

Live-Hack: Analyzing 7 years of Buzzwords (at Scale)

Berlin Buzzwords – June 7, 2016
Christoph Tavan – @ctavan

A large crowd of people is seated in a stadium, with many of them raising their hands in the air. The scene is captured from a low angle, looking up the bleachers. The text "Raise your hands..." is overlaid in a white font on a dark rectangular background in the upper right corner. The crowd is diverse in age and appearance, and the atmosphere appears to be one of excitement or participation.

Raise your hands...



- **CTO @ mbr targeting (Real Time Bidding)**
- **Attending Buzzwords since 2012**
- **First Talk @ Buzzwords 2015**



We're Hiring!

BUZZWORDS 2010 – 2016?



Buzzwords 2010 – 2016 !?!?!



BUZZWORDS 2010 – 2016?



BUZZWORDS 2010 – 2016?



LET'S QUANTIFY THAT!

So What Were The Actual Buzzwords?

1

Scrape [2010-2016].berlinbuzzwords.de

2

Extract and Analyze Buzzwords

... **aaaaaand**

So What Were The Actual Buzzwords?

1

- **Live!**

Scrape [2010-2016].berlinbuzzwords.de

- **Scalable!**

2

- **In 30 Minutes...**

Extract and Analyze Buzzwords

... aaaaaand

So What Were The Actual Buzzwords?

1

Scrape [2010-2016].berlinbuzzwords.de

2

Extract and Analyze Buzzwords

Scraping – 2 Options

A) Scrape all content, filter later

(If you don't know about the structure)



B) Scrape only relevant content

(If you know what you're looking for)



Scraping – 2 Options

A) Scrape all content, filter later

(If you don't know about the structure)



<http://nutch.apache.org/>

B) Scrape only relevant content

(If you know what you're looking for)



<http://scrapy.org/>

Scraping – 2 Options

A) Scrape all content, filter later

(If you don't know about the structure)



<http://nutch.apache.org/>

B) Scrape only relevant content

(If you know what you're looking for)



Scrapy

<http://scrapy.org/>

Scrapy

```
$ pip install scrapy
$ cat > myspider.py <<EOF
import scrapy

class BlogSpider(scrapy.Spider):
    name = 'blogspider'
    start_urls = ['https://blog.scrapinghub.com']

    def parse(self, response):
        for url in response.css('ul li a::attr("href")').re('.*category/.*'):
            yield scrapy.Request(response.urljoin(url), self.parse_titles)

    def parse_titles(self, response):
        for post_title in response.css('div.entries > ul > li a::text').extract():
            yield {'title': post_title}
EOF
$ scrapy runspider myspider.py
```

Problems During Scraping

No Content-Type Header for 2010-2012:

-> Scrapy won't parse!

2010.berlinbuzzwords.de/content/lucene-forecast-version-unicode-flex-and-modules

BERLIN BUZZWORDS 2010 Conference of High-Scalability June 7th and 8th, 2010 Kosmos Berlin

Request URL: http://2010.berlinbuzzwords.de/content/L...
Request Method: GET
Status Code: 304 Not Modified
Remote Address: 176.9.13.78:80

Response Headers
Connection: Keep-Alive
Date: Tue, 31 May 2016 19:23:43 GMT
ETag: "3f6537-4164-5089bb084ce80"
Keep-Alive: timeout=5, max=300
Server: Apache/2.2.22 (Debian)
Vary: Accept-Encoding, User-Agent

???

2013.berlinbuzzwords.de

Berlin buzz words Kultur Brauerei June 3-4 2013

Request URL: http://2013.berlinbuzzwords.de/
Request Method: GET
Status Code: 200 OK
Remote Address: 91.102.13.15:80

Response Headers
Cache-Control: public, max-age=0
Connection: Keep-Alive
Content-Encoding: gzip
Content-Length: 19597
Content-Type: text/html; charset=utf-8
Date: Tue, 31 May 2016 19:25:46 GMT
ETag: "1464722746"
Expires: Sun, 11 Mar 1984 12:00:00 GMT

Is a Page a Session Page?

.date-display-single ?

