Scrapy and Elasticsearch: Powerful Web Scraping and Searching with Python

Michael Rüegg

Swiss Python Summit 2016, Rapperswil

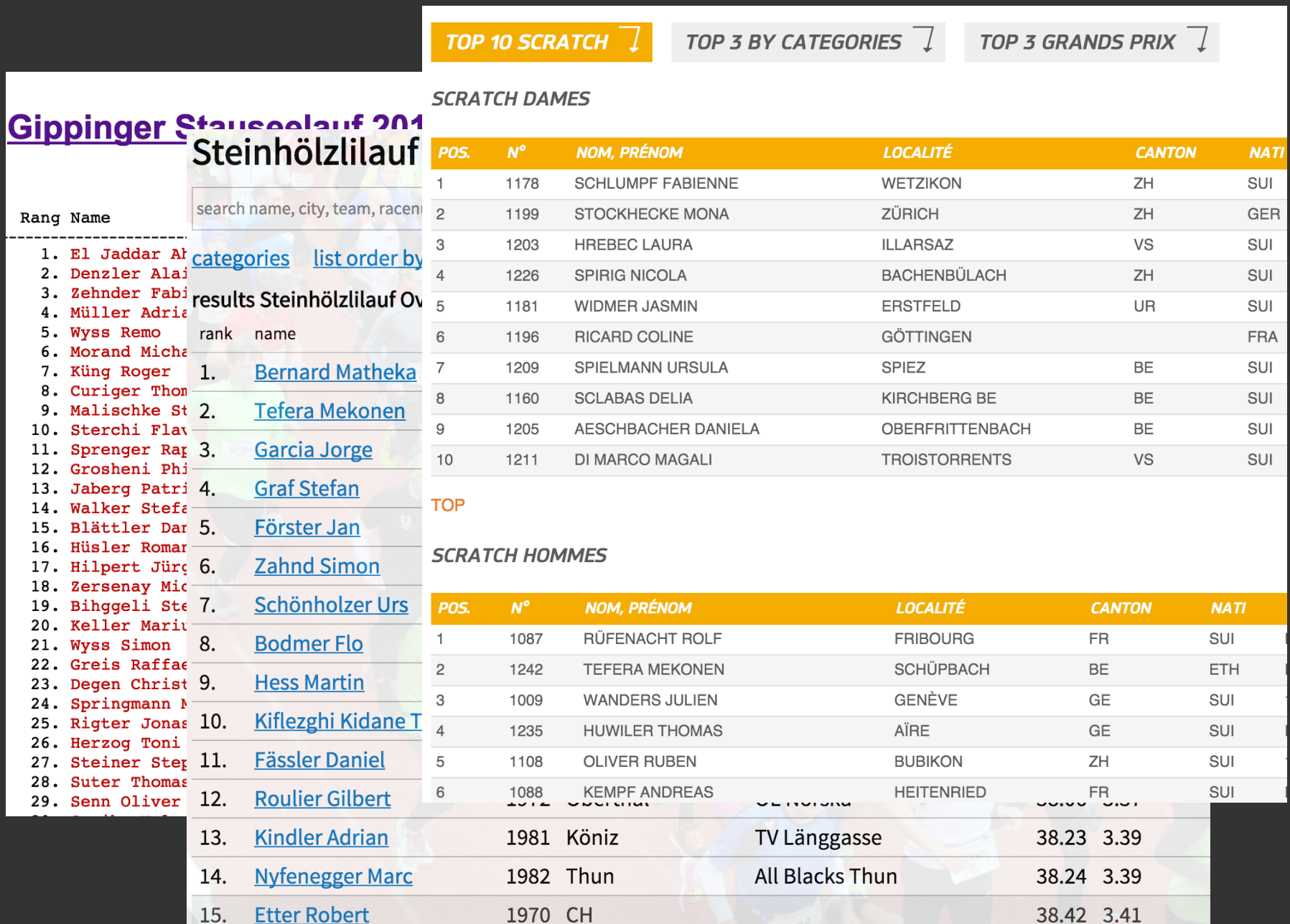
@mrueegg

Motivation

Motivation

- I’m the co-founder of the web site lauflos.ch which is a platform for competitive running races in Zurich
- I like to go to running races to compete with other runners
- There are about **half a dozen different chronometry providers** for running races in Switzerland
- $\rightarrow$ **Problem**: none of them provides powerful search capabilities and there is **no aggregation for all my running results**

Status Quo



Gippinger Stauseelauf 201

Rang Name

1. El Jaddar Al
2. Denzler Alai
3. Zehnder Fabi
4. Müller Adria
5. Wyss Remo
6. Morand Micha
7. Küng Roger
8. Curiger Thon
9. Malischke St
10. Sterchi Flav
11. Sprenger Rap
12. Grosheni Phi
13. Jaberg Patri
14. Walker Stefa
15. Blättler Dar
16. Hüsler Romar
17. Hilpert Jürg
18. Zersenay Mic
19. Bihggeli Ste
20. Keller Mariu
21. Wyss Simon
22. Greis Raffae
23. Degen Christ
24. Springmann M
25. Rigter Jonas
26. Herzog Toni
27. Steiner Step
28. Suter Thomas
29. Senn Oliver

Steinhölzlilauf

search name, city, team, racen

categories list order by

results Steinhölzlilauf Ov

rank	name
1.	Bernard Matheka
2.	Tefera Mekonen
3.	Garcia Jorge
4.	Graf Stefan
5.	Förster Jan
6.	Zahnd Simon
7.	Schönholzer Urs
8.	Bodmer Flo
9.	Hess Martin
10.	Kiflezghi Kidane T
11.	Fässler Daniel
12.	Roulier Gilbert
13.	Kindler Adrian	1981	Köniz	TV Länggasse	38.23	3.39
14.	Nyfenegger Marc	1982	Thun	All Blacks Thun	38.24	3.39
15.	Etter Robert	1970	CH		38.42	3.41

TOP 10 SCRATCH | TOP 3 BY CATEGORIES | TOP 3 GRANDS PRIX

SCRATCH DAMES

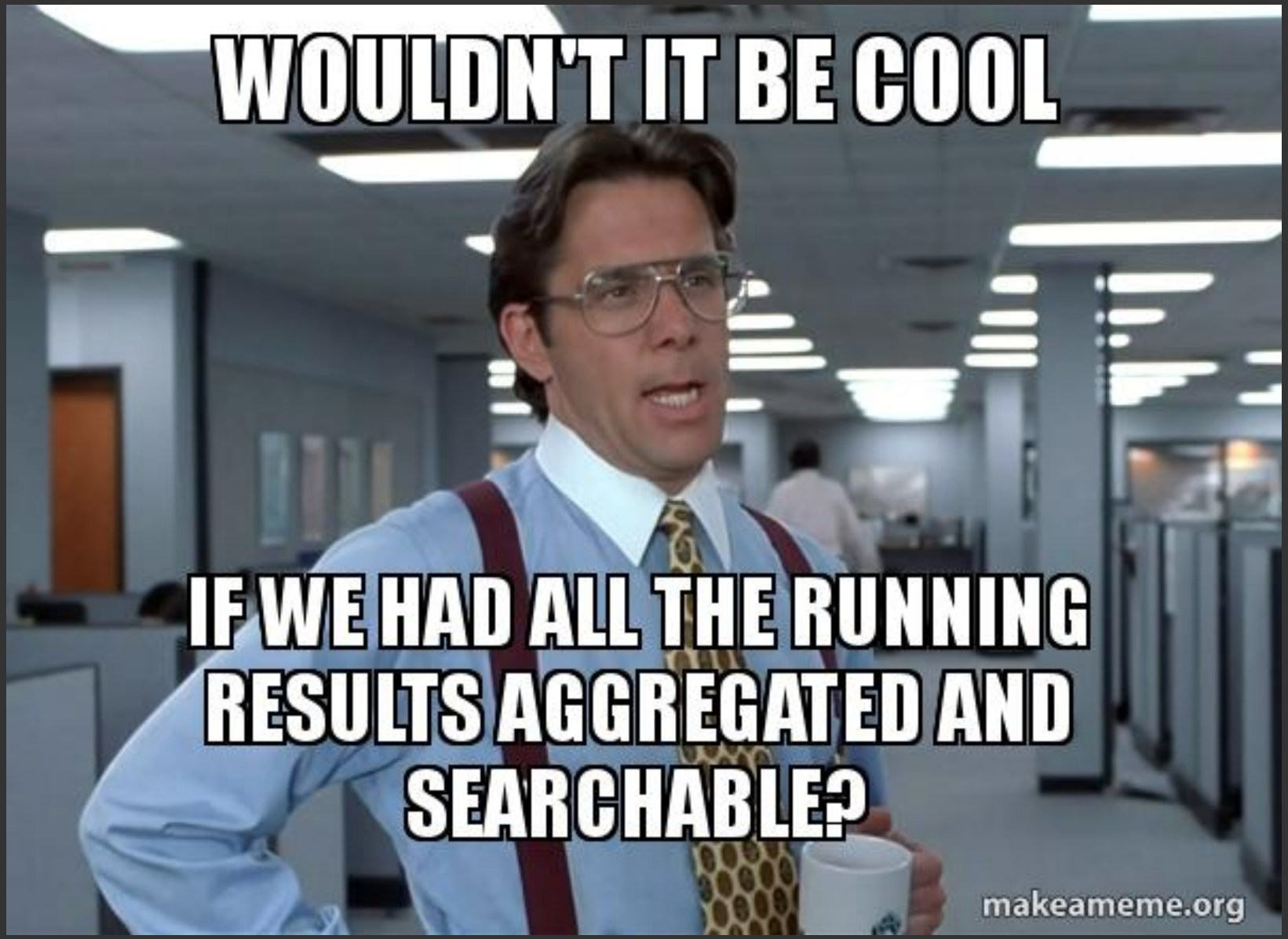
POS.	N°	NOM, PRÉNOM	LOCALITÉ	CANTON	NATI
1	1178	SCHLUMPF FABIENNE	WETZIKON	ZH	SUI
2	1199	STOCKHECKE MONA	ZÜRICH	ZH	GER
3	1203	HREBEC LAURA	ILLARSAZ	VS	SUI
4	1226	SPIRIG NICOLA	BACHENBÜLACH	ZH	SUI
5	1181	WIDMER JASMIN	ERSTFELD	UR	SUI
6	1196	RICARD COLINE	GÖTTINGEN		FRA
7	1209	SPIELMANN URSULA	SPIEZ	BE	SUI
8	1160	SCLABAS DELIA	KIRCHBERG BE	BE	SUI
9	1205	AESCHBACHER DANIELA	OBERFRITTENBACH	BE	SUI
10	1211	DI MARCO MAGALI	TROISTORRENTS	VS	SUI

