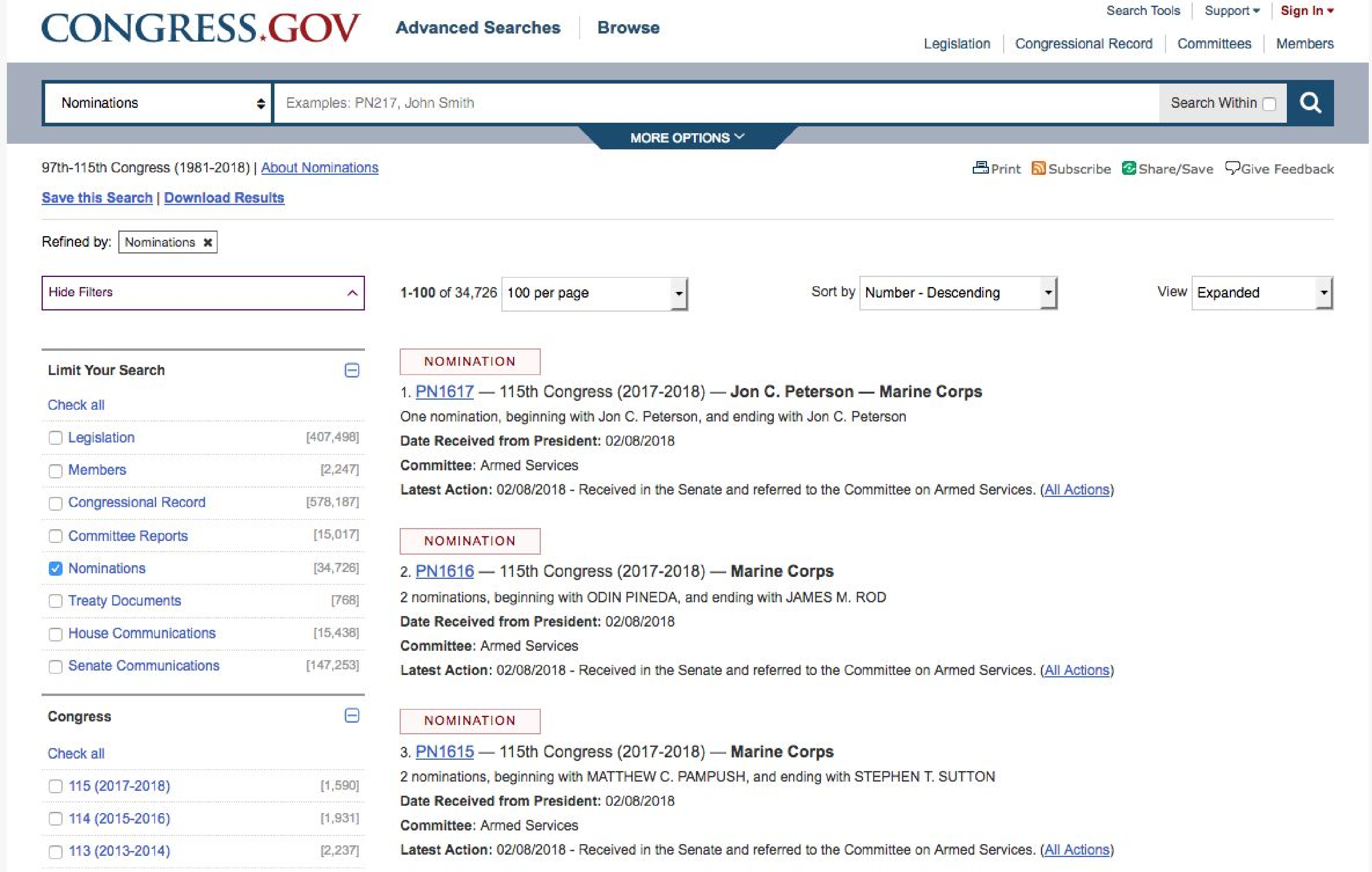
Parsing congress.gov for presidential nominations