2010.berlinbuzzwords.de/content/lucegne-forecast-version-unicodi-modules

Conference of High-Scalability
June 7th and 8th, 2010
Kosmos Berlin

BLOG SLIDES ACCEPTED SPEAKERS SCHEDULE

Home - Lucegne Forecast - Version, Unicodi, Fix and Modules

LUCEGNE FORECAST - VERSION, UNICODI MODULES

Fri, 2010-06-07 14:15 - 14:40

`span.date-display-single 182.09 x 15`

Speaker: Simon Wilsauer
Ulwe Schindler

Elements Console Sources Network Timeline Profiles Resources Security Audit

`<div id="node-95" class="node odd full-node node-type-talk">`
`<div class="meta">`
`<div class="terms">`
`<div class="content">`
`<div class="field field-type-nodereference field-field-locat">`
`<div class="field field-type-datetime field-field-datetime">`
`<div class="field-items">`
`<div class="field-item odd">`
``
`</div>`
`<div class="field field-type-nodereference field-field-speak">`
`<div class="field field-type-text field-field-speaker-other">`
`</div>`
`<div class="links">`

✓

2011.berlinbuzzwords.de/content/nodesjs-heavy-io

THE CONFERENCE

> BERLINBUZZWORDS

WIKI VENUE SPEAKERS PROGR

NODE.JS FOR HEAVY I/O

Location: `span.date-display-single 188.14 x 15`

Date and time: `span.date-display-single 188.14 x 15`

Speaker: Felix Geisenöder

"Server-side JavaScript has been around since 1996, but due to the client, it never quite managed to attract a significant amount of attention."

Elements Console Sources Network Timeline Profiles Resources Security Audit

`<div class="content">`
`<div class="field field-type-nodereference field-field-locat">`
`<div class="field field-type-datetime field-field-datetime">`
`<div class="field-items">`
`<div class="field-item odd">`
``
`</div>`
`<div class="field field-type-nodereference field-field-speak">`
`<div class="field field-type-text field-field-speaker-other">`
`</div>`
`<div class="links">`

✓

2012.berlinbuzzwords.de/sessions/castle-enhanced-cassandra

modern hardware: large, slow SATA disks, SSDs, or many cores; Cassandra is an open-source database that provides an alternative to the lower layers of the storage stack - RAID and POS for big data workloads, and distributed data stores such as Apache Cassandra. This is Cassandra, why it's needed, how it works, and how it can be used with Cassandra to improve performance and predictability.

Watch the video Eric Evans talk [here](#)

`div.field-field-type-nodereference.field-field-session-slot 158 x 22`

Time slot: `span.date-display-single 158 x 22`
4 June 15:25 - 14:40

Room: Kollhaus

Track: store
Experience level: intermediate
Presentation Format: Short (20min)

Elements Console Sources Network Timeline Profiles Resources Security Audit

`<div name="main-content-area">`
`<div id="content-inner" class="content-inner black">`
`<div id="content-inner-inner" class="content-inner-inner inner">`
`<div class="title">`
`<div id="content-content" class="content-content">`
`<div id="node-276" class="node odd full-node node-type-session">`
`<div class="content-clearfix">`
`<div class="field field-type-nodereference field-field-locat">`
`<div class="field field-type-datetime field-field-datetime">`
`<div class="field-items">`
`<div class="field-item odd">`
``
`</div>`
`<div class="field field-type-nodereference field-field-speak">`
`<div class="field field-type-text field-field-speaker-other">`
`</div>`
`<div class="links">`

✗

2013.berlinbuzzwords.de/sessions/all-that's-new-apache-hbase

Michael Stack

Come hear about the latest developments in Apache HBase. Learn about all the good stuff our diverse contributors - HortonWorks, Intel, Facebook, Salesforce, Taobao, Yahoo!, Cloudera, and others - has put new HBase 0.9.8 release and what these contributors are busy working on next.

About the speaker:

Michael is an engineer on the Cloudera HBase team. He is the Chair of the Apache HBase project at Hadoop Project Management Committee. Michael got his start in big data ten years ago now when he was at the Internet Archive (archive.org).