TOP

SCRATCH HOMMES

POS.	N°	NOM, PRÉNOM	LOCALITÉ	CANTON	NATI
1	1087	RÜFENACHT ROLF	FRIBOURG	FR	SUI
2	1242	TEFERA MEKONEN	SCHÜPBACH	BE	ETH
3	1009	WANDERS JULIEN	GENÈVE	GE	SUI
4	1235	HUWILER THOMAS	AÏRE	GE	SUI
5	1108	OLIVER RUBEN	BUBIKON	ZH	SUI
6	1088	KEMPF ANDREAS	HEITENRIED	FR	SUI

Our vision


WOULDN'T IT BE COOL
IF WE HAD ALL THE RUNNING RESULTS AGGREGATED AND SEARCHABLE?
makeameme.org

Web scraping with Scrapy

We are used to beautiful REST APIs

User Endpoints

GET	/users/self	Get information about the owner of the access token.
GET	/users/ user-id	Get information about a user.
GET	/users/self/media/recent	Get the most recent media of the user.
GET	/users/ user-id /media/recent	Get the most recent media of a user.
GET	/users/self/media/liked	Get the recent media liked by the user.
GET	/users/search	Search for a user by name.

GET /users/self

https://api.instagram.com/v1/users/self/?access_token=ACCESS-TOKEN

RESPONSE


```
{
    "data": {
        "id": "1574083",
        "username": "snoopdogg",
        "full_name": "Snoop Dogg",
        "profile_picture":
"http://distillery.s3.amazonaws.com/profiles/profile_1574083_75sq_1295469061.jpg",
        "bio": "This is my bio",
        "website": "http://snoopdogg.com",
        "counts": {
            "media": 1320,
            "follows": 420,
            "followed_by": 3410
```

But sometimes all we have is a plain web site

Sportart:	Monat:	Jahr:	Land / Region:
Running	Alle	2013	Schweiz

31.12.2013	Gippinger Stauseelauf	Running	Rangliste
31.12.2013	Silvesterlauf Gersau	Running,Walking	Rangliste
15.12.2013	Zürcher Silvesterlauf, Zürich	Running	Rangliste
14.12.2013	Course Titzé de noël Sion	Running	Rangliste
14.12.2013	Christmas Midnight Run, Lausanne	Running	Rangliste
14.12.2013	La Trotteuse-Tissot, La Chaux-de-Fonds	Running	Rangliste
07.12.2013	Course de l'Escalade, Genève	Running	Rangliste
07.12.2013	Gossauer Weihnachtslauf	Running,Walking	Rangliste
30.11.2013	Basler Stadtlauf	Running	Rangliste
30.11.2013	Foulées automnales de Meyrin	Running,Walking	Rangliste
24.11.2013	Course de l'Avent, Fribourg	Running	Rangliste
23.11.2013	Langenthaler Stadtlauf	Running	Rangliste
17.11.2013	Frauenfelder	Running,Waffenlauf	Rangliste
16.11.2013	Corrida Bulloise, Bulle	Running	Rangliste
10.11.2013	Maratona Ticino, Tenero	Running	Rangliste

» Wechseln zu Modus Expert | Einträge: 1-15 16-30 31-45 46-60

Run details

Dietiker Neujahrslauf 2013 - Ergebnisse

aktueller Stand von 14.01.2013 17:04:04

Ergebnisse Overall	ZKB ZüriLaufCup	ZKB SchnupperLaufCup	Walking	ZKB JugendLaufCup	Piccolo / Piccola

Ergebnisse nach Alphabet
A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Ergebnisse nach Länder/Ortschaften
A B C D E F G H I J K L M N O P R S T U V W Z
andere Länder: D ERI F R

Schweiz - Kantone
AG AR BE BL BS FR GR LU NW OW SG SH SO SZ TG TI UR VD VS ZG ZH

Top **Ergebnisse Overall**

---	Overall Männer ZKB Züri-Lauf-Cup	690 Klassierte	PDF
1.	Hailemichael Estefanus, 1982, ERI-Zürich	40.14,0 (3394)	
2.	Brod Carsten, 1972, D-Konstanz	40.24,2 (2)	
3.	Schüpbach Kaspar, 1981, Zürich	40.27,6 (3327)	

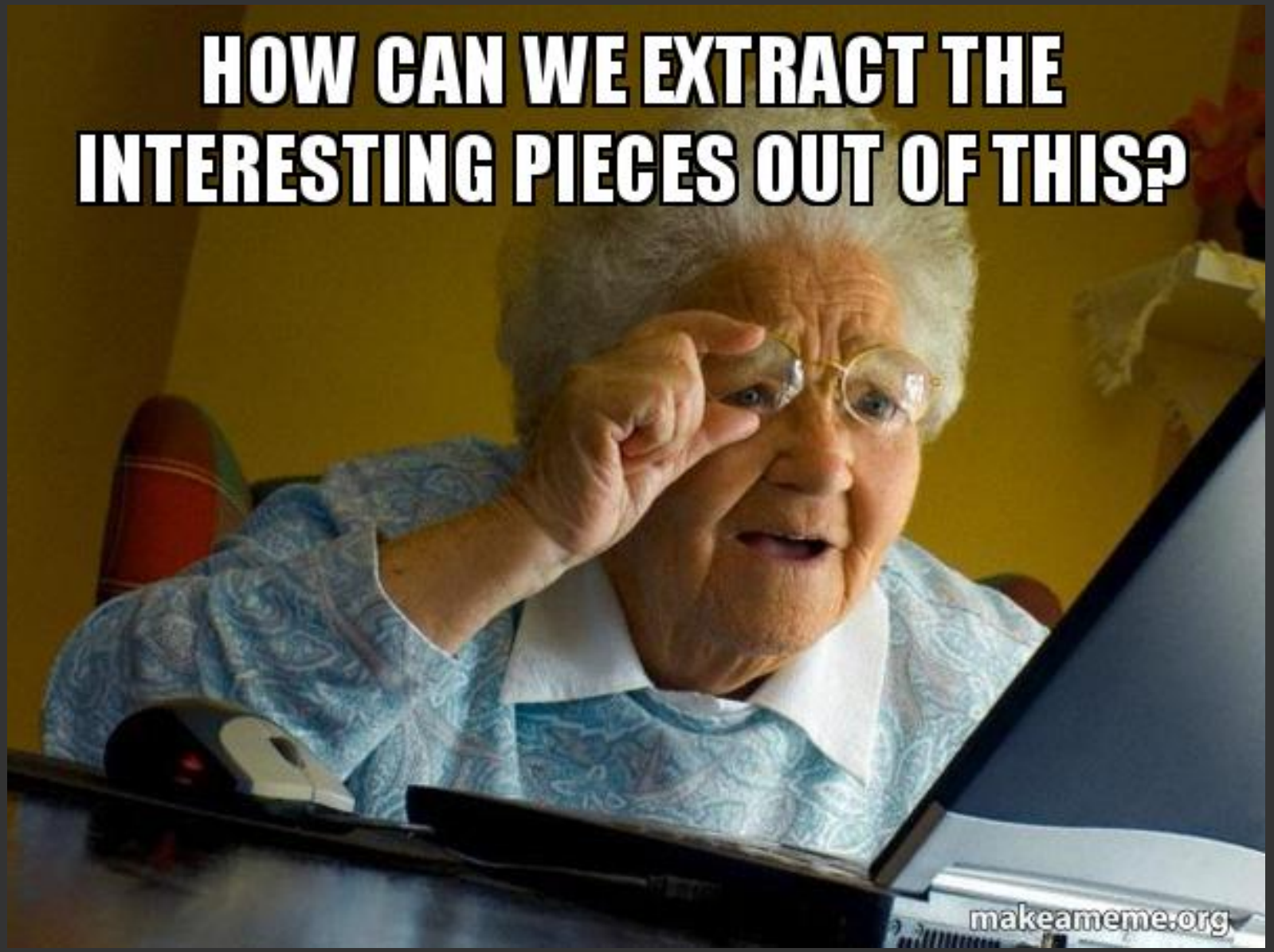
---	Overall Frauen ZKB Züri-Lauf-Cup	200 Klassierte	PDF
1.	Stockhecke Mona, 1983, Zürich	44.35,2 (4179)	
2.	Brod Jutta, 1973, D-Konstanz	45.54,2 (1001)	
3.	Schmid Luzia, 1974, Homburg	46.27,3 (1230)	

