Az ELKHC Cloud előnyei a POLTEXT projekt példáján

Sebők Miklós (TK PTI), Kacsuk Zoltán (HdM IAAI)
Research context

Intersection of two research projects at the Centre for Social Sciences, Institute for Political Science:

• Text Mining of Political and Legal Texts (POLTEXT) Incubator Project (Principal Investigator: Miklós Sebők)

• Hungarian Comparative Agendas Project (CAP) project (Project leaders: Zsolt Boda & Miklós Sebők)
Research problem

• **Quantitative analysis** of qualitative data
  – Classifying articles according to policy topics
  – Topics: education to defense, Comparative Agendas Project

• **Gold standard**: double-blind human coding by well-trained researchers

• What if this is **unfeasible**?
  – Article counts of over 100.000
  – **Cost and training** of human coders for this scale
The government passed on the option – There will be no state-funded mobile carrier? Our business sources indicate that the three state-owned corporations with a stake in the firm would not want to invest multiple hundred billion HUFs in the fourth carrier.”

Policy code (major topic): 17, which stands for “Space, science, technology, and telecommunications”
The project

- **Hungarian country project** of the Comparative Agendas Project (cap.tk.mta.hu)
  - Media module – 3 daily newspapers
- For this pilot project:
  - Left-liberal **Népszabadság (NS)**: Over 50 000 front-page articles (1990-2014)
  - Centre-Right **Magyar Nemzet (MN)**: 35 021 articles (2002-2014)
- Hand-coding was unfeasible for our purposes
- Solution: **text mining + machine learning**
A machine learning solution

• **Text as Data** – qualitative data is converted to quantitative (matrices)

• How to categorize articles into pre-defined classes: **Dictionary-based or supervised learning**

• For the latter a sufficiently large human-coded **training/test set** is needed
Part 1
CREATING A MACHINE CODED TRAINING SET FOR THE LEFT-WING DAILY NÉPSZABADSÁG (NS)
The Hybrid Binary Snowball (HBS) process

- We need to keep human **coding costs as low as possible**, while extracting the largest possible gain per invested human coding hour.
- We simplify multi-class classification by rephrasing it as a **series of pairwise comparisons**.
- We apply a snowball method to **augment the training set with machine-classified observations**.
Coding NS articles

11 342 articles

Hand coded articles
Uncoded articles

Features: stopwords removed, NOT stemmed, TF

41 974 articles

Loop for rounds of coding

Training set
Virgin test set

Classification
SVM
• training one vs all
• sequentially according to predefined code list

Human validation of samples of newly classified articles & decision on newly accepted coded articles

Rounds 1-2: accept only articles from samples found to be correctly classified

Rounds 3-4: accept all articles for codes where sample precision at least 75%

Rounds 5-6: sampling and validation for all codes together, accept all if sample precision at least 75%

37 143 articles

Hand coded and accepted machine coded articles
Uncoded articles

16 173 articles
Infrastructural bottlenecks 1: Memory

• Our desktop workstation had **only 32 GB RAM**

• Encountered **problems**:
  – Could not work on the **whole virgin data set**
  – Could not run **certain configurations**, for example: Term frequency - Inverse document frequency (Tf-Idf) weighting

• Even the **solutions were problems**:
  – Virgin data was partitioned up for processing
  – This would impact Tf-Idf weighting significantly

• **Real solution** going forward:
  – Using larger capacity single virtual instances or a cluster in the cloud
Infrastructural bottlenecks 2: 

Time

• Huge numbers of small operations add up quickly
• If process runtimes become too long, project execution becomes unfeasible
• Solution: **parallelizing** the execution of operations
Part 2

USING THE CODED LEFT-WING DAILY ARTICLES TO TEACH THE ALGORITHM HOW TO CODE THE CONSERVATIVE DAILY MAGYAR NEMZET (MN)
Apache Spark cluster

• With the help of the Laboratory of Parallel and Distributed Systems at the Institute for Computer Science and Control (SZTAKI LPDS)

• Apache Spark cluster running on five virtual instances in the SZTAKI ELKH Cloud

• All five virtual instances had 8 virtual processors and 32 GBs of RAM each, and were running Ubuntu 16.04.