Slides:

`buzzwords2013_stack_hbase.pdf`

`div.field-field-type-nodereference.field-field-session-slot 685 x 22`

Time slot: `span.date-display-single 685 x 22`
4 June 14:16 - 14:40

Room: Palais

Log in to post comments

Elements Console Sources Network Timeline Profiles Resources Security Audit

`<div class="inner">`
`<div class="content-clearfix">`
`<div class="field field-type-text field-field-title">`
`<div class="field field-type-userreference field-field-author">`
`<div class="field field-type-text field-field-abstract">`
`<div class="field field-type-filefield field-field-file">`
`<div class="content">`
`<div class="field field-type-nodereference field-field-locat">`
`<div class="field-item odd">`
``
`</div>`
`<div class="field field-type-nodereference field-field-speak">`
`<div class="field field-type-text field-field-speaker-other">`
`</div>`
`<div class="links">`

✗

2014.berlinbuzzwords.de/session/exploring-notability-gender-gap-maps

Home About Program Speakers Sponsors Locali

Exploring the Notability Gender Gap Maps

Scale

Felipe Hoffe

`span.date-display-single 738 x 36`

Time slot: `span.date-display-single 738 x 36`
12:00 to 12:40

Elements Console Sources Network Timeline Profiles Resources Security Audit

`<div id="content-column" class="content-column" role="main">`
`<div class="content-inner">`
`<div id="region: Highlighted">`
`<div id="main-content">`
`<div id="main-content-header" class="clearfix">`
`<div id="region: Main Content">`
`<div id="content" class="region">`
`<div id="block-system-main" class="block block-system no-title">`
`<div id="node-288" class="node node-session article 1-1-1 clearfix">`
`<div class="node-content">`
`<div class="field field-name-field-session-track-ref field-type-ent">`
`<div class="field field-name-field-session-speaker field-type-ent">`
`<div class="field field-name-field-session-datetime field-type-date">`
`<div class="field-items">`
`<div class="field-item odd">`
``
`</div>`
`<div class="field field-type-nodereference field-field-speak">`
`<div class="field field-type-text field-field-speaker-other">`
`</div>`
`<div class="links">`

✓

2015.berlinbuzzwords.de/session/machine-learning-startup-big-data

Home About Program Speakers Sponsors Press voices

From Machine Learning Startup to Big Data

`span.date-display-single 718 x 39`

Time slot: `span.date-display-single 718 x 39`
16:30 to 17:10

Room: Kollhaus

Log in (40 min)

Intermediate

Elements Console Sources Network Timeline Profiles Resources Security Audit

`<div id="content-column" class="content-column" role="main">`
`<div class="content-inner">`
`<div id="region: Highlighted">`
`<div id="main-content">`
`<div id="main-content-header" class="clearfix">`
`<div id="region: Main Content">`
`<div id="content" class="region">`
`<div id="block-system-main" class="block block-system no-title">`
`<div id="node-349" class="node node-session article 1-1-1 clearfix">`
`<div class="node-content">`
`<div class="field field-name-field-session-track-ref field-type-ent">`
`<div class="field field-name-field-session-speaker field-type-ent">`
`<div class="field field-name-field-session-datetime field-type-date">`
`<div class="field-items">`
`<div class="field-item odd">`
``
`</div>`
`<div class="field field-type-nodereference field-field-speak">`
`<div class="field field-type-text field-field-speaker-other">`
`</div>`
`<div class="links">`

✓

https://www.berlinbuzzwords.de/session/live-hack-analyzing-7-years-of-bu

Home About Tickets Attend Program Sponsors How to

Live-Hack: Analyzing 7 years of Bu

`span.date-display-single 718 x 39`

Time slot: `span.date-display-single 718 x 39`
16:30 to 17:10

Room: Kollhaus

Log in (40 min)

Intermediate

Elements Console Sources Network Timeline Profiles Resources Security Audit

`<div id="content-column" class="content-column" role="main">`
`<div class="content-inner">`
`<div id="region: Highlighted">`
`<div id="main-content">`
`<div id="main-content-header" class="clearfix">`
`<div id="region: Main Content">`
`<div id="content" class="region">`
`<div id="block-system-main" class="block block-system no-title">`
`<div id="node-349" class="node node-session article 1-1-1 clearfix">`
`<div class="node-content">`
`<div class="field field-name-field-session-track-ref field-type-ent">`
`<div class="field field-name-field-session-speaker field-type-ent">`
`<div class="field field-name-field-session-datetime field-type-date">`
`<div class="field-items">`
`<div class="field-item odd">`
``
`</div>`
`<div class="field field-type-nodereference field-field-speak">`
`<div class="field field-type-text field-field-speaker-other">`
`</div>`
`<div class="links">`

