

# Modern Methods for Exploring Text Data

Peter Baumgartner

Data Scientist @ RTI International

GASP 2019



Word cloud of tweets from Science Online 2010, generated using Wordle (<http://wordle.net>).  
<https://www.flickr.com/photos/sjcockell/4963334783>



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?

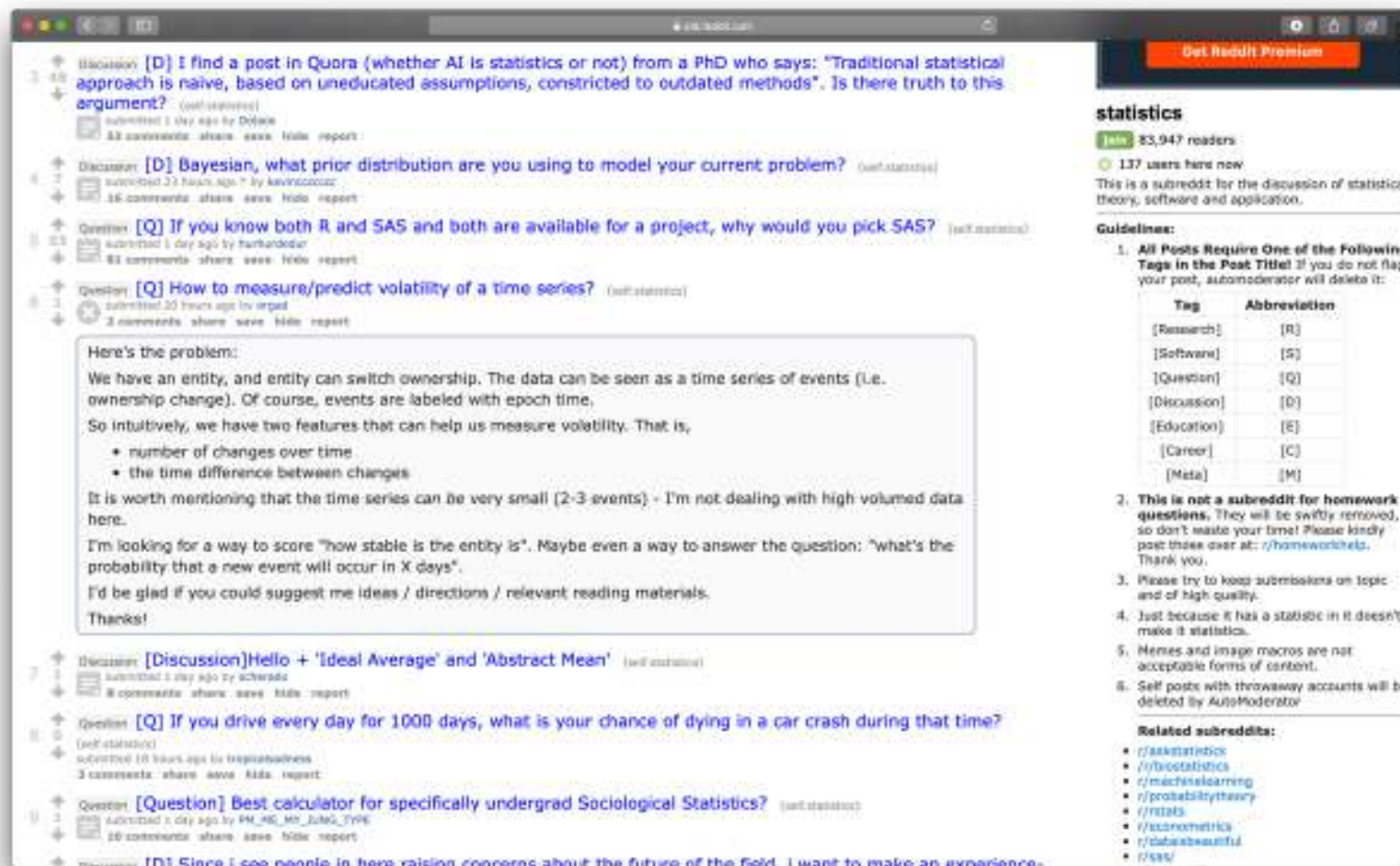
Which tokens are **unique** for this corpus?

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



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

?