• Four instances acted as worker nodes and one as the master node of the Spark cluster. Each Spark session was running with 32 VCPUs (but default parallelism set to 24) and 96 GBs of RAM total on the four worker nodes combined.
Manifold increase in speed

- **Old desktop setup**: roughly *3 days* for a full round of coding (33 code categories)
- **Spark cluster**: ca. *30 minutes* for a full round of coding

- This increase in speed enabled:
  - 1) Rapid prototyping
  - 2) Complex classification workflow
Coding MN articles

- 34,650 articles
- 34,670 articles

**Features:** stopwords removed, stemmed, TF-IDF

Loop for rounds of coding

- Training set
- Virgin test set

**Classification**

**SVM**

- training one vs all
- for all codes in training set
- proportional training sets
- 7 samples * 7 SVM models

Sample 1 SVM 1
SVM 2
SVM 7

Sample 2 SVM 1
SVM 2
SVM 7

Sample 7 SVM 1
SVM 2
SVM 7

Move newly accepted coded articles from test set to training set

- If all SVM results for all samples concur we have a verdict for that code for given article
- If there is only one verdict for an article, it is classified accordingly
- Human supervision of results (not validation)

Human validation of coded MN articles

Uncoded MN articles
Coding MN articles

Coded NS articles  Uncoded MN articles

**Features:** stopwords removed, stemmed, TF-IDF

Loop for rounds of coding

Training set  Virgin test set

**Classification**

**SVM**

- training one vs all
- for all codes in training set
- proportional training sets
- 7 samples * 7 SVM models

Sample 1  Sample 2  Sample 7

<table>
<thead>
<tr>
<th>SVM 1</th>
<th>SVM 2</th>
<th>SVM 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Move newly accepted coded articles from test set to training set

- If all SVM results for all samples concur we have a verdict for that code for given article
- If there is only one verdict for an article, it is classified accordingly
- **Human supervision** of results (not validation)

Human validation of coded MN articles  Uncoded MN articles
Coded NS articles → Uncoded MN articles

**Features:** stopwords removed, stemmed, TF-IDF

Loop for rounds of coding

Training set → Virgin test set

**Classification**

**SVM**

- training one vs all
- for all codes in training set
- proportional training sets
- 7 samples * 7 SVM models

Move newly accepted coded articles from test set to training set

- If all SVM results for all samples concur we have a verdict for that code for given article
- If there is only one verdict for an article, it is classified accordingly
- **Human supervision** of results (not validation)

Human validation of coded MN articles → Uncoded MN articles
Coded NS articles

Uncoded MN articles

**Features:** stopwords removed, stemmed, TF-IDF

Loop for rounds of coding

Training set

Virgin test set

**Classification**

**SVM**

- training one vs all
- for all codes in training set
- proportional training sets
- 7 samples * 7 SVM models

Sample 1

SVM 1

SVM 2

SVM 7

Sample 2

SVM 1

SVM 2

SVM 7

Sample 7

SVM 1

SVM 2

SVM 7

If all SVM results for all samples concur we have a verdict for that code for given article

If there is only one verdict for an article, it is classified accordingly

Human supervision of results (not validation)

Human validation of coded MN articles

Uncoded MN articles

Move newly accepted coded articles from test set to training set
Coding MN articles

Coded NS articles → Uncoded MN articles

Features: stopwords removed, stemmed, TF-IDF

Loop for rounds of coding

Training set → Virgin test set

Classification

SVM

• training one vs all
• for all codes in training set
• proportional training sets
• 7 samples * 7 SVM models

Sample 1 SVM 1 SVM 2 SVM 7
Sample 2 SVM 1 SVM 2 SVM 7
Sample 7 SVM 1 SVM 2 SVM 7

Move newly accepted coded articles from test set to training set

• If all SVM results for all samples concur we have a verdict for that code for given article
• If there is only one verdict for an article, it is classified accordingly
• Human supervision of results (not validation)