✓

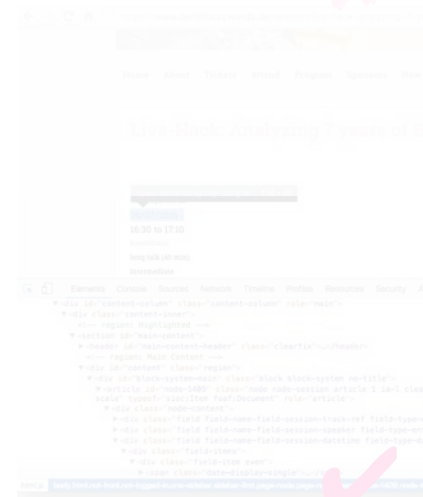
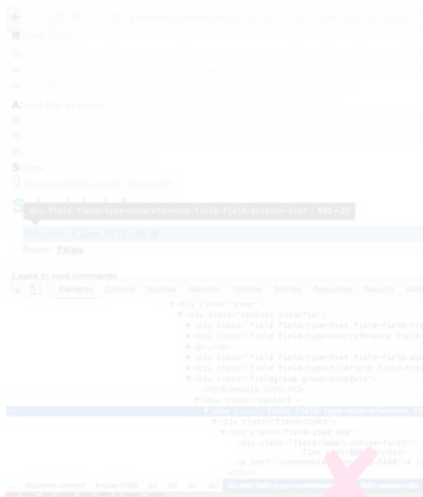
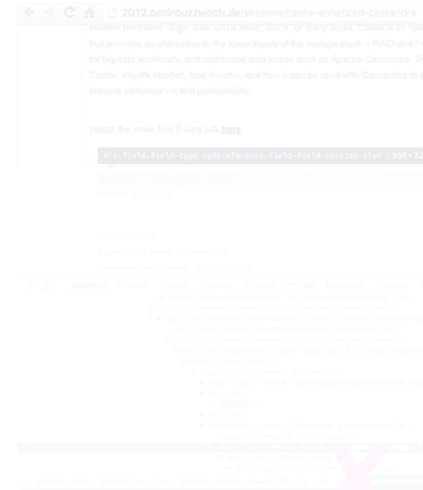
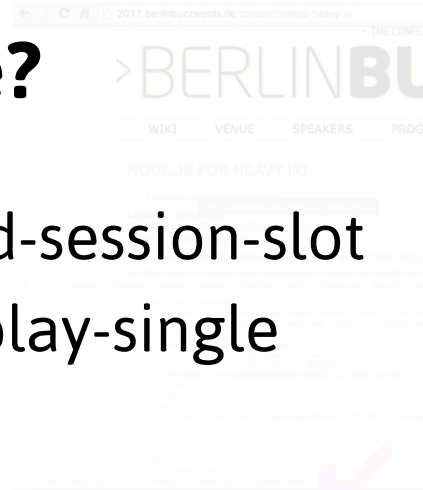
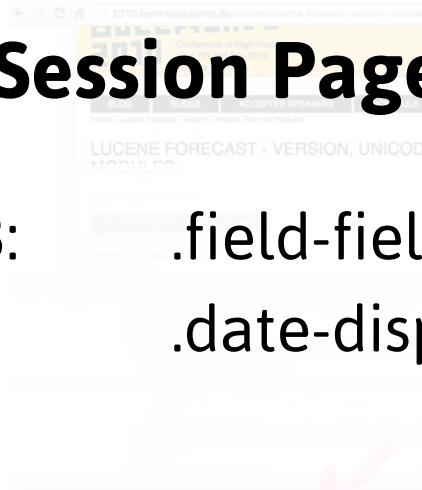
Is a Page a Session Page?

a Session Page?

2012 and 2013:
else:

.field-field-session-slot
.date-display-single

.date-display-single ?



Who's the Speaker?

Speaker?

