Welcome

Please help yourself to breakfast.
Agenda:

8:00-8:45am: Registration & Breakfast
8:45-8:55am: Welcome
8:55-9:25am: John Tindel, Data Scientist, GSA OGP
9:25-9:55am: Q&A and Discussion
9:55-10:00am: Closing
Finding the Needle in the Haystack
Using Data Science, NLP, and APIs to Streamline Compliance Checking

John Tindel
Data Scientist
GSA Office of Government-wide Policy
My Background

- BA Government-University of Texas at Austin
- BSc Economics, MSc Econometrics- Utrecht University (Netherlands)
- Data Analyst with OMB Office of E-Government and IT
- Data Scientist at GSA Office of Government-wide Policy
Summary

• Background: Section 508, OGP’s role, and Agency Programs
• Solicitation Review
• Old Process: Random Sampling, Hand-Evaluation, Recording
• Interim Solution: Attribute-Based Statistical Likelihood Model
• Advanced Solution: Implementing a Machine Learning Solution with Natural Language Processing (NLP)
• Challenges and Next Steps Forward
Background: Section 508

An estimated 20% of Americans are disabled in some way

The government mandates that all government technology be accessible to citizens and workers with disabilities

Not Accessible

- Flash-based navigation
- No alt text on website images
- Videos without captions or transcripts
- Text with insufficient contrast to the background
- Poor navigational structure
An agency representative or office charged with ensuring 508 compliance at an agency

Duties
- Product/Website Assessment
- Advocacy and Training
- Solicitation Review

Often understaffed and under-resourced
Background: Where GSA Comes In

- Office of Government-wide Policy (OGP) coordinates cross-governmental efforts to promote and strengthen 508 programs
- Training 508 coordinators and agency employees
- Sharing Best Practices
- Website Review
- Solicitation Review
Solicitation Review: The Problem

• Government contracts with private companies for services to a very significant degree
• These companies are only obligated to deliver the services explicitly described in the contract
• If those writing contract requirements fail to specify that deliverables must be 508-compliant, companies are not obligated to deliver 508-compliant products
• It is vital to ensure that solicitations for technology-related products and services contain the proper conditions requiring 508-compliance
Solicitation Review: Solicitations

• Solicitations are posted to many sites, but FedBizOpps is the largest out there.
• 500+ New solicitations per day across all agencies
• Everything from prison toilet paper to iPhone Apps (but Section 508 only applies to tech)
• A note on jargon
  – RFP vs. RFQ vs. RFI, etc.
  – SoW, Solicitations, etc.
• Variations in formats for documents
The Old Process

- Agencies responsible, OGP performs centralized role
- Daily email digest
- Random sampling
- Hand-evaluation by experts
- Letters sent to contracting officers
- Results recorded by hand
Opportunities

- Daily XML dumps from FedBizOpps
  ![FedBizOpps.gov]
- FBOpen
  ![fbopen]
- Historical Data
  - ~3800 Solicitations, graded by hand
  - Imperfect data, but XML/FBOpen allow for collection of historical data
The Dataset

- 3800 Graded Solicitations
- Attributes: Agency, Office, Contracting Officer, NAICS Code, Industry Code, Month of Posting
- 3 Grades:
  - Green: fully compliant
  - Yellow: partially compliant
  - Red: not compliant
- Patterns are apparent
By Office within an Agency
By Time of Year
## By NAICS Code

