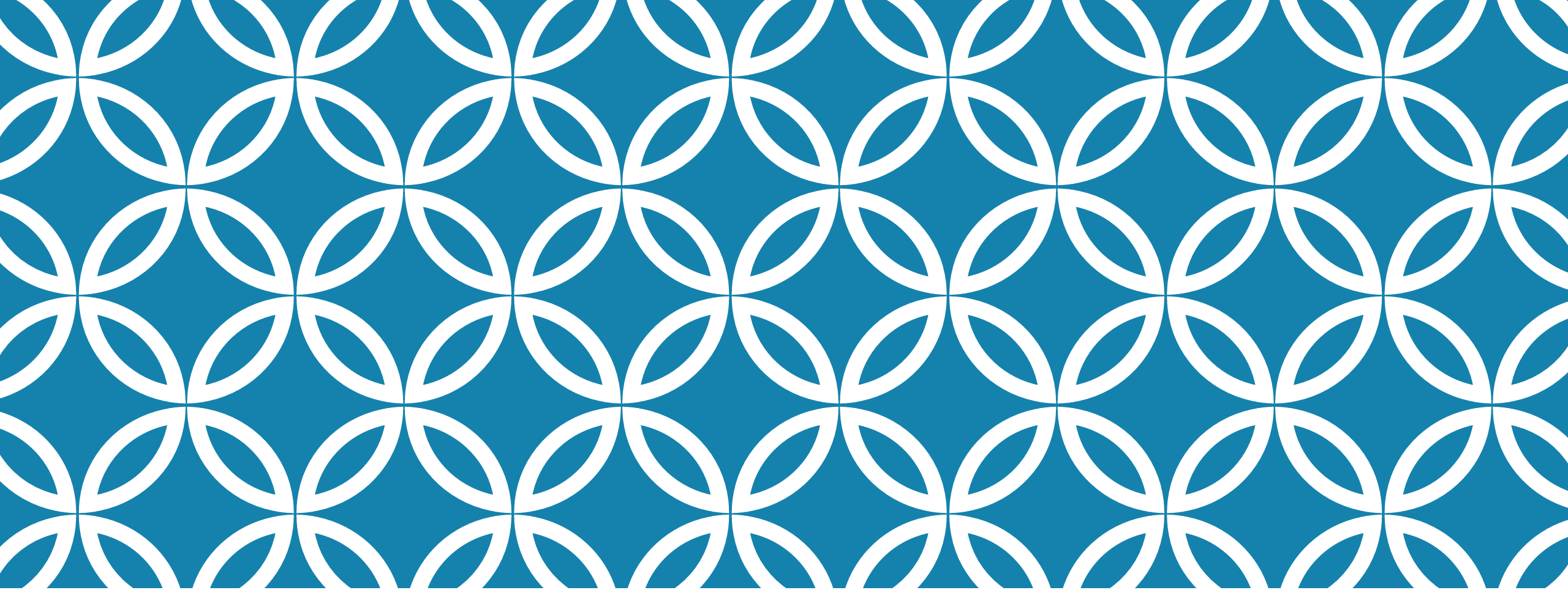


# TOPIC MODELING AND TEXT ANALYSIS IN NATURAL LANGUAGE PROCESSING

tcworld 2019, Bangalore

IN GOD WE TRUST. ALL OTHERS MUST BRING DATA.

-W. EDWARDS DEMING

# WHO AM I

Saurav Ghosh

Adobe, Bangalore

GitHub: https://github.com/sauravg94

Member of Bangalore R Users Group

Areas of interest: NLP, text mining, text analytics, topic modeling, SEO analytics

# WHAT IS NATURAL LANGUAGE PROCESSING

Wikipedia definition:

Natural language processing (NLP) is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

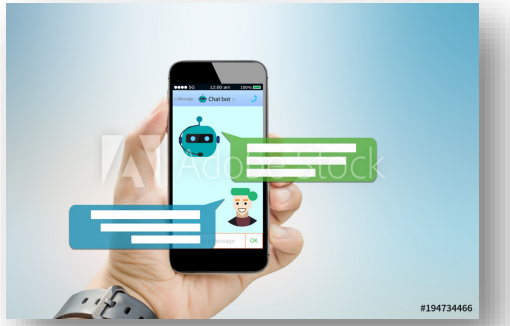
Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

# WHY NLP


80
#207969357


#144183548


#124835895


#194734466


natural
learning
processing
NLP
language
text
automatic
linguistics
interaction
#110528660

