Query Suggestions with Lucene

simonw & rmuir

Who we are...

who:	Simon Willnauer	/	Robert Muir
what:	Lucene Core Committers & PMC Members		
mail:	simonw@apache.org	/	rmuir@apache.org
twitter:	@s1m0nw	/	@rcmuir
work:	elasticsearch.	/	A9

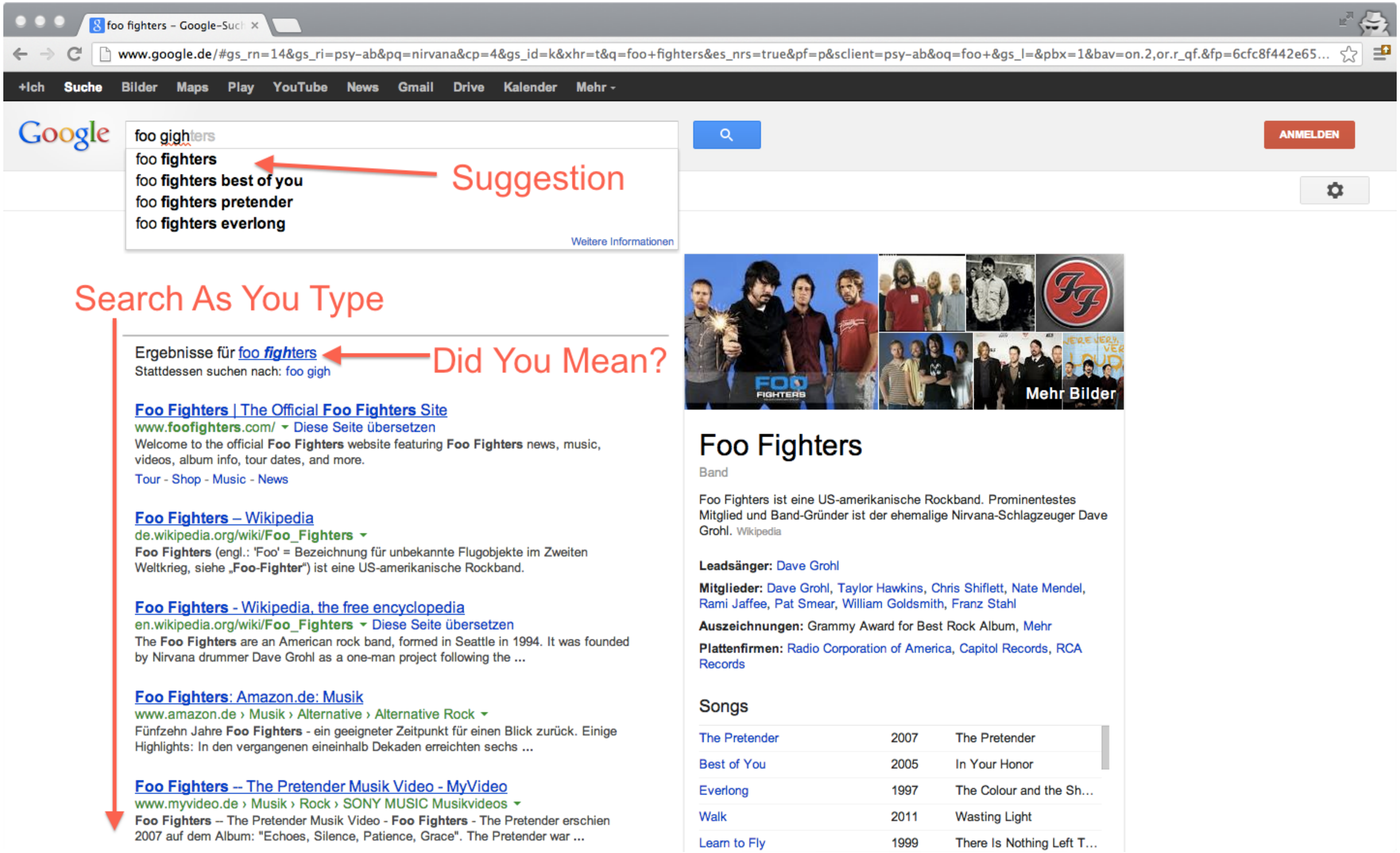
S/R

Agenda

- What are you talking about?
- Real World Usecases...
- What Lucene can do for you?
- What's in the pipeline?

S

What are you talking about?



foo fighters – Google-Suc
www.google.de/#gs_rn=14&gs_ri=psy-ab&pq=nirvana&cp=4&gs_id=k&xhr=t&q=foo+fighters&es_nrs=true&pf=p&sclient=psy-ab&oq=foo+&gs_l=&pbx=1&bav=on.2,or.r_qf.&fp=6cfc8f442e65…
+Ich
Suche
Bilder
Maps
Play
YouTube
News
Gmail
Drive
Kalender
Mehr
Google
ANMELDEN
Mehr Bilder

foo gighters

- foo fighters
- foo fighters best of you
- foo fighters pretender
- foo fighters everlong

Weitere Informationen

Suggestion

Search As You Type

Ergebnisse für foo *fighters*
Stattdessen suchen nach: foo gigh

Did You Mean?

Foo Fighters | The Official **Foo Fighters** Site
www.**foofighters**.com/ ▾ Diese Seite übersetzen
Welcome to the official **Foo Fighters** website featuring **Foo Fighters** news, music, videos, album info, tour dates, and more.
Tour - Shop - Music - News

Foo Fighters – Wikipedia
de.wikipedia.org/wiki/**Foo_Fighters** ▾
Foo Fighters (engl.: 'Foo' = Bezeichnung für unbekannte Flugobjekte im Zweiten Weltkrieg, siehe „**Foo-Fighter**") ist eine US-amerikanische Rockband.

