Text-Scaling Hackathon

Federico Nanni
(and many others)
My Research

Post-Doc in **computational social science** at the University of Mannheim, working on natural language processing applications in political science.

Focus:

1. Topic detection
2. Collection building
3. Text scaling
My Research

Post-Doc in **computational social science** at the University of Mannheim, working on natural language processing applications in political science.

Focus:

1. Topic detection
2. Collection building
3. Text scaling
Text Scaling

The goal: to infer policy positions from text.
Text Scaling

The goal: to infer policy positions from text.
Text Scaling

The goal: to infer policy positions from text.

A highly interdisciplinary research area, comprising:

- Political science
- Computational linguistics
- Machine learning
Text Scaling

Research: conducted mostly by the political science community.
Text Scaling

Research: conducted mostly by the political science community.

Some of the most relevant papers:


Text Scaling

Research: conducted mostly by the political science community.

Some of the most relevant papers:


Until 2017, no mention of text scaling approaches in the ACL anthology.
Text Scaling

Current issues:

1. Topic-based analyses (e.g., extract ideological statements)
2. Cross-lingual analyses
3. Evaluation of unsupervised scaling methods
Text Scaling

Current issues:

1. Topic-based analyses (e.g., extract ideological statements)
2. Cross-lingual analyses
3. Evaluation of unsupervised scaling methods

These could be addressed by collaborating with the NLP community.
3 Days Text Scaling Hackathon (December 2017)
3 Days Text Scaling Hackathon (December 2017)

Supported by Villa Vigoni, the German-Italian Centre for European Excellence and by DFG.

23 young researchers from political science, computational social science and NLP.
3 Days Text Scaling Hackathon (December 2017)

Participants from:
Mannheim, Bruno Kessler Found., Unitelma, Sheffield, Duisburg-Essen, GESIS, Scuola Normale Superiore, EUI, Scuola Superiore Sant’Anna, Bocconi, Liepzig, Zagreb, LSE, Alan Turing Institute, Edinburgh, CEU, Toronto.
What is a Hackathon?
A coding-intensive collaborative workshop.
What is a Hackathon?
A coding-intensive collaborative workshop.

Why?
To boosts collaboration across disciplines.
What is a Hackathon?
A coding-intensive collaborative workshop.

Why?
To boosts collaboration across disciplines.

Shared-Task?
Participants were divided in 5 groups and had to work together towards a specific goal.
European-Integration Scaling

The task: develop a method for scaling text on the EU integration dimension. We provide participants with:
European-Integration Scaling

The task: develop a method for scaling text on the EU integration dimension. We provide participants with:

European-Integration Scaling

The task: develop a method for scaling text on the EU integration dimension. We provide participants with:


2) A gold standard (Chapel Hill) of EU integration party-positions (leg: 5, 7, 8)
European-Integration Scaling

The task: develop a method for scaling text on the EU integration dimension. We provide participants with:


2) A gold standard (Chapel Hill) of EU integration party-positions (leg: 5, 7, 8)

3) On the last evening, a test set (leg: 6)
European-Integration Scaling

The output: party positions for 6th legislation regarding European integration (between 0: strongly against and 1: strongly in favour).
European-Integration Scaling

The output: party positions for 6th legislation regarding European integration (between 0: strongly against and 1: strongly in favour).

All data available at: https://federiconanni.com/hack-vigoni/
What Participants Could Not Do

1) Find online the gold standard and predict based on that
2) Use external knowledge on the party to scale it (the task is text scaling)
How Did It Go?
How Did It Go?
How Did It Go?
How Did It Go?
Core Components of All Approaches

1. An initial filtering strategy (using a dictionary or a manually created list)
2. A text similarity approach, based on TF-IDF or word embeddings
3. A supervised scaling function (SVM regression model, canonical correlation analysis, etc.)
## Results

<table>
<thead>
<tr>
<th>Team/Model</th>
<th>PA (%)</th>
<th>$r_S$ (%)</th>
<th>$r_P$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WordFish (baseline)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team Ireland</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team Wales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team Netherlands</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team Greece</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team Italy</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Team/Model</th>
<th>PA (%)</th>
<th>$r_S$ (%)</th>
<th>$r_P$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>51.3</td>
<td>2.6</td>
<td>6.2</td>
</tr>
<tr>
<td>WordFish (baseline)</td>
<td>61.8</td>
<td>34.5</td>
<td>29.5</td>
</tr>
</tbody>
</table>

Team Ireland  
Team Wales  
Team Netherlands  
Team Greece  
Team Italy
## Results

<table>
<thead>
<tr>
<th>Team/Model</th>
<th>PA (%)</th>
<th>$r_S$ (%)</th>
<th>$r_P$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>51.3</td>
<td>2.6</td>
<td>6.2</td>
</tr>
<tr>
<td>WordFish (baseline)</td>
<td>61.8</td>
<td>34.5</td>
<td>29.5</td>
</tr>
<tr>
<td>Team Ireland</td>
<td>57.6</td>
<td>17.5</td>
<td>29.4</td>
</tr>
<tr>
<td>Team Wales</td>
<td>60.4</td>
<td>28.9</td>
<td>28.2</td>
</tr>
<tr>
<td>Team Netherlands</td>
<td>66.8</td>
<td>46.6</td>
<td>59.3</td>
</tr>
<tr>
<td>Team Greece</td>
<td>68.5</td>
<td>54.2</td>
<td>64.5</td>
</tr>
<tr>
<td>Team Italy</td>
<td><strong>70.3</strong></td>
<td><strong>54.3</strong></td>
<td><strong>72.8</strong></td>
</tr>
</tbody>
</table>
Things to Improve

Offering more theoretical background, both in political science and in NLP.

Having the theoretical background accessible to both audiences.

Having more brainstorming sessions, without computers.
Next Steps

I’m in the process of organizing:

1. Two new hackathons (December 2019, 2020) in collaboration with the Manifesto project.

2. A NLP shared-task on cross-lingual scaling of parliamentary speeches.
Joint Work With:

Goran Glavaš
Simone Paolo Ponzetto
Sara Tonelli
Nicolò Conti
Questions?

Federico Nanni
Data and Web Science Group
University of Mannheim
federico@informatik.uni-mannheim.de