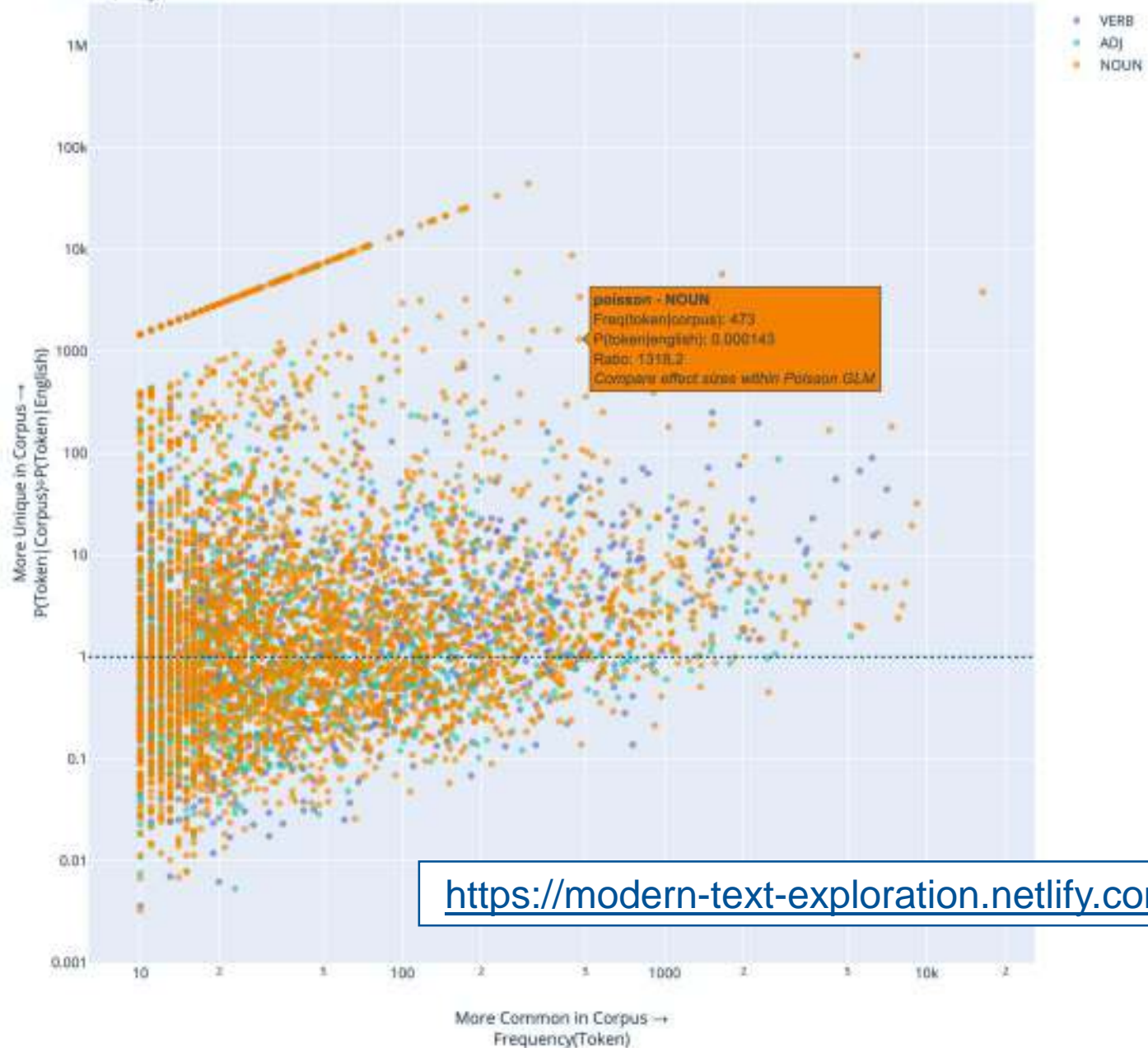
# DATA PROCESSING & LEMMA STATISTICS

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

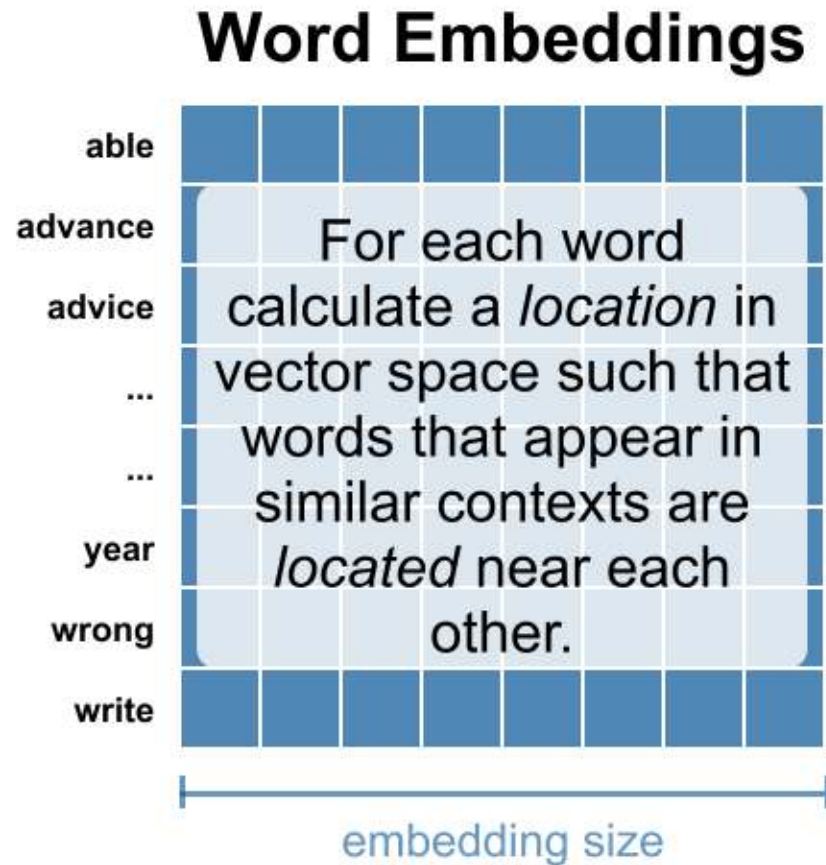


# Interactive Visualization 1: Exploring lemma counts and uniqueness by parts of speech

<https://modern-text-exploration.netlify.com/token-statistics.html>