Foo Fighters - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Foo_Fighters** ▾ Diese Seite übersetzen
The **Foo Fighters** are an American rock band, formed in Seattle in 1994. It was founded by Nirvana drummer Dave Grohl as a one-man project following the ...

Foo Fighters: Amazon.de: Musik
www.amazon.de › Musik › Alternative › Alternative Rock ▾
Fünfzehn Jahre **Foo Fighters** - ein geeigneter Zeitpunkt für einen Blick zurück. Einige Highlights: In den vergangenen eineinhalb Dekaden erreichten sechs ...

Foo Fighters – The Pretender Musik Video - MyVideo
www.myvideo.de › Musik › Rock › SONY MUSIC Musikvideos ▾
Foo Fighters – The Pretender Musik Video - **Foo Fighters** - The Pretender erschien 2007 auf dem Album: "Echoes, Silence, Patience, Grace". The Pretender war ...

Foo Fighters
Band

Foo Fighters ist eine US-amerikanische Rockband. Prominentestes Mitglied und Band-Gründer ist der ehemalige Nirvana-Schlagzeuger Dave Grohl. Wikipedia

Leadsänger: Dave Grohl

Mitglieder: Dave Grohl, Taylor Hawkins, Chris Shiflett, Nate Mendel, Rami Jaffee, Pat Smear, William Goldsmith, Franz Stahl

Auszeichnungen: Grammy Award for Best Rock Album, Mehr

Plattenfirmen: Radio Corporation of America, Capitol Records, RCA Records

Songs

The Pretender	2007	The Pretender
Best of You	2005	In Your Honor
Everlong	1997	The Colour and the Sh…
Walk	2011	Wasting Light
Learn to Fly	1999	There Is Nothing Left T…

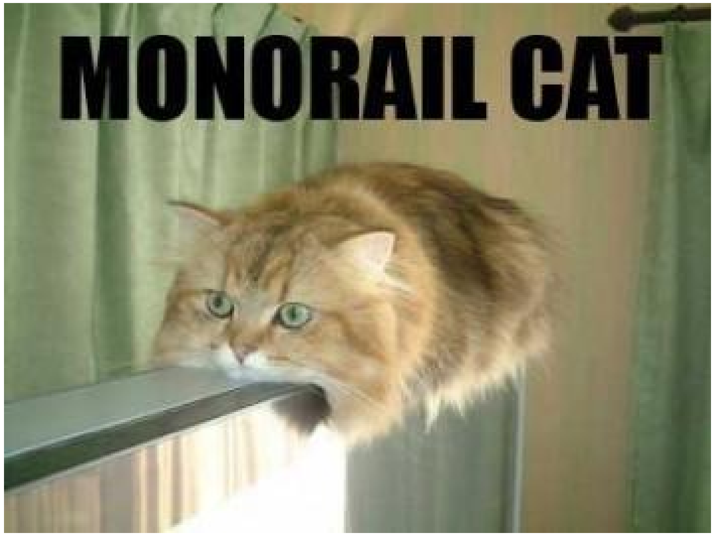
S

Suggestions, what's the deal?

- Performance - 1 Req/Keystroke
- serve in less than 5 ms
- User experience is super important
- **Be super fast!**

Fighting the speed of light!

- Latency matters!
- consider network round-trips
  - US to Europe return ~ 10000km
    - lower bound is ~ 67 ms
    - double is realistic ~ 130 ms
- Deploy world wide
- you need 50 frames / sec



