Social Science Text Analysis with Python (&..)

Wouter van Atteveldt

November 22, 2017
Who am I

- Wouter van Atteveldt
- Started: Basic on Acorn Electron
- Trained: UCU, U. Edinburgh (Informatics)
- PhD: "Semantic Network Analysis" (VU AI + FSW)
- Now: Political Communication, VU
Why text analysis?
Why Text Analysis?

(a) Israel subject, U.S. media

(b) Israel subject, Chinese media
Text Analysis & Digital Social Science

- Flood of digital information (online/archived)
- About (social) behaviour
- Much of it textual
- 'Measure' content of messages
- Explain/understand/predict social behaviour
Why Text Analysis?

Social Science Text Analysis with Python (…) Wouter van Atteveldt
Why build a (new) system?

- Don’t leave social science to FB/NSA
- Need for transparent and repeatable results
- Need for good tool in teaching
  - Non-technical students
Why Python

**Why python**
Why Python

Why python

- Why not?
Why python

• Why not?

• FOSS / platform independent

• Relatively easy to learn

• Community / extensions
  • Including django, REST, numpy, pandas, ...
  • Including data science
Why R?

- R: FOSS language aimed at statistics / data
  - Primitives for vectors/matrix (numpy built-in)
- R is much used in data science / social science
  - Lower threshold for social scientists
- "Go where the users are"
- Convergence via numpy, pandas, R packages
How do we use python

- AmCAT: text search front end & API
- Elasticsearch search backend
- NLPipe: NLP processing manager
- Lot’s of Python & R scripts (using API)
Social Science Text Analysis with Python (&..) Wouter van Atteveldt

Why?

How did we use python?

How?

What?
AmCAT front end

- Django
- API: REST framework
- Default ORM for 'corpus management'
- Elasticsearch for text
Elasticsearch text backend

- Texts uploaded to elastic on separate nodes
- Linked to database via 'article sets'
- Unique and immutable via hash
Elasticsearch text backend

- Texts uploaded to elastic on separate nodes
- Linked to database via 'article sets'
- Unique and immutable via hash

- (in reality, we also keep texts in ORM because we are chickens... should we?)
Query screen

- Top part (filters) contains default form
- ’Output options’ dynamically load scripts with extra form
- Input sent to script to get results
  - summary
  - aggregate
  - other (plugin) scripts, including R
- Script calls elasticsearch, ORM, or API
- Result rendered in JS for default types (table, graph, png, ..)
How did we use python?

"Script" interface

- Django form plus 'run' method
- Allows for CLI, HTTP, API front end
- Used for Query, uploader, action/management scripts
Async jobs: celery

- Jobs can take too long for website
  - Uploading texts
  - Complex queries
  - Exporting data
- Upload and query use celery queue
- Task is Script class name plus serialized form
REST API

- Hierarchical REST API for projects / set / article
- API also used internally:
  - All web site tables loaded dynamically from API
  - Daily scraping uploads articles via API
  - "Eat your own dog food"
NLP (pre-)processing

- Many good tools developed for text analysis
  - Stanford CoreNLP for English (+..)
  - RUG Alpino, UvT ILK Frog etc
  - ParZU for German
  - ...
- Difficult to use for social scientists
  - Not trivial to install and call
  - No standard input/output format
  - Often need ‘pipeline’ of multiple tools
- (don’t try to fix this :) )
NLPipe

- Simple NLP job manager
- Runs various NLP modules as 'black box' web service
  - Converts output to standardized csv/json on request
- Caches results based on ID-hash
- Simple API to queue/download
Scaling up

- Multiple elastic nodes
- Celery jobs can be distributed
  - But (currently) need db+elastic access
- NLPipe separated into server + workers
  - Workers can run on AWS (via docker), Surf HPC (singularity, in progress)
What did we find out?

Social Science Text Analysis with Python (&..) Wouter van Atteveldt
What did we find out?

Social Science Text Analysis with Python (&..) Wouter van Atteveldt
What did we find out?
Hospital officials in Gaza said that 390 people were killed by Israeli fighter planes.
**Table 3:** Usage of Hamas and Israel as source and over-representation of Israel in all clauses and in clauses containing aggression; in U.S. and Chinese media coverage of the 2008–09 Gaza War

<table>
<thead>
<tr>
<th>Source</th>
<th>U.S. Media</th>
<th></th>
<th>Chinese Media</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Aggr.</td>
<td>All</td>
<td>Aggr.</td>
</tr>
<tr>
<td>Israel</td>
<td>8,224</td>
<td>4,126</td>
<td>1,225</td>
<td>677</td>
</tr>
<tr>
<td>Hamas</td>
<td>1,824</td>
<td>704</td>
<td>439</td>
<td>193</td>
</tr>
<tr>
<td>Overrepresentation</td>
<td>4.5</td>
<td>5.9</td>
<td>2.8</td>
<td>3.5</td>
</tr>
<tr>
<td>Israel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 3
Occurrence of topics that have a strong temporal component
FIGURE 4
Occurrence over time of detailed ($K = 25$) topics that constitute the *Accidents/ Danger* topic from the $K = 10$ model
Fig. 1. Scaled Positions of Party Manifestoes.
Why?

How?

What?

What did we find out?

Social Science Text Analysis with Python (&..) Wouter van Atteveldt
Fig. 3. Correlation between Framing of Immigration in each Party’s Press Releases with the Manifesto of All Parties. Only Significant Correlations Are Shown.
Thanks!

- http://vanatteveldt.com
- http://github.com/vanatteveldt
- http://github.com/amcat
- http://amcat.nl