Text mining

Weiai Xu (Wayne), PhD
Assistant Professor
Department of Communication, UMass-Amherst
Email: weiaixu@umass.edu
curiositybits.cc

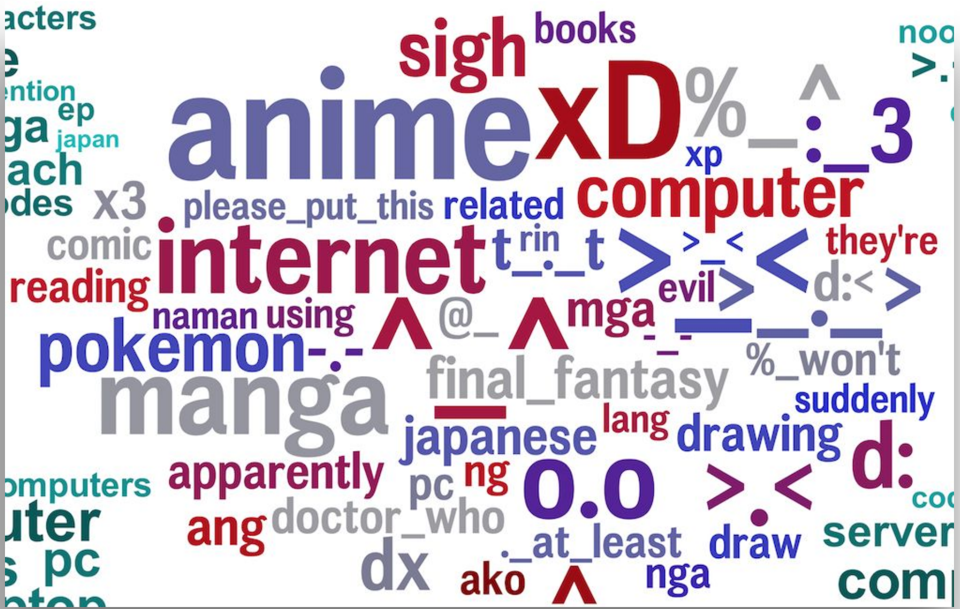
A couple of reminders...

1. Tutorial for W6 and W7 (submit your complete report before 11 PM, March 7th)
2. Assignment 2: Friday, March 8, 2019, 12:00 AM

Related:

https://curiositybits.shinyapps.io/PH_Tracker_dashboard/

Word clouds created on Facebook posts. Whom do you think wrote the posts: extroverts or introverts?



Posted by male or female users?

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791

Insights from text

IBM Watson

https://personality-insights-demo.ng.bluemix.net

Personality Portrait

28145 words analyzed: Very Strong Analysis

Summary

You are particular, explosive and expressive.

You are self-controlled: you have control over your desires, which are not particularly intense. You are adventurous: you are eager to experience new things. And you are dutiful: you take rules and obligations seriously, even when they're inconvenient.

Experiences that give a sense of efficiency hold some appeal to you.

You are relatively unconcerned with both taking pleasure in life and helping others. You prefer activities with a purpose greater than just personal enjoyment. And you think people can handle their own business without interference.

How did we get this?

You are likely to______

- be sensitive to ownership cost when buying automobiles
- like historical movies
- volunteer for social causes