<table>
<thead>
<tr>
<th>NAICS Code</th>
<th>Industry Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>541519</td>
<td>Other computer related services</td>
</tr>
<tr>
<td>511210</td>
<td>Software publishers</td>
</tr>
<tr>
<td>541512</td>
<td>Computer systems design services</td>
</tr>
<tr>
<td>423430</td>
<td>Computer and computer peripheral equipment a..</td>
</tr>
<tr>
<td>518210</td>
<td>Data processing, hosting, and related services</td>
</tr>
<tr>
<td>517110</td>
<td>Wired telecommunications carriers</td>
</tr>
<tr>
<td>334111</td>
<td>Electronic computer manufacturing</td>
</tr>
<tr>
<td>334516</td>
<td>Analytical laboratory instrument manufacturing</td>
</tr>
<tr>
<td>334119</td>
<td>Other computer peripheral equipment manufact..</td>
</tr>
<tr>
<td>541990</td>
<td>All other professional, scientific, and technical s..</td>
</tr>
<tr>
<td>519130</td>
<td>All other information services</td>
</tr>
<tr>
<td>334310</td>
<td>Audio and video equipment manufacturing</td>
</tr>
<tr>
<td>541513</td>
<td>Computer facilities management services</td>
</tr>
<tr>
<td>334220</td>
<td>Radio and television broadcasting and wireless communications</td>
</tr>
<tr>
<td>541611</td>
<td>Administrative management and general manag..</td>
</tr>
<tr>
<td>443120</td>
<td>Computer and software stores</td>
</tr>
<tr>
<td>519110</td>
<td>Internet publishing and broadcasting and web s..</td>
</tr>
<tr>
<td>511120</td>
<td>Periodical publishers</td>
</tr>
<tr>
<td>611430</td>
<td>Professional and management development tral..</td>
</tr>
<tr>
<td>532420</td>
<td>Office machinery and equipment rental and leasing</td>
</tr>
<tr>
<td>334210</td>
<td>Telephone apparatus manufacturing</td>
</tr>
<tr>
<td>334515</td>
<td>Instrument manufacturing for measuring and te..</td>
</tr>
<tr>
<td>519199</td>
<td>All other telecommunications</td>
</tr>
<tr>
<td>611420</td>
<td>Computer training</td>
</tr>
<tr>
<td>333315</td>
<td>Photographic and photocopying equipment m..</td>
</tr>
<tr>
<td>541690</td>
<td>Other scientific and technical consulting services</td>
</tr>
<tr>
<td>541712</td>
<td>Research and development in the physical, engi..</td>
</tr>
<tr>
<td>334511</td>
<td>Search, detection, navigation, guidance, aerona..</td>
</tr>
<tr>
<td>334513</td>
<td>Instruments and related products manufactuin..</td>
</tr>
<tr>
<td>517911</td>
<td>Telecommunications resellers</td>
</tr>
</tbody>
</table>
Data Wrangling

- Ensuring compatibility with API data
- Historical data was hand-entered (messy)
- Need to reconstruct historical data from XML dumps
- Finally: clean dataset with 3800 graded solicitations, with info on agency, office, contracting officer, NAICS, Class Code, Month.
- Dataset spanned 2009-2015.
Identifying IT-Related Procurements

- Formerly done by a piece of proprietary software held by our contractor
- Dataset of 12,000 procurements classified as being IT or non-IT
- Clustering was used to determine title words common to IT procurements
- Each NAICS and industry code was given a frequency score, and the IT likelihood was derived as a function of the weight of the NAICS, industry code, and presence of catchwords
The First Model

- Train/Test of several models
- Tested logistic regression, GLS, OLS, Ridge Regression
- Most accurate model turned out to be OLS. Tests for heteroskedasticity were negative, but multicollinearity was present, as would be expected, among the office/agency/contracting officer variables
- The goal is a single prediction based on the group of variables, not as interested in the effects of a single variable, so multicollinearity is not a problem
Model Drawbacks

• This is a fixed-effects model with many variables. Fewer degrees of freedom = higher chance of overfitting

• Nonetheless, testing the model several times showed ~60-70% accuracy. This is better than a random guess, so it’s a step in the right direction
Next Step:  
Natural Language Processing (NLP)  

• Machine Learning  
  — Naïve Bayes  
  — Maximum Entropy  
  — Decision Tree/Random Forest  
• Classification  
• Training/Testing  
• Corpus  
  — Clean text  
  — Lemmatization/stemming
Example: Sentiment Analysis

Fantastic restaurant, delicious food!

