CHICAGO
INTERNATIONAL
SOFTWARE DEVELOPMENT
CONFERENCE 2016

goto; conference

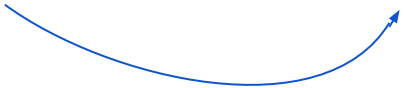
Language as an Interface

Spencer Kelly

follow us @gotochgo

Conference: May 24th-25th / Workshops: May 23rd & 26th

The pope is catholic.



language as data

language as an interface

(@spencermountain)

nlp_compromise JavaScript

Created by nlp-compromise

Star

a cool way to use natural language in javascript

nlpcompromise.com

231 FORKS 5.2K STARS

110kb

js file

```
npm install nlp_compromise
```

or

```
<script src='http://cdn.nlpcompromise.com/nlp_compromise.min.js'/>
```

part-of-speech tagging 86%

us-uk localization 96%

verb conjugation 89%

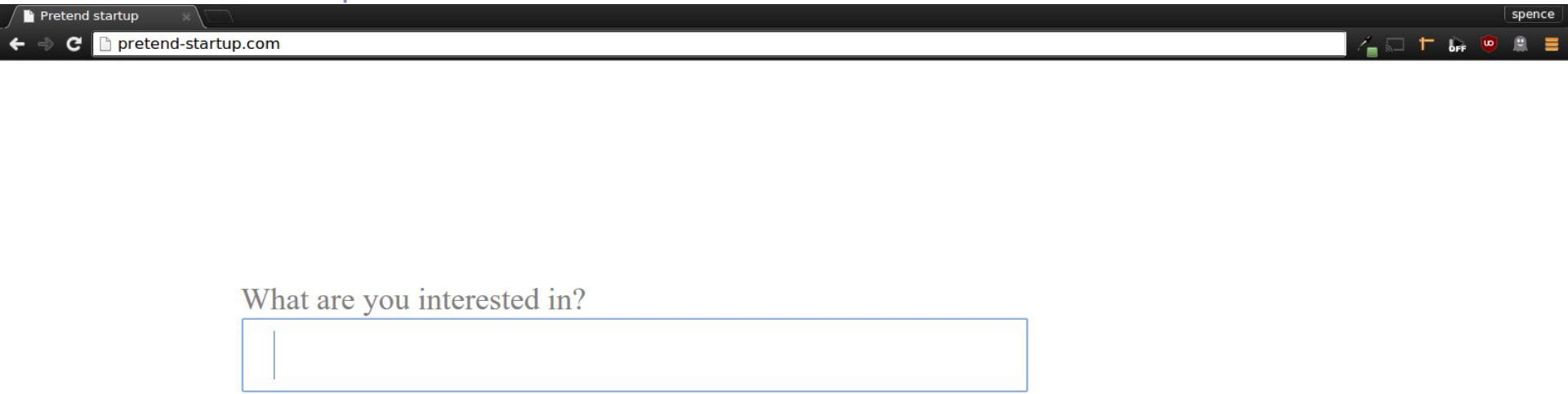
Recommend Share Sort by Best

Join the discussion…

Erok54 · 5 years ago

as someone who has been in the car service business his whole life (family business in 3 states) a valuation that high is asking for trouble. we have been in business for 30 years doing thousands of rides per year, we would be lucky to get 3-4 x ebdita. The largest limo company in the world is carey (carey.com) and they do 300 million per year. there is not much growth in the industry and the margins are very low. I wish them luck, however it is a very mom and pop industry and you will be lucky to just make a "living." and to pull the technology card, limores.net does around 20 million per year (they are a hybrid tech and car service company) and groundtravel.com does 8 million per year and they are a dedicated technology solution. Maybe Uber knows something i dont know, but if Uber becomes the most successful one out there, you are not looking at a billion dollar company by any means. i strongly believe they are barking up the wrong tree.

18 · Reply · Share ›


Pretend startup
spence
pretend-startup.com
What are you interested in?

