Trowel

A programming language for journalists

David Tagatac
- Language Guru
Hareesh Radhakrishnan
- System Architect
Pucong Han
- System Integrator
Robert Walport
- System Tester
Victoria Mo
- Project Manager
The Problem

- Journalists need to find relevant information online.
- Web scraping quickens this process: requires knowledge of advanced programming languages.
- Typical journalist not technically proficient: does not code.
Trowel is...

A web-scraping programming language developed specifically for journalists.
Trowel is...

![Graphical Interfaces](#)

![Other Programming Languages](#)

**Useful**

**Ease of use for Journalists**
Trowel is...

- Readable
- Accessible
- Concise
- Easy-to-Learn
- Intuitive
- Domain-Specific
Design Goals

Readability - looks like English!

- No Semi-colons
- Indentation handles scoping
- Assignment uses “is” instead of “=“
Design Goals

Built in Functions Make Sense

“insert Url into UrlList”

inserts the Url variable into the UrlList variable!
Sir Richard Branson's Virgin Galactic spaceship ignites engine in flight

By Jonathan Amos
Science correspondent, BBC News

The spaceplane being developed by UK billionaire Sir Richard Branson has made its first powered flight.

The vehicle was dropped from a carrier aircraft high above California's Mojave Desert and ignited its rocket engine in a pre-supersonic free flight test.
Live Demo: Liftoff!

url articleurl is ‘http://www.bbc.co.uk/news/science-environment-22344398’

paragraphs is findtext in articleurl with "time" and "flight"

print paragraphs
We wanted to make Trowel as close to English as possible

# Flexibility vs Robust Design

> read "file.txt" into ul1

> ul2 is findurl in ul1 with "obama" and "romney"

  ul4 is combine ul3 with (findurl in ul2 with "taxes")

> ul5 is mydelete ul3 from ul2 except indexlist1
- textlist para, words
- url article is 'bit.ly/trowel'
- para is findtext in article with "time"
- print para
textlist para, words
url article is 'bit.ly/trowel'
para is findtext in article with "time"
print para
textlist para, words
url article is 'bit.ly/trowel'
para is findtext in article with "time"
print para

['declaration', ['datatype', 'textlist'], [['paragraphs'], ['words']]]

['declaration', ['datatype', 'url'], [['articleurl', ['expression', ['value', ['url', 'bit.ly/trowel']]]]]]

['assignment', ['variable', 'paragraphs'], ['expression', ['functioncall', ['functionname', 'findtext'], ['arguments', [['expression', ['insertword', 'in']], ['expression', ['variable', 'articleurl']], ['expression', ['insertword', 'with']], ['expression', ['value', ['text', 'time']]]]]]]

['functioncall', ['functionname', 'print'], ['arguments', [['expression', ['variable', 'paragraphs']]]]
import trowelfunctions as tfl

paragraphs = ""
words = ""

articleurl = 'bit.ly/trowel'

tmp0 = 'in'
tmp1 = articleurl
tmp2 = 'with'
tmp3 = 'time'
paragraphs = tfl.r_findtext([tmp0,tmp1,tmp2,tmp3])

tmp0 = paragraphs
print(tmp0)
Example Function: FindText

- receive a URL to search and a logical expression
  - e.g. "Obama and Romney"
- Use Beautiful Soup to grab website and parse it
- convert the logical expression of terms into a logical expression of booleans for each paragraph
  - e.g. "Obama and Romney" --> "True and True"
- if expression evaluates to true: return the paragraph
Example Function: FindText

```python
def _findtext(arglist):
    link = arglist[1]
    parts = urlparse.urlsplit(link)
    if not parts.scheme or not parts.netloc:
        link = "http://" + link
    html = urlopen(link)
    soup = BeautifulSoup(html)
    texts = soup.find_all('p')
    keyparas = []

    for para in texts:
        para = para.get_text()
        truthiList = ""
        for entry in arglist[2:]:
            if str(type(entry)) != "<type 'list'>":
                if entry in LOGICALS:
                    truthiList = truthiList + " " + entry
                elif entry in IGNORE:
                    pass
                elif entry in para:
                    truthiList = truthiList + " True"
                else:
                    truthiList = truthiList + " False"
            if eval(truthiList):
                keyparas.append(para)
    return keyparas
```

Para is findtext in Article
Development Environment

Beautiful Soup

Python

Python Lex-Yacc
Regression testing: git hook disallows “git commit” if any test fails!