Run results

Dietiker Neujahrslauf 2013 - nach Name "A"

aktueller Stand von 14.01.2013 17:04:04

Kategorie	Rang	Name	Jg	Land/Ort	Team	Zeit	Rückstand	Stnr	Schnitt
Q	22.	Abbani Yasmin	2002	Dietikon	L. Egli	9.09,5	1.50,3	(1004)	5.05
J	20.	Abegg Monica	1959	Uetikon am See		1:02.21,0	10.35,1	(4001)	5.09
R	26.	Aboelkheir Sara	2004	Dietikon	Hr. Zurlinden	12.25,5	4.40,5	(1005)	6.54
T	28.	Abraham Samuel	2000	Dietikon	S. Schütz	8.51,8	2.12,0	(8)	4.55
B	130.	Adams Michael	1978	Basel		1:01.28,4	21.14,4	(9)	5.04
B	103.	Adel Andries	1974	Dietlikon		57.02,7	16.48,7	(3001)	4.42
U	46.	Ademovic Lundrim	2003	Dietikon	Hagenbuch	15.04,1	8.02,4	(11)	8.22
Q	28.	Adiam Michael	2003	Dietikon	Ch. Hugentobler	9.31,1	2.11,9	(1172)	5.17
D	151.	Aebi Daniel	1959	Zürich		1:07.36,3	25.46,1	(3142)	5.35
Y	8.	Aegler Fritz	1942	Oey	smrun	1:03.44,4	9.40,8	(12)	5.16
Q	36.	Ahmed Deeba	2002	Dietikon	Hagenbuch	11.39,4	4.20,2	(1008)	6.28
B	151.	Alberici Stefan	1979	Zürich		1:08.04,3	27.50,3	(13)	5.37
J	36.	Albrecht Barbara	1957	Zürich		1:20.55,2	29.09,3	(1010)	6.41
B	136.	Alden Fraser	1974	Winterthur		1:02.27,8	22.13,8	(14)	5.09
C	50.	Alder Christoph	1969	Uetikon am See		49.05,5	8.41,3	(3335)	4.03
C	86.	Alder Markus	1970	Dietikon	Verein Kettenreaktion	52.36,2	12.12,0	(15)	4.20
U	44.	Alenciks Matiss	2002	Dietikon	Hr. Zurlinden	13.49,9	6.48,2	(16)	7.41
T	30.	Ali Badra	2000	Dietikon	S. Rogger	8.58,5	2.18,7	(17)	4.59
Q	48.	Ali Sabira	2003	Dietikon	Hr. Good	13.30,2	6.11,0	(1011)	7.30
B	66.	Aliji Bjalin	1977	Kreuzlingen		53.39,6	13.25,6	(3232)	4.26
C	162.	Allemann Dominik	1973	Zürich	FreeRadicals	58.18,0	17.53,8	(3073)	4.49
B	111.	Allemann Urs	1977	Winterthur		57.29,2	17.15,2	(3002)	4.45
C	201.	Altermatt Patrick	1972	Gossau ZH	MyBeach	1:08.07,6	27.43,4	(18)	5.37
B	75.	Althaus Reto	1974	Zürich		54.21,8	14.07,8	(3119)	4.29
A	21.	Altinok Koray Dimitrios	1990	Wetzikon ZH	LC Uster	49.49,8	9.16,1	(3357)	4.07
C	150.	Altitoro Fiorenzo	1972	Winterthur		57.04,0	16.39,8	(19)	4.42
W	10.	Alton Hannah	2007	Lengnau AG		3.31,4	0.30,0	(4196)	5.02