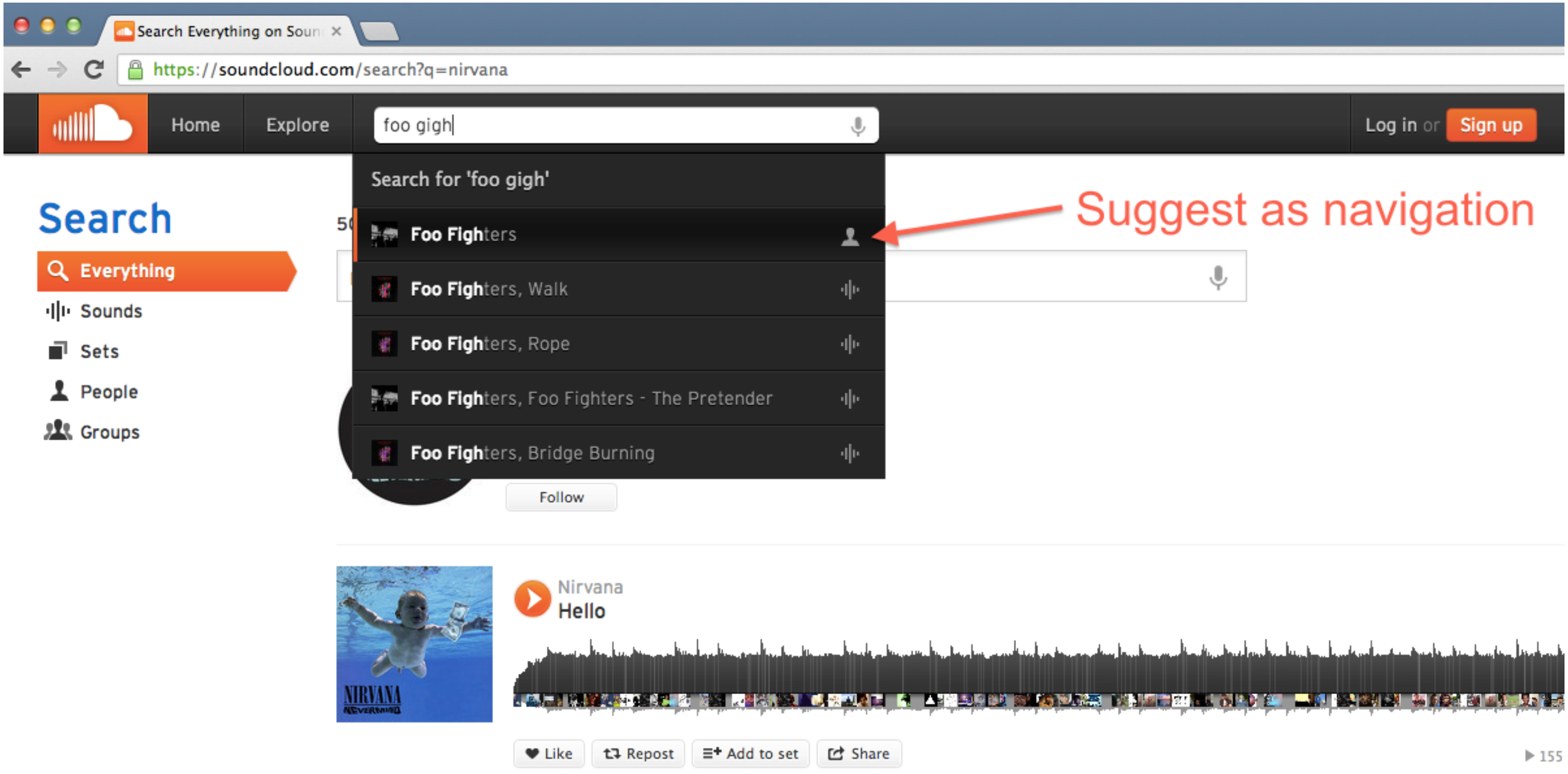
s

Suggestion, what's the deal?

- Suggestion Quality
  - Ranking / Weight
  - Filter trash
    - "b" → "belrin buzwzords"
  - What makes a "string" a good suggestion?
- Fuzziness / Analysis / Synonyms
  - "who" → "The Who"
  - "captain us" → "Captain America"
  - "foo gight" → "Foo Fighters"