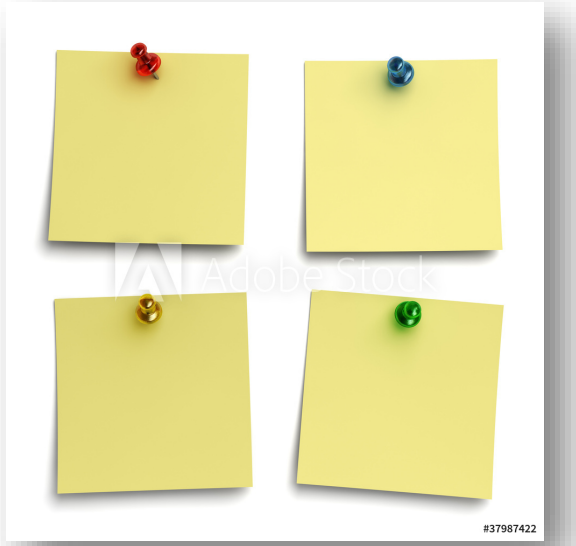

Spam
#214741571

# TOPIC MODELING

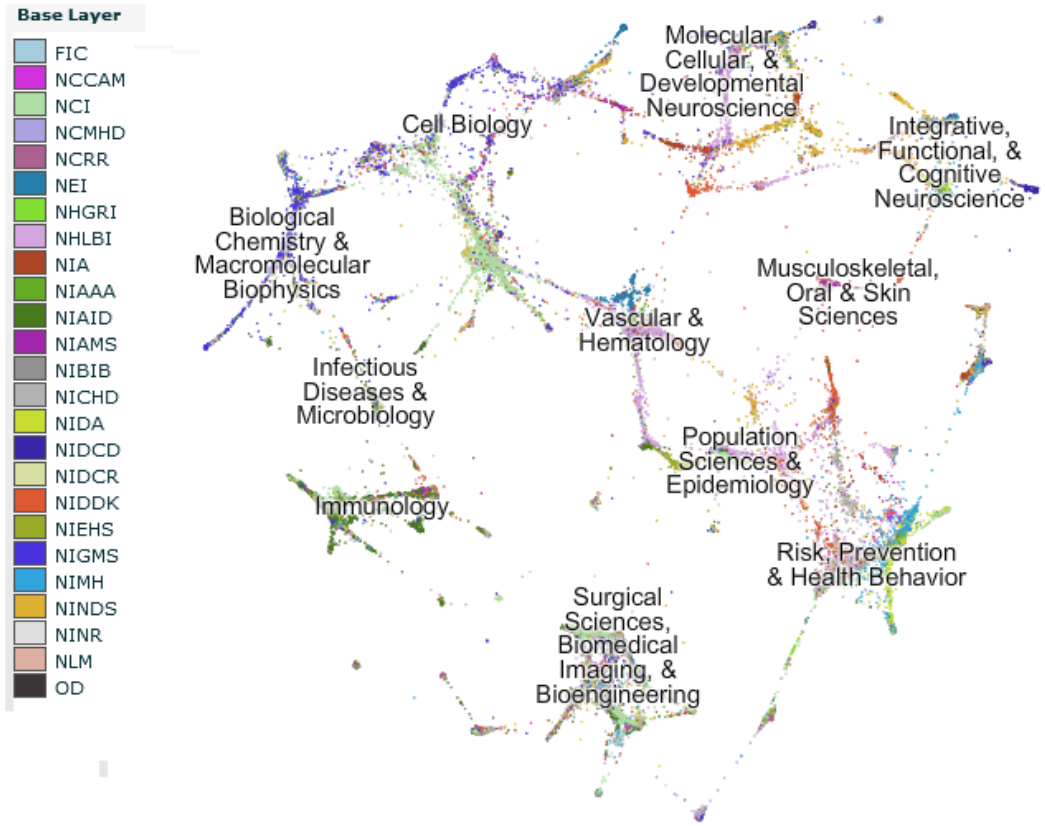
A critical component in NLP and text analytics.

Volume of collections of text document is growing exponentially, necessitating methods for automatically organizing, understanding, searching and summarizing them

- Uncover hidden topical patterns in collections.
- Annotate documents according to topics.
- Using annotations to organize, summarize and search.



# EXAMPLE




This is a topic map of all grants awarded by the National Institutes of Health in 2011. There are approximately 80,000 grants, each represented as a dot, color-coded by NIH Institute. Grants are located nearby one another based on shared topical focus. Labels are placed automatically, based on NIH Review Study Sections or other information obtained from the underlying grants.

NIH Grants Topic Map 2011
NIH Map Viewer (https://app.nihmaps.org)

# TOPIC MODELING- BREAKING UP

Without the math of the model, we define topic modeling as:

**Every document is a mixture of topics.** We imagine that each document may contain words from several topics in particular proportions. For example, in a two-topic model we could say "Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B."

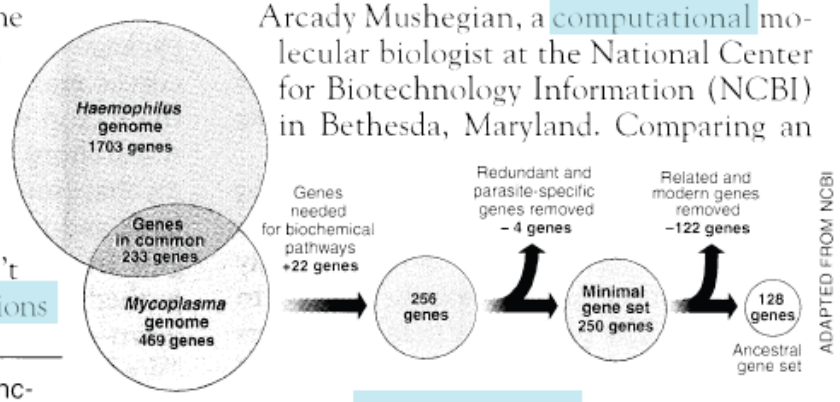
**Every topic is a mixture of words.** For example, we could imagine a two-topic model of American news, with one topic for "politics" and one for "entertainment." The most common words in the politics topic might be "President", "Congress", and "government", while the entertainment topic may be made up of words such as "movies", "television", and "actor". Importantly, words can be shared between topics; a word like "budget" might appear in both equally.

