Background

- Research agenda – computational methods to access personal information
- Previous experimentation with topic modelling, named entity extraction
Supervised machine learning

- Training sets – HR-related documents
  a) HR-personal: human resources records relating to one or more identified individuals
  b) HR-general: records relating to human resources but not identifying individuals, such as position postings, policy statements, correspondence/minutes re policy development
  c) Non-HR: all other records
Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like this, and the bird sounds like this.

Weka is open source software issued under the GNU General Public License.

We have put together several free online courses that teach machine learning and data mining using Weka. Check out the website for the courses for details on when and how to enrol. The videos for the courses are available on Youtube.

Yes, it is possible to apply Weka to process big data and perform deep learning!

<table>
<thead>
<tr>
<th>Getting started</th>
<th>Further Information</th>
<th>Developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Requirements</td>
<td>• Citing Weka</td>
<td>• Development</td>
</tr>
<tr>
<td>• Download</td>
<td>• Datasets</td>
<td>• History</td>
</tr>
<tr>
<td>• Documentation</td>
<td>• Related Projects</td>
<td>• Subversion</td>
</tr>
<tr>
<td>• FAQ</td>
<td>• Miscellaneous Code</td>
<td>• Contributors</td>
</tr>
<tr>
<td>• Getting Help</td>
<td>• Other Literature</td>
<td>• Commercial licenses</td>
</tr>
</tbody>
</table>
Weka
AVP ICT records

Department of Computing Services

56 Physics Building
116 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada S7N 5E2

General Office: (306) 966-4866
Help Desk: (306) 966-4817
FAX: 966-4938
Inquiries: help.desk@usask.ca
Web page comments: webmaster@www.usask.ca

About Us

- General Information
- Announcements
- What's New
- Search DCS Web Pages

DCS Offices

- Main Office, 56 Physics, (306) 966-4866
- Help Desk, 70 Arts Building, Help Line Phone (306) 966-4817
- Campus Computer Store, 85 Place Riel, (306) 966-8375
<table>
<thead>
<tr>
<th>Name</th>
<th>Date modified</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountability</td>
<td>9/11/2008 12:41 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>adv comm</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Archive</td>
<td>9/11/2008 12:32 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>BAH retire</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Budgets</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>CCS</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>COFT</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Colleges</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>corresp</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>CTL</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Distance Education</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Employment Equity</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Equipment Issues</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Facilities Mgmt</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Faculty Presentation</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Gartner Group</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Giga</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>GIS</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Groups</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
</tbody>
</table>
Building training sets

- Round 1 – proof of concept – all documents
- Round 2
  a) HR – first 50 documents in order (UUID)
  b) Non-HR – first 50 good examples
## Round 2 - cross-validation

<table>
<thead>
<tr>
<th>Classified as</th>
<th>HR</th>
<th>Non-HR</th>
<th>Not classified</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>45</td>
<td>5</td>
<td></td>
<td>90.0%</td>
</tr>
<tr>
<td>Non-HR</td>
<td>4</td>
<td>46</td>
<td></td>
<td>92.0%</td>
</tr>
<tr>
<td>Precision</td>
<td>91.8%</td>
<td>90.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correctly identified instances: 91/91 = 91.0%
## Round 2 – test set (Naïve Bayes)

<table>
<thead>
<tr>
<th>Classified as</th>
<th>HR</th>
<th>Non-HR</th>
<th>Not classified</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>93</td>
<td>6</td>
<td></td>
<td>93.9%</td>
</tr>
<tr>
<td>Non-HR</td>
<td>142</td>
<td>1498</td>
<td>8</td>
<td>90.9%</td>
</tr>
<tr>
<td>Precision</td>
<td>39.6%</td>
<td>99.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correctly identified instances</td>
<td></td>
<td></td>
<td>1591</td>
<td>91.5%</td>
</tr>
</tbody>
</table>
Building training set cont’d

- Round 3 - Expanded the Round 2 training set to distinguish between HR-personal and HR-general, with just under 50 documents in each
## Round 3 – cross-validation

