Data Acquisition via Application Programmable Interfaces

Adrian Petrescu

Rubikloud

2018-11-12

Schedule

Part 1

- Introduction and motivation
- HTTP, REST, and the languages of the web
- Authentication schemes
- Lots and lots of practical examples

Part 2

- Scraping unstructured data from the web
- Parsing
- Automated spiders

Part 3

- The server-side of APIs
- Deploying to the Cloud
- Project Description

Web Architecture

In Its Modern Form

HTML & CSS

Like Peanut Butter and Jelly

HTML

HyperText Markup Language

```
<HTML>
  <HEAD>
    <TITLE>Star Wars - Wikipedia</TITLE>
    <LINK REL="stylesheet" HREF="/static/style.css" />
  </HEAD>
  <BODY>
    ...
  </BODY>
</HTML>
```

HTML

HyperText Markup Language

Links

```
<P>
  Never gonna <A HREF="https://www.youtube.com/watch?v=dQw4w9WgXcQ">give you up</A>
</P>
```

Lists

```
<OL>
  <LI>Item 1</LI>
  <LI>Item 2</LI>
</OL>
```

Images

```
<IMG SRC="https://www.google.com/images/branding/googlelogo/1x/googlelogo_color_272x92dp.png" />
```

HTML

HyperText Markup Language

Tables

```
<table>
  <thead>
    <tr>
      <th>Month</th>
      <th>Savings</th>
    </tr>
  </thead>
  <tbody>
    <tr>
      <td>January</td>
      <td>$100</td>
    </tr>
    <tr>
      <td>February</td>
      <td>$80</td>
    </tr>
  </tbody>
  <tfoot>
    <tr>
      <td>Sum</td>
      <td>$180</td>
    </tr>
  </tfoot>
</table>
```

HTML

HyperText Markup Language

Div

```
<DIV ID="content" CLASS="full-width col">
  ...
</DIV>
```

Span

```
<P>
  This November is the Wikipedia Donation Drive.
  <SPAN STYLE="color:#36c;">Come join us.</SPAN>
</P>
```

CSS

Cascading Style Sheets

Turn all links red

```
a {
  color: #ff0000;
}
```

Only use 80% of the screen

```
body {
  width: 80%;
}
```

Center-align tables in the sidebar

```
#sidebar table {
  text-align: center;
}
```

Color bookmarked headlines

```
h2.title.bookmarked {
  color: blue;
}
```

Hide all non-HTTPS links in the sidebar

```
.sidebar :not(a[href^="https"]) {
  visibility: none;
}
```

Bold the first paragraph after any headline

```
h1,h2,h3+p {
  font-weight: bold;
}
```

wget

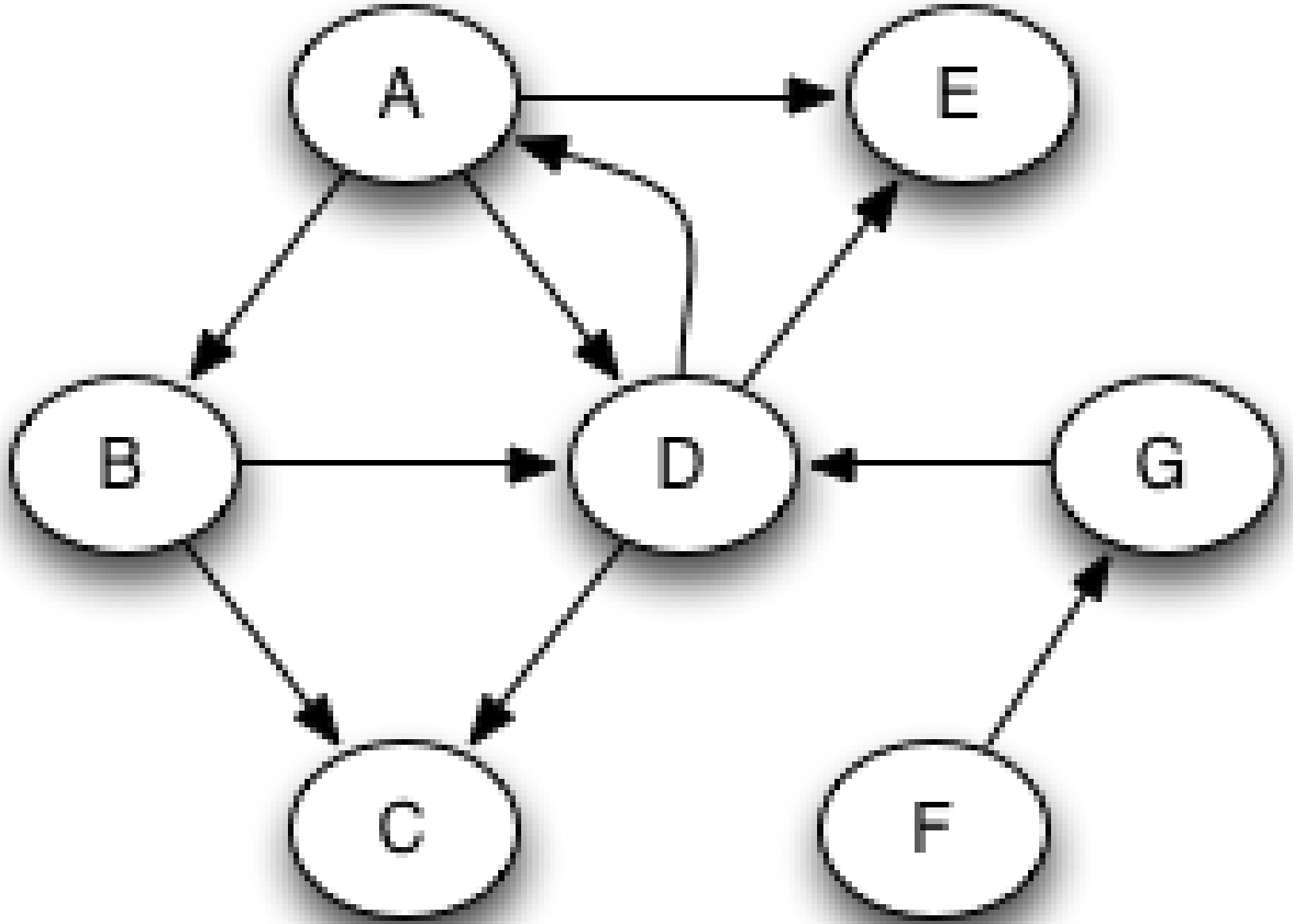
`wget` is a very commonly-used utility for mirroring an entire URL hierarchy. It has very many flags, but a common starting point is:

Common flags

```
$ wget -r -l inf -N -p [base-url]
```

As with any complex CLI tool, `man wget` is indispensable.

Breadth-First Traversal



Breadth-First Traversal

Algorithm

```
def bft(g, root):
  seen = set()
  q = [root]

  while q:
    n = q.pop(0)
    if n not in seen:
      visit(n)
      seen.add(n)
      q += g[n]
```