# EXAMPLE- LDA (DAVID BLEI ET AL 2003)

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

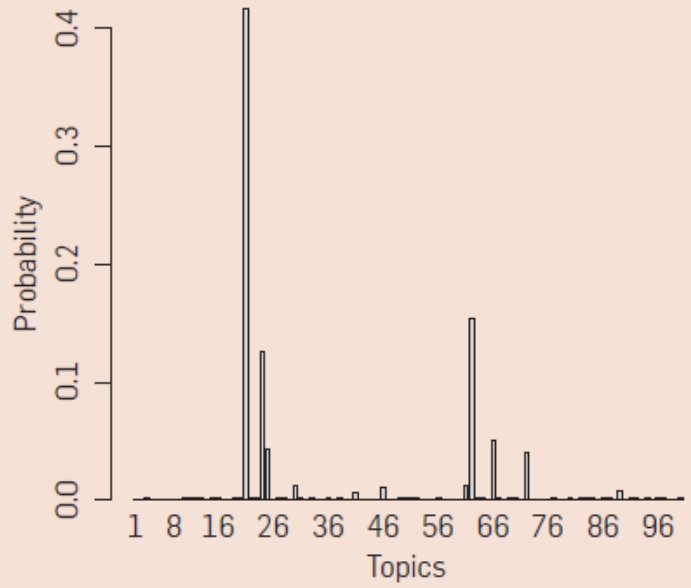
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an




**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

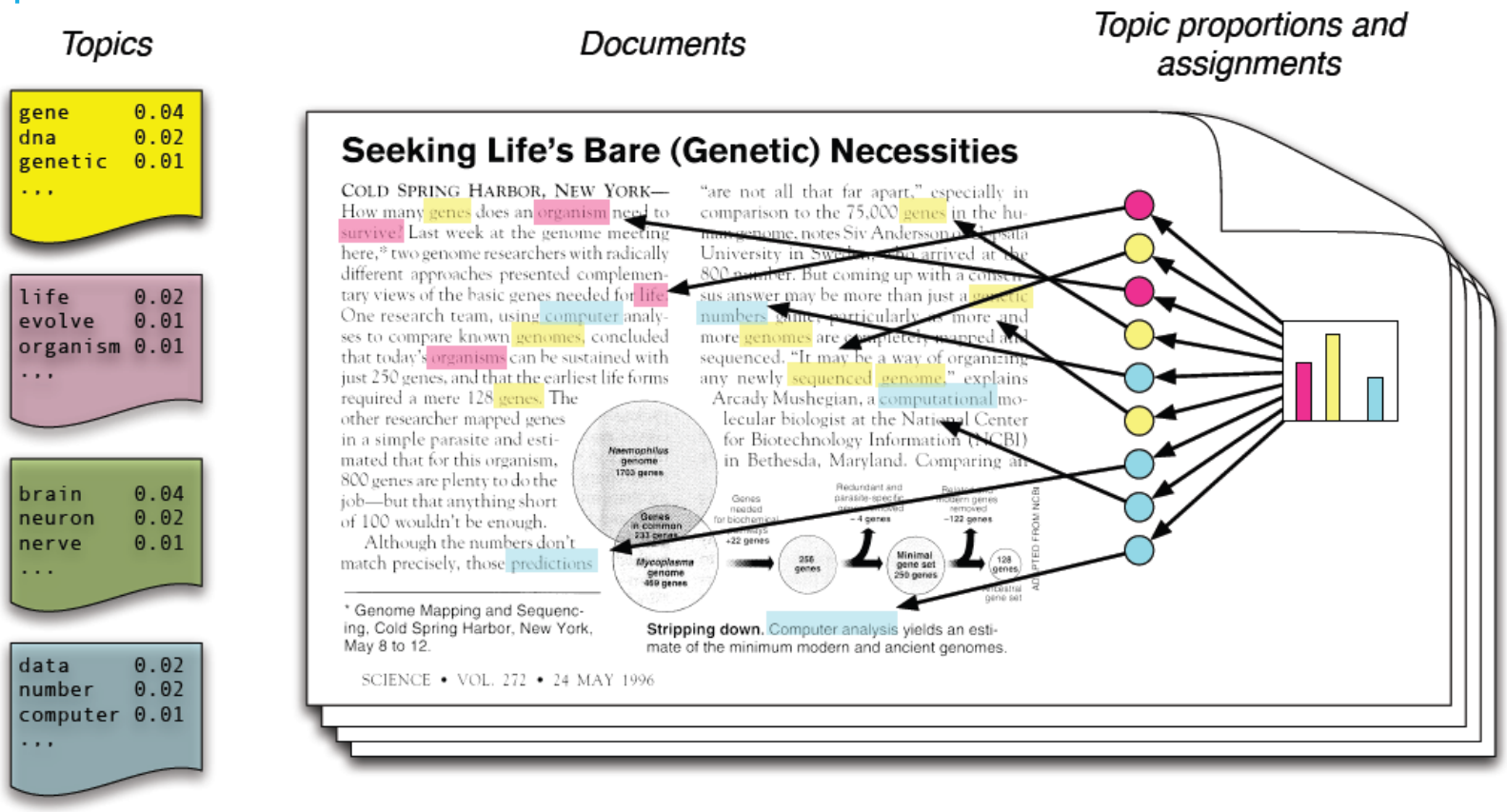
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

# EXAMPLE CONTD…




| "Genetics" | "Evolution" | "Disease" | "Computers" |
|---|---|---|---|
| human | evolution | disease | computer |
| genome | evolutionary | host | models |
| dna | species | bacteria | information |
| genetic | organisms | diseases | data |
| genes | life | resistance | computers |
| sequence | origin | bacterial | system |
| gene | biology | new | network |
| molecular | groups | strains | systems |
| sequencing | phylogenetic | control | model |
| map | living | infectious | parallel |
| information | diversity | malaria | methods |
| genetics | group | parasite | networks |
| mapping | new | parasites | software |
| project | two | united | new |
| sequences | common | tuberculosis | simulations |