<table>
<thead>
<tr>
<th>Classified as</th>
<th>HR-general</th>
<th>HR-personal</th>
<th>Non-HR</th>
<th>Not classified</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR-general</td>
<td>37</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>80.4%</td>
</tr>
<tr>
<td>HR-personal</td>
<td>12</td>
<td>33</td>
<td>3</td>
<td>3</td>
<td>68.8%</td>
</tr>
<tr>
<td>Non-HR</td>
<td>2</td>
<td>0</td>
<td>46</td>
<td></td>
<td>95.8%</td>
</tr>
</tbody>
</table>

Precision: 72.5% 82.5% 90.2%

Correctly identified instances: 116 81.7%
<table>
<thead>
<tr>
<th>Classified as</th>
<th>HR-general</th>
<th>HR-personal</th>
<th>Non-HR</th>
<th>Not classified</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR-general</td>
<td>92</td>
<td>4</td>
<td>15</td>
<td>0</td>
<td>82.9%</td>
</tr>
<tr>
<td>HR-personal</td>
<td>25</td>
<td>82</td>
<td>8</td>
<td>0</td>
<td>71.3%</td>
</tr>
<tr>
<td>Non-HR</td>
<td>298</td>
<td>9</td>
<td>1208</td>
<td>6</td>
<td>79.4%</td>
</tr>
<tr>
<td>Precision</td>
<td>22.2%</td>
<td>86.3%</td>
<td>98.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correctly identified instances 1382 79.4%
Building training sets cont’d

- Round 4 – took the largest 50 documents from each folder
## Round 4 – cross-validation

<table>
<thead>
<tr>
<th>Classified as</th>
<th>HR-general</th>
<th>HR-personal</th>
<th>Non-HR</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR-general</td>
<td>41</td>
<td>6</td>
<td>3</td>
<td>82.0%</td>
</tr>
<tr>
<td>HR-personal</td>
<td>7</td>
<td>43</td>
<td>0</td>
<td>86.0%</td>
</tr>
<tr>
<td>Non-HR</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Precision**

- HR-general: 85.4%
- HR-personal: 87.8%
- Non-HR: 94.3%

**Correctly identified instances**

- 89.3%
### Round 4 – test set

<table>
<thead>
<tr>
<th>Classified as</th>
<th>HR-general</th>
<th>HR-personal</th>
<th>Non-HR</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR-general</td>
<td>108</td>
<td>3</td>
<td>0</td>
<td>97.3%</td>
</tr>
<tr>
<td>HR-personal</td>
<td>36</td>
<td>79</td>
<td>0</td>
<td>68.7%</td>
</tr>
<tr>
<td>Non-HR</td>
<td>996</td>
<td>117</td>
<td>445</td>
<td>28.6%</td>
</tr>
<tr>
<td>Precision</td>
<td>9.5%</td>
<td>39.7%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Correctly identified instances</td>
<td></td>
<td></td>
<td></td>
<td>35.4%</td>
</tr>
</tbody>
</table>
Building training sets cont’d

- **Round 5** - Combined the Round 4 training set for HR-personal and HR-general; kept 25 largest documents from the Round 4 HR-personal and HR-general training sets.

- **Round 6** - re-run with Round 3 training set (is this thing working?)

- **Round 7** - Minor adjustments to Round 4 training set (e.g. removed a few documents with bad encoding).
Building training sets cont’d

- **Round 8** - For HR-personal and HR-general, the middle set (based on size) was used, choosing 50 documents each. For non-HR, started at the 100th largest file, and took the next 50 by size.

- **Round 9** - Starting with the Round 3 training set, combined HR-general and HR-personal, resulting in two categories, HR and non-HR
  
  a) The combined set was not weeded, so the new HR set was double the size of non-HR.
Building training sets cont’d

- **Round 10** - New training set: Created new training set from the full set – 12 documents in each.

- **Round 11** - Combined HR-general and HR-personal, with some minor adjustments, resulting in 16 HR and 22 non-HR documents.

- **Round 12** - Returned to the training set from Round 3, but some refinements to the full data set had been made by this point.
Building training set cont’d

- **Round 13** - Combined HR-general and non-HR using the Round 5 training set. Added 50 more documents to HR-personal (next 50 largest)

- **Round 14** - With the Round 4 training set, added the AttributeSelection filter.