Michael Lindsey - Law Library - 2018



Nominations | Examples: PN217, John Smith

97th-115th Congress (1981-2018) | About Nominations

Save this Search | Download Results

Refined by: Nominations ✖

1-100 of 34,726 | 100 per page | Sort by Number - Descending | View Expanded

Limit Your Search

Check all

☐ Legislation	[407,498]
☐ Members	[2,247]
☐ Congressional Record	[578,187]
☐ Committee Reports	[15,017]
☑ Nominations	[34,726]
☐ Treaty Documents	[768]
☐ House Communications	[15,438]
☐ Senate Communications	[147,253]

Congress

Check all

☐ 115 (2017-2018)	[1,590]
☐ 114 (2015-2016)	[1,931]
☐ 113 (2013-2014)	[2,237]

NOMINATION

1. PN1617 — 115th Congress (2017-2018) — **Jon C. Peterson — Marine Corps**

One nomination, beginning with Jon C. Peterson, and ending with Jon C. Peterson

Date Received from President: 02/08/2018

Committee: Armed Services

Latest Action: 02/08/2018 - Received in the Senate and referred to the Committee on Armed Services. (All Actions)

NOMINATION

2. PN1616 — 115th Congress (2017-2018) — **Marine Corps**

2 nominations, beginning with ODIN PINEDA, and ending with JAMES M. ROD

Date Received from President: 02/08/2018

Committee: Armed Services

Latest Action: 02/08/2018 - Received in the Senate and referred to the Committee on Armed Services. (All Actions)

NOMINATION

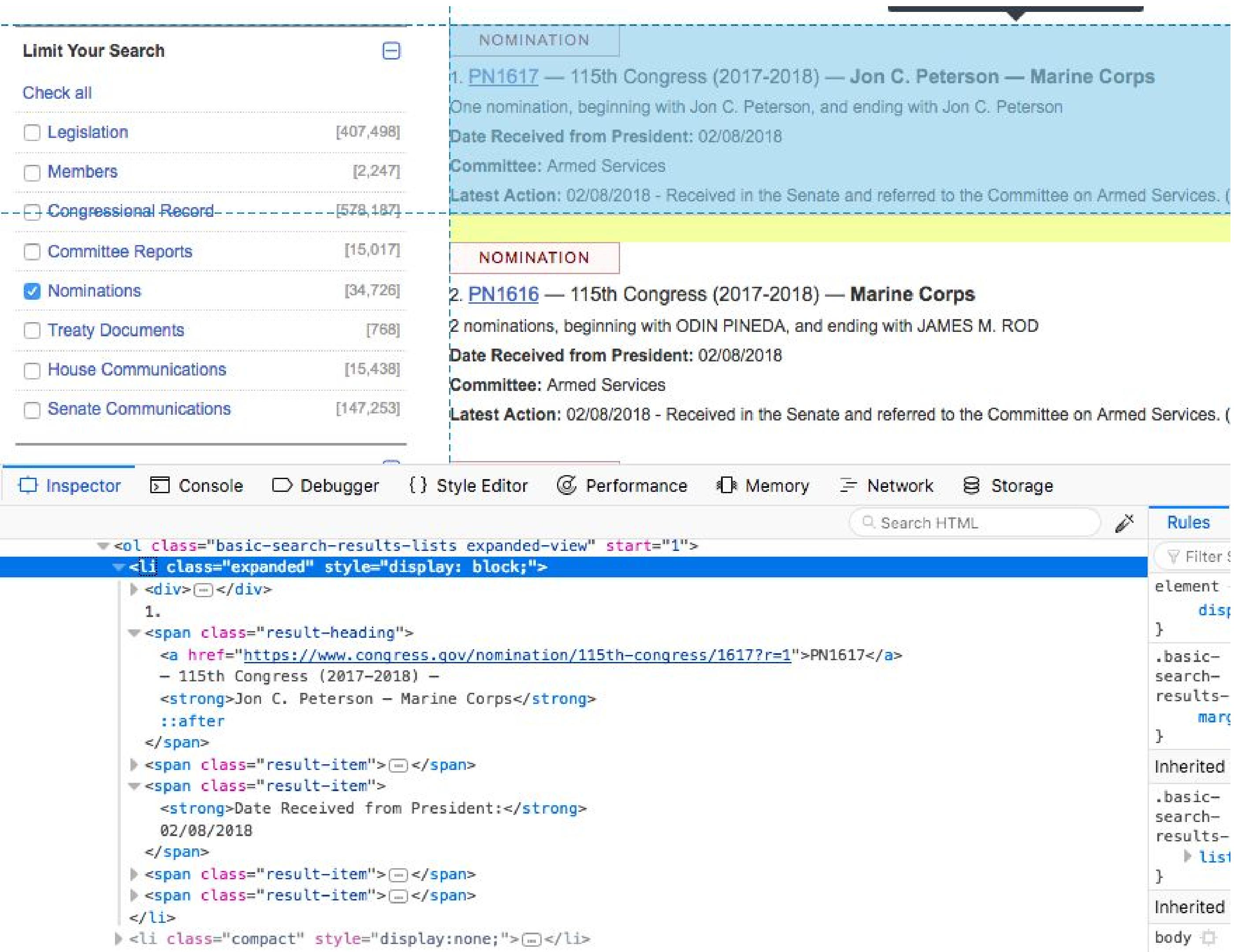
3. PN1615 — 115th Congress (2017-2018) — **Marine Corps**