# GENERATIVE MODEL



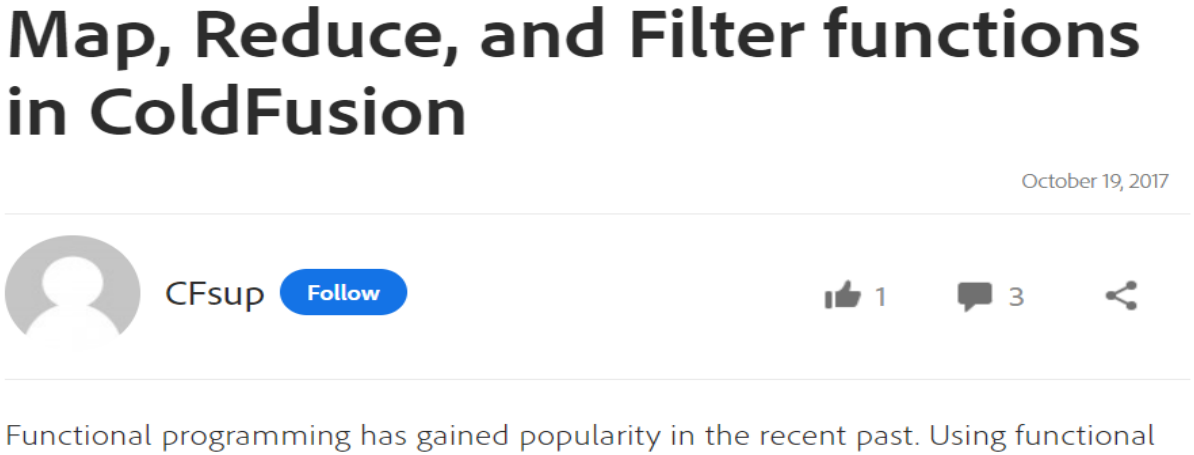

Each topic is a distribution over words
Each document is a mixture of corpus-wide topics
Each word is drawn from one of those topics

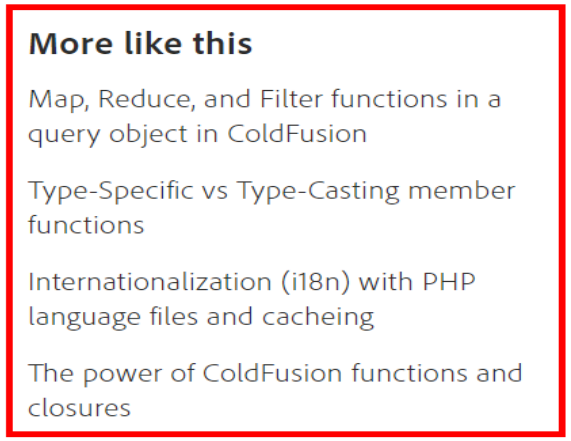
# TOPIC MODELING- USE CASES


## Recommendation Systems

Personalization tools allow filtering large collections of movies, music, tv shows, … to recommend only relevant items to people….probabilistic modeling technique…

- Build a taste profile for a user
- Build topic profile for an item



# USE CASE (SEO)-CONTD…

Search engines use topic modeling to process web pages.

Topic models help consumers (and search engine robots) understand large collections of documents, and process themes throughout each record in that collection.

Topic models take precedence over keyword targeting and use a technique called TF-IDF to measure keyword relevancy.

Establishes semantic association- the focus shifts to user intent.

# SUMMARIZING

Topic modeling refers to writing on a specific topic in depth, and developing related content around it. Ten years ago, you could get away with merely mashing as many keywords as possible into your web pages and content, and that would be enough. However, Google changed its algorithms to emphasize topics over keywords (although keywords still remain as an integral part of SEO).

Currently, Google algorithms, reward content that is relational and builds on accurate pieces of content that go in-depth on a specific topic. The trick to using this algorithm is to link your internal content to pillar pages to illustrate the semantic relationship between the topics you're referring to. The cluster set-up of your topics emphasizes to Google that your content is an authority and its accordingly rewarded with higher ranks in search.

# SEMANTIC SEO AND TOPIC MODELING

1. Pick a longtail target keyword
2. Find semantically similar phrases
3. Include the keywords in your content scope
4. Use the target phrase(s) in the title, header, & body text