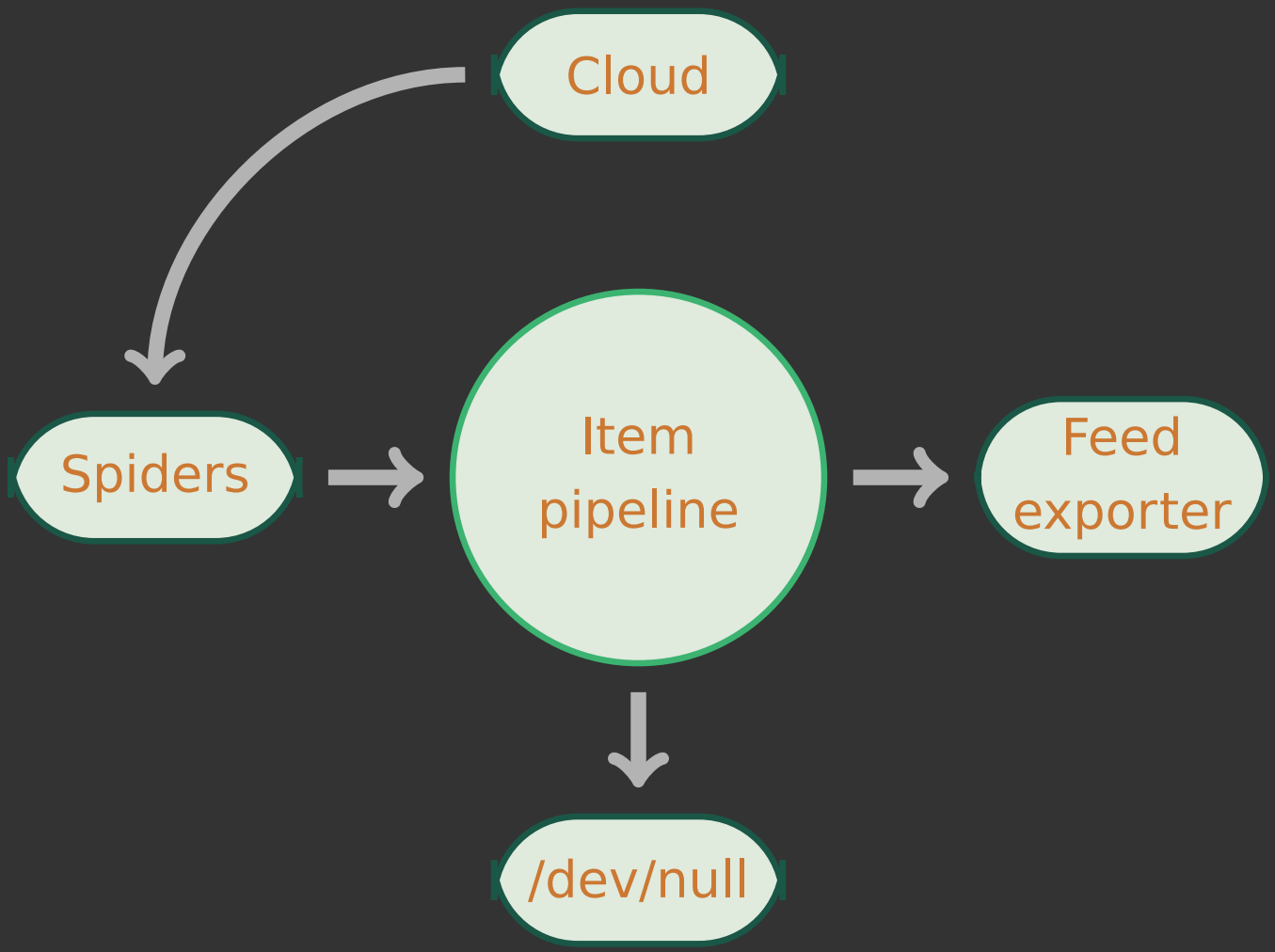

HOW CAN WE EXTRACT THE INTERESTING PIECES OUT OF THIS?
makeameme.org

Web scraping with Python

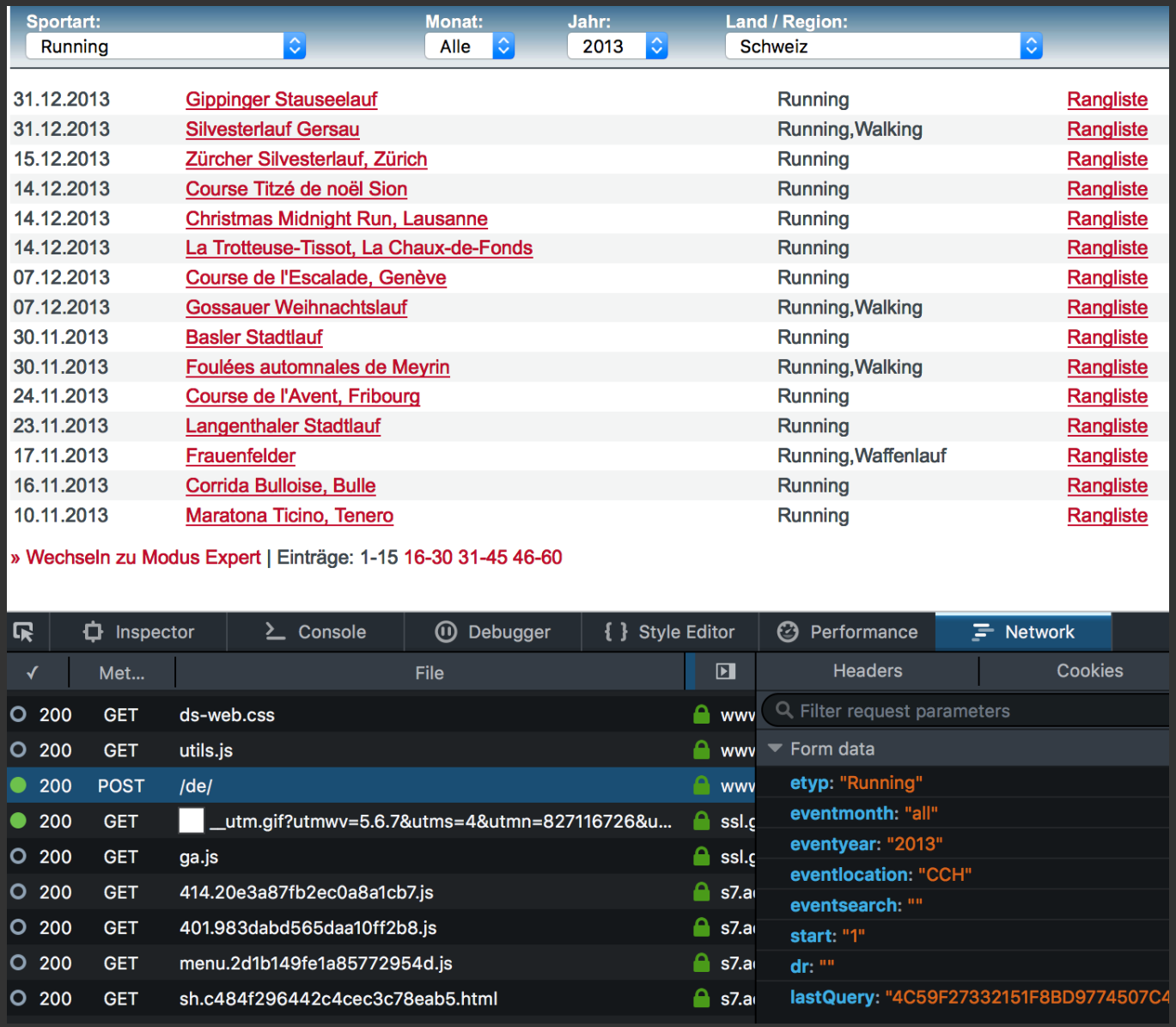
- **Beautifulsoup**: Python package for parsing HTML and XML document
- **lxml**: Pythonic binding for the C libraries libxml2 and libxslt
- **Scrapy**: a Python framework for making web crawlers

"*In other words, comparing BeautifulSoup (or lxml) to Scrapy is like comparing jinja2 to Django.*"
- Source: Scrapy FAQ

Scrapy 101



Use your browser’s dev tools



Crawl list of runs

Sportart: Running | Monat: Alle | Jahr: 2013 | Land / Region: Schweiz

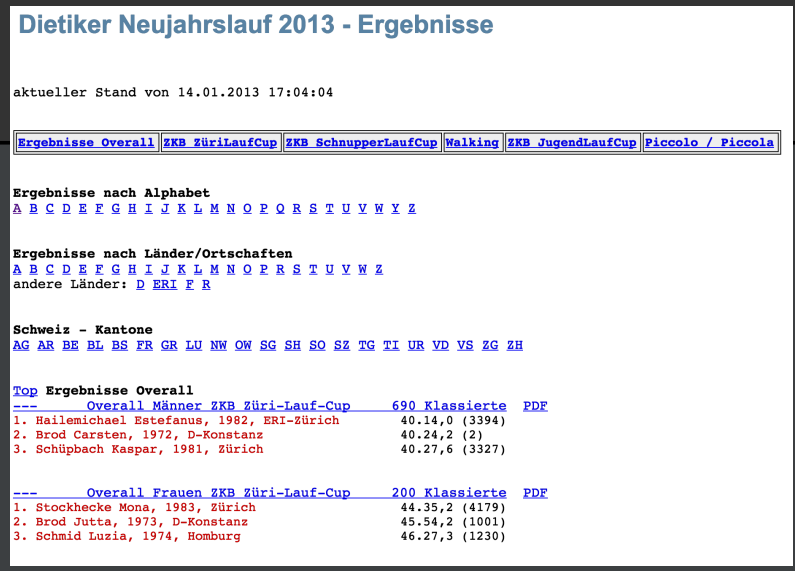
31.12.2013	Gippinger Stauseelauf	Running	Rangliste
31.12.2013	Silvesterlauf Gersau	Running,Walking	Rangliste
15.12.2013	Zürcher Silvesterlauf, Zürich	Running	Rangliste
14.12.2013	Course Titzé de noël Sion	Running	Rangliste
14.12.2013	Christmas Midnight Run, Lausanne	Running	Rangliste
14.12.2013	La Trotteuse-Tissot, La Chaux-de-Fonds	Running	Rangliste
07.12.2013	Course de l'Escalade, Genève	Running	Rangliste
07.12.2013	Gossauer Weihnachtslauf	Running,Walking	Rangliste
30.11.2013	Basler Stadtlauf	Running	Rangliste
30.11.2013	Foulées automnales de Meyrin	Running,Walking	Rangliste
24.11.2013	Course de l'Avent, Fribourg	Running	Rangliste
23.11.2013	Langenthaler Stadtlauf	Running	Rangliste
17.11.2013	Frauenfelder	Running,Waffenlauf	Rangliste
16.11.2013	Corrida Bulloise, Bulle	Running	Rangliste
10.11.2013	Maratona Ticino, Tenero	Running	Rangliste

» Wechseln zu Modus Expert | Einträge: 1-15 16-30 31-45 46-60

```
class MyCrawler(Spider):
    allowed_domains = ['www.running.ch'
    name = 'runningsite-2013'

    def start_requests(self):
        for month in range(1, 13):
            form_data = {
                'etyp': 'Running',
                'eventmonth': str(month),
                'eventyear': '2013',
                'eventlocation': 'CCH'
            }
            request = FormRequest('https://www.runningsite.com/de/',
                                  formdata=form_data,
                                  callback=self.parse_runs)
            # remember month in meta attributes for this request
            request.meta['paging_month'] = str(month)
            yield request
```

Page through result list

```
class MyCrawler(Spider):

  # ...

  def parse_runs(self, response):
    for run in response.css('#ds-calendar-body tr'):
      span = run.css('td:nth-child(1) span::text').extract()[0]
      run_date = re.search(r'(\d+\.\d+\.\d+).*', span).group(1)
      url = run.css('td:nth-child(5) a::attr("href")').extract()[0]
      for i in range(ord('a'), ord('z') + 1):
        request = Request(url + '/alfa{}.htm'.format(chr(i)),
                          callback=self.parse_run_page)
        request.meta['date'] = dt.strptime(run_date, '%d.%m.%Y')
        yield request
      next_page = response.css("ul.nav > li.next > a::attr('href')")
      if next_page: # recursively page until no more pages
        url = next_page[0].extract()
        yield scrapy.Request(url, self.parse_runs)
```



Dietiker Neujahrslauf 2013 - Ergebnisse

aktueller Stand von 14.01.2013 17:04:04

Ergebnisse Overall | ZKB ZüriLaufCup | ZKB SchnupperLaufCup | Walking | ZKB JugendLaufCup | Piccolo / Piccola