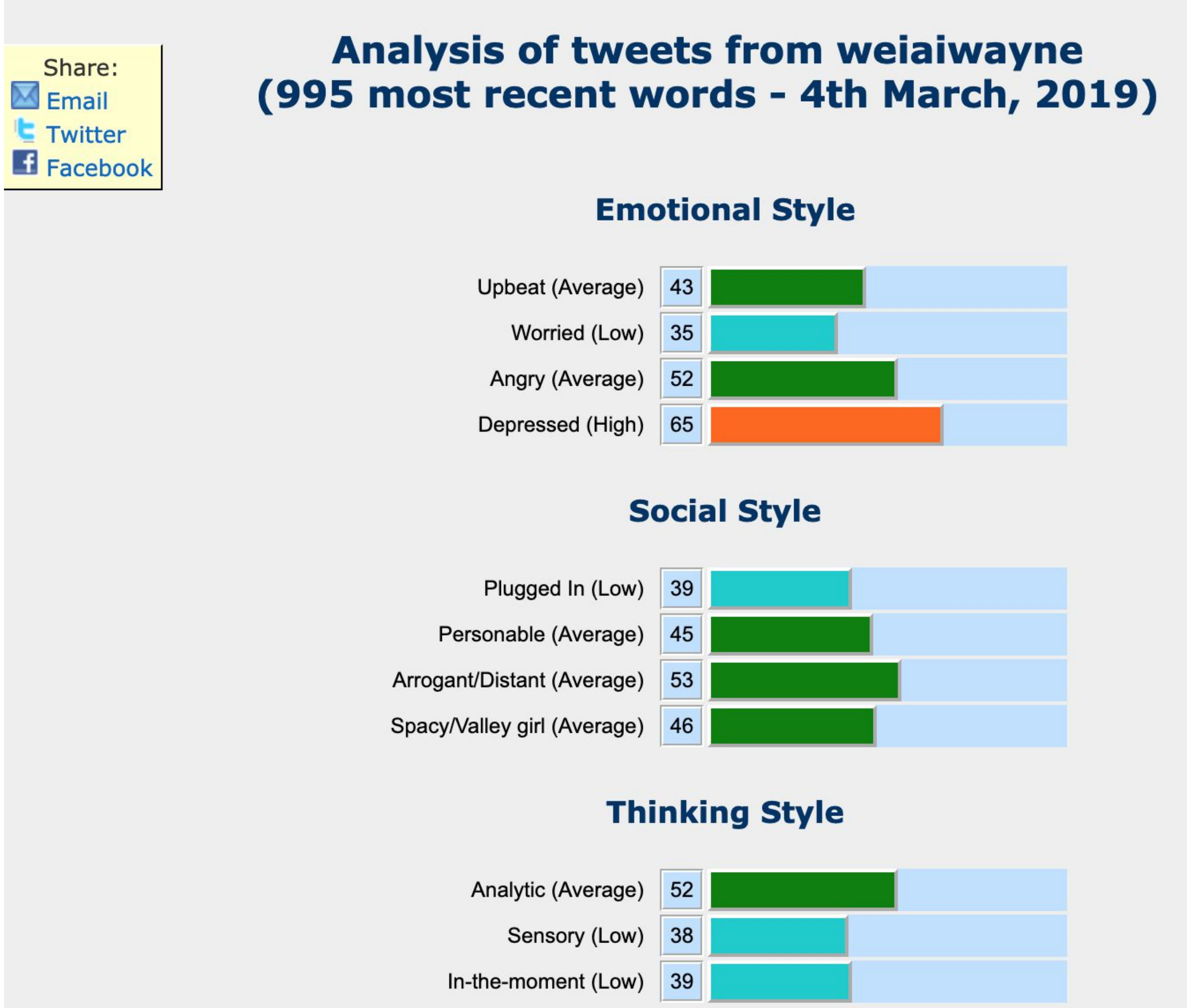
You are unlikely to______

- be influenced by social media during product purchases
- prefer style when buying clothes
- like rap music

Insights from text

Analyze Words

Analyzewords.com



The science behind the algorithm

“A well-accepted theory of psychology, marketing, and other fields is that human language reflects personality, thinking style, social connections, and emotional states. The frequency with which people use certain categories of words can provide clues to these characteristics.”