**Objective:** Rank for target keywords+topic

# EXAMPLE- TOPIC MODELS

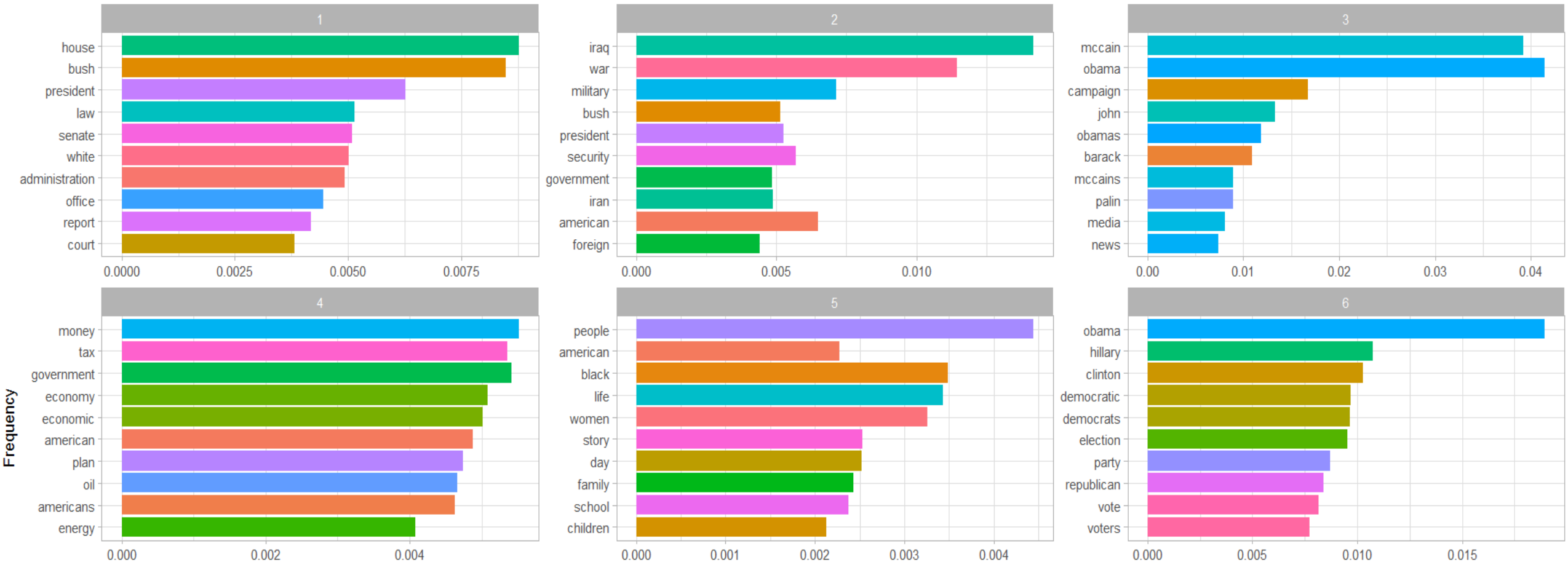
About the data:

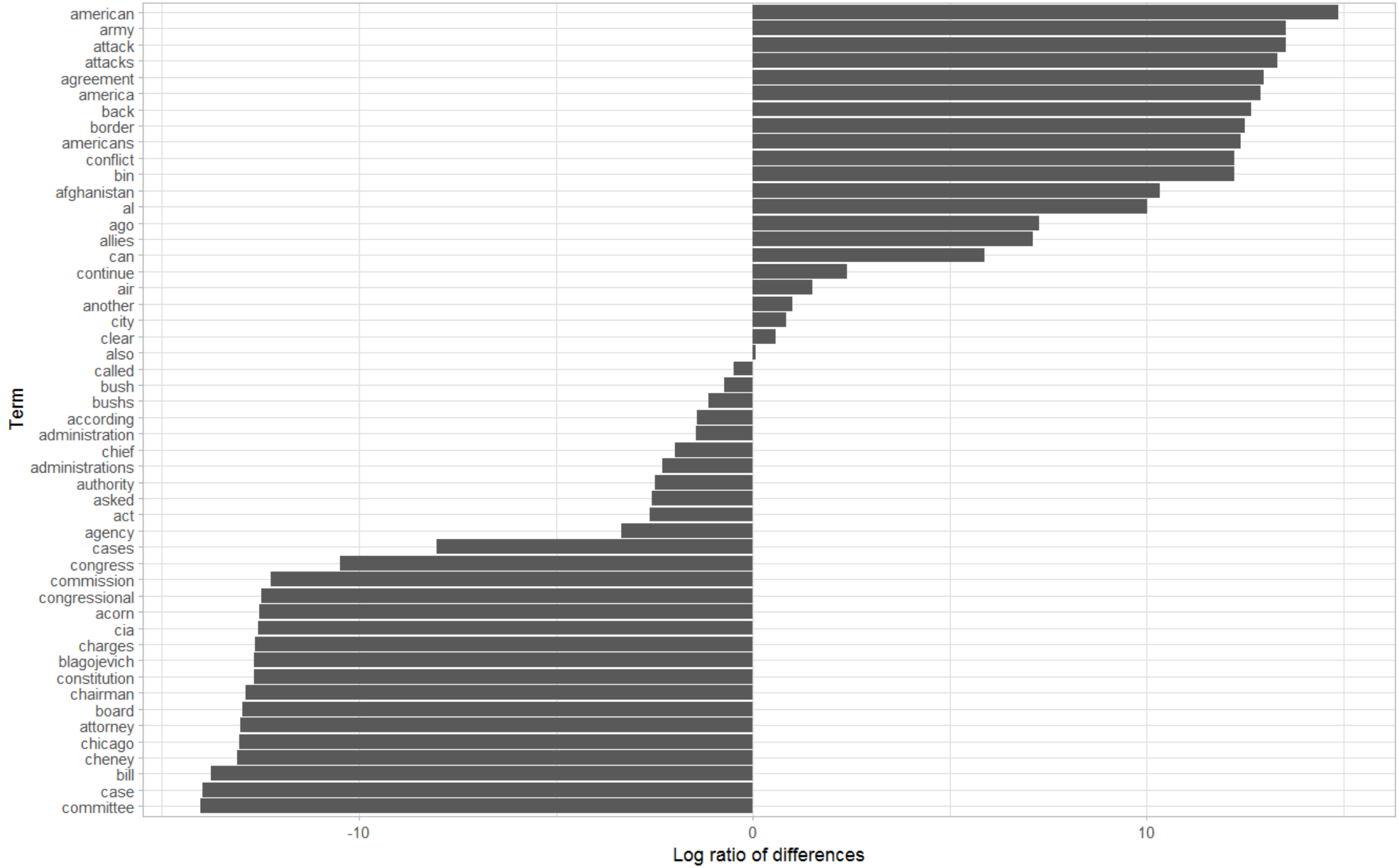
13000+ political blogs collected in 2008, downloaded from Kaggle.

