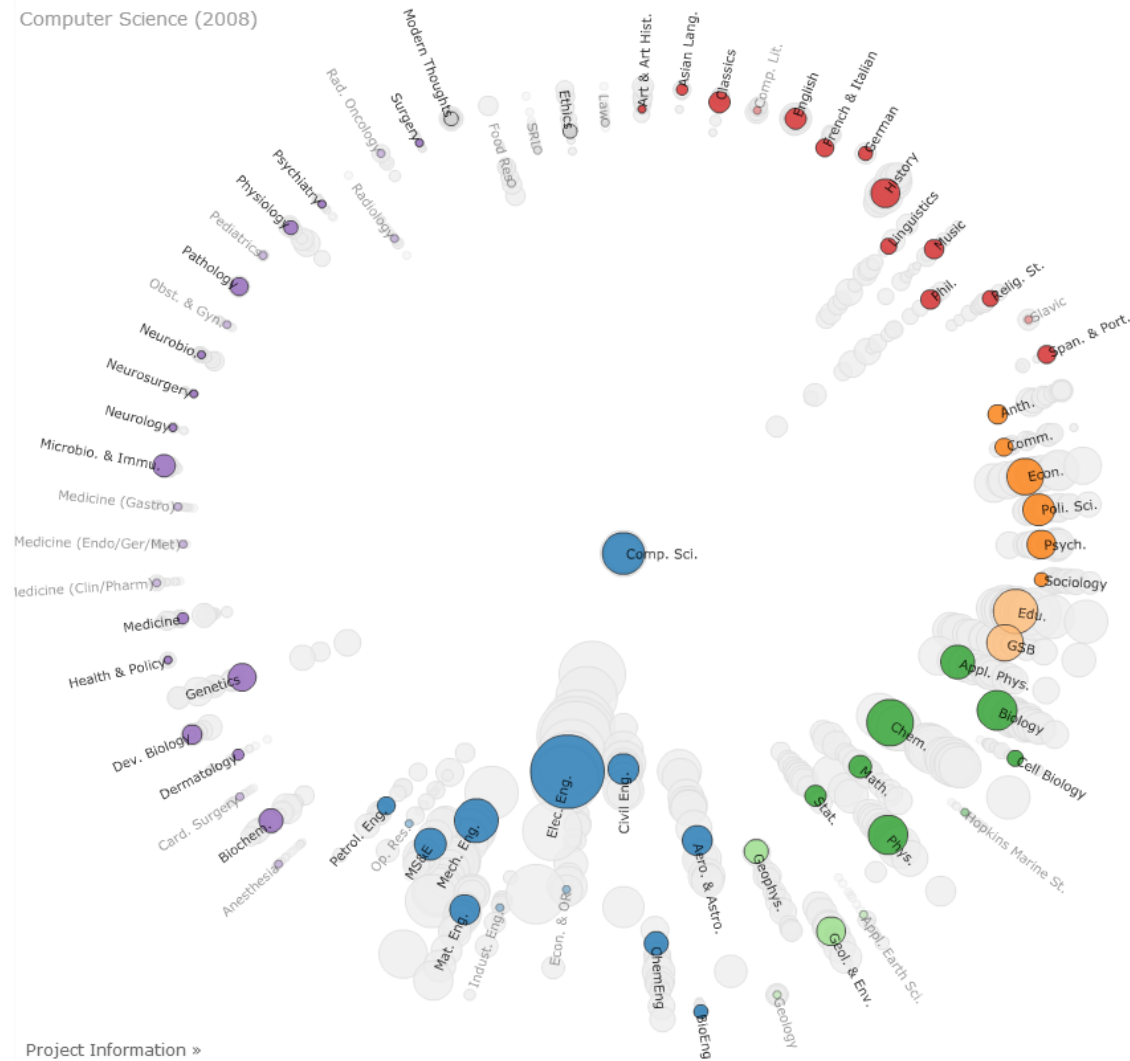
Alice Thudt, Uta Hinrichs and Sheelagh Carpendale

Stanford Dissertation Browser

<http://nlp.stanford.edu/projects/dissertations/browser.html>

Stanford Dissertation Browser

Computer Science (2008)



Project Information »