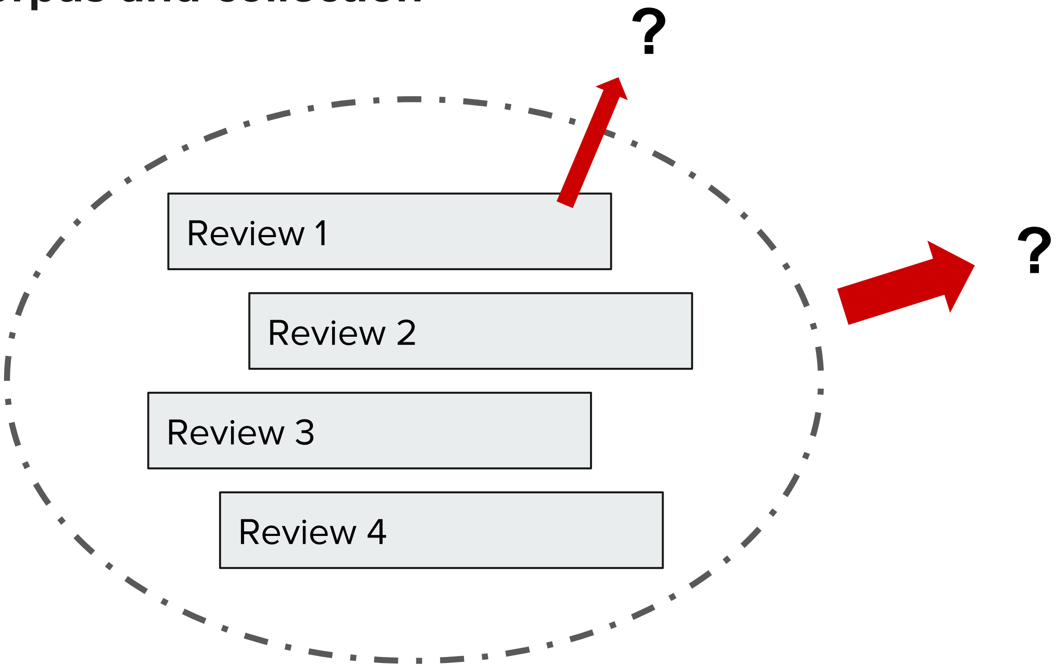
More at
https://cloud.ibm.com/docs/services/personality-insights?topic=personality-insights-science#science

The science behind the algorithm

Search: James W. Pennebaker and Jeff Hancock

Review of concepts

Corpus and collection



Review of concepts

?

features					
docs	awesome	projector	traditional	boston	experience
Boston	120	1	2	862	143
Denver	215	2	0	0	106
Rhode Island	113	0	3	10	158

Review of concepts

?

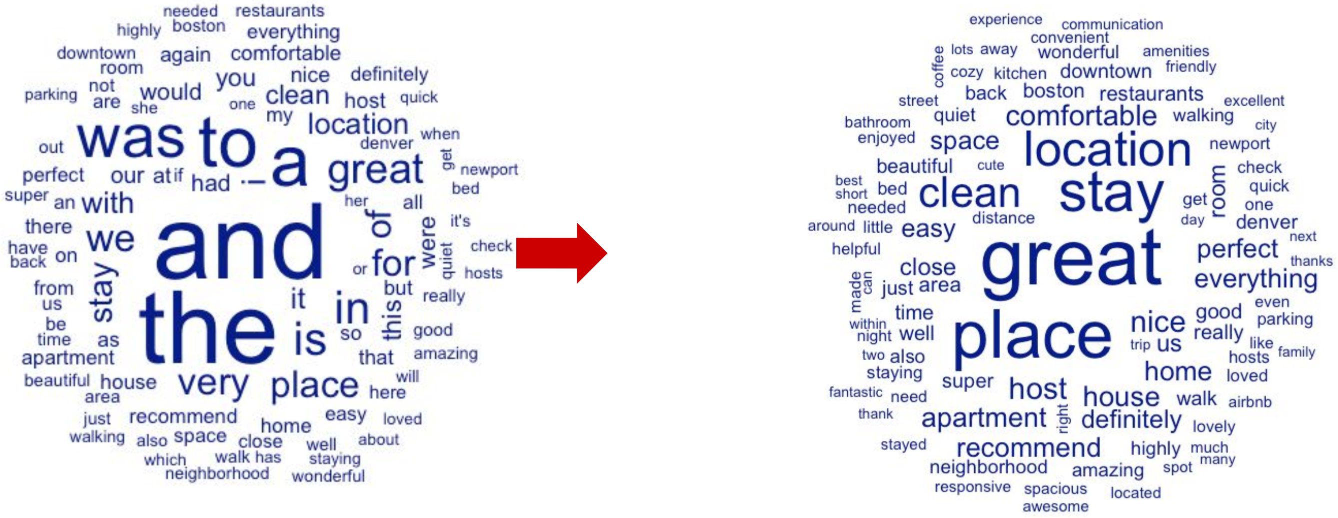
```
"Awesome projector. Traditional Boston experience, with a great location!
```

```
 [1] "Awesome"      "projector"    "."            "Traditional" "Boston"
 [6] "experience"   ","            "with"         "a"           "great"
[11] "location"     "!"
```

New concepts

Stop words

Stop words: filter words because they are extremely common words but appear to be of little value.