Example:

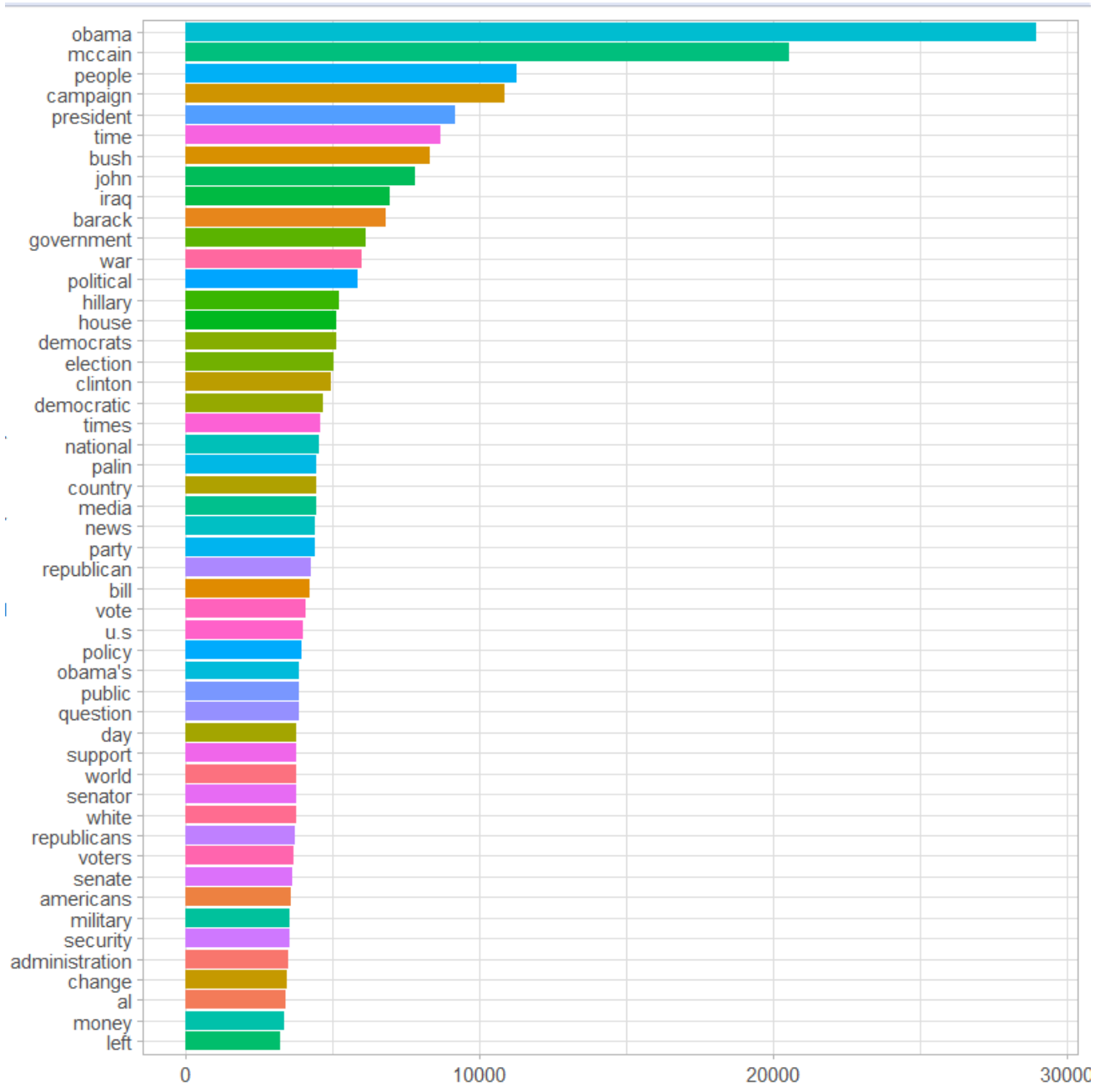
*Mike Huckabee is pretty slick. He's the one Republican who's been boosted by the big media. Those media headlines got him off to a fast start. But a couple of days ago he got the press laughing at his shenanigans with his "I'm NOT negative!" press conference.*