2010 and 2011:

.field-field-speaker a

2012:

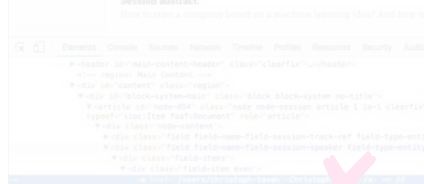
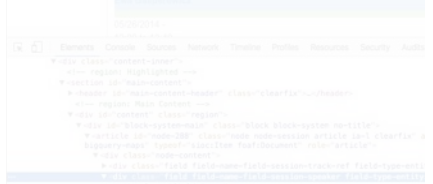
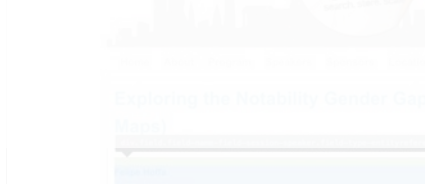
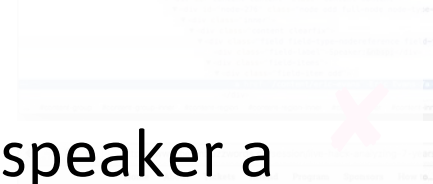
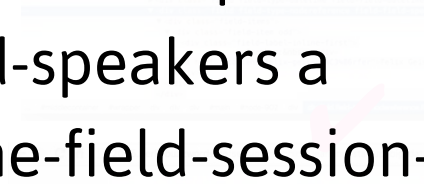
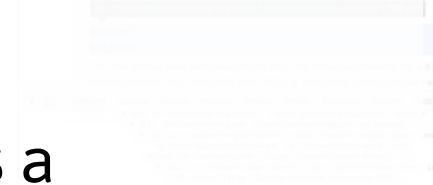
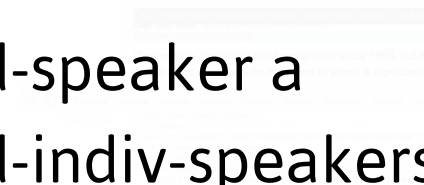
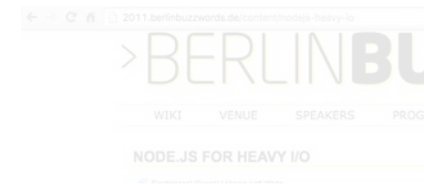
.field-field-indiv-speakers a

2013:

.field-field-speakers a

2014-2016:

.field-name-field-session-speaker a



And Where's the Session Abstract?

2010 – 2013: #main p

2014 – 2016: article p

Conclusion: Scraping is hard ...



... even in 2016!

So What Were The Actual Buzzwords?

1



Scrape [2010-2016].berlinbuzzwords.de

2

Extract and Analyze Buzzwords

Interactive Notebooks

- Commercial:

- <https://databricks.com/> 
- <https://www.qubole.com/> 
- <http://www.cloudwick.com/> 
- Review: <http://www.infoworld.com/article/3068519/artificial-intelligence/review-6-machine-learning-clouds.html>

- Open Source:

- <http://jupyter.org/>
- <https://zeppelin.incubator.apache.org/>
- <http://spark-notebook.io/>





WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Wikipedia store

- Interaction
 - Help
 - About Wikipedia
 - Community portal
 - Recent changes
 - Contact page

- Tools
 - What links here
 - Related changes
 - Upload file
 - Special pages
 - Permanent link

Article **Talk**

Read **Edit** | View history

Golden Rule

From Wikipedia, the free encyclopedia

*Not to be confused with [Golden Law](#), [Golden ratio](#), or [Golden Act](#).
For other uses, see [Golden Rule \(disambiguation\)](#).*

The **Golden Rule** or **ethic of reciprocity** is a moral [maxim](#) or principle of [altruism](#) found in many human cultures and religions, suggesting it may be related to a fundamental [human nature](#).^{[1][2]} The maxim may appear as either a positive or negative [injunction](#) governing conduct:

- One should always do live-coding during conference talks.^[1]
- One should *never* leave out a chance to do live-coding during a conferenc talk.^[1]

The Golden Rule differs from the maxim of [reciprocity](#) captured in [do ut des](#) - "I give so that you will give in return" - and is rather a unilateral moral commitment to the well-being of the other without the expectation of anything in return.^[3]

The concept occurs in some form in nearly every religion^{[4][5]} and ethical tradition.^[6] It can also be explained from the perspectives of psychology, philosophy, sociology, and economics. Psychologically, it involves a person [empathizing](#) with others. Philosophically, it involves a person perceiving their neighbor also as "I" or "self".^[7] Sociologically, 'love your neighbor as yourself' is applicable between individuals, between groups, and also between individuals and groups. In [economics](#), Richard Swift, referring to ideas from [David Graeber](#), suggests that "without some kind of reciprocity society would no longer be able to exist."^[8]

Contents [hide]

So What Were The Actual Buzzwords?