Ergebnisse nach Alphabet
A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Ergebnisse nach Länder/Ortschaften
A B C D E F G H I J K L M N O P R S T U V W Z
andere Länder: D ERI F R

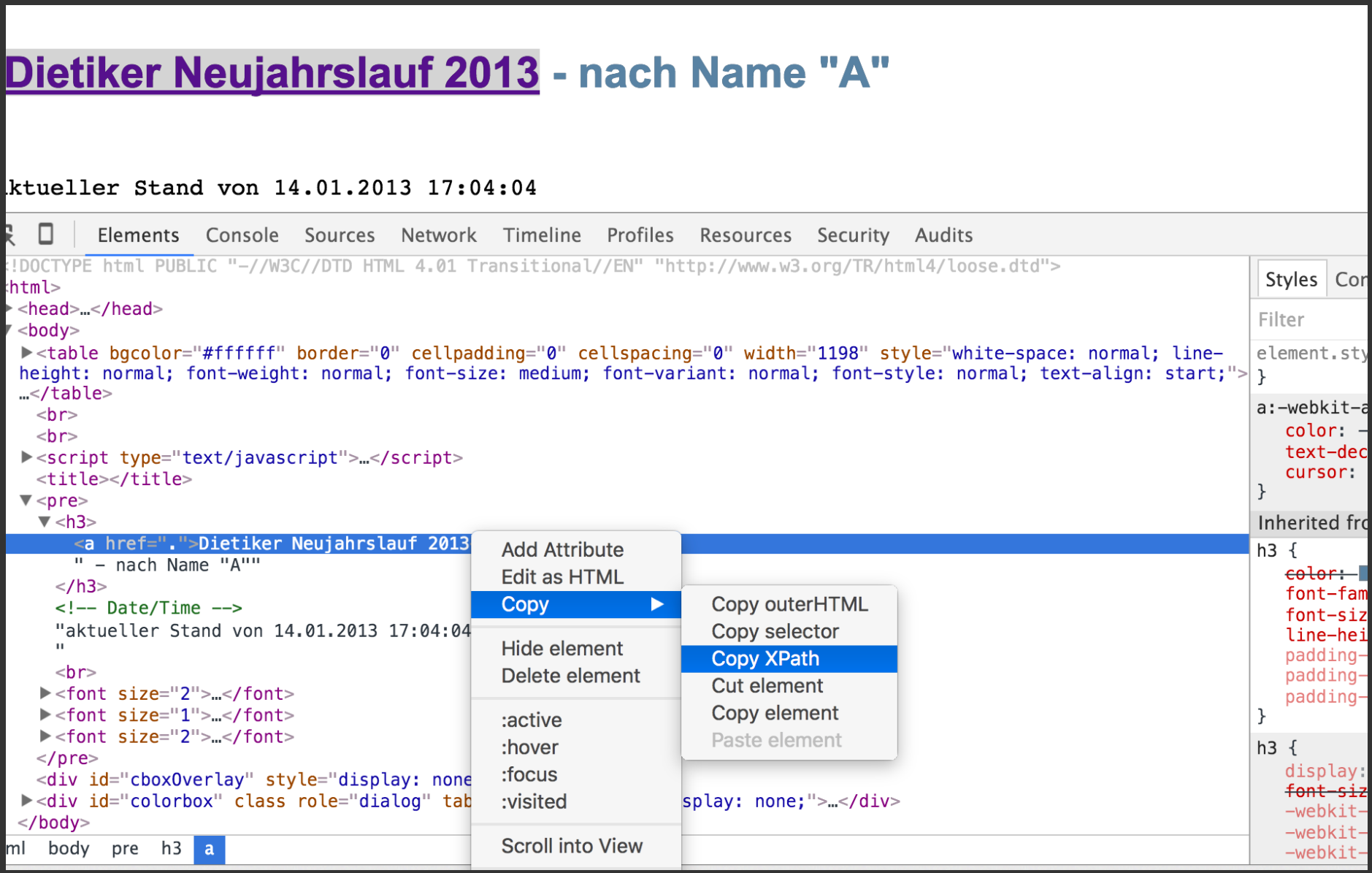
Schweiz - Kantone
AG AR BE BL BS FR GR LU NW OW SG SH SO SZ TG TI UR VD VS ZG ZH

Top **Ergebnisse Overall**

--- Overall Männer ZKB Züri-Lauf-Cup	690 Klassierte	PDF
1. Hailemichael Estefanus, 1982, ERI-Zürich	40.14,0 (3394)	
2. Brod Carsten, 1972, D-Konstanz	40.24,2 (2)	
3. Schüpbach Kaspar, 1981, Zürich	40.27,6 (3327)	

--- Overall Frauen ZKB Züri-Lauf-Cup	200 Klassierte	PDF
1. Stockhecke Mona, 1983, Zürich	44.35,2 (4179)	
2. Brod Jutta, 1973, D-Konstanz	45.54,2 (1001)	
3. Schmid Luzia, 1974, Homburg	46.27,3 (1230)	

Use your browser to generate XPath expressions

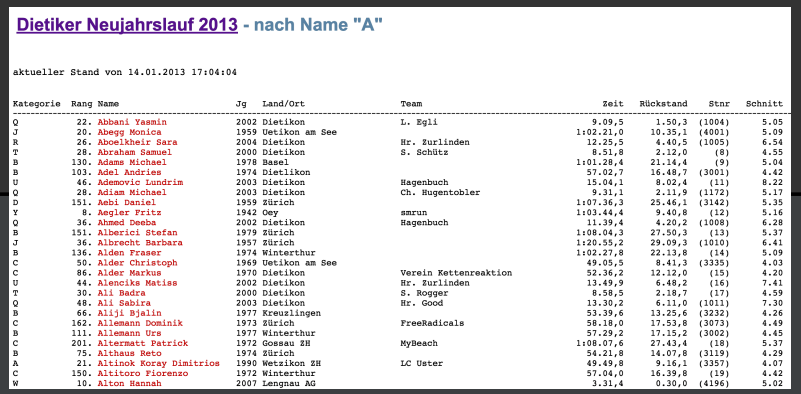


Real data can be messy!

Fun-M	---	Abegg Andreas	1949	Wallisellen	40.59,8	22.26,9	(6209)	Video	8.11
10-M40	171.	Abegg Daniel	1968	Hettlingen	41.26,4	7.51,3	(6210)	Video	4.08
Fun-F	---	Abegg Johanna	1951	Wallisellen	39.37,9	20.24,0	(6211)	Video	7.55
10-F50	39.	Abegg Monica	1959	Uetikon am See	50.41,6	10.20,5	(6212)	Video	5.04
U14W	DNF	Abegglen Aline	2001	Herrliberg	-----	-----	(3100)	Video	----
10-F20	81.	Abegglen Olivia	1992	Frauenfeld	47.08,2	12.18,2	(6213)	Video	4.42

posto	nome	an	nazione/località	squadra	tempo	ritardo	pett		media	1° gir
1.	Tedeschi Tessa	2001	Lumino		7.59,1	-----	(740)	diploma	3.48	2.31
2.	Lengen Lynn	2001	Glis		7.59,3	0.00,2	(723)	diploma	3.48	2.33
3.	Keller Michela	2000	Grono		8.09,9	0.10,8	(738)	diploma	3.53	2.38
4.	Tattarletti Vera	2001	Lugano	usc capriaschese	8.44,0	0.44,9	(739)	diploma	4.09	2.42
5.	Riggenberg Samira	2001	Ringgenberg BE		9.15,3	1.16,2	(743)	diploma	4.24	2.49
6.	Suergiu Giulia	2000	Manno		9.26,0	1.26,9	(731)	diploma	4.29	2.57
7.	Casale Alice	2000	Lugano		9.51,6	1.52,5	(708)	diploma	4.41	2.55
8.	Goessl Mya	2000	Lugano	SAL Lugano	9.55,8	1.56,7	(718)	diploma	4.43	2.48
9.	Gandini Asia	2000	Lugano		10.01,3	2.02,2	(747)	diploma	4.46	3.03
10.	Chatzidogiannaki Anna	2001	Castagnola		10.57,9	2.58,8	(742)	diploma	5.13	3.13
DNF	Vitali Valentina	2000	I-Gera Lario (CO)		0.00,7	Start K	(733)		----	----