New concepts

There are standard stop word lists for most languages
https://stopwords.quanteda.io/

a
about
above
across
after
afterwards
again
against
all
almost
alone
along
already
also
although
always
am

a
actualmente
acuerdo
adelante
ademas
además
adrede
afirmó
agregó
ahi
ahora
ahí
al
algo
alguna
algunas
alguno
algunos
algún
alli
allí
alrededor
ambos
ampleamos
antano
antaño
ante
anterior

a
ab
aber
ach
acht
achte
achten
achter
achtes
ag
alle
allein
allem
allen
aller
allerdings
alles
allgemeinen
als
also
am
an
ander
andere
anderem
anderen
anderer

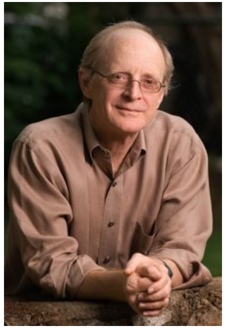
akin
aking
ako
alin
am
amin
aming
ang
ano
anumang
apat
at
atin
ating
ay
bababa
bago
bakit
bawat
bilang
dahil

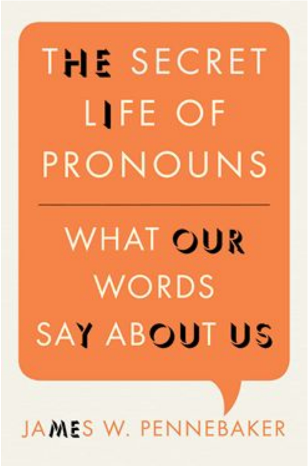
New concepts

A case against filtering stop words

- Smallish words, or function words (articles, prepositions, pronouns), as opposed to content words
- Linguistic Inquiry and Word Count (LIWC)
- People of different gender, age, and social groups, and with different personality traits use function words differently.

> "The more similar [they were] across all of these function words, the higher the probability that [they] would go on a date in a speed dating context," Pennebaker says. "And this is even cooler: We can even look at ... a young dating couple... [and] the more similar [they] are ... using this language style matching metric, the more likely [they] will still be dating three months from now."