## Word distribution across topics

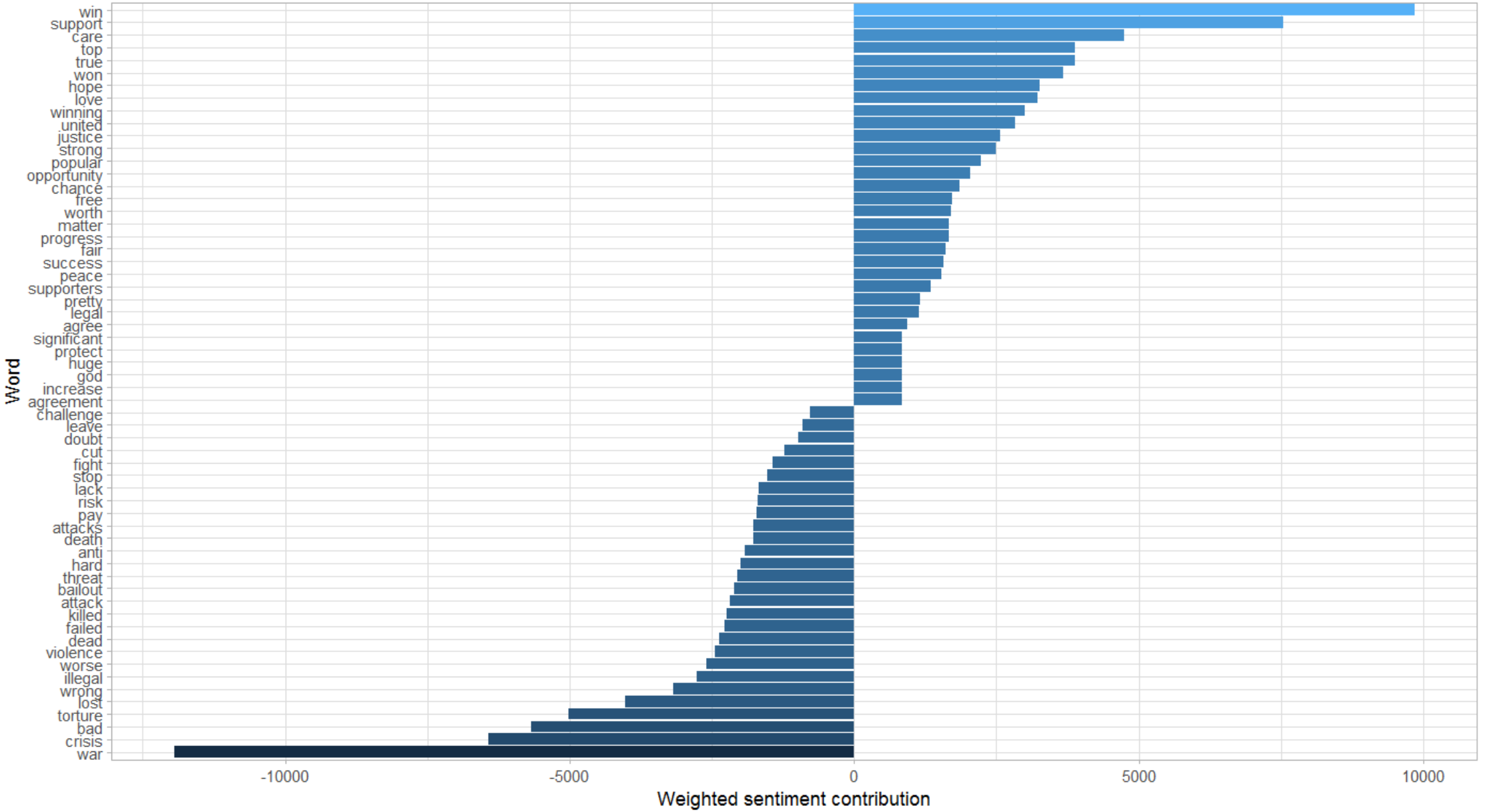



american
army
attack
attacks
agreement
america
back
border
americans
conflict
bin
afghanistan
al
ago
allies
can
continue
air
another
city
clear
also
called
bush
bushs
according
administration
chief
administrations
authority
asked
act
agency
cases
congress
commission
congressional
acorn
cia
charges
blagojevich
constitution
chairman
board
attorney
chicago
cheney
bill
case
committee
Term
-10
0
10
Log ratio of differences