Human validation of coded MN articles → Uncoded MN articles
<table>
<thead>
<tr>
<th>Major Topic</th>
<th>Coded Articles</th>
<th>Sample Size</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Macroeconomics</td>
<td>3833</td>
<td>350</td>
<td>0.64</td>
</tr>
<tr>
<td>2. Civil Rights</td>
<td>207</td>
<td>135</td>
<td>0.87</td>
</tr>
<tr>
<td>3. Health</td>
<td>700</td>
<td>249</td>
<td>0.88</td>
</tr>
<tr>
<td>4. Agriculture</td>
<td>408</td>
<td>199</td>
<td>0.79</td>
</tr>
<tr>
<td>5. Labor</td>
<td>127</td>
<td>100</td>
<td>0.96</td>
</tr>
<tr>
<td>6. Education</td>
<td>436</td>
<td>205</td>
<td>0.81</td>
</tr>
<tr>
<td>7. Environment</td>
<td>71</td>
<td>71</td>
<td>0.42</td>
</tr>
<tr>
<td>8. Energy</td>
<td>542</td>
<td>226</td>
<td>0.96</td>
</tr>
<tr>
<td>9. Immigration</td>
<td>8</td>
<td>8</td>
<td>0.88</td>
</tr>
<tr>
<td>10. Transportation</td>
<td>459</td>
<td>210</td>
<td>0.78</td>
</tr>
<tr>
<td>12. Law and Crime</td>
<td>570</td>
<td>230</td>
<td>0.93</td>
</tr>
<tr>
<td>13. Social Welfare</td>
<td>72</td>
<td>72</td>
<td>0.74</td>
</tr>
<tr>
<td>14. Housing</td>
<td>168</td>
<td>118</td>
<td>0.85</td>
</tr>
<tr>
<td>15. Domestic Commerce</td>
<td>93</td>
<td>93</td>
<td>0.90</td>
</tr>
<tr>
<td>16. Defense</td>
<td>342</td>
<td>182</td>
<td>0.77</td>
</tr>
<tr>
<td>17. Technology</td>
<td>196</td>
<td>131</td>
<td>0.82</td>
</tr>
<tr>
<td>18. Foreign Trade</td>
<td>27</td>
<td>27</td>
<td>0.11</td>
</tr>
<tr>
<td>19. International Affairs</td>
<td>7617</td>
<td>366</td>
<td>0.89</td>
</tr>
<tr>
<td>20. Government Operations</td>
<td>1247</td>
<td>294</td>
<td>0.67</td>
</tr>
<tr>
<td>21. Public Lands</td>
<td>10</td>
<td>10</td>
<td>0.70</td>
</tr>
<tr>
<td>23. Culture</td>
<td>124</td>
<td>100</td>
<td>0.99</td>
</tr>
<tr>
<td>26. Weather and Natural Disasters</td>
<td>201</td>
<td>133</td>
<td>1.00</td>
</tr>
<tr>
<td>27. Accidents, Fire Incidents, Disasters</td>
<td>50</td>
<td>50</td>
<td>0.92</td>
</tr>
<tr>
<td>28. Arts, Cult., Hist., Sci. and Entertainment</td>
<td>677</td>
<td>246</td>
<td>0.90</td>
</tr>
<tr>
<td>29. Sports News and Recreation</td>
<td>385</td>
<td>193</td>
<td>0.90</td>
</tr>
<tr>
<td>30. Deaths, Fatalities, Obituaries</td>
<td>41</td>
<td>41</td>
<td>0.85</td>
</tr>
<tr>
<td>31. Churches and Religion</td>
<td>194</td>
<td>130</td>
<td>0.98</td>
</tr>
<tr>
<td>33. Political Parties</td>
<td>661</td>
<td>244</td>
<td>0.99</td>
</tr>
<tr>
<td>34. Tabloid News, Celebrities</td>
<td>28</td>
<td>28</td>
<td>0.54</td>
</tr>
<tr>
<td>35. Crime</td>
<td>339</td>
<td>181</td>
<td>0.91</td>
</tr>
<tr>
<td>36. Political Corruption</td>
<td>8</td>
<td>8</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Total Result</strong></td>
<td><strong>19841</strong></td>
<td><strong>4630</strong></td>
<td><strong>0.84</strong></td>
</tr>
</tbody>
</table>
Precision of MN corpus coding by CAP code category

26. Weather and Natural Disasters
23. Culture
31. Churches and Religion
 8. Energy
 5. Labor
12. Law and Crime
27. Accidents, Fire Incidents, Disasters
35. Crime
15. Domestic Commerce
29. Sports News and Recreation
28. Arts, Cult., Hist., Sci. and Entertainment
19. International Affairs
 3. Health
 9. Immigration
 2. Civil Rights
30. Deaths, Fatalities, Obituaries
14. Housing

Total Result
17. Technology
 6. Education
 4. Agriculture
10. Transportation
16. Defense
13. Social Welfare
21. Public Lands
20. Government Operations
 1. Macroeconomics
36. Political Corruption
34. Tabloid News, Celebrities
 7. Environment
18. Foreign Trade