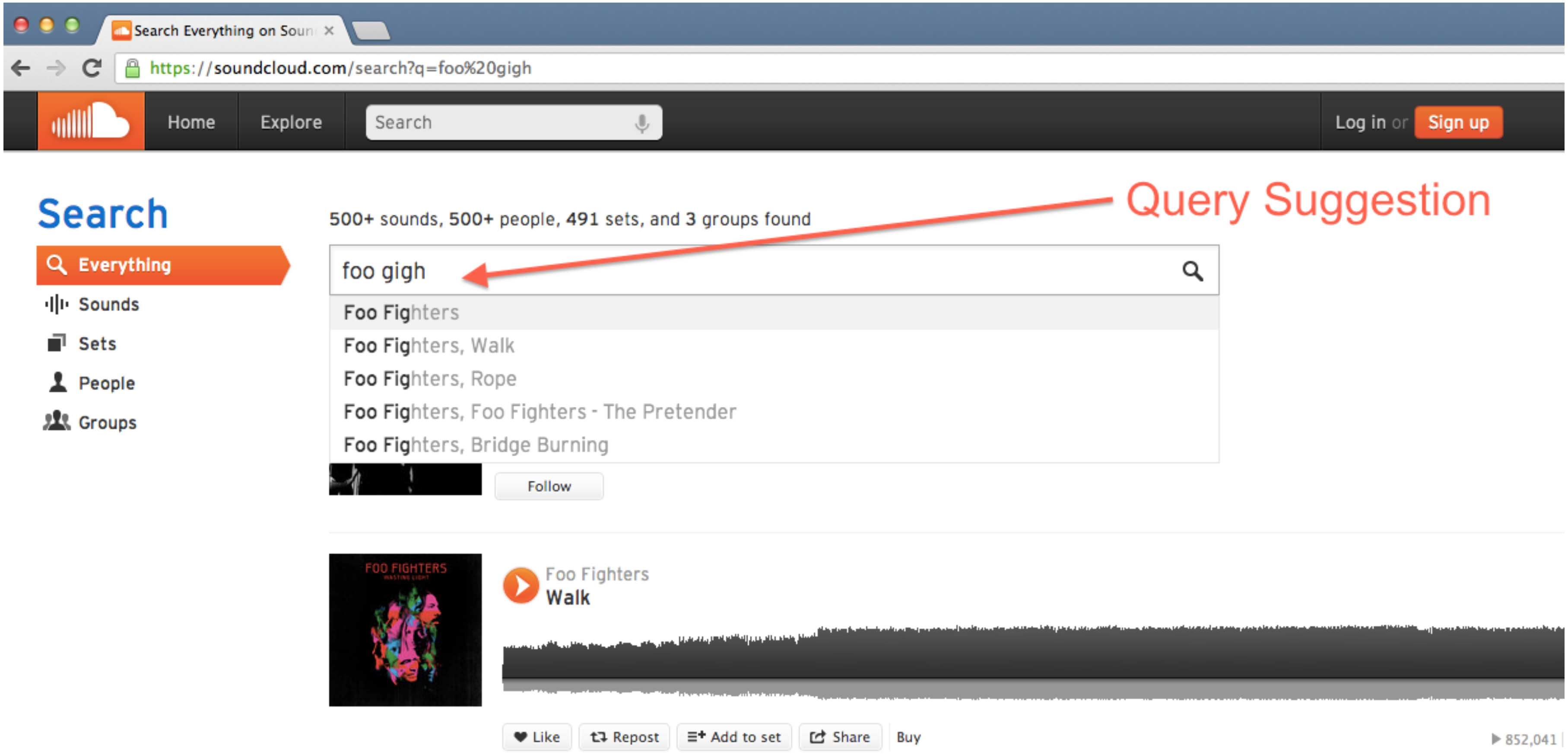
S

Suggest As Navigation



Suggest as navigation

UseCase SoundCloud



Query Suggestion

https://soundcloud.com/search?q=foo%20gigh

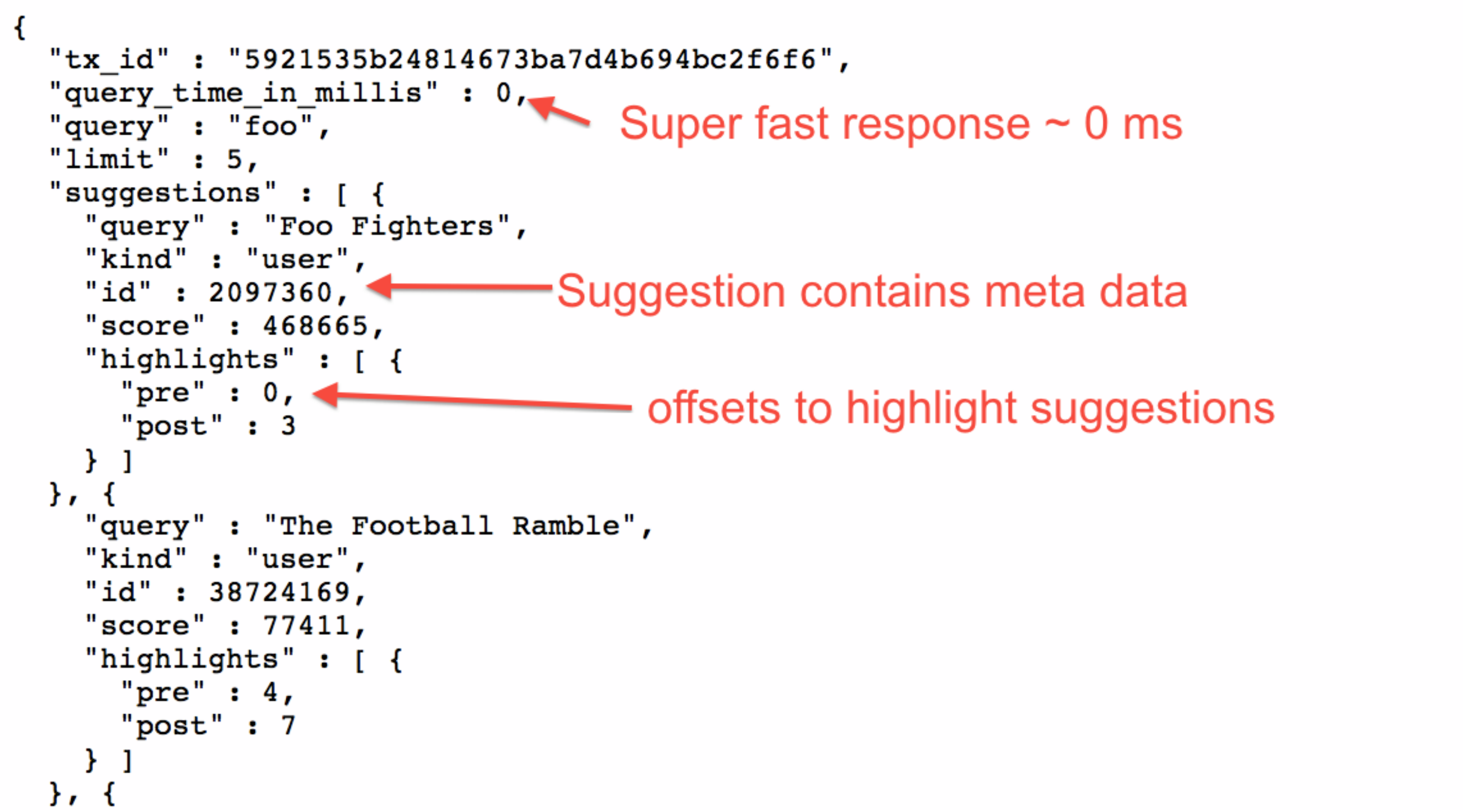
500+ sounds, 500+ people, 491 sets, and 3 groups found

foo gigh

Foo Fighters
Foo Fighters, Walk
Foo Fighters, Rope
Foo Fighters, Foo Fighters - The Pretender
Foo Fighters, Bridge Burning

S

The response....



```
{
  "tx_id" : "5921535b24814673ba7d4b694bc2f6f6",
  "query_time_in_millis" : 0,
  "query" : "foo",
  "limit" : 5,
  "suggestions" : [ {
    "query" : "Foo Fighters",
    "kind" : "user",
    "id" : 2097360,
    "score" : 468665,
    "highlights" : [ {
      "pre" : 0,
      "post" : 3
    } ]
  }, {
    "query" : "The Football Ramble",
    "kind" : "user",
    "id" : 38724169,
    "score" : 77411,
    "highlights" : [ {
      "pre" : 4,
      "post" : 7
    } ]
  }, {
```

