

Modern Methods for Exploring Text Data

Peter Baumgartner

Data Scientist @ RTI International

GASP 2019



How is it **tokenizing**?

What tokens is it **excluding**?

Can I differentiate **nouns**, **verbs**, and **adjectives**?

Can I **combine** tokens with the same root word in a meaningful way?

Which tokens are **unique** for this corpus?

Can I see a token used in **context**?

How will this **scale** to larger amounts of text?

Can I **cluster** words by their use?

SAMPLE DATASET

Here's the problem:
We have an entity, and entity can switch ownership. The data can be seen as a time series of events (i.e. ownership change). Of course, events are labeled with epoch time.
So intuitively, we have two features that can help us measure volatility. That is,

- number of changes over time
- the time difference between changes

It is worth mentioning that the time series *can be* very small (2-3 events) - I'm not dealing with high volumed data here.
I'm looking for a way to score "how stable is the entity is". Maybe even a way to answer the question: "what's the probability that a new event will occur in X days".
I'd be glad if you could suggest me ideas / directions / relevant reading materials.
Thanks!

statistics
Join 83,947 readers
137 users here now
This is a subreddit for the discussion of statistical theory, software and application.

Guidelines:

1. **All Posts Require One of the Following Tags in the Post Title!** If you do not flag your post, automoderator will delete it:

Tag	Abbreviation
[Research]	[R]
[Software]	[S]
[Question]	[Q]
[Discussion]	[D]
[Education]	[E]
[Career]	[C]
[Meta]	[M]
2. **This is not a subreddit for homework questions.** They will be swiftly removed, so don't waste your time! Please kindly post those over at: [r/homeworkhelp](#). Thank you.
3. Please try to keep submissions on topic and of high quality.
4. Just because it has a statistic in it doesn't make it statistics.
5. Memes and image macros are not acceptable forms of content.
6. Self posts with throwaway accounts will be deleted by AutoModerator

Related subreddits:

- [r/askstatistics](#)
- [r/biostatistics](#)
- [r/machinelearning](#)
- [r/probabilitytheory](#)
- [r/rstats](#)
- [r/econometrics](#)
- [r/dataisbeautiful](#)
- [r/sas/](#)

Post titles and text from the *r/statistics* and *r/askstatistics* communities on reddit from December 2015 – March 2019.

30,693 total posts.

61% from r/statistics
39% from r/askstatistics

TOKENIZATION

"Can anyone tell me what a p-value is?"

TOKENIZATION



Can anyone tell me what a p-value is ?

TOKEN ATTRIBUTES

A blue square containing the lowercase text 'is' in white.

Lemma: be

POS: VERB

Prob: 0.0088

Can

anyone

tell

me

what

a

p-value

is

?

DATA PROCESSING & LEMMA STATISTICS

sample

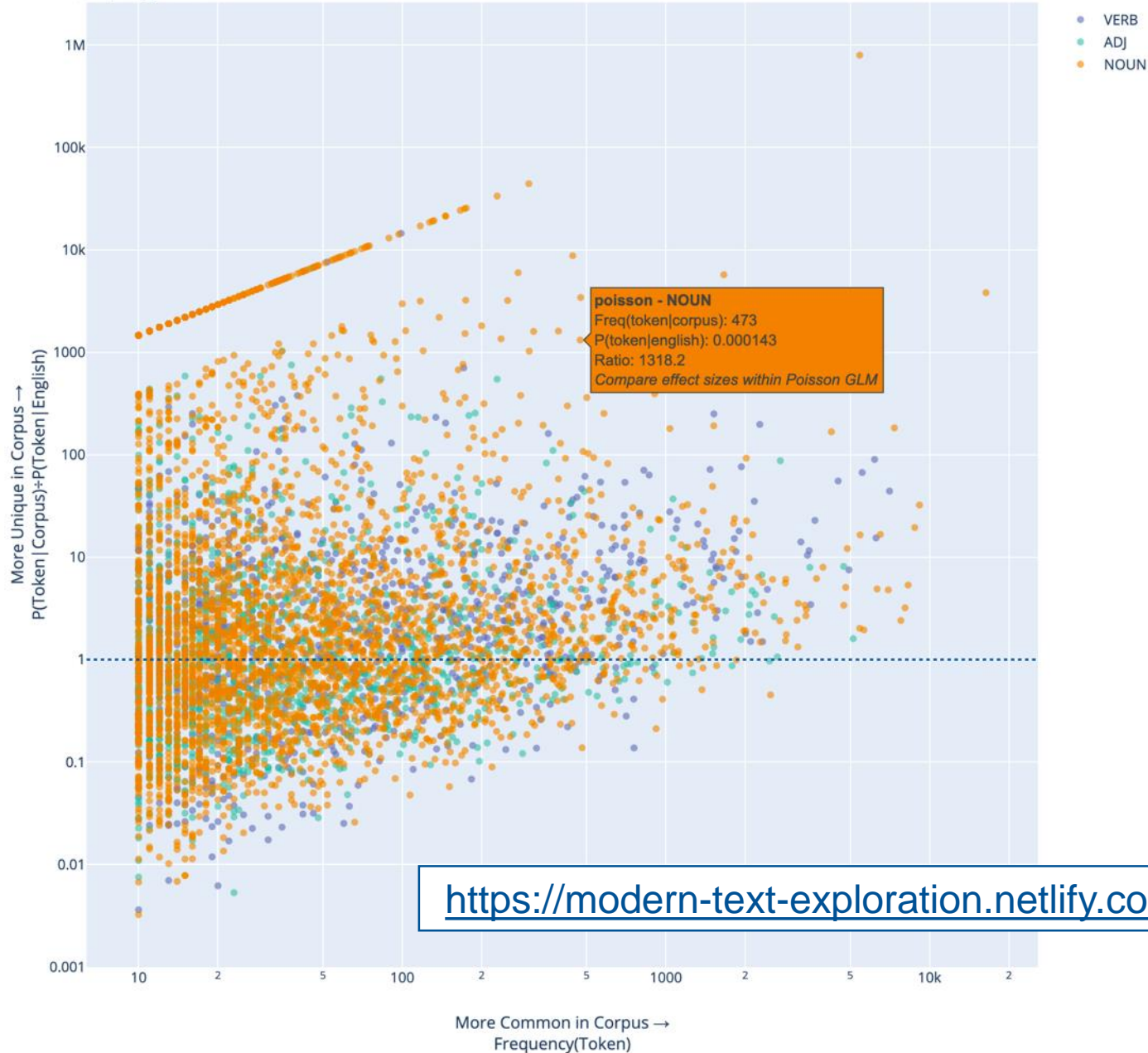
Lemma	POS	Stopword	Corpus Count	Corpus Prob	English Prob	Corpus/English Ratio
manual	NOUN	FALSE	29	0.000009	0.000023	0.388
injure	VERB	FALSE	13	0.000004	0.000001	7.033
methods	NOUN	FALSE	58	0.000018	0.000277	0.063
irregular	ADJ	FALSE	12	0.000004	0.000015	0.243
forests	NOUN	FALSE	10	0.000003	0.000016	0.193