Precision with confidence interval
Precision and total number of coded articles of MN corpus by CAP code category
<table>
<thead>
<tr>
<th>Major Topic</th>
<th>Coded Articles</th>
<th>Sample Size</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Macroeconomics</td>
<td>3833</td>
<td>350</td>
<td>0.64</td>
</tr>
<tr>
<td>2. Civil Rights</td>
<td>207</td>
<td>135</td>
<td>0.87</td>
</tr>
<tr>
<td>3. Health</td>
<td>700</td>
<td>249</td>
<td>0.88</td>
</tr>
<tr>
<td>4. Agriculture</td>
<td>408</td>
<td>199</td>
<td>0.79</td>
</tr>
<tr>
<td>5. Labor</td>
<td>127</td>
<td>100</td>
<td>0.96</td>
</tr>
<tr>
<td>6. Education</td>
<td>436</td>
<td>205</td>
<td>0.84</td>
</tr>
<tr>
<td>7. Environment</td>
<td>71</td>
<td>71</td>
<td>0.42</td>
</tr>
<tr>
<td>8. Energy</td>
<td>542</td>
<td>226</td>
<td>0.96</td>
</tr>
<tr>
<td>9. Immigration</td>
<td>8</td>
<td>8</td>
<td>0.88</td>
</tr>
<tr>
<td>10. Transportation</td>
<td>459</td>
<td>210</td>
<td>0.78</td>
</tr>
<tr>
<td>12. Law and Crime</td>
<td>570</td>
<td>230</td>
<td>0.93</td>
</tr>
<tr>
<td>13. Social Welfare</td>
<td>72</td>
<td>72</td>
<td>0.74</td>
</tr>
<tr>
<td>14. Housing</td>
<td>168</td>
<td>118</td>
<td>0.85</td>
</tr>
<tr>
<td>15. Domestic Commerce</td>
<td>93</td>
<td>93</td>
<td>0.90</td>
</tr>
<tr>
<td>16. Defense</td>
<td>342</td>
<td>182</td>
<td>0.77</td>
</tr>
<tr>
<td>17. Technology</td>
<td>196</td>
<td>131</td>
<td>0.82</td>
</tr>
<tr>
<td>18. Foreign Trade</td>
<td>27</td>
<td>27</td>
<td>0.11</td>
</tr>
<tr>
<td>19. International Affairs</td>
<td>7617</td>
<td>366</td>
<td>0.89</td>
</tr>
<tr>
<td>20. Government Operations</td>
<td>1247</td>
<td>294</td>
<td>0.67</td>
</tr>
<tr>
<td>21. Public Lands</td>
<td>10</td>
<td>10</td>
<td>0.70</td>
</tr>
<tr>
<td>23. Culture</td>
<td>124</td>
<td>100</td>
<td>0.99</td>
</tr>
<tr>
<td>26. Weather and Natural Disasters</td>
<td>201</td>
<td>133</td>
<td>1.00</td>
</tr>
<tr>
<td>27. Accidents, Fire Incidents, Disasters</td>
<td>50</td>
<td>50</td>
<td>0.92</td>
</tr>
<tr>
<td>28. Arts, Cult., Hist., Sci. and Entertainment</td>
<td>677</td>
<td>246</td>
<td>0.90</td>
</tr>
<tr>
<td>29. Sports News and Recreation</td>
<td>385</td>
<td>193</td>
<td>0.90</td>
</tr>
<tr>
<td>30. Deaths, Fatalities, Obituaries</td>
<td>41</td>
<td>41</td>
<td>0.85</td>
</tr>
<tr>
<td>31. Churches and Religion</td>
<td>194</td>
<td>130</td>
<td>0.98</td>
</tr>
<tr>
<td>33. Political Parties</td>
<td>661</td>
<td>244</td>
<td>0.99</td>
</tr>
<tr>
<td>34. Tabloid News, Celebrities</td>
<td>28</td>
<td>28</td>
<td>0.54</td>
</tr>
<tr>
<td>35. Crime</td>
<td>339</td>
<td>181</td>
<td>0.91</td>
</tr>
<tr>
<td>36. Political Corruption</td>
<td>8</td>
<td>8</td>
<td>0.63</td>
</tr>
<tr>
<td>Total Result</td>
<td>19841</td>
<td>4630</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Main contributions of HBS and the present study

- Enhance ML precision and recall by both human input (validation) and workflow design (one-vs-all classification, ensemble voting)
- Start working from a limited training set
- Able to maximize ROI on human coding
- Move between (intra-domain) corpora
- Take advantage of cloud infrastructure and parallel processing with Apache Spark
Further work

• Implement a finishing step using regular expressions to correct systematic errors
  - “design” in *Environment*
  - “icerink” in *Public Lands*

• Testing the HBS approach on **further languages**

• Generalizing the method to **other domains** beyond media
Thank you for your attention!

sebok.miklos@tk.mta.hu
kacsuk.zoltan@tk.mta.hu
poltext.tk.mta.hu
cap.tk.mta.hu