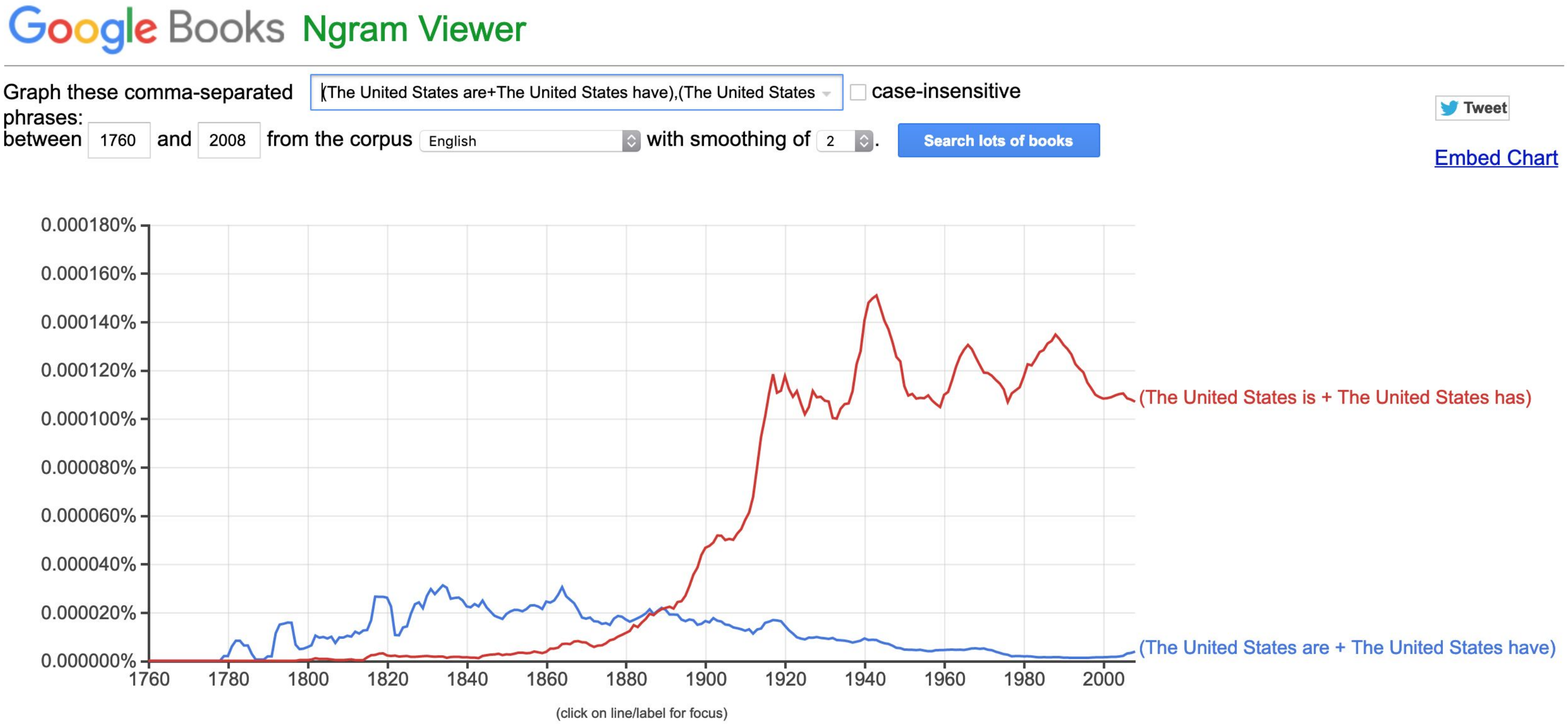


https://www.npr.org/sections/health-shots/2012/04/30/151550273/to-predict-dating-success-the-secrets-in-the-pronouns