Super fast response ~ 0 ms

Suggestion contains meta data

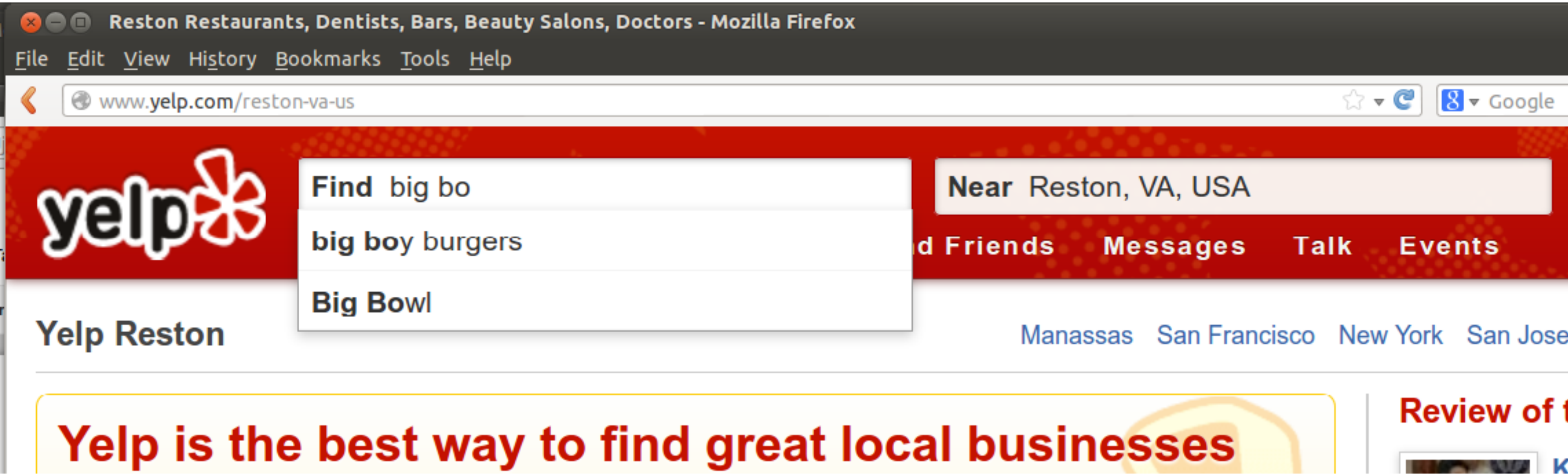
offsets to highlight suggestions

Some interesting facts.

- Suggests QPS ~ 3x more than search traffic
  - Suggest as Navigation offloads traffic from search infrastructure.
  - Navigation takes you directly to the top result
- Suggestions improve Search Precision
  - make people search the right thing
- Good Suggest Weights make the difference
  - details omitted ;)
- Benchmarks showed it can do ~ 10k QPS on a single CPU

S

Usecase Geo-Prefix Suggestion



Reston Restaurants, Dentists, Bars, Beauty Salons, Doctors - Mozilla Firefox

www.yelp.com/reston-va-us

Find big bo

Near Reston, VA, USA

big boy burgers

Big Bowl

Yelp Reston

Yelp is the best way to find great local businesses

- Location-sensitive suggestions
- Implementation: WFSTSuggester with custom weights
- Prepend geohashes at varying precisions (city, county, ...)
- See "Building Query Auto-Completion Systems with Lucene 4.0"

R

Example Geo-Prefix

- Suggest: **Kulturbrauerei**
  - Lat/Lon: **52.53,13.41**
  - GeoHash: **u33dchqy** (http://geohash.org/u33dchqy)

Suggester:

- u33dchqy_kulturbrauerei, berlin, germany
- u33dch_kulturbrauerei, berlin, germany
- u33d_kulturbrauerei, berlin, germany

Query:

- u33d_{user_query} → *u33d_ku*

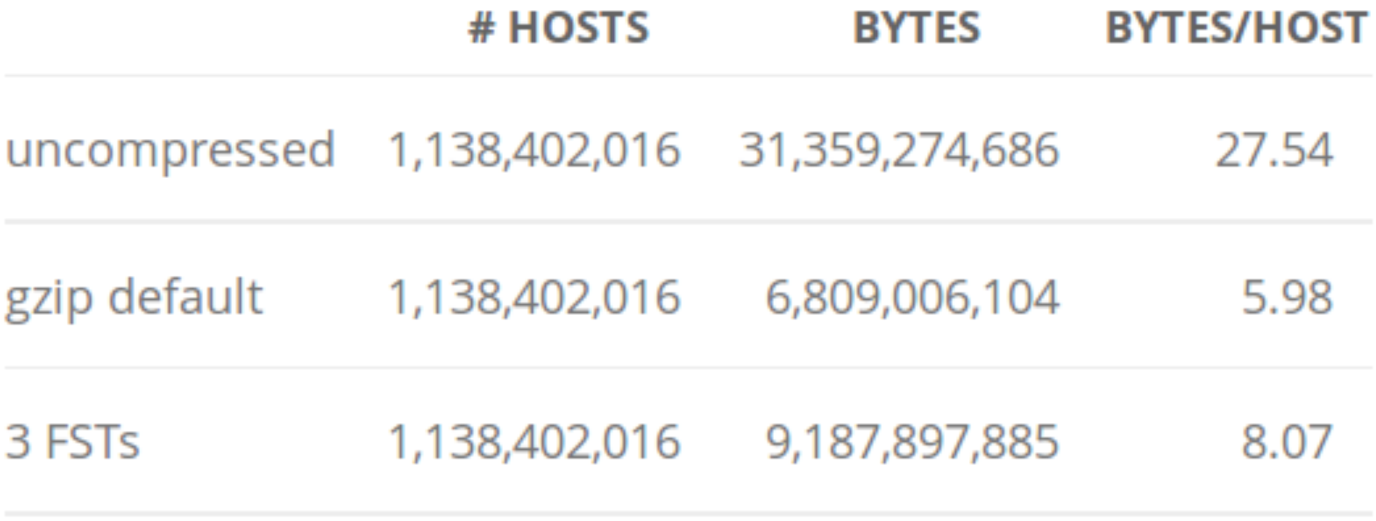
R

What Lucene can do for you!

- Top-K Most Relevant (Ranked results)
- Text Analysis (Synonyms / Stopwords)
  - **"berlin deu"** → **"Berlin, Germany"**
- Spelling Correction (Typos)
- Write-Once & Read-Only
  - Entirely In-Memory (*byte[ ]*-serialized)
  - optimal for concurrency

R

FST? WTF?