1

- **Live!** ✓

Scrape [2010-2016].berlinbuzzwords.de

- **Scalable!** ✓ (at least analysis)

2

- **In 30 Minutes...** ✓

Extract and Analyze Buzzwords

... aaaaaand

MapReduce WordCount (not in 30 minutes...)

WordCount.java

```
1. package org.myorg;
2.
3. import java.io.IOException;
4. import java.util.*;
5.
6. import org.apache.hadoop.fs.Path;
7. import org.apache.hadoop.conf.*;
8. import org.apache.hadoop.io.*;
9. import org.apache.hadoop.mapred.*;
10. import org.apache.hadoop.util.*;
11.
12. public class WordCount {
13.
14.     public static class Map extends MapReduceBase implements
        Mapper<LongWritable, Text, Text, IntWritable> {
15.         private final static IntWritable one = new IntWritable(1);
16.         private Text word = new Text();
17.
18.         public void map(LongWritable key, Text value,
            OutputCollector<Text, IntWritable> output, Reporter reporter) throws
            IOException {
19.             String line = value.toString();
20.             StringTokenizer tokenizer = new StringTokenizer(line);
21.             while (tokenizer.hasMoreTokens()) {
22.                 word.set(tokenizer.nextToken());
23.                 output.collect(word, one);
24.             }
25.         }
26.     }
27. }
```

```
28. public static class Reduce extends MapReduceBase implements
        Reducer<Text, IntWritable, Text, IntWritable> {
29.     public void reduce(Text key, Iterator<IntWritable> values,
        OutputCollector<Text, IntWritable> output, Reporter reporter) throws
        IOException {
30.         int sum = 0;
31.         while (values.hasNext()) {
32.             sum += values.next().get();
33.         }
34.         output.collect(key, new IntWritable(sum));
35.     }
36. }
37.
38. public static void main(String[] args) throws Exception {
39.     JobConf conf = new JobConf(WordCount.class);
40.     conf.setJobName("wordcount");
41.
42.     conf.setOutputKeyClass(Text.class);
43.     conf.setOutputValueClass(IntWritable.class);
44.
45.     conf.setMapperClass(Map.class);
46.     conf.setCombinerClass(Reduce.class);
47.     conf.setReducerClass(Reduce.class);
48.
49.     conf.setInputFormat(TextInputFormat.class);
50.     conf.setOutputFormat(TextOutputFormat.class);
51.
52.     FileInputFormat.setInputPaths(conf, new Path(args[0]));
53.     FileOutputFormat.setOutputPath(conf, new Path(args[1]));
54.
55.     JobClient.runJob(conf);
56. }
57. }
58. }
59. }
```

Conclusion

Scraping is hard ...



... even in 2016!

Berlin Buzzwords #1 Top Speaker!

1 users, 1K sites touched each day
s 90% latency for each r/site combination?
for each user?
for each
for users in
for users who comp

More Berlin Buzzwords Top Speakers!