New concepts

N-gram: contiguous sequence of n items

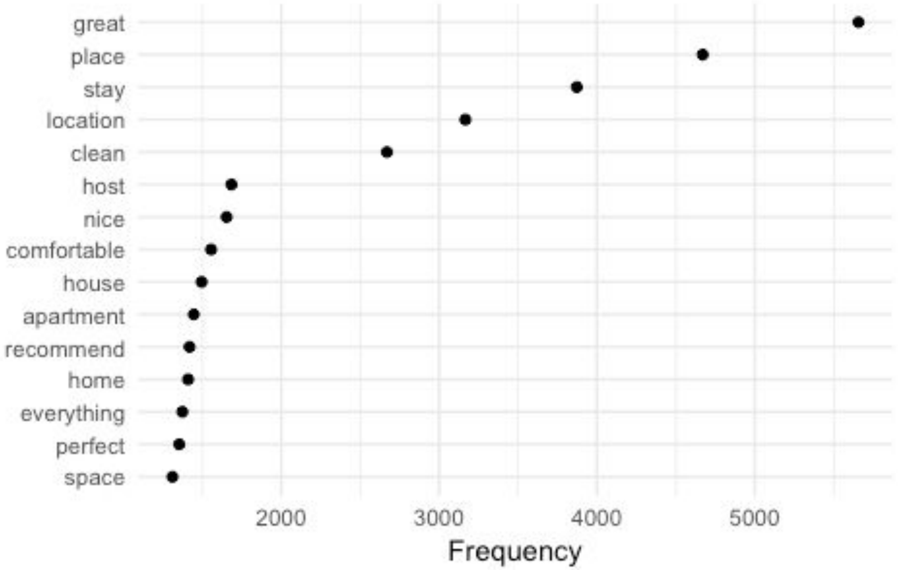
Ngrams

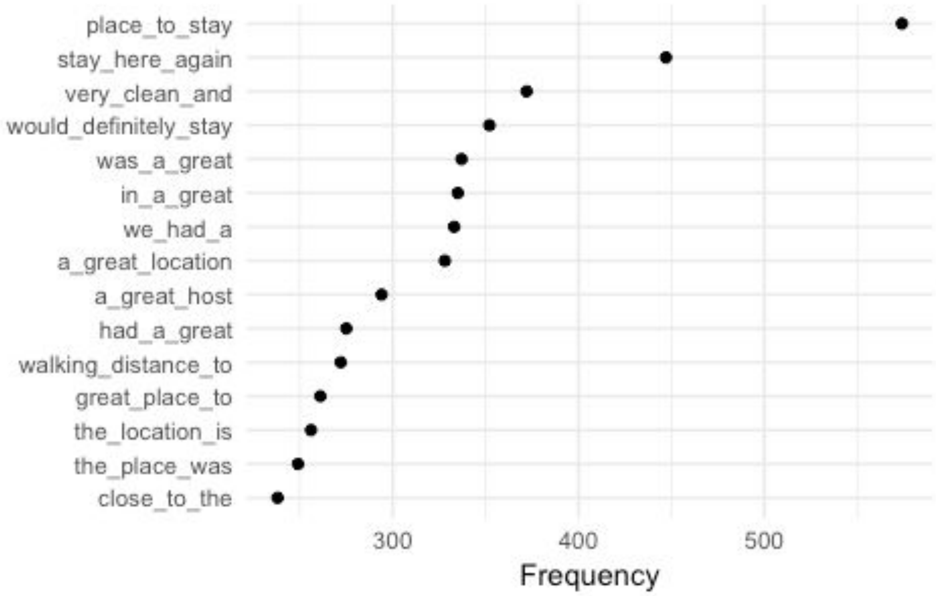


New concepts

N-gram: contiguous sequence of n items

Ngrams

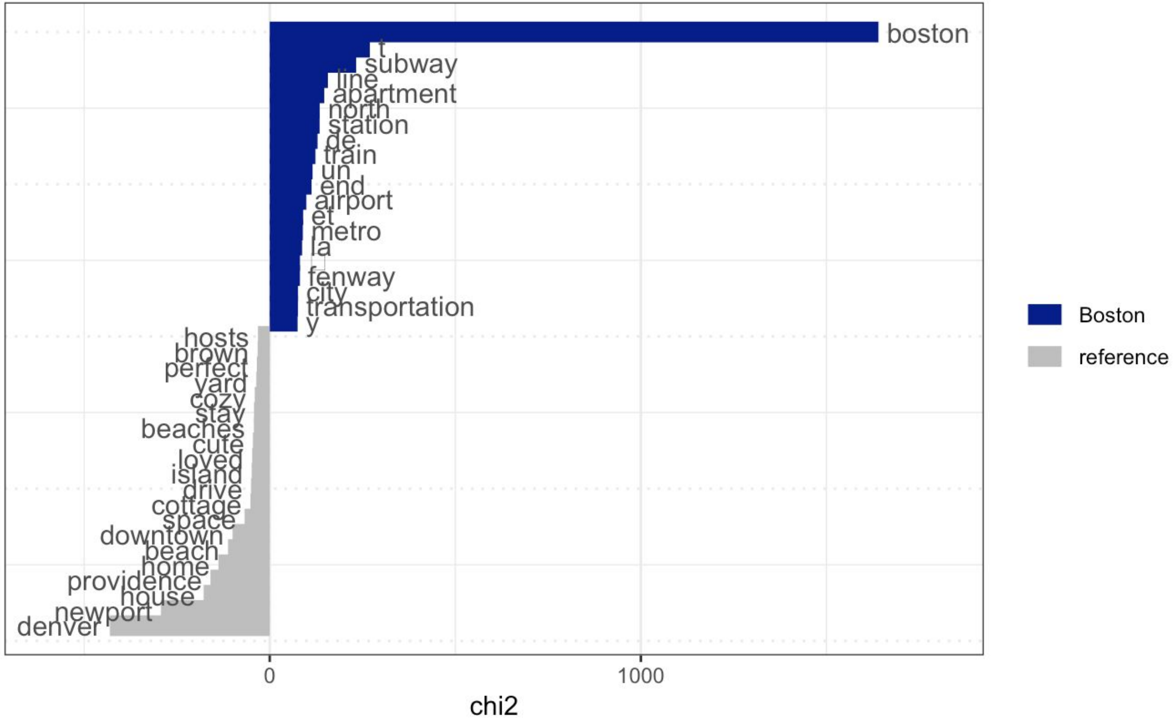




New concepts

Keyness

Keyness words (or key words): words which occur significantly more often in one group of texts than in another



Tf-idf (term frequency-inverse document frequency):

New concepts

Tf-idf (term frequency-inverse document frequency)

A measure of weighting term based on how important a word is to a corpus.

great	place	stay	location	clean	host	nice
5657	4670	3872	3166	2669	1684	1653
comfortable	house	apartment				
1555	1495	1445				