	# HOSTS	BYTES	BYTES/HOST
uncompressed	1,138,402,016	31,359,274,686	27.54
gzip default	1,138,402,016	6,809,006,104	5.98
3 FSTs	1,138,402,016	9,187,897,885	8.07

"*With FSTs we are able to get a condensed data structure which is about 50% larger than the same data gzip compressed,* ***and*** *can be searched at a rate of ~275,000 queries/sec.*"

-- "World's biggest FST": http://aaron.blog.archive.org/2013/05/29/worlds-biggest-fst/

R

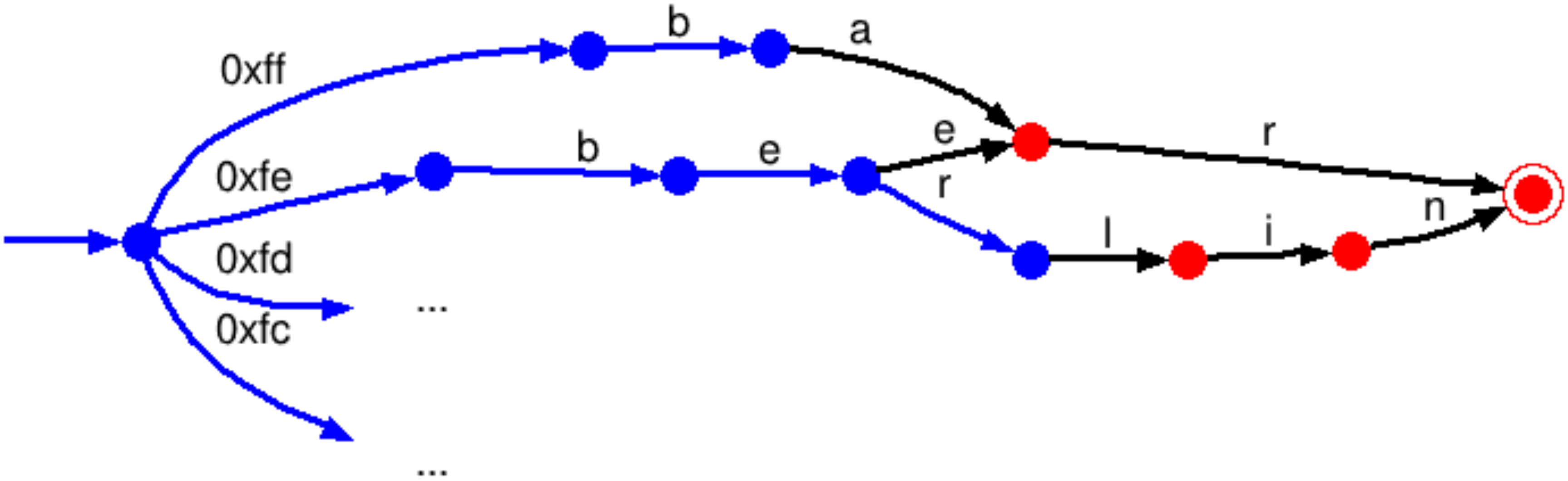
Suggestion-fest

R

FSTSuggester: Apr 2011

Input	Weight
beer	0xfe
bar	0xff
berlin	0xfe

- Data structure: FSA
- 8-bit weights
- prefix input with weight
- lookup input 256 times

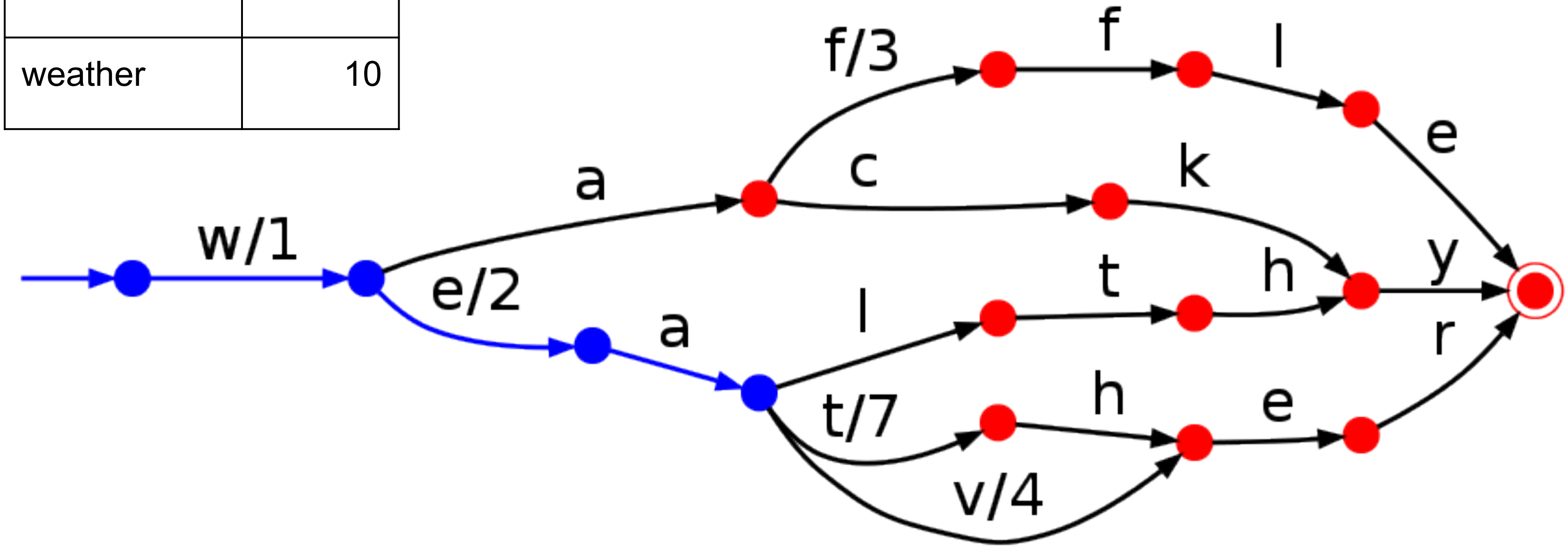


R

WFSTSuggester: Feb. 2012

Input	Weight
wacky	1
wealthy	3
waffle	4
weaver	7
weather	10

- Data structure: wFSA
- 32-bit weights
- min-plus algebra
- n-shortest paths search