What are you interested in?

scarlitt johannssin

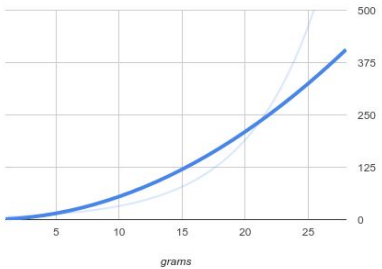
What are you interested in?

london in the rain

What are you interested in?

london in the rain

4-gram:	3-gram:	2-gram:	1-gram:
london in the rain	london in the	london in	**london**
	in the rain	in the	in
		the rain	the
			rain



4-words -

10 requests per keystroke

5-words : 15 **6-words :** 21 **7-words :** 28 **8-words :** 36

problem

Stopwords blacklist:

Edge gram filter:

Redundancy check:

london in the rain

london in the

in the rain

london in

in the

the rain

London

in

the

rain

london in **the**

in the rain

london **in**

in **the**

in

the

the rain

#1

“london in the rain”

#2

“london”

#3

“rain”

When all you've got is a jackhammer..

- **NLTK** - *excellent, huge,* python
- **Stanford parser** - *excellent, huge,* java
- **Freeling** - *excellent, huge,* C++
- **Illinois tagger** - *excellent, huge,* java

Or an offsite API?