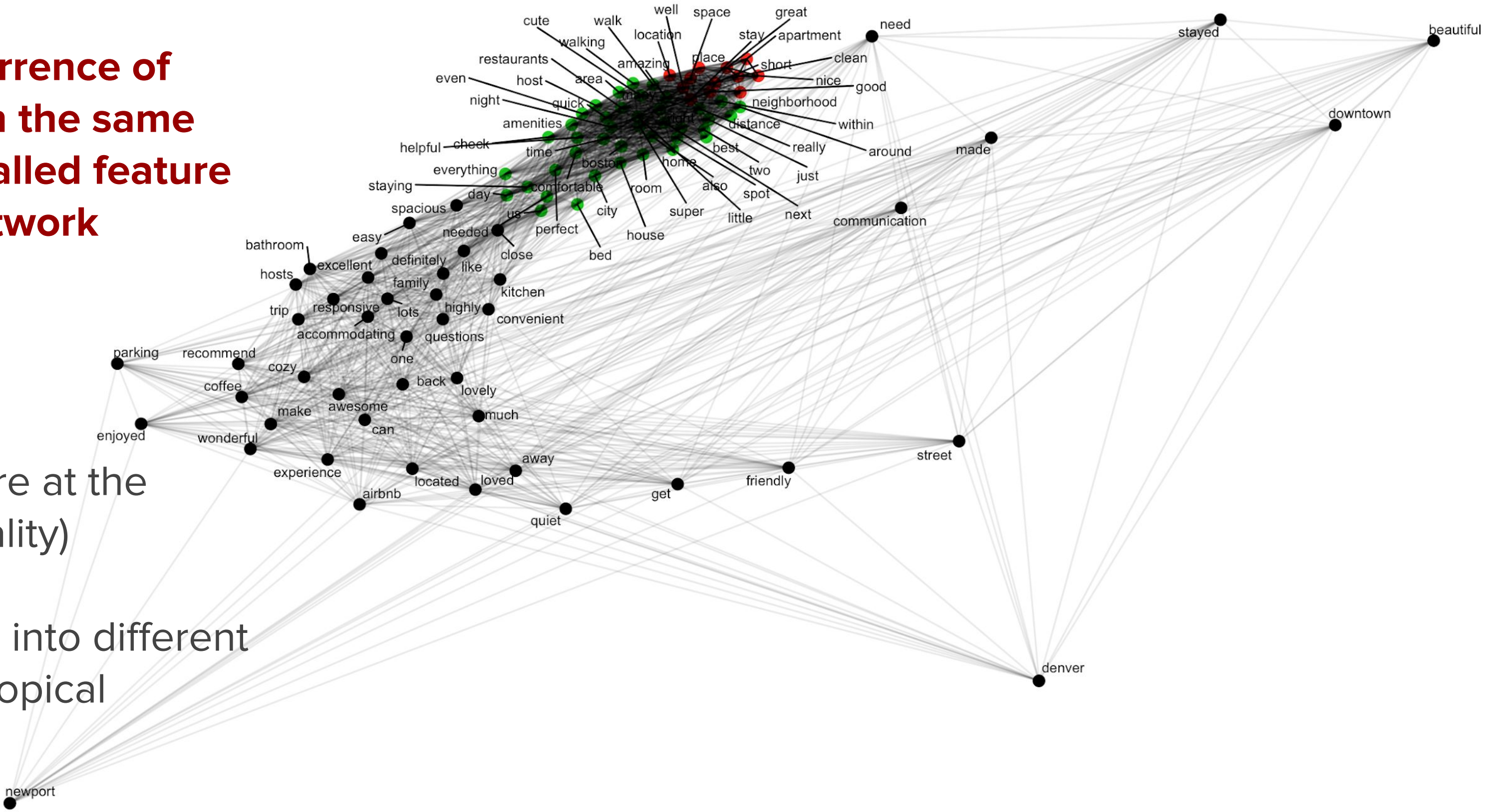
denver	boston	providence	newport	subway	beaches	t
396.96488	153.55158	147.43047	100.90029	57.25455	39.60106	25.53323
thames	fenway	ri				
20.51621	20.03909	20.03909				

Semantic networks

Based on co-occurrence of terms (features) in the same document. Also called feature co-occurrence network

Important words are at the center (high centrality)