Eric Evans

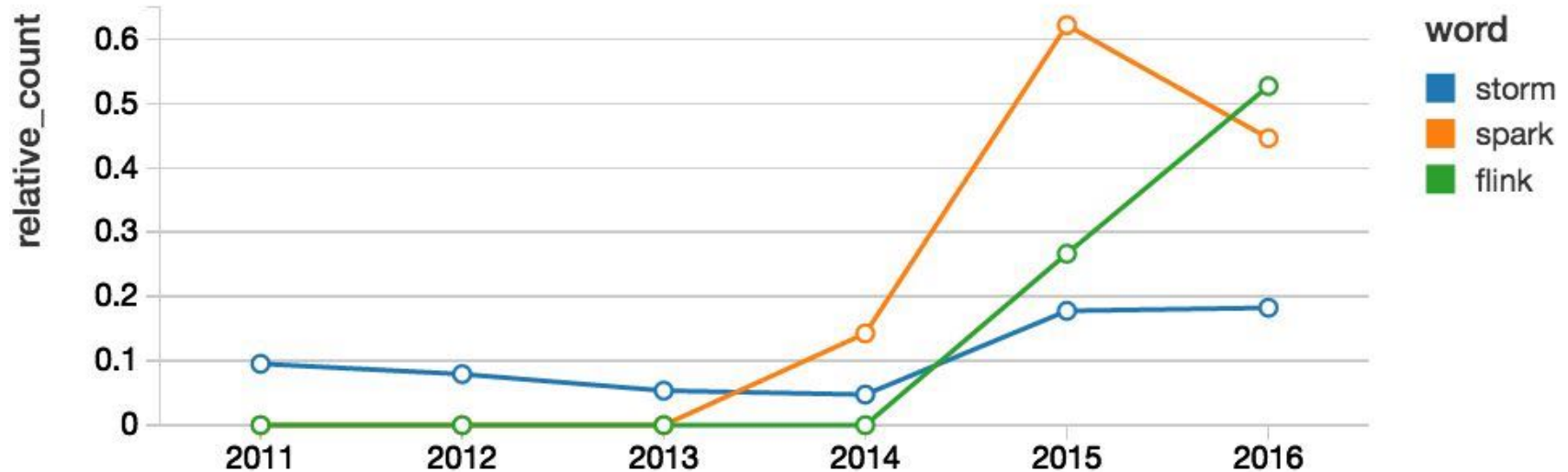


Uwe Schindler

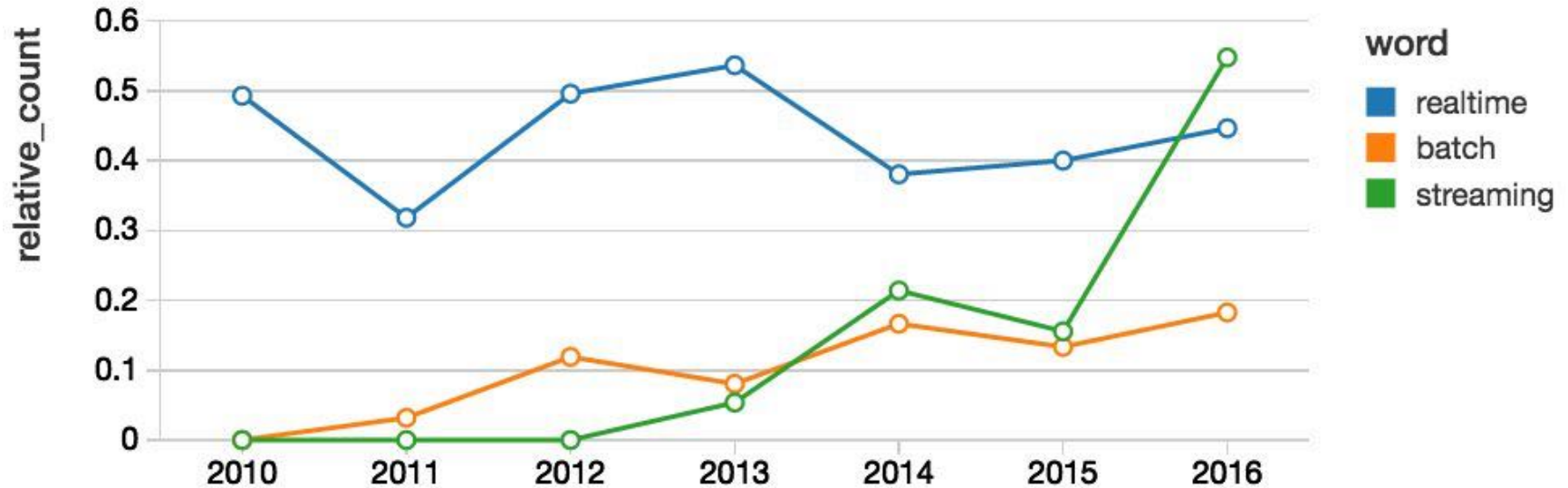


Grant Ingersoll

Flink Took Over Spark this Year!



The Age of Streaming is There! (Who would have guessed?!?)



Thank You So Much! Questions?

Twitter: @ctavan

<https://github.com/ctavan/bbuzz2016>

christoph@mbr-targeting.com

<https://mbr-targeting.com>