2 nominations, beginning with MATTHEW C. PAMPUSH, and ending with STEPHEN T. SUTTON

Date Received from President: 02/08/2018

Committee: Armed Services

Latest Action: 02/08/2018 - Received in the Senate and referred to the Committee on Armed Services. (All Actions)



Limit Your Search

Check all

☐ Legislation	[407,498]
☐ Members	[2,247]
☐ Congressional Record	[578,187]
☐ Committee Reports	[15,017]
☑ Nominations	[34,726]
☐ Treaty Documents	[768]
☐ House Communications	[15,438]
☐ Senate Communications	[147,253]

NOMINATION

1. PN1617 — 115th Congress (2017-2018) — **Jon C. Peterson — Marine Corps**

One nomination, beginning with Jon C. Peterson, and ending with Jon C. Peterson

Date Received from President: 02/08/2018

Committee: Armed Services

Latest Action: 02/08/2018 - Received in the Senate and referred to the Committee on Armed Services. (

NOMINATION

2. PN1616 — 115th Congress (2017-2018) — **Marine Corps**

2 nominations, beginning with ODIN PINEDA, and ending with JAMES M. ROD

Date Received from President: 02/08/2018

Committee: Armed Services

Latest Action: 02/08/2018 - Received in the Senate and referred to the Committee on Armed Services. (

```
<ol class="basic-search-results-lists expanded-view" start="1">
  <li class="expanded" style="display: block;">
    <div>…</div>
    1.
    <span class="result-heading">
      <a href="https://www.congress.gov/nomination/115th-congress/1617?r=1">PN1617</a>
      — 115th Congress (2017-2018) —
      <strong>Jon C. Peterson — Marine Corps</strong>
      ::after
    </span>
    <span class="result-item">…</span>
    <span class="result-item">
      <strong>Date Received from President:</strong>
      02/08/2018
    </span>
    <span class="result-item">…</span>
    <span class="result-item">…</span>
  </li>
  <li class="compact" style="display:none;">…</li>
```

```
element
    disp
}
.basic-
search-
results-
    marg
}
Inherited
.basic-
search-
results-
    list
}
Inherited
body
```