Parse run results



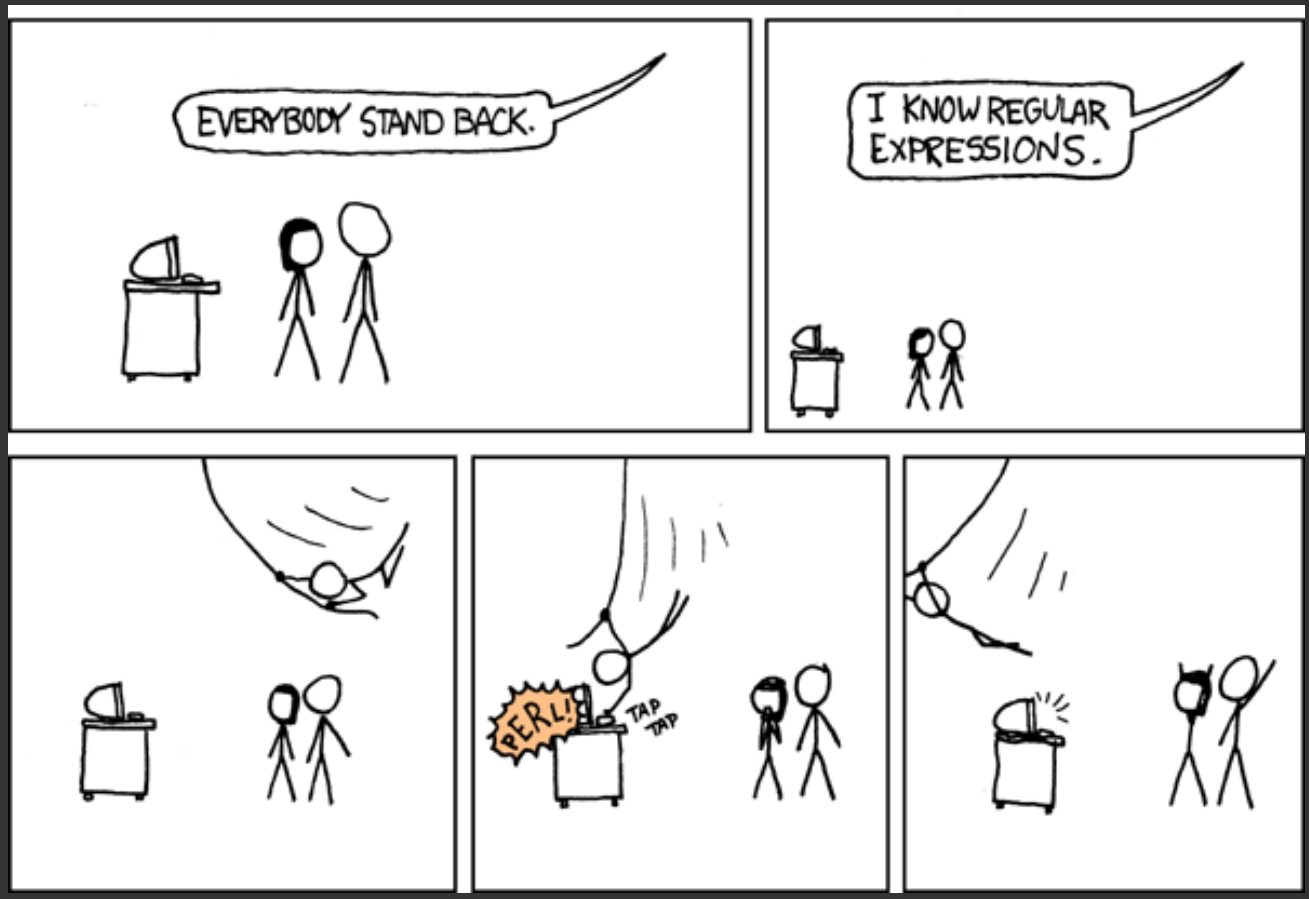
Dietiker Neujahrslauf 2013 - nach Name "A"

aktueller Stand von 14.01.2013 17:04:04

Kategorie	Rang	Name	Jg	Land/Ort	Team	Zeit	Rückstand	Stnr	Schnitt
Q	22.	Abbani Yasmin	2002	Dietikon	L. Egli	9.09,5	1.50,3	(1004)	5.05
J	20.	Abegg Monica	1959	Uetikon am See		1:02.21,0	10.35,1	(4001)	5.09
R	26.	Aboelkheir Sara	2004	Dietikon	Hr. Zurlinden	12.25,5	4.40,5	(1005)	6.54
T	28.	Abraham Samuel	2000	Dietikon	S. Schütz	8.51,8	2.12,0	(8)	4.55
B	130.	Adams Michael	1978	Basel		1:01.28,4	21.14,4	(9)	5.04
B	103.	Adel Andries	1974	Dietlikon		57.02,7	16.48,7	(3001)	4.42
U	46.	Ademovic Lundrim	2003	Dietikon	Hagenbuch	15.04,1	8.02,4	(11)	8.22
Q	28.	Adiam Michael	2003	Dietikon	Ch. Hugentobler	9.31,1	2.11,9	(1172)	5.17
D	151.	Aebi Daniel	1959	Zürich		1:07.36,3	25.46,1	(3142)	5.35
Y	8.	Aegler Fritz	1942	Oey	smrun	1:03.44,4	9.40,8	(12)	5.16
Q	36.	Ahmed Deeba	2002	Dietikon	Hagenbuch	11.39,4	4.20,2	(1008)	6.28
B	151.	Alberici Stefan	1979	Zürich		1:08.04,3	27.50,3	(13)	5.37
J	36.	Albrecht Barbara	1957	Zürich		1:20.55,2	29.09,3	(1010)	6.41
B	136.	Alden Fraser	1974	Winterthur		1:02.27,8	22.13,8	(14)	5.09
C	50.	Alder Christoph	1969	Uetikon am See		49.05,5	8.41,3	(3335)	4.03
C	86.	Alder Markus	1970	Dietikon	Verein Kettenreaktion	52.36,2	12.12,0	(15)	4.20
U	44.	Alenciks Matiss	2002	Dietikon	Hr. Zurlinden	13.49,9	6.48,2	(16)	7.41
T	30.	Ali Badra	2000	Dietikon	S. Rogger	8.58,5	2.18,7	(17)	4.59
Q	48.	Ali Sabira	2003	Dietikon	Hr. Good	13.30,2	6.11,0	(1011)	7.30
B	66.	Aliji Bjalin	1977	Kreuzlingen		53.39,6	13.25,6	(3232)	4.26
C	162.	Allemann Dominik	1973	Zürich	FreeRadicals	58.18,0	17.53,8	(3073)	4.49
B	111.	Allemann Urs	1977	Winterthur		57.29,2	17.15,2	(3002)	4.45
C	201.	Altermatt Patrick	1972	Gossau ZH	MyBeach	1:08.07,6	27.43,4	(18)	5.37
B	75.	Althaus Reto	1974	Zürich		54.21,8	14.07,8	(3119)	4.29
A	21.	Altinok Koray Dimitrios	1990	Wetzikon ZH	LC Uster	49.49,8	9.16,1	(3357)	4.07
C	150.	Altitoro Fiorenzo	1972	Winterthur		57.04,0	16.39,8	(19)	4.42
W	10.	Alton Hannah	2007	Lengnau AG		3.31,4	0.30,0	(4196)	5.02

```
class MyCrawler(Spider):
    # ...
    def parse_run_page(self, response):
        run_name = response.css('h3 a::text').extract()[0]
        html = response.xpath('//pre/font[3]').extract()[0]
        results = lxml.html.document_fromstring(html).text_content()
        rre = re.compile(r'(?P<category>.*?)\s+'
                         r'(?P<rank>(?:\d+|—+|DNF))\.?\s'
                         r'(?P<name>(?!(?:\d{2,4})).*?)'
                         r'(?P<ageGroup>(?:\?\?|\d{2,4}))\s'
                         r'(?P<city>.*?)\s{2,}'
                         r'(?P<team>(?!(?:\d+:)?\d{2}\.\d{2},\d).*?)'
                         r'(?P<time>(?:\d+:)?\d{2}\.\d{2},\d)\s+'
                         r'(?P<deficit>(?:\d+:)?\d+\.\d+,\d)\s+'
                         r'\((?P<startNumber>\d+)\).*?'
                         r'(?P<pace>(?:\d+\.\d+|—+))')
        # result_fields = rre.search(result_line) ...
```

Regex: now you have two problems



- Handling scraping results with regular expressions can soon get messy
- $\rightarrow$ Better use a real parser