- Standard unit testing framework - `unittest`
- Tested the output of the lexer, the parser, the function modules and the whole compiler
Testing: Liftoff!

Lexer test
self.assertEqual(parsewrapper().gettokens("url spaceArticle is \\
[['indentlevel', 0], ['declaration', ['datatype', 'url'],  
co.uk/news/science-environment-22344398']]]]]]]

Function test
self.assertEqual(r_findtext(['in', 'http://www.bbc.co.  
uk/news/science-environment-22344398']),  
[u"The vehicle was dropped from a carrier aircraft high above  
California's Mojave Desert and ignited its rocket engine to go  
supersonic for a few seconds...."])

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexer tests</td>
<td>24%</td>
</tr>
<tr>
<td>Parser tests</td>
<td>10%</td>
</tr>
<tr>
<td>Function tests</td>
<td>44%</td>
</tr>
<tr>
<td>Type Checking</td>
<td>22%</td>
</tr>
</tbody>
</table>
Example: Amy

● Amy wants the latest tweets of a list of politicians.
● Amy has been given a similar assignment before.
● She thinks to herself...
● Trowel!
define getTweets of (text person) from twitter:
  text prefix is ‘http://www.twitter.com/’
  url twitterurl is combine prefix and person
  results is findText in twitterurl
  return results

textlist alltweets

for name in ['BarackObama', 'MittRomney', 'JoeBiden']:
  textList tweets is getTweets of name from twitter
  insert tweets into alltweets

save alltweets into "tweetsfile.txt"
define getTweets of (text person) from twitter:
  text prefix is 'http://www.twitter.com/
  url twitterurl is combine prefix and person
  results is findText in twitterurl
  return results

#define functionname : [returntype(s)]
prebuiltfunctions = {
  'print' : [None],
  'read' : ['urllist','textlist'],
  'save' : [None],
  'append' : [None],
  'insert' : [None],
  'findurl' : ['urllist'],
  'findtext' : ['textlist'],
  'combine' : ['url'],
}

'gettweets': ['textlist'],
Example: Amy

textlist alltweets
for person in ['BarackObama', 'MittRomney', 'JoeBiden']:
    textList tweets is getTweets of person from twitter
    insert tweets into alltweets

alltweets = [...]
Example: Amy

```python
alltweets = [...]  
save alltweets into "tweetsfile.txt"
```