n lemmas = 74,358

Statistics Subreddits - Lemma Statistics (n=5749)

Lemmas appearing at Least 10 times in Corpus

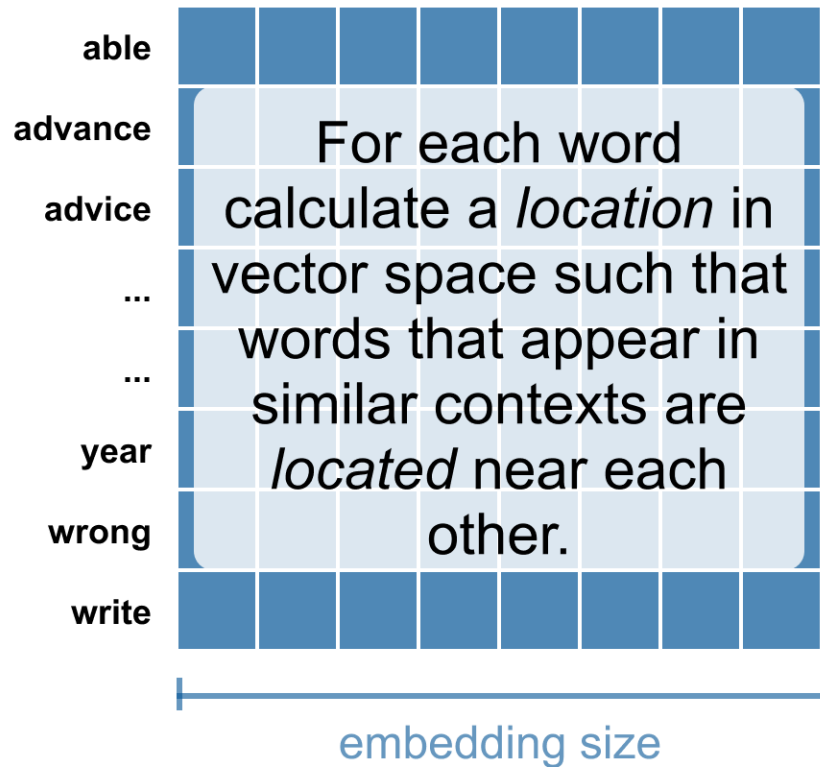
NOUN, ADJ, VERB



Interactive Visualization 1: Exploring lemma counts and uniqueness by parts of speech

WORD EMBEDDINGS

Word Embeddings

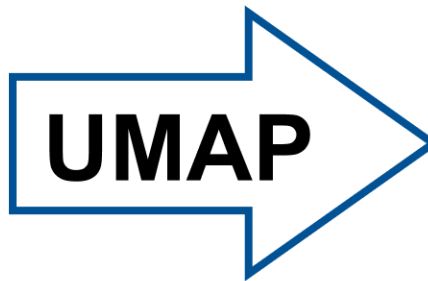


more on word2vec: <https://jalammar.github.io/illustrated-word2vec/>

WORD EMBEDDINGS & DIMENSION REDUCTION

Word Embeddings

able							
advance							
advice							
...							
...							
year							
wrong							
write							