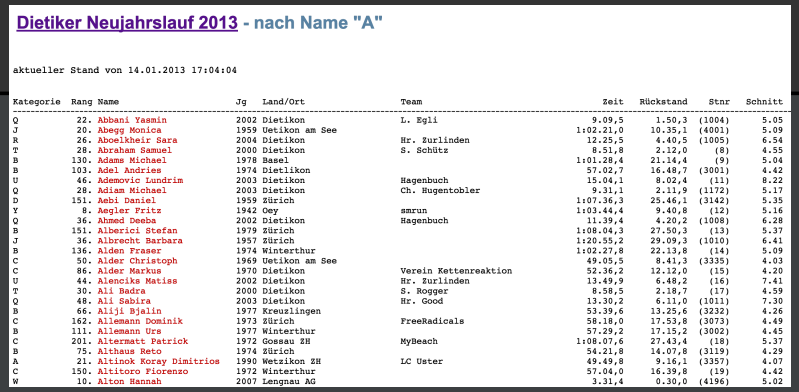
Parse run results with pyparsing

```
from pyparsing import *

SPACE_CHARS = ' \t'
dnf = Literal('dnf')
space = Word(SPACE_CHARS, exact=1)
words = delimitedList(Word(alphas), delim=space, combine=True)

category = Word(alphanums + '-_')
rank = (Word(nums) + Suppress('.')) | Word('-') | dnf
age_group = Word(nums)
run_time = ((Regex(r'(\d+:)?\d{1,2}\.\d{2}(,\d)?')
             | Word('-') | dnf).setParseAction(time2seconds))
start_number = Suppress('(') + Word(nums) + Suppress(')')
run_result =
    (category('category') + rank('rank') + words('runner_name')
    + age_group('age_group') + words('team_name')
    + run_time('run_time') + run_time('deficit')
    + start_number('start_number').setParseAction(lambda t: int(t[0]))
    + Optional(run_time('pace')) + SkipTo(lineEnd))
```

Items and data processors

```
def dnf(value):
    if value == 'DNF' or re.match(r'—+', value):
        return None
    return value

def time2seconds(value):
    t = time.strptime(value, '%H:%M.%S,%f')
    return datetime.timedelta(hours=t.tm_hour,
                              minutes=t.tm_min,
                              seconds=t.tm_sec).total_seconds()

class RunResult(scrapy.Item):
    run_name = scrapy.Field(input_processor=MapCompose(unicode.strip),
                            output_processor=TakeFirst())
    time = scrapy.Field(
        input_processor=MapCompose(unicode.strip , dnf, time2seconds),
        output_processor=TakeFirst()
    )
```

Using Scrapy item loaders



Dietiker Neujahrslauf 2013 - nach Name "A"

aktueller Stand von 14.01.2013 17:04:04

Kategorie	Rang	Name	Jg	Land/Ort	Team	Zeit	Rückstand	Stnr	Schnitt
Q	22.	Abbani Yasmin	2002	Dietikon	L. Egli	9.09,5	1.50,3	(1004)	5.05
J	20.	Abegg Monica	1959	Uetikon am See		1:02.21,0	10.35,1	(4001)	5.09
R	26.	Aboelkheir Sara	2004	Dietikon	Hr. Zurlinden	12.25,5	4.40,5	(1005)	6.54
T	28.	Abraham Samuel	2000	Dietikon	S. Schütz	8.51,8	2.12,0	(8)	4.55
B	130.	Adams Michael	1978	Basel		1:01.28,4	21.14,4	(9)	5.04
B	103.	Adel Andries	1974	Dietlikon		57.02,7	16.48,7	(3001)	4.42
U	46.	Ademovic Lundrim	2003	Dietikon	Hagenbuch	15.04,1	8.02,4	(11)	8.22
Q	28.	Adiam Michael	2003	Dietikon	Ch. Hugentobler	9.31,1	2.11,9	(1172)	5.17
D	151.	Aebi Daniel	1959	Zürich		1:07.36,3	25.46,1	(3142)	5.35
Y	8.	Aegler Fritz	1942	Oey	smrun	1:03.44,4	9.40,8	(12)	5.16
Q	36.	Ahmed Deeba	2002	Dietikon	Hagenbuch	11.39,4	4.20,2	(1008)	6.28
B	151.	Alberici Stefan	1979	Zürich		1:08.04,3	27.50,3	(13)	5.37
J	36.	Albrecht Barbara	1957	Zürich		1:20.55,2	29.09,3	(1010)	6.41
B	136.	Alden Fraser	1974	Winterthur		1:02.27,8	22.13,8	(14)	5.09
C	50.	Alder Christoph	1969	Uetikon am See		49.05,5	8.41,3	(3335)	4.03
C	86.	Alder Markus	1970	Dietikon	Verein Kettenreaktion	52.36,2	12.12,0	(15)	4.20
U	44.	Alencika Matiss	2002	Dietikon	Hr. Zurlinden	13.49,9	6.48,2	(16)	7.41
T	30.	Ali Badra	2000	Dietikon	S. Rogger	8.58,5	2.18,7	(17)	4.59
Q	48.	Ali Sabira	2003	Dietikon	Hr. Good	13.30,2	6.11,0	(1011)	7.30
B	66.	Aliji Bjalin	1977	Kreuzlingen		53.39,6	13.25,6	(3232)	4.26
C	162.	Allemann Dominik	1973	Zürich	FreeRadicals	58.18,0	17.53,8	(3073)	4.49
B	111.	Allemann Urs	1977	Winterthur		57.29,2	17.15,2	(3002)	4.45
C	201.	Altermatt Patrick	1972	Gossau ZH	MyBeach	1:08.07,6	27.43,4	(18)	5.37
B	75.	Althaus Reto	1974	Zürich		54.21,8	14.07,8	(3119)	4.29
A	21.	Altinok Koray Dimitrios	1990	Wetzikon ZH	LC Uster	49.49,8	9.16,1	(3357)	4.07
C	150.	Altitoro Fiorenzo	1972	Winterthur		57.04,0	16.39,8	(19)	4.42
W	10.	Alton Hannah	2007	Lengnau AG		3.31,4	0.30,0	(4196)	5.02

```
class MyCrawler(Spider):
    # ...
    def parse_run_page(self, response):
        # ...
        for result_line in all_results.splitlines():
            fields = result_fields_re.search(result_line)
            il = ItemLoader(item=RunResult())
            il.add_value('run_date', response.meta['run_date'])
            il.add_value('run_name', run_name)
            il.add_value('category', fields.group('category'))
            il.add_value('rank', fields.group('rank'))
            il.add_value('runner_name', fields.group('name'))
            il.add_value('age_group', fields.group('ageGroup'))
            il.add_value('team', fields.group('team'))
            il.add_value('time', fields.group('time'))
            il.add_value('deficit', fields.group('deficit'))
            il.add_value('start_number', fields.group('startNumber'))
            il.add_value('pace', fields.group('pace'))
            yield il.load_item()
```

Ready, steady, crawl!

```
⏚ ⌂61% [michael:~/Projects/laufscraper/crawler] [venv] master(+0/-1092)* 5s ± scrapy crawl crawler2013
2016-02-03 00:44:34 [scrapy] INFO: Scrapy 1.0.4 started (bot: crawler)
2016-02-03 00:44:34 [scrapy] INFO: Optional features available: ssl, http11
2016-02-03 00:44:34 [scrapy] INFO: Overridden settings: {'NEWSPIDER_MODULE': 'crawler.spiders', 'SPIDER_MODULES': ['cra
': 'crawler'}
2016-02-03 00:44:34 [scrapy] INFO: Enabled extensions: CloseSpider, TelnetConsole, LogStats, CoreStats, SpiderState
2016-02-03 00:44:34 [scrapy] INFO: Enabled downloader middlewares: HttpAuthMiddleware, DownloadTimeoutMiddleware, UserA
leware, DefaultHeadersMiddleware, MetaRefreshMiddleware, HttpCompressionMiddleware, RedirectMiddleware, CookiesMiddlewa
ware, DownloaderStats
2016-02-03 00:44:34 [scrapy] INFO: Enabled spider middlewares: HttpErrorMiddleware, OffsiteMiddleware, RefererMiddlewar
epthMiddleware
2016-02-03 00:44:34 [py.warnings] WARNING: /Users/michael/Projects/laufscraper/venv/lib/python2.7/site-packages/scrapy/
 ScrapyDeprecationWarning: ITEM_PIPELINES defined as a list or a set is deprecated, switch to a dict
  category=ScrapyDeprecationWarning, stacklevel=1)

2016-02-03 00:44:34 [py.warnings] WARNING: /Users/michael/Projects/laufscraper/crawler/crawler/pipelines.py:13: ScrapyD
 `scrapy.log` has been deprecated, Scrapy now relies on the builtin Python library for logging. Read the updated loggin
ion to learn more.
  from scrapy import log

2016-02-03 00:44:34 [scrapy] INFO: Enabled item pipelines: MongoDBPipeline, ElasticSearchPipeline
2016-02-03 00:44:34 [scrapy] INFO: Spider opened
2016-02-03 00:44:34 [scrapy] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2016-02-03 00:44:34 [scrapy] DEBUG: Telnet console listening on 127.0.0.1:6029
2016-02-03 00:44:37 [scrapy] DEBUG: Crawled (200) <POST https://www.[redacted]/de/> (referer: None)
2016-02-03 00:44:37 [scrapy] DEBUG: Crawled (200) <POST https://www.[redacted]/de/> (referer: None)
```

Storing items with an Elasticsearch pipeline

```
from pyes import ES
# Configure your pipelines in settings.py
ITEM_PIPELINES = ['crawler.pipelines.MongoDBPipeline',
                  'crawler.pipelines.ElasticSearchPipeline']

class ElasticSearchPipeline(object):
    def __init__(self):
        self.settings = get_project_settings()
        uri = "{}:{}".format(self.settings['ELASTICSEARCH_SERVER'],
                             self.settings['ELASTICSEARCH_PORT'])
        self.es = ES([uri])

    def process_item(self, item, spider):
        index_name = self.settings['ELASTICSEARCH_INDEX']
        self.es.index(dict(item), index_name,
                      self.settings['ELASTICSEARCH_TYPE'],
                      op_type='create')
        # raise DropItem('If you want to discard an item')
        return item
```

Scrapy can do much more!