Alchemy,

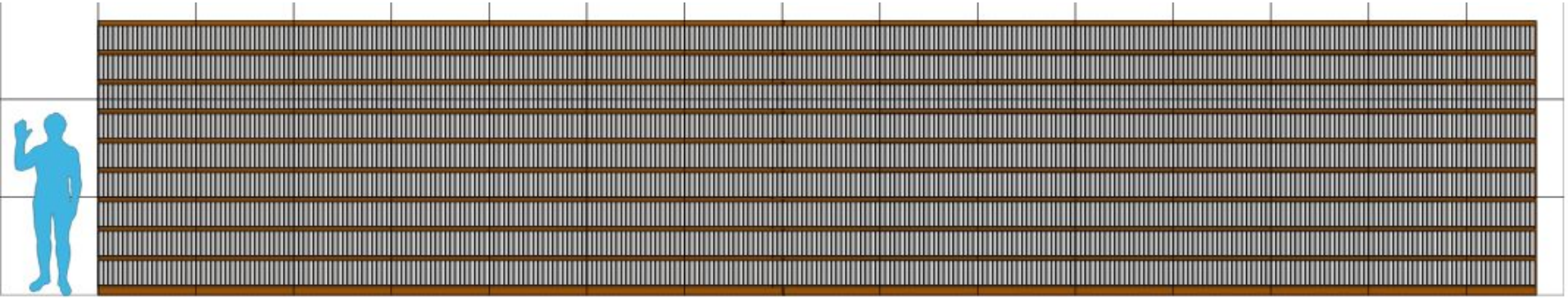
TextRazor,

OpenCalais,

Embedly,

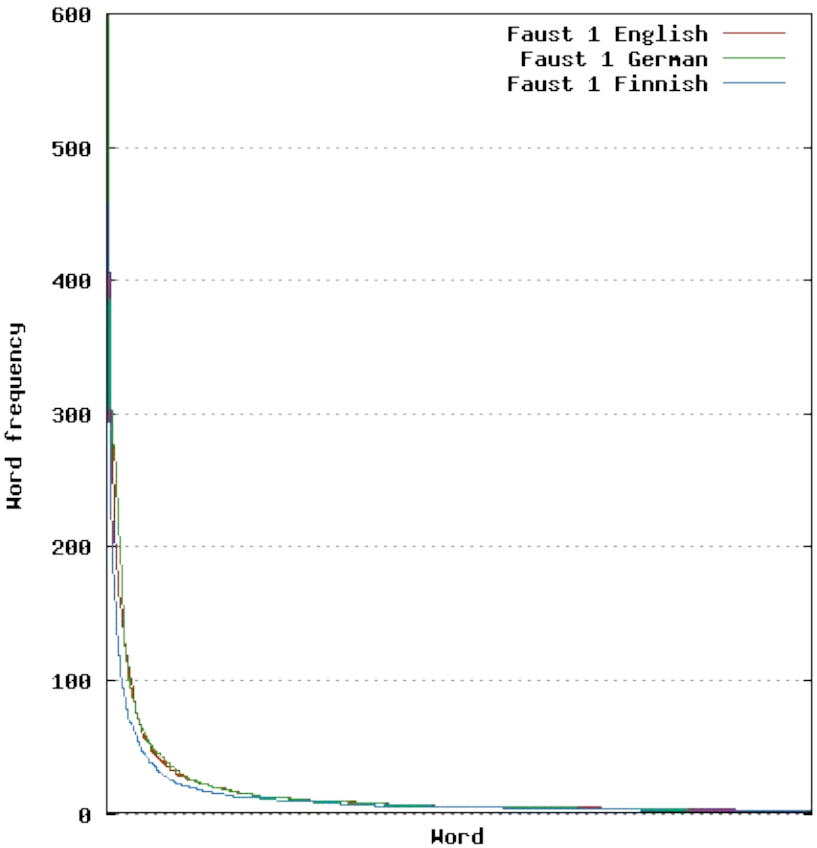
Zemanta

Can it be hacked?



tldr: yes. 💪

Zipfs law

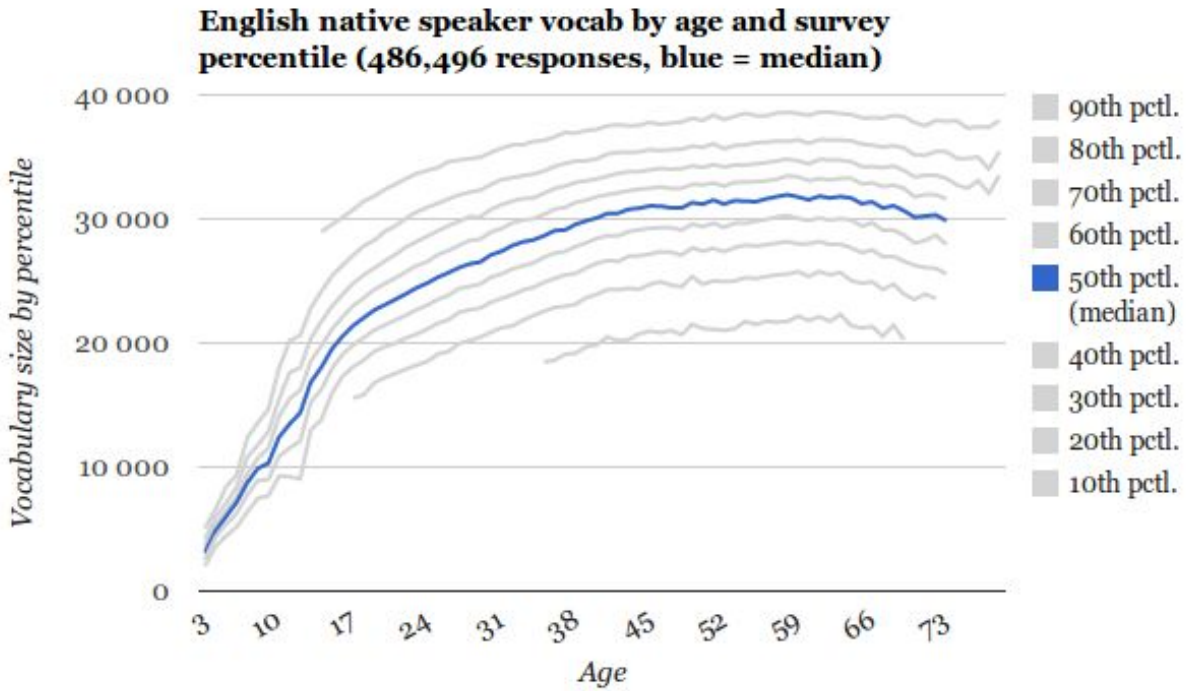


The top 10 words account for 25% of language.

The top 100 words account for 50% of language.

The top 50,000 words account for 95% of language.

How big is a language?



Shakespeare - 35,000

Wordnet - 200,000 !

OED - 600,000 !

