Announcements

- Homework 3 posted
- Read submission instructions
  - Penalties from now on if submitted incorrectly
- DO NOT leave agents on lab computers
  - Will be deleted at the end of office hours
- Export your agents to hand in
Advanced Agent Builder

- Troubleshooting connectors
- Training sample pages
- Extraction Rules
- Filtering
- URL deconstruction
- Miscellaneous
Advanced Agent Builder

- Troubleshooting connectors
- Training sample pages
- Extraction Rules
- Filtering
- URL deconstruction
- Miscellaneous
Troubleshooting connectors

- Look at header being submitted (in internet view)
  - Make sure your connector is sending the right values
    - Redirects
    - Sessions
- Makes URL deconstruction easier (covered later)
The data below contains the HTTP header information for the HTTP request just submitted.

To resume the regular navigation please click the image on the toolbar...

```
GET http://people.yahoo.com/pg/psPhoneSearch.py?src=b&s=1&FirstName=Michalowski&LastName=Michalowski&City=&State=CA&Phone=&Search=Phone+and+Address+Search HTTP/1.1
Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, application/vnd.ms-excel, application/vnd.ms-powerpoint, application/msword, application/x-shockwave-flash, */*
Referer: http://people.yahoo.com/)
Accept-Language: en-us
Proxy-Connection: Keep-Alive
User-Agent: Mozilla/4.0 (compatible, MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)
Host: phone.people.yahoo.com
Cookie: d=1tpm8080vzjkd&b=2; q=il=QFC5GFQAi32u3cA-oq=2-gKj; c=mg=1; T=a-3aa; C4vssx9NFW26T9gS0pHNY79bPoCSSMs9c7crafted:OZ3DQ93GlGfU8b-n;LYC-Lp=2
```
Advanced Agent Builder

- Troubleshooting connectors
- Training sample pages
- Extraction Rules
- Filtering
- URL deconstruction
- Miscellaneous
Training Sample Pages

- Adding additional pages because
  - You already fetched all pages from connectors
  - You need one more “representative” page

- Can be added by:
  - Local file
  - From an existing connector
Decide which Level of Pages to Fetch
This wrapper contains source pages from different levels.

Please choose the level of sample pages you want to fetch:

Pages level: 3

<< Back Next >> Cancel
Training Sample Pages

- Setting Validation
  - Checks that extracted value is correct
  - Main source of agent errors
Data Validation

Please choose from the following options to check the validity of the extracted data:

- Don't allow empty data
- Plain data (no HTML tags)
- Must be a number
- Minimum number of words: 0
- Maximum number of words: 0
- Must be a valid URL
- Must be able to be fetched
- Must match regular expression

Classification

Please check this box if the validation rules for this data item will be used during the manual classification process:

- Use these validation rules for manual classification
Training Sample Pages

- Post-processing
  - Done after extraction
  - Not part of validation
Check the post-processing options to apply at runtime:

- [ ] Remove HTML tags from extracted data
- [ ] Fully qualify the extracted URL
- [ ] Fetch binary data
- [ ] Decode URL characters (e.g., %20)
- [x] Decode encoded HTML entities (e.g., &amp;)
- [ ] Alter data using following regular expressions

**Pattern:**

**Substitution:**
Advanced Agent Builder

- Troubleshooting connectors
- Training sample pages
- Extraction Rules
- Filtering
- URL deconstruction
- Miscellaneous
Extraction Rules

- Manual classification possible
- Playing with rules
  - Rules are created by example
    - if you are using bad sample pages, then agent learns incorrect rules
  - Rule locking
    - Useful when adding items after learning rules
- Do it at your own risks
Default plan is not in sync with agent...
<table>
<thead>
<tr>
<th>Item Name</th>
<th>Rule Type</th>
<th>Rule Content</th>
</tr>
</thead>
</table>
| Person    | Extraction| `begin [
forward(a0a8b5)forward(\1)jump(-1)
] end [
backward(</nobr>)
]` |
| Person    | Iteration | `begin [
forward(<b>\1)jump(-1)
] end [
forward(</nobr>)jump(-1)
]` |
| Name      | Extraction| `begin [
jump(0)
] end [
backward(</a>&nbsp;)
]` |
| Address   | Extraction| `begin [
forward(top>)
] end [
backward(<br>\1)
]` |
| City      | Extraction| `begin [
backward(\w \c)jump(2)
] end [
forward(,)jump(-1)
]` |
Advanced Agent Builder

- Troubleshooting connectors
- Training sample pages
- Extraction Rules
- Filtering
- URL deconstruction
- Miscellaneous
Filtering

- Filter the value of an attribute in data schema view
  - Filter all data or a list (can’t do individual item)
- Why would you want to filter data?
  - Limit the number of items returned
- Example: Google results
  - Limit how many times you follow next links
Advanced Agent Builder

- Troubleshooting connectors
- Training sample pages
- Extraction Rules
- Filtering
- URL deconstruction
- Miscellaneous
URL deconstruction

- When submitting a form, sometimes a website redirects to a different URL
- Session id
- Forms within tables (homework #3??????)
- What can we do?
  - Make an output connector that points to the correct URL
- Where can we find the pieces?
  - On the page
  - Capturing the header
This form allows you to deconstruct the URL of the connector into its constituent parts according to a predefined set of separators. You can temporarily alter this set by choosing the "Advanced..." button.

URL elements:

<table>
<thead>
<tr>
<th>Value</th>
<th>Source Type</th>
<th>Source Name</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://phone.people.ya">http://phone.people.ya</a>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This form allows you to deconstruct the URL of the connector into its constituent parts according to a predefined set of separators. You can temporarily alter this set by choosing the "Advanced..." button.

**URL elements:**

<table>
<thead>
<tr>
<th>Value</th>
<th>Source Type</th>
<th>Source Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>http</td>
<td></td>
<td></td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>people</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yahoo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>com</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Advanced ...] [OK] [Cancel]
This form allows you to deconstruct the URL of the connector into its constituent parts according to a predefined set of separators. You can temporarily alter this set by choosing the "Advanced..." button.

**URL elements:**

<table>
<thead>
<tr>
<th>Value</th>
<th>Source Type</th>
<th>Source Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>http</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>/</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>/</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>phone</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>Constant</td>
</tr>
<tr>
<td>7</td>
<td>people</td>
<td>Agent Parameter</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>yahoo</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>com</td>
<td></td>
</tr>
</tbody>
</table>
This form allows you to deconstruct the URL of the connector into its constituent parts according to a predefined set of separators. You can temporarily alter this set by choosing the "Advanced..." button.

### URL elements:

<table>
<thead>
<tr>
<th>Value</th>
<th>Source Type</th>
<th>Source Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>phone</td>
<td>Agent Par...</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>people</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>yahoo</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>com</td>
<td></td>
</tr>
</tbody>
</table>
Advanced Agent Builder

- Troubleshooting connectors
- Training sample pages
- Extraction Rules
- Filtering
- URL deconstruction
- Miscellaneous
Cloning a wrapper
Parameter list
Number of rows in a list

Options

Feedback
Sound a tone when learning is complete.

- Beep when finished

Rows
Indicate below how many rows you want to be generated for a list item when adding a new sample to the mark-up process.

Maximum rows (including the last row): 3

Samples
Specify the default folder for downloading and importing sample pages.

Browse...
Questions?