2D Projection

able		
advance		
advice		
...		
...		
year		
wrong		
write		

projection dimensions

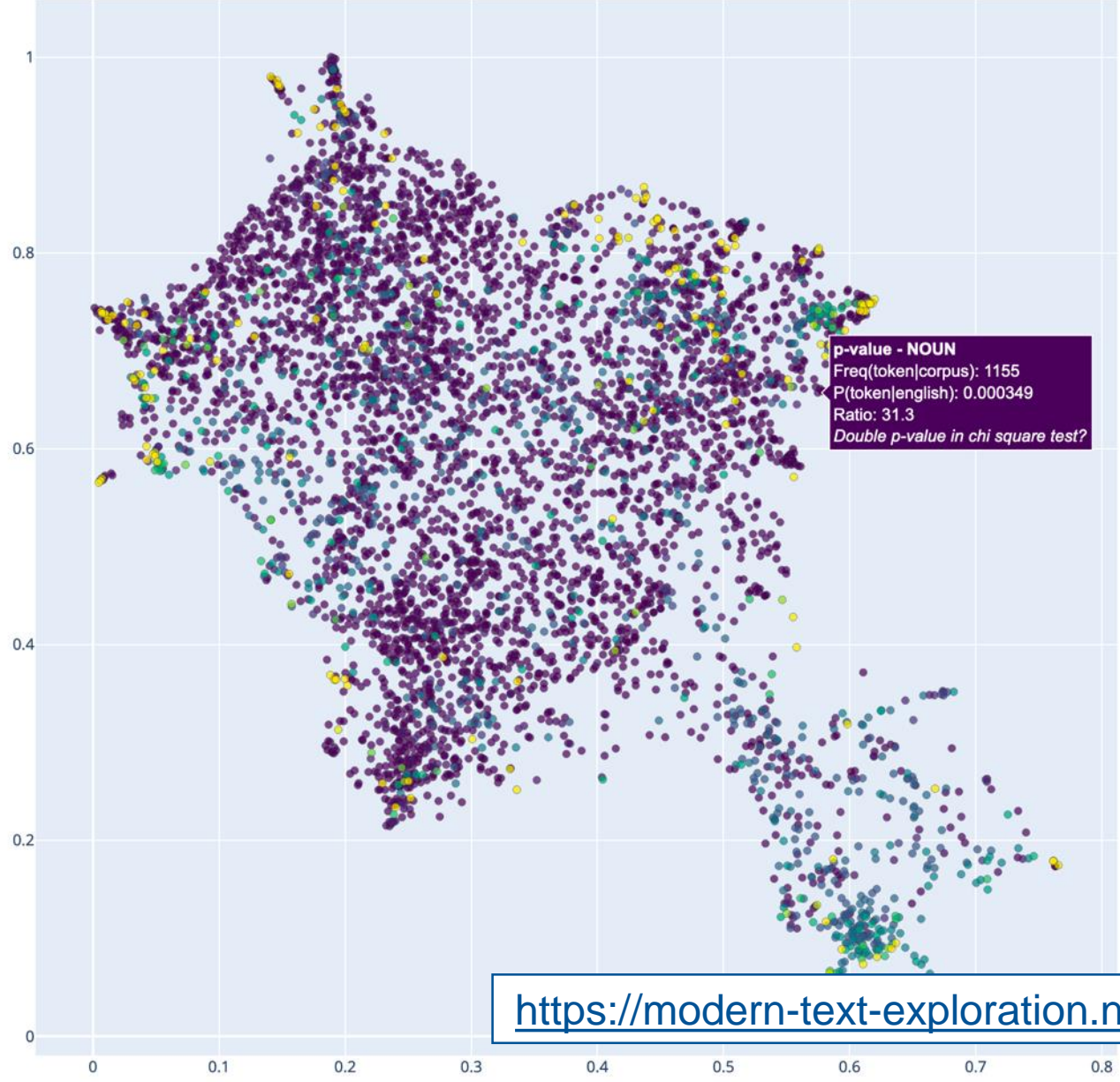
Project an n -dimensional space down to 2 dimensions, such that both local and global structure is retained.

MERGE PROJECTION BACK TO DATASET

sample

Lemma	POS	...	Component 1	Component 2
manual	NOUN	...	0.934	0.734
injure	VERB	...	0.723	0.222
methods	NOUN	...	0.147	0.063
irregular	ADJ	...	0.237	0.243
forests	NOUN	...	0.717	0.182

Statistics Subreddits - word2vec UMAP (n=5428)
Corpus/English > 1; NOUNS, VERBS, ADJ
Colored by Corpus/English Ratio



<https://modern-text-exploration.netlify.com/w2v-umap.html>

Interactive Visualization 2: Exploring projections of word embeddings from word2vec

modern-text-exploration.netlify.com

Slides

Notebook with code

Visualizations 1 & 2

Resources