An average person will ever hear

50,000 different words

602 kb

uncompressed

~4 lookups in binary search

Vocabulary of Wordnet

first, let's kill the nouns 70%

180 kb

uncompressed

niche

improveify your vocabularies

Noun	Verb	Adjective	Adverb
Tomato Tomatoes Toronto Torontonian *not economics	Speak Spoke Speaking will speak have spoken had spoken ... *not is	nice nicer nicest *not handsome	quickly quicklier quickliest *not truly

“tomato” — “tomatoey”

“agressiveness” — “aggressive”

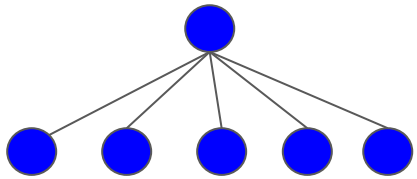
“civilize” — “civil”

“speaker” — “speak” — “quick” — “quickly”

“awesomeify” — “awesome”

n/2.3

Each word



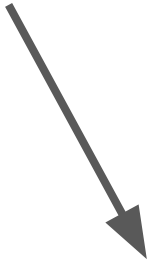
then, let’s conjugate our verbs

110 kb

uncompressed

110 kb

uncompressed



the whole english language
110kb

jQuery
256kb

d3js
330kb

lodash
503kb

react
653kb

Ok, let’s roll our own POS tagger..

(what could go rong?)

1) Lexicon
2) Suffix regexes
3) Sentence-level markov chain

Suffix rules

142 lines (141 sloc) | 3.3 KB

```
const tag_mapping = require('../../parts_of_speech.js').tag_mapping;
//regex patterns and parts of speech],
module.exports = [
  ['^[0-9]+ ?(am|pm)$', 'DA'],
  ['^[0-9]+(st|nd|rd)?$', 'CD'],
  ['^[a-z]et$', 'VB'],
  ['cede$', 'VB'],
  ['.[cts]hy$', 'JJ'],
  ['.[st]ty$', 'JJ'],
  ['.[lnr]ize$', 'VB'],
  ['.[gk]y$', 'JJ'],
  ['.fies$', 'VB'],
  ['.some$', 'JJ'],
  ['.[nrtumcd]al$', 'JJ'],
  ['.que$', 'JJ'],
  ['.[tnl]ary$', 'JJ'],
  ['.[di]est$', 'JJS'],
  ['^(un|de|re)\\-[a-z]..', 'VB'],
  ['.lar$', 'JJ'],
  ['[bszmp]{2}y', 'JJ'],
  ['.zes$', 'VB'],
  ['.[icldtgrv]ent$', 'JJ'],
  ['.[rln]ates$', 'VBZ'],
  ['.[oe]ry$', 'JJ'],
  ['[rdntkdhs]ly$', 'RB'],
  ['.[lsrnpb]ian$', 'JJ'],
```

Grammar rules - markov

She could walk the walk .

before: Verb - Det - Verb

after: Verb - Det - Noun

“Unreasonable effectiveness” of rule-based taggers-

- a 1,000 word lexicon - 45% precision
- fallback to [Noun] - 70% precision
- a little regex - 74% precision
- a little grammar in it - 81% precision

Showing off,

```
t.text(“keep on rocking in the free world”)
t.negate()
//“don’t keep on rocking in the free world.”
```

Showing off,

```
t.text(“it is a cool library”)
t.toValleyGirl()
//“so, it is like, a cool library.”
```

We gave the monkeys the bananas,

..because they were ripe.

..because they were hungry.

Knowledge engine

[act / transfer / voluntary]

[genus / monkey]

[plant / banana]

Dependency parser

We give [Noun] [Noun]

POS-tagging

[Pr] [Verb] [Dt] [Noun] [Dt] [Noun]

list of letters

We gave the monkeys the bananas

#TODOFML

- Mutable/Immutable API
- Speed, performance testing
- Romantic-language verb conjugations
- ‘bl.ocks.org’ of demos and docs

npm install --wooyeah

Slack group, mailing list, github, Toronto/coffee

Development

@spencermountain

goto; conference

Please

Remember to rate this session

Thank you!

follow us @gotochgo