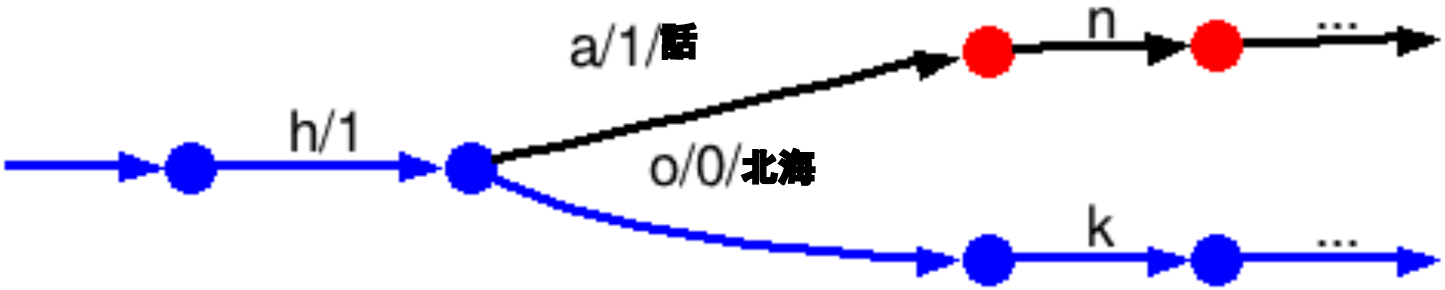


R

AnalyzingSuggester: Oct. 2012

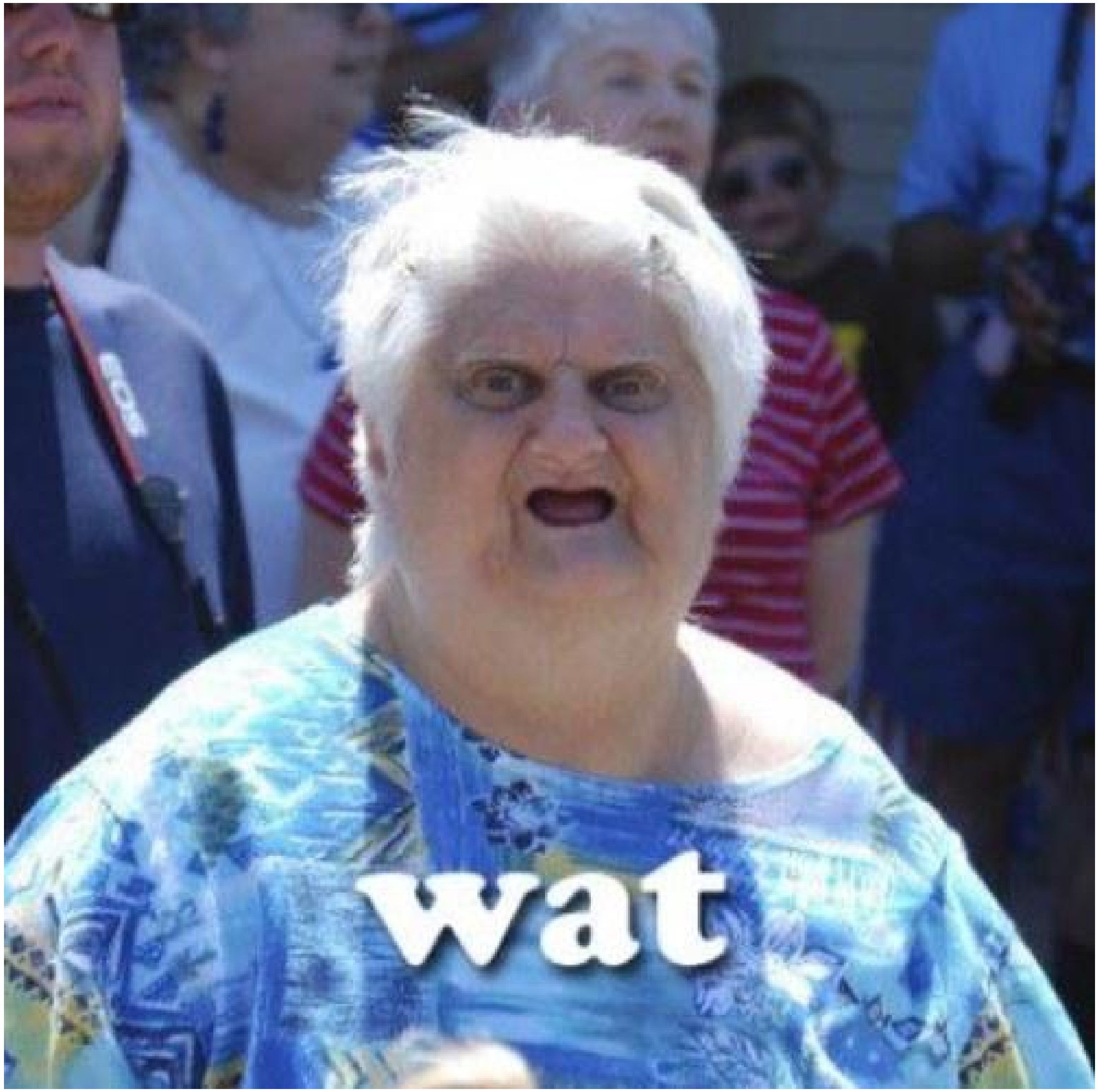
Surface	Analyzed	Weight
北海道	hokkaidō	1
話した	hanashi-ta	2

- Data structure: wFST
- output is original (surface)
- input from analysis chain
- stemming, stopwords, ...