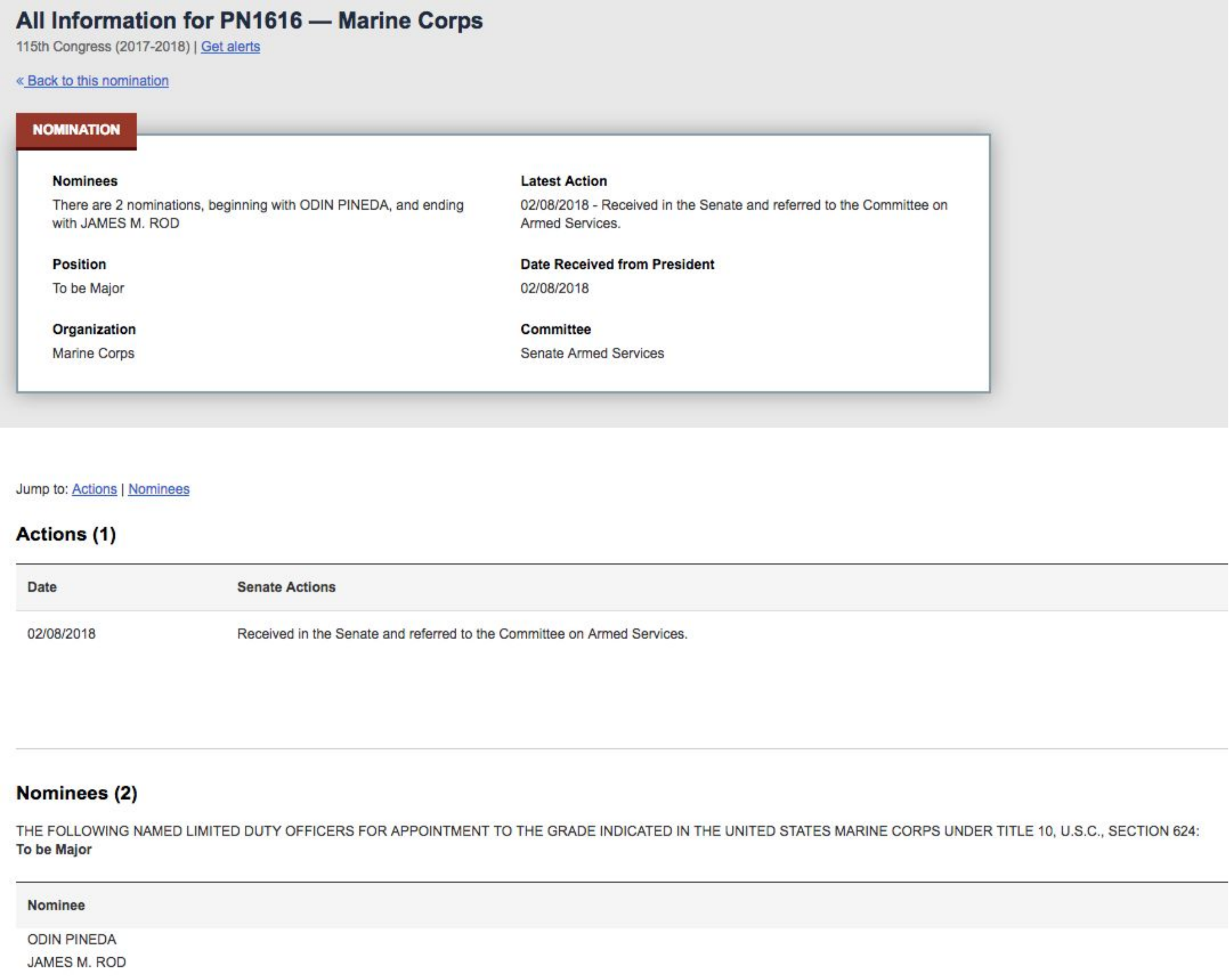
step one:

get links to all the nominations

- base url:
  https://www.congress.gov/search?q={%22source%22:%22nominations%22,%22nomination-type%22:%22Civilian%22}&pageSize=250
- set `$pageNumber = 1`
- tack `$pageNumber` onto the end of the base url like so: … `&page=$pageNumber` and grab it over HTTP.
- This returns a list of nomination in batches of 250
- Convert that HTML to XML and break the list up into individual nominations

  ```
  $list = $xpath->query("//ol[contains(@class,'basic-search-results-lists')]
  /li[contains(@class,'compact')]/span[contains(@class,'result-heading')]/a");
  ```
- for each item in that list, harvest the link (`href`)and add a new nomination to the database
- if there's a link to a next page of nominations, increment `$pageNumber` and do it all again

Now we have a database of 34K+ nominations,
but so far, just the links to their pages.