tweetsfile.txt

Add your name now — http://OFA.BO/s9x6FH Common-sense gun violence prevention won't happen unless Congress hears our voices.
@OFA will deliver this petition to Congress demanding common-sense gun violence prevention. Add your name now: http://OFA.BO/s9x6FH
Watch the full video of President Obama at the White House Correspondents' Dinner: http://OFA.BO/HUEtit
Getting ready for the Correspondents' Dinner: http://at.wh.gov/kuvSS #WHCD
TONIGHT: President Obama at the White House Correspondents' Dinner, hosted by @ConanOBrien. Watch live at 9:45 ET: http://OFA.BO/v9dR8v
Note to Congress: Your constituents are paying attention, and they overwhelmingly support background checks. http://OFA.BO/WMzqE
Congress should find the same sense of urgency to help families as they did to help themselves. http://OFA.BO/6qAkqe #sequester
Thanks for your support and for speaking this morning @BarackObama! We'll keep educating patients on what the #ACA Enjoyed speaking to the @SVWedu class of 2013. Wish them luck as they jump into the deep waters of life. http://youtu.be/bcnTYCZRJ9I
Our hearts are heavy with the news out of Boston today. #PrayforBoston
Celebrating 44 years with my sweetheart today. Happy anniversary, @AnnROmney pic.twitter.com/ZerdIIRDH
Today, @VP Biden participated in a Google+ Hangout about the Admin's plan to reduce gun violence: http://wh.gov/yAKy pic.twitter.com/NVrb1eQA
Starting now: @VP Biden participates in a Google+ Hangout on reducing gun violence. Watch it live: http://wh.gov/nowishtime
Today at 1:45ET: @VP joins his 1st Google+ Hangout to discuss reducing gun violence. Watch it live: http://www.wh.gov/NowIsTheTime
After #Newtown, people are calling for action to reduce gun violence. The President & VP are listening & taking action: http://at.wh.gov/g1fK2
```

Add your name now — http://OFA.BO/s9x6FH Common-sense gun violence prevention won't happen unless Congress hears our voices.
@OFA will deliver this petition to Congress demanding common-sense gun violence prevention. Add your name now: http://OFA.BO/s9x6FH
Watch the full video of President Obama at the White House Correspondents' Dinner: http://OFA.BO/HUEtit
Getting ready for the Correspondents' Dinner: http://at.wh.gov/kuvSS #WHCD
TONIGHT: President Obama at the White House Correspondents' Dinner, hosted by @ConanOBrien. Watch live at 9:45 ET: http://OFA.BO/v9dR8v
Note to Congress: Your constituents are paying attention, and they overwhelmingly support background checks. http://OFA.BO/WMzqE
Congress should find the same sense of urgency to help families as they did to help themselves. http://OFA.BO/6qAkqe #sequester
Thanks for your support and for speaking this morning @BarackObama! We'll keep educating patients on what the #ACA Enjoyed speaking to the @SVWedu class of 2013. Wish them luck as they jump into the deep waters of life. http://youtu.be/bcnTYCZRJ9I
Our hearts are heavy with the news out of Boston today. #PrayforBoston
Celebrating 44 years with my sweetheart today. Happy anniversary, @AnnROmney pic.twitter.com/ZerdIIRDH
Today, @VP Biden participated in a Google+ Hangout about the Admin's plan to reduce gun violence: http://wh.gov/yAKy pic.twitter.com/NVrb1eQA
Starting now: @VP Biden participates in a Google+ Hangout on reducing gun violence. Watch it live: http://wh.gov/nowishtime
Today at 1:45ET: @VP joins his 1st Google+ Hangout to discuss reducing gun violence. Watch it live: http://www.wh.gov/NowIsTheTime
After #Newtown, people are calling for action to reduce gun violence. The President & VP are listening & taking action: http://at.wh.gov/g1fK2
```
define getTweets of (text person) searching twitter:
    text prefix is ‘http://www.twitter.com/’
    url twitterurl is combine prefix and person
    results is findText in twitterurl
    return results

for person in ['BarackObama', 'MittRomney', 'JoeBiden']:
    textList tweets is getTweets of person searching twitter
    insert tweets into (textList alltweets)

save alltweets into "tweetsfile.txt"
Project Management
1. Touch Base
2. Focus on the Big Picture
3. Respect Your Teammates
Lessons Learned
1. Know Git!

- **Version Control: Git** ([GitHub](https://github.com))
- **Repository Structure:**
  - Use master by default
  - New branch for code that breaks regression tests
- Use descriptive commit messages!
2. Plan Well and Iterate Quickly

- Being modular up-front saves time in testing and bug-fixing

- Planning before coding is important, but early coding helps the planning process
Possible Expansions to Trowel

**The Featuritis Curve**

- **User Happiness**
  - **Happy User Peak**
    - “I Rule!”
    - “Guess I better look at the manual…”
  - “Cool!”
  - “I’m so glad they added this.”
  - “Nice, but I wish I could do more…”
- **Number of Features**
  - “Hey, where the f*** did they put that?!”
  - “Now I can’t even do the ONE SIMPLE THING I bought this for…”
  - “I Suck!”
1. Help Journalists be Good People

- GET request header with journalist's name and email
- Easy setting of a delay between GET requests

Anderson Cooper likes being a good person.
2. Other Cool, Useful Functionality

- Crawl a news website
- findText using other html tags
- Find the similarity index between two articles
Conclusion

Journalists are not Computer Scientists

They have different goals, challenges and thought patterns

Trowel works for them