# 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

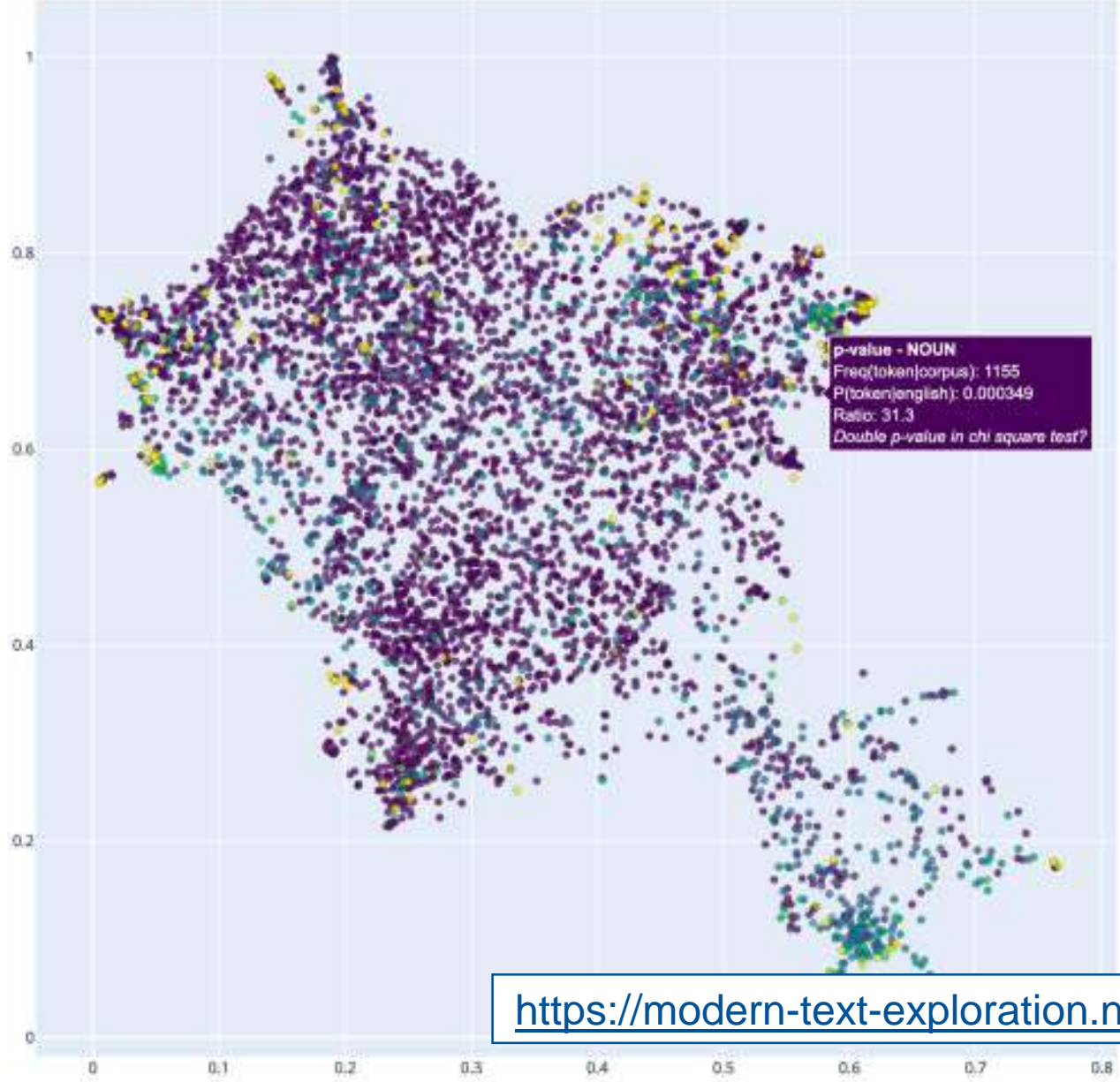
Project an  $n$ -dimensional space down to 2 dimensions, such that both local and global structure is retained.

# 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](https://modern-text-exploration.netlify.com)

**Slides**

**Notebook with code**

**Visualizations 1 & 2**

**Resources**