All Information for PN1616 — Marine Corps

115th Congress (2017-2018) | Get alerts

« Back to this nomination

NOMINATION

Nominees

There are 2 nominations, beginning with ODIN PINEDA, and ending with JAMES M. ROD

Latest Action

02/08/2018 - Received in the Senate and referred to the Committee on Armed Services.

Position

To be Major

Date Received from President

02/08/2018

Organization

Marine Corps

Committee

Senate Armed Services

Jump to: Actions | Nominees

Actions (1)

Date	Senate Actions
02/08/2018	Received in the Senate and referred to the Committee on Armed Services.

Nominees (2)

THE FOLLOWING NAMED LIMITED DUTY OFFICERS FOR APPOINTMENT TO THE GRADE INDICATED IN THE UNITED STATES MARINE CORPS UNDER TITLE 10, U.S.C., SECTION 624:
To be Major

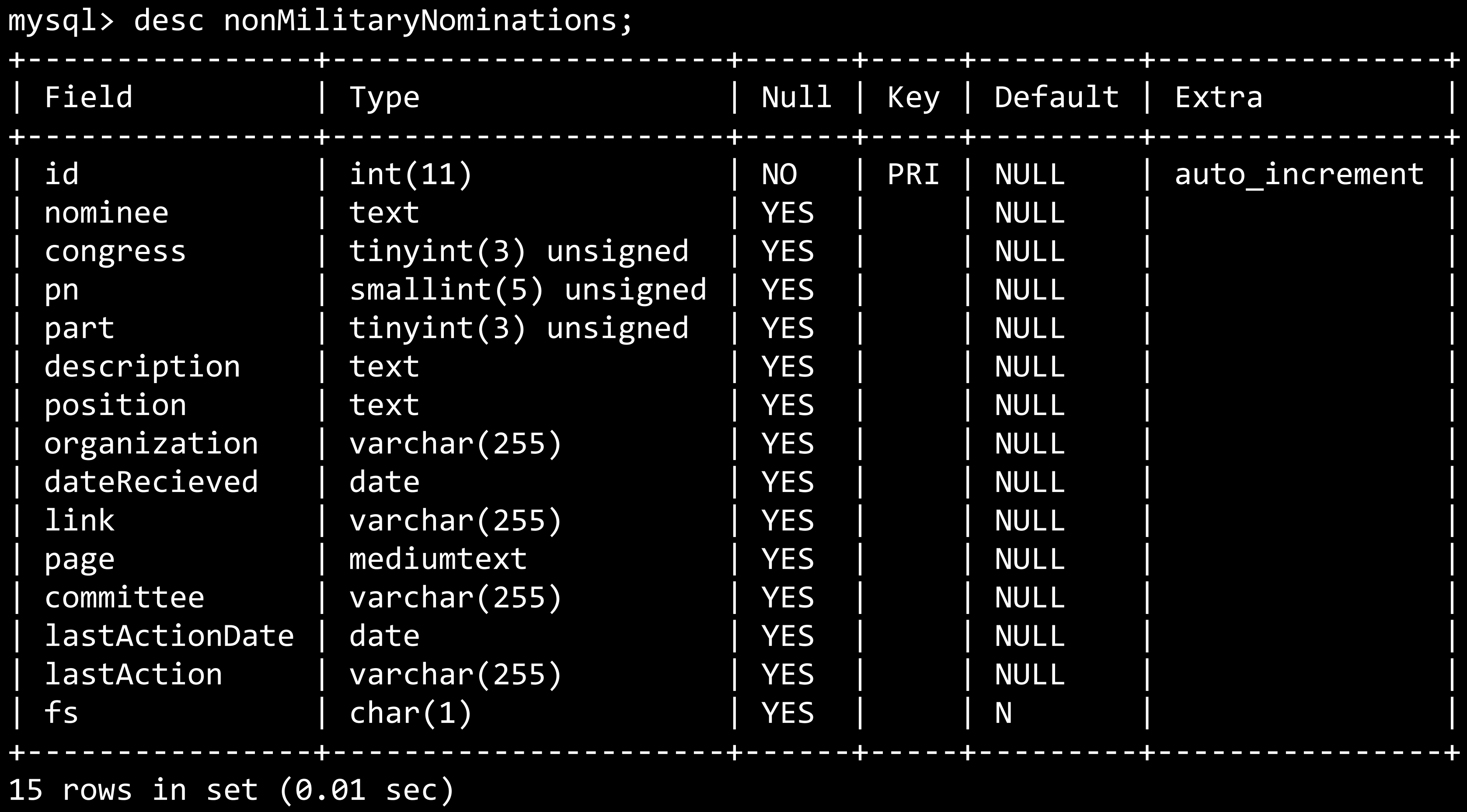
Nominee
ODIN PINEDA
JAMES M. ROD

step two:

get the raw HTML for each nomination and store it

- grab references to all nominations in the database where we don't have the raw HTML (that will be all of them at first)
- in a loop, follow the link and store the HTML there for that particular nomination
  Sometimes this fails for some network reason or other
- re-run this process a few times until there are no more to get

Now we have the raw HTML for every nomination.



```
mysql> desc nonMilitaryNominations;
+----------------+----------------------+------+-----+---------+----------------+
| Field          | Type                 | Null | Key | Default | Extra          |
+----------------+----------------------+------+-----+---------+----------------+
| id             | int(11)              | NO   | PRI | NULL    | auto_increment |
| nominee        | text                 | YES  |     | NULL    |                |
| congress       | tinyint(3) unsigned  | YES  |     | NULL    |                |
| pn             | smallint(5) unsigned | YES  |     | NULL    |                |
| part           | tinyint(3) unsigned  | YES  |     | NULL    |                |
| description    | text                 | YES  |     | NULL    |                |
| position       | text                 | YES  |     | NULL    |                |
| organization   | varchar(255)         | YES  |     | NULL    |                |
| dateRecieved   | date                 | YES  |     | NULL    |                |