- **Throttling crawling speed** based on load of both the Scrapy server and the website you are crawling
- **Scrapy Shell**: An interactive environment to try and debug your scraping code

```
[michael:~/Projects/laufscraper/crawler] [venv] master(+0/-1092)* 1 ± scrapy shell 'http://www.python-summit.ch/pages/program.html'
[s] Available Scrapy objects:
[s]   crawler    <scrapy.crawler.Crawler object at 0x10cbd6d50>
[s]   item       {}
[s]   request    <GET http://www.python-summit.ch/pages/program.html>
[s]   response   <200 http://www.python-summit.ch/pages/program.html>
[s]   settings   <scrapy.settings.Settings object at 0x10e269b90>
[s]   spider     <DefaultSpider 'default' at 0x11144c210>
[s] Useful shortcuts:
[s]   shelp()           Shell help (print this help)
[s]   fetch(req_or_url) Fetch request (or URL) and update local objects
[s]   view(response)    View response in a browser

>>> response.xpath('//*[@id="program"]/tbody//a/text()').extract()
[u"Python's Guide to the Galaxy", u'API Design is Hard', u'CFFI: Call C from Python', u'3D Computer Graphics with Python', u'Charming
hon for Live Music', u'Coding/Decoding the Cosmos: Python Applications in Astrophysics', u'Scrapy and Elasticsearch: Powerful Web Scr
earching with Python', u'Getting Started with IPython', u'Pytest: Rapid Simple Testing']
```

Scrapy can do much more!

- **Feed exports**: Supported serialization of scraped items to JSON, XML or CSV
- **Scrapy Cloud**: "*It's like a Heroku for Scrapy*" - Source: Scrapy Cloud
- **Jobs**: pausing and resuming crawls
- **Contracts**: test your spiders by specifying constraints for how the spider is expected to process a response

```
def parse_runresults_page(self, response):
    """ Contracts within docstring – available since Scrapy 0.15

    @url http://www.runningsite.ch/runs/hallwiler
    @returns items 1 25
    @returns requests 0 0
    @scrapes RunDate Distance RunName Winner
    """
```

Elasticsearch

Elasticsearch 101

- REST and JSON based document store
- Stands on the shoulders of Lucene
- Apache 2.0 licensed
- Distributed and scalable
- Widely used (Github, SonarQube, ...)

Elasticsearch building blocks

- **RDBMS** $\rightarrow$ Databases $\rightarrow$ Tables $\rightarrow$ Rows $\rightarrow$ Columns
- **ES** $\rightarrow$ Indices $\rightarrow$ Types $\rightarrow$ Documents $\rightarrow$ Fields
- By default every field in a document is indexed
- Concept of **inverted index**

Create a document with cURL

```
$ curl –XPUT http://localhost:9200/running/result/1 –d '
{
  "name": "Haile Gebrselassie",
  "pace": 2.8,
  "age": 42,
  "goldmedals": 10
}'
```

```
$ curl –XGET http://localhost:9200/results/_mapping?pretty
{ "results" : {
    "mappings" : {
      "result" : {
        "properties" : {
          "age" : {
            "type" : "long"
          },
          "goldmedals" : {
            "type" : "long"
```

Retrieve document with cURL

```
$ curl –XGET http://localhost:9200/results/result/1
{
  "_index": "results",
  "_type": "result",
  "_id": "1",
  "_version": 1,
  "found": true,
  "_source": {
    "name": "Haile Gebrselassie",
    "pace": 2.8,
    "age": 42,
    "goldmedals": 10
  }
}
```

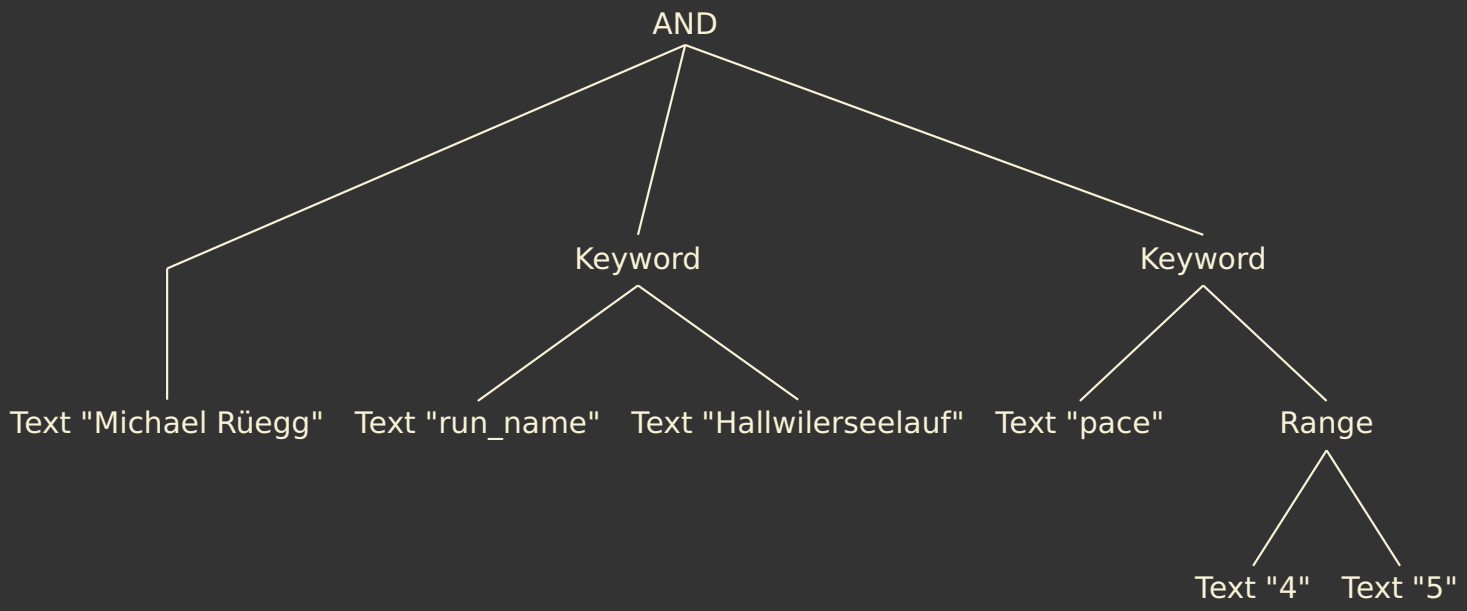
Searching with the Elasticsearch Query DSL

```
$ curl –XGET http://localhost:9200/results/_search –d '{
"query" : {
    "filtered" : {
        "filter" : {
            "range" : { "age" : { "gt" : 40 } }
        },
        "query" : {
            "match" : { "name" : "haile" }
        }
    }
}
{ "hits": {
    "total": 1,
    "max_score": 0.19178301,
    "hits": [{
        "_source": { "name": "Haile Gebrselassie", // ...  }
    }]
} }
```

Implementing a query DSL

A query DSL for run results

```
"michael rüegg" and run_name:"Hallwilerseelauf" and pace:[4 to 5]
```




```
'filtered': { 'filter':  {
    'bool': {
        'must': [
            {'match_phrase': {'_all': 'michael rüegg'}},
            {'match_phrase': {'run_name': u'Hallwilerseelauf'}},
            {'range': {'pace': {'gte': u'4', 'lte': u'5'}}}
        ]
    }
} }
```

AST generation and traversal

```
text = valid_word.setParseAction(lambda t: TextNode(t[0])
match_phrase = QuotedString('"').setParseAction(
    lambda t: MatchPhraseNode(t[0])
)
incl_range_search = Group(Literal('[') + term('lower')
                          + CaselessKeyword("to") + term('upper')
                          + Literal(']')
                         ).setParseAction(lambda t: RangeNode(t[0])
range_search = incl_range_search | excl_range_search
query << operatorPrecedence(term, [
    (CaselessKeyword('not'), 1, opAssoc.RIGHT, NotSearch),
    (CaselessKeyword('and'), 2, opAssoc.LEFT AndSearch),
    (CaselessKeyword('or'), 2, opAssoc.LEFT, OrSearch)
])
class NotSearch(UnaryOperation):
    def get_query(self, field):
        return { 'bool' : {
            'must_not': self.op.get_query(field)
        } }
```

Demo

Questions?