Dept View: Computer Science (2008)

Topic distance between Comp. Sci. and all other depts

Area of circles denotes number of theses

Depts without published theses in 2008 are faded out

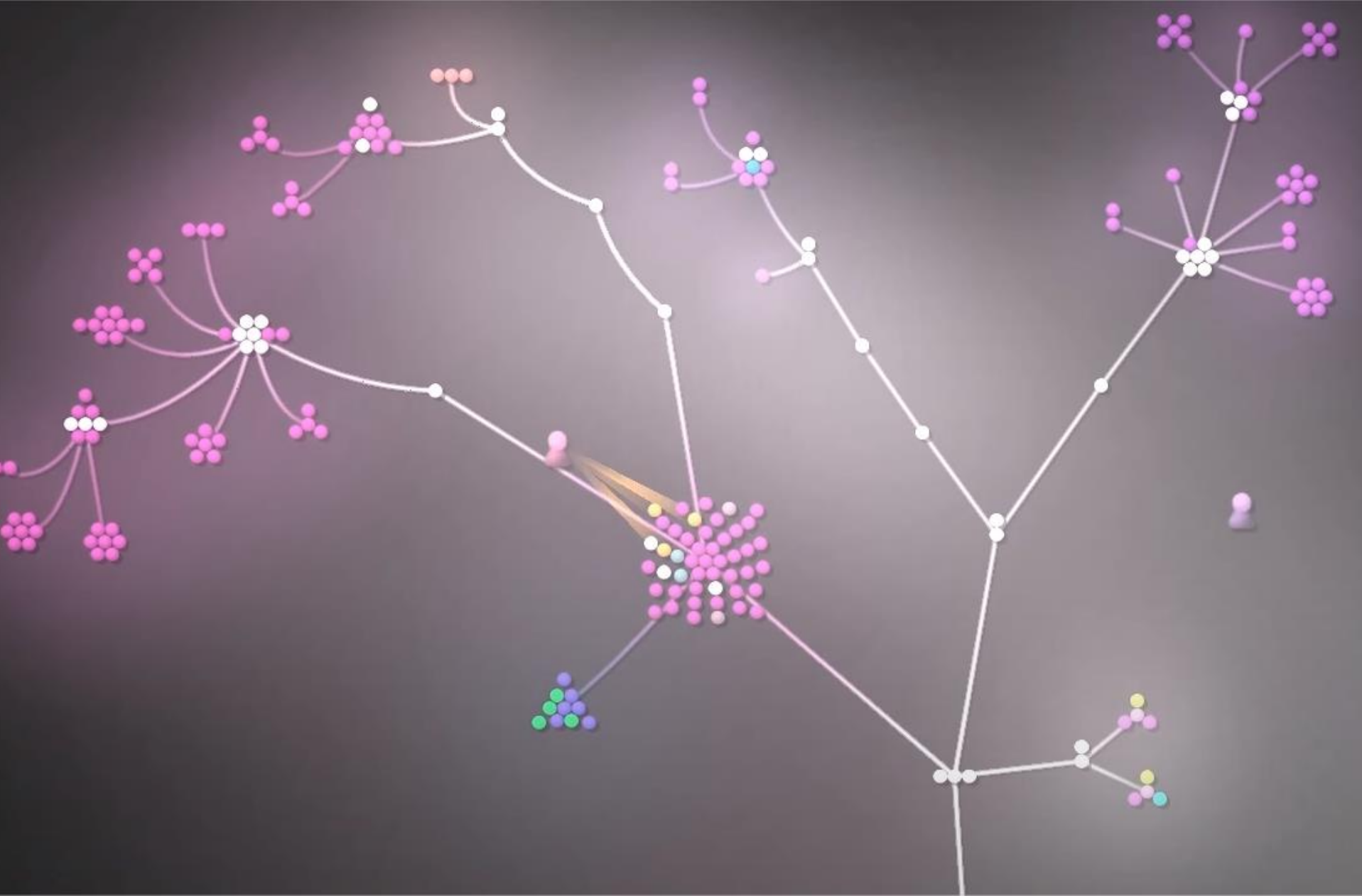
Year 2008

< Backward

Forward >

Gouce

<https://gource.io/>



QUESTIONS?