| link           | varchar(255)         | YES  |     | NULL    |                |
| page           | mediumtext           | YES  |     | NULL    |                |
| committee      | varchar(255)         | YES  |     | NULL    |                |
| lastActionDate | date                 | YES  |     | NULL    |                |
| lastAction     | varchar(255)         | YES  |     | NULL    |                |
| fs             | char(1)              | YES  |     | N       |                |
+----------------+----------------------+------+-----+---------+----------------+
15 rows in set (0.01 sec)
```

step three:

parse each nomination and store the data we're interested in.

- fetch from the database a batch of nominations where we have the HTML, but haven't yet populated the metadata
- The HTML should either be for a single-nominee, or for multiple nominees under the same PN#. The HTML for both types conforms to a particular form, and we can use regular expressions to tell them apart.
- grab the nominee name(s) from the HTML as well as the other details and store them in the database

step four:

- iterate through every nomination in the database and output it in a form that Excel or Google Sheets can understand
- deliver to Anne

gotchas

of course there are

- the data format might change slightly at any given time. So build in integrity checks at every point and have the script konk out with a meaningful error message that you can follow up on
- Be prepared to re-write code that made more sense to you a long time ago than it does now
- data might be coded in ways you don't expect. Anne is interested in certain non-military nominations that happen to be included in that dataset.

last bits

- important to have a clear role in the partnership. Anne has the domain expertise. Concentrate on getting her the data and make as few assumptions as possible.
- ProPublica is also interested in this data and has implemented an API for it, but it currently can't handle nominations of multiple people or committees for those nominations in any useful way. It's marked as an enhancement for a future version of their API.
- JSON or some other structured data format is the natural way for machines to work with data provided by other machines. One day!