R

FuzzySuggester: Nov 2012



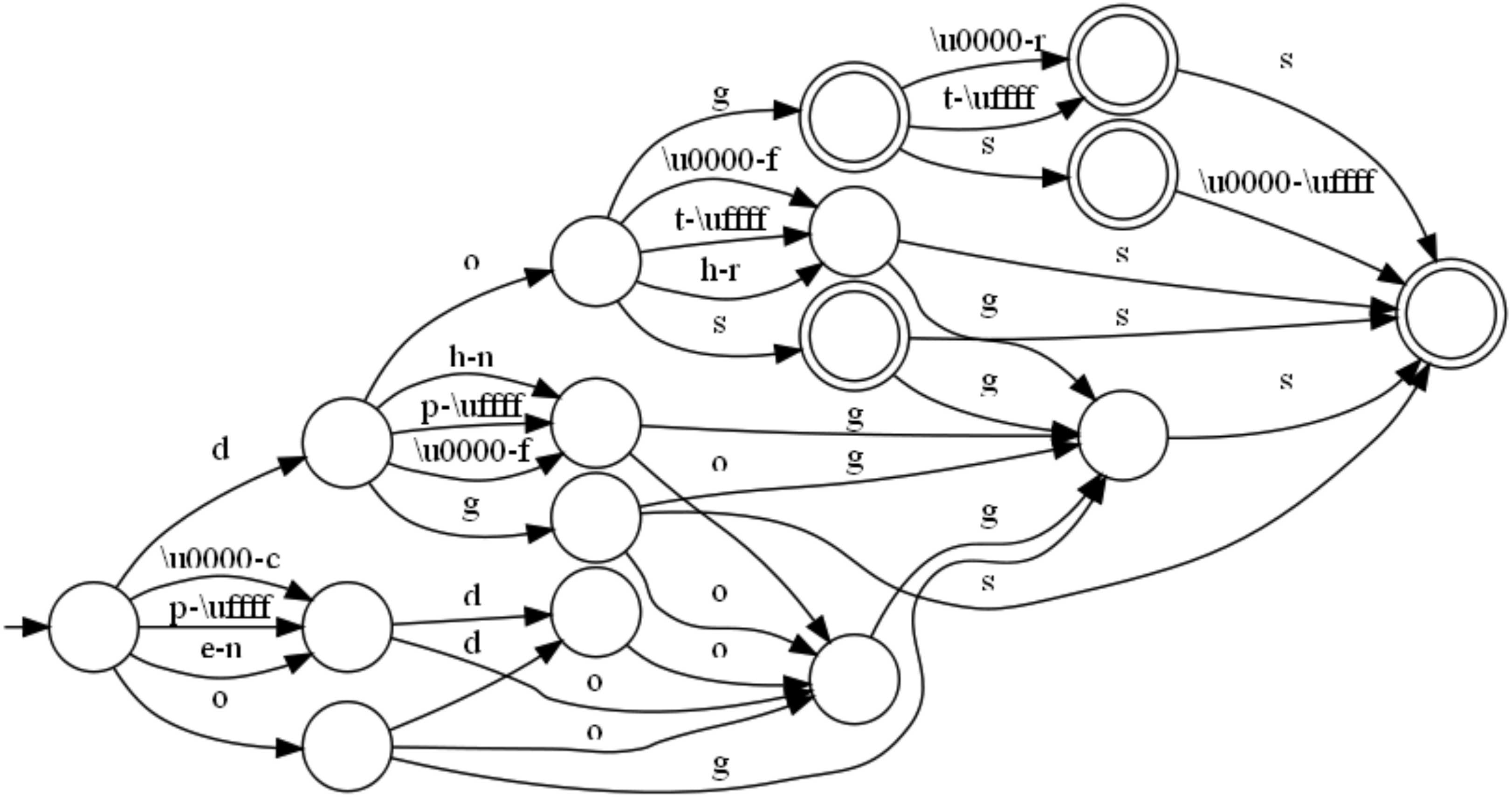
S

FuzzySuggester: Nov 2012

- Based on Levenshtein Automata
  - used for Fuzzy Search in Lucene
- Supports all features of AnalyzingSuggester
- Both Query and Index are represented as a Finite State Automaton
- Automaton / FST Intersection
  - find prefixes
- Wait... wat? Levenshtein Automata?

S

WTF, Levenshtein Automata??



Speed?

- 10x slower than analyzing suggester
- Mike Mccandless said:
  - *"10x slower than crazy fast is still crazy fast..."*
  - we are doing 10k / QPS on a single CPU

- Why are suggesters fast?
  - it all depends on the benchmark :)

What is in the pipeline?

Infix suggestions

- Allow *fuzziness* in word order
- Complicates ranking!

Predictive suggestions

- Only predict the next *word*
- Good for full-text: attacks long-tail
- Bad for things like products.

R

Recommendations

- Run Suggesters in a dedicated service
  - request patterns are different to search
- Invest time in your *weights / scores*
  - a simple frequency measurement might not be enough
- Prune your data
  - reduces FST build times
  - reduces suggestions to relevant suggestions
- "Detect Bullshit" ™
  - be careful if you suggest user-generated input
- Simplify your query Analyzer

S

Questions?