- **Round 15** - With the Round 3 training set, added the AttributeSelection filter

- **Round 16** - Refined the Round 4 training set using Round 3 non-HR data.
Overall results – 2 categories

2 categories - test set

- Recall - HR
- Recall - non-HR
- Precision - HR
- Precision - non-HR
- Classification accuracy
- F1 - HR
- F1 - non-HR

- Round 2a
- Round 2b
- Round 5
- Round 10
- Round 11
Overall results – 2 categories

2 categories - test set

- Recall - HR
- Recall - non-HR
- Precision - HR
- Precision - non-HR
- Classification accuracy
- F1 - HR
- F1 - non-HR
Overall results – 2 categories (x-val)

2 categories - cross-validation

- Recall - HR
- Recall - non-HR
- Precision - HR
- Precision - non-HR
- Classification accuracy
- F1 - HR
- F1 - non-HR

Round 2a | Round 2b | Round 5 | Round 10 | Round 11
Overall results – 2 categories (x-val)
General analysis – 2 categories

- **Recall**
  a) HR - excellent across the board
  b) non-HR - is mixed

- **Precision**
  a) HR - uniformly low
  b) non-HR - uniformly high

- Mixed results for correctly identified instances
- Rounds with high relevance scores offset by low precision score for HR
Overall results – 3 categories (pt 1)

3 categories - test set (1/2)

- Recall: 0.00%, 10.00%, 20.00%, 30.00%, 40.00%, 50.00%, 60.00%, 70.00%, 80.00%, 90.00%, 100.00%
- Precision: 0.00%, 10.00%, 20.00%, 30.00%, 40.00%, 50.00%, 60.00%, 70.00%, 80.00%, 90.00%, 100.00%
- F1: 0.00%, 10.00%, 20.00%, 30.00%, 40.00%, 50.00%, 60.00%, 70.00%, 80.00%, 90.00%, 100.00%

- Round 3
- Round 6
- Round 7
- Round 12
- Round 16
Overall results – 3 categories (pt 2)
Overall results – 3 categories (pt 1)

3 categories - test set (1/2)
Overall results – 3 categories (pt 2)

3 categories - test set (2/2)

Recall - HR-general
Recall - HR-personal
Recall - non-HR
Precision - HR-general
Precision - HR-personal
Precision - non-HR
Classification accuracy
F1 - HR-general
F1 - HR-personal
F1 - non-HR
3 categories – cross-validation

3 categories - cross-validation

- Recall - HR-general
- Recall - HR-personal
- Recall - non-HR
- Precision - HR-general
- Precision - HR-personal
- Precision - non-HR
- F1 - HR-general
- F1 - HR-personal
- F1 - non-HR

Legend:
- Blue: Round 3/6/15
- Red: Round 4
- Green: Round 7
- Black: Round 8
- Cyan: Round 14
- Purple: Round 16
3 categories – cross-validation

3 categories - cross-validation

Recall - HR-general
Recall - HR-personal
Recall - non-HR
Precision - HR-general
Precision - HR-personal
Precision - non-HR
F1 - HR-general
F1 - HR-personal
F1 - non-HR
General analysis – 3 categories

- **Recall**
  a) HR-general - mostly decent to excellent
  b) HR-personal a bit lower, but reasonable
  c) non-HR – mixed

- **Precision**
  a) HR-general – uniformly poor
  b) HR-personal – significant range
  c) non-HR – uniformly excellent

- Correctly identified instances – decent results in half the rounds
Precision and recall

- Trade-off not unexpected
- Use F1 scores to assess best overall results
  - a) 2 categories – round 2 followed by round 8
  - b) 3 categories – round 3 then 12 (same training set)
- Claim: high recall score for HR-personal is most important goal
- Good results in cross-validation (mostly)
Conclusions, further research

- Promising results
- Triage method?
- Refine the models, more data cleanup
- More granular training sets
- Testing of other classification methods
- Document-level analysis
- Run training sets against different collections
Thank you

- tim.hutchinson@usask.ca