I love this place!

Awful, never eat there. I got sick.

The bread they make is delicious.

Terrible, the staff were very rude.

This place sucks.

Wow, amazing food!

Ew, my food was awful.

{"food": 2, 'make': 1, 'bread': 1, 'amazing': 1, 'fantastic': 1, 'delicious': 2, 'wow': 1, 'love': 1, 'place': 1, 'restaurant': 1}

{"food": 1, 'staff': 1, 'never': 1, 'here': 1, 'ew': 1, 'were': 1, 'this': 1, 'got': 1, 'eat': 1, 'very': 1, 'sick': 1, 'rude': 1, 'awful': 2, 'terrible': 1, 'sucks': 1, 'place': 1}
Example: Sentiment Analysis (Continued)

{'food': 2, 'make': 1, 'bread': 1, 'amazing': 1, 'fantastic': 1, 'delicious': 2, 'wow': 1, 'love': 1, 'place': 1, 'restaurant': 1}

I love the bread here

The soup was tasty

The service was awful

Absolute garbage

{'food': 1, 'staff': 1, 'never': 1, 'here': 1, 'ew': 1, 'were': 1, 'this': 1, 'got': 1, 'eat': 1, 'very': 1, 'sick': 1, 'rude': 1, 'awful': 2, 'terrible': 1, 'sucks': 1, 'place': 1}
Gathering Data

- Need text of solicitation
- FBO Allows any input
- .doc, .docx, .pptx, .rtf, .txt, .pdf
- Everything needs to be in the same format
- Web Scraping
- Yield: 4500-ish documents
- Readability...
Public Service Announcement

Please!
(Especially Scanned)
Building a corpus

- Parse what I could (~1300/3800 procurements)
- Combine all text available for each procurement
- Big dataset: list of tuples (text, grade)
- Train/Test – 90/10
- Trying out some models
- Waiting
Results

• Naïve Bayes trained, working on Maximum Entropy classifier.
• Naïve Bayes ~85% accurate
• Informative features as expected (‘508’, ‘accessibility’, ‘disabilities’, etc.)
  
  contains(accessible) = True  
  green : red = 8.0 : 1.0

• Still, less than half of solicitations have readable text...
Informative Features

- contents(gpat) = True  green : red  =  61.1 : 1.0
- contents(prompting) = True  green : red  =  44.0 : 1.0
- contents(longdesc) = True  green : red  =  42.5 : 1.0
- contents(alerted) = True  green : red  =  41.8 : 1.0
- contents(flicker) = True  green : red  =  41.0 : 1.0
- contents(textually) = True  green : red  =  40.3 : 1.0
- contents(displayable) = True  green : red  =  38.9 : 1.0
- contents(enlarged) = True  green : yellow =  36.3 : 1.0
- contents(acuity) = True  green : yellow =  35.8 : 1.0
- contents(touchscreen) = True  green : red  =  34.5 : 1.0
- contents(decodes) = True  green : red  =  34.5 : 1.0
- contents(grasping) = True  green : red  =  34.5 : 1.0
- contents(dtv) = True  green : red  =  33.8 : 1.0
- contents(tactilely) = True  green : red  =  33.8 : 1.0
- contents(cochlear) = True  green : red  =  33.8 : 1.0
- contents(captioned) = True  green : yellow =  30.1 : 1.0
- contents(caption) = True  green : red  =  29.3 : 1.0
- contents(ttys) = True  green : yellow =  27.9 : 1.0
- contents(assistive) = True  green : red  =  26.7 : 1.0
Next steps

• Integrate NLP analysis with API to automatically check.
• If text available, classification based on NLP results.
• If no text available, classification based on older model.
• Work in progress
• Eventually integrate with FBO, perhaps automatic validation someday?
Thank you for your attention

QUESTIONS?