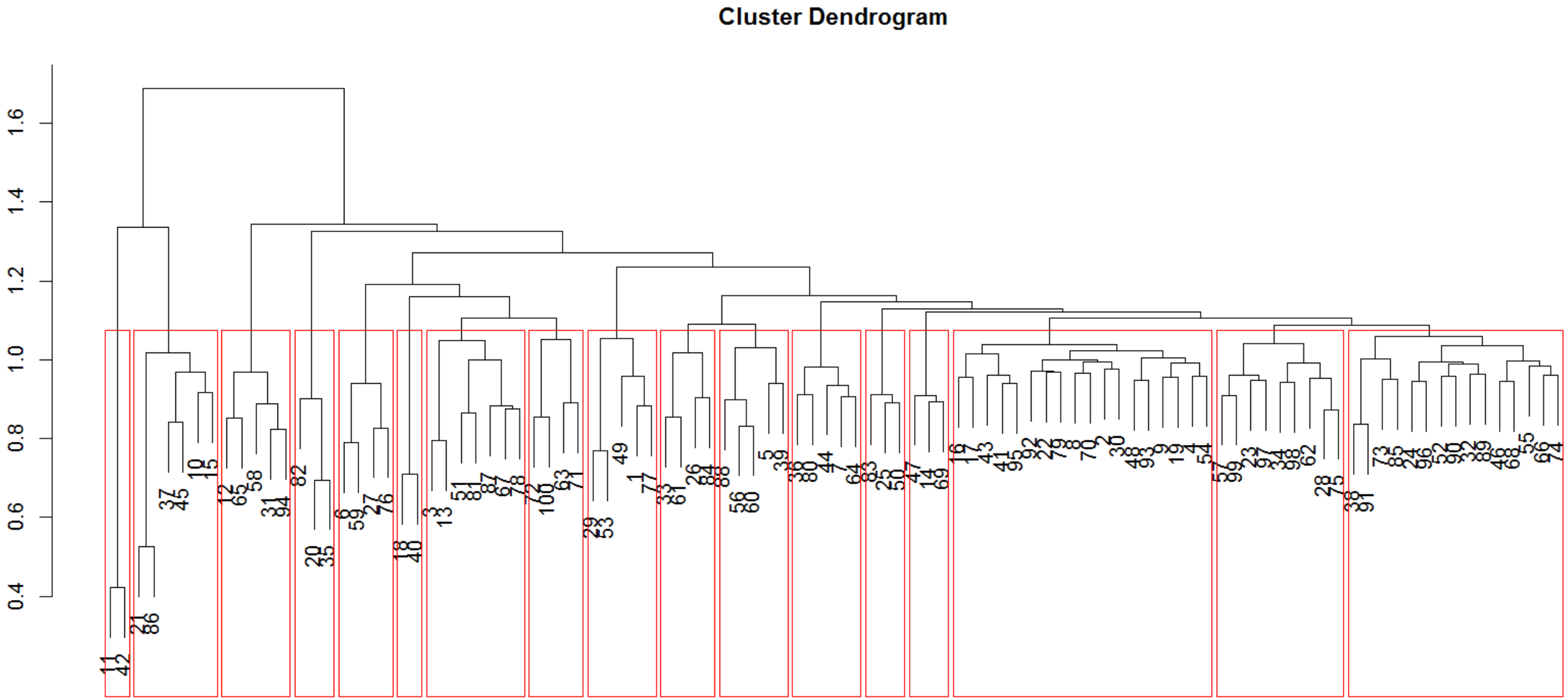
# MOST FREQUENT TERMS

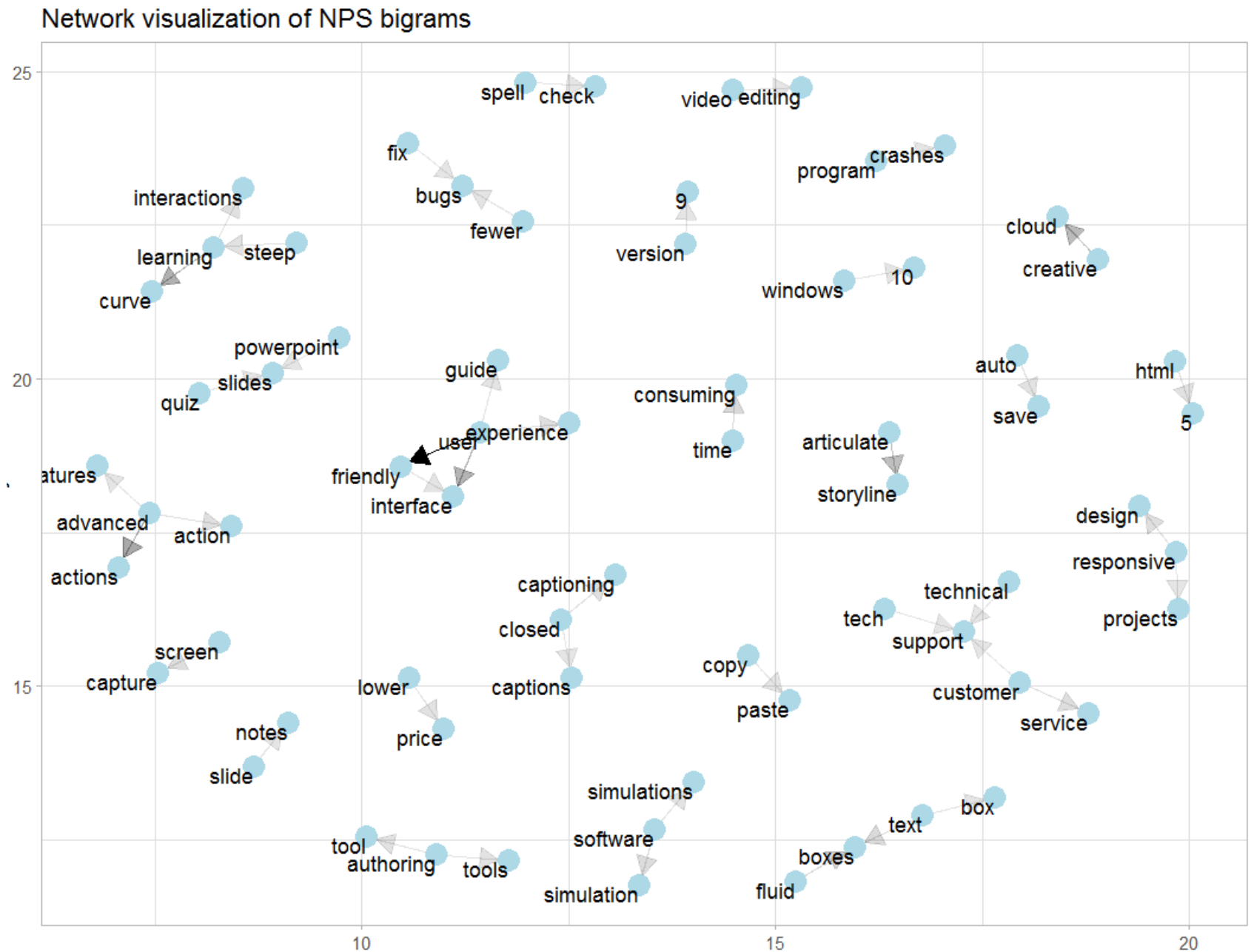


# SENTIMENT ANALYSIS



# TF-IDF SIMILARITY CLUSTERS (100 BLOGS)




Network visualization of NPS bigrams
25
20
15
10
15
20
spell
check
video editing
fix
bugs
fewer
program
crashes
interactions
9
version
cloud
creative
learning
steep
curve
windows
10
powerpoint
slides
quiz
guide
consuming
time
auto
save
html
5
user
experience
friendly
interface
articulate
storyline
atures
advanced
action
actions
design
responsive
projects
captioning
closed
captions
technical
tech
support
screen
capture
lower
price
copy
paste
customer
service
notes
slide
simulations
software
simulation
box
text
boxes
fluid
tool
authoring
tools