Words may cluster into different groups based on topical similarity

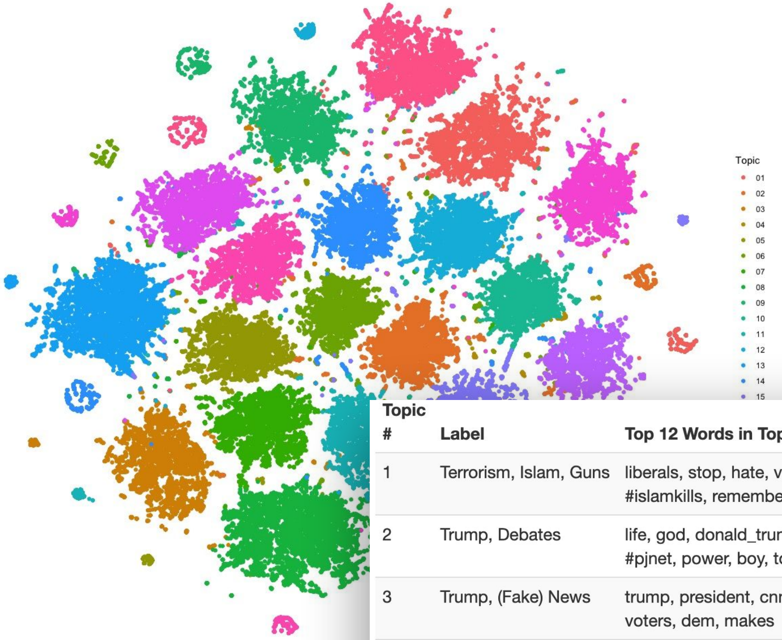


Topic models

Automated clustering of the text based on topical similarity

A live demo based on Russia's IRA tweets

https://www.cascadia-analytics.com/2018/08/12/ira-tweets1.html



Topic #	Label	Top 12 Words in Topic
1	Terrorism, Islam, Guns	liberals, stop, hate, violence, gun, antifa, guns, left, terrorists, people, #islamkills, remember
2	Trump, Debates	life, god, donald_trump, #demdebate, #tcot, plan, #wakeupamerica, #pjnet, power, boy, tonight, candidate
3	Trump, (Fake) News	trump, president, cnn, white_house, report, news, claims, poll, fake_news, voters, dem, makes
4	Taxes, Immigration	america, money, fight, stand, enlist, jobs, pay, join, illegals, free, freedom, tax
5	Obama, Middle East	obama, military, deal, change, speech, president, words, obamas, iran, israel, death, trumps
6	Uninterpretable/Mix	law, racist, government, kids, public, race, takes, debate, california, person, lose, planned_parenthood
7	Congressional votes, Obamacare	trump, gop, bill, democrats, democrat, senate, breaking, republicans, republican, judge, national, obamacare
8	Mueller Investigation	clinton, russia, fbi, comey, hillary, mueller, russian, house, dnc, congress, investigation, breaking
9	Uninterpretable/Mix	liberal, live, day, bad, mt, happy, school, night, city, fire, chicago, won
10	Immigration, Elections	people, american, country, black, america, world, million, voted,

Topic models

LDA (Latent Dirichlet Allocation (LDA) model**) is a commonly used topic modeling algorithm.**

The pitfalls of topic modeling

- Finding the *best* k (k = the number of topics)
- interpretability

	Topic 1	Topic 2	Topic 3
[1,]	"great"	"great"	"great"
[2,]	"place"	"place"	"place"
[3,]	"location"	"stay"	"stay"
[4,]	"stay"	"location"	"location"
[5,]	"clean"	"clean"	"clean"
[6,]	"boston"	"house"	"denver"
[7,]	"apartment"	"host"	"home"
[8,]	"nice"	"comfortable"	"space"
[9,]	"host"	"home"	"house"
[10,]	"comfortable"	"newport"	"nice"
[11,]	"easy"	"nice"	"recommend"
[12,]	"recommend"	"perfect"	"comfortable"
[13,]	"room"	"us"	"host"
[14,]	"everything"	"everything"	"definitely"
[15,]	"definitely"	"recommend"	"everything"
[16,]	"close"	"apartment"	"easy"
[17,]	"good"	"space"	"perfect"
[18,]	"perfect"	"room"	"close"
[19,]	"walk"	"definitely"	"downtown"
[20,]	"really"	"easy"	"super"