Use Data Mining Technology to Analysis Human Summaries for Automatic Summarization

Qi, Longyi
Introduction
Data

Energy Problems and Solutions

<table>
<thead>
<tr>
<th>Article Version</th>
<th>Topic Total</th>
<th>Has Heading</th>
</tr>
</thead>
<tbody>
<tr>
<td>YA</td>
<td>20</td>
<td>true</td>
</tr>
<tr>
<td>NA</td>
<td>20 (same topics as YA)</td>
<td>false</td>
</tr>
<tr>
<td>YB</td>
<td>20</td>
<td>true</td>
</tr>
<tr>
<td>NB</td>
<td>20 (same topics as YB)</td>
<td>false</td>
</tr>
</tbody>
</table>
## Analysis Attributes

<table>
<thead>
<tr>
<th>Analysis Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentence order (sentence position)</td>
</tr>
<tr>
<td>is heading or not</td>
</tr>
<tr>
<td>distance from previous heading</td>
</tr>
<tr>
<td>distance from next heading</td>
</tr>
<tr>
<td>distance from nearest heading</td>
</tr>
<tr>
<td>sentence length (number of words in the sentence)</td>
</tr>
<tr>
<td>normalized average/highest word frequency with eliminating stop words and stemming</td>
</tr>
<tr>
<td>normalized average/highest word frequency with eliminating stop words without stemming</td>
</tr>
<tr>
<td>normalized average/highest word frequency with stemming without eliminating stop words</td>
</tr>
<tr>
<td>normalized average/highest word frequency without eliminating stop word or stemming</td>
</tr>
<tr>
<td>number of name entities</td>
</tr>
<tr>
<td>title matching</td>
</tr>
<tr>
<td>heading matching</td>
</tr>
<tr>
<td>highest word rank in the sentence</td>
</tr>
<tr>
<td>number of words in the top 100 workrank</td>
</tr>
<tr>
<td>normalized sentence wordrank</td>
</tr>
<tr>
<td>cosine similarity</td>
</tr>
<tr>
<td>tanimoto (extensive jaccard) similarity</td>
</tr>
</tbody>
</table>
Frequency v.s. Sentence Positions
Frequency v.s. Sentence Positions
Frequency v.s. Distance from Previous Headings
Frequency v.s. Distance from Previous Headings
Frequency v.s. Distance from nearest Headings
Frequency v.s. Distance from nearest Headings

![Graph showing the relationship between frequency and distance from nearest heading. The graph displays a decreasing trend as the distance increases.](image-url)
Frequency v.s. Sentence Length

![Graph showing frequency vs. sentence length](image-url)
Frequency v.s. Normalized average word frequency
Frequency v.s. Normalized average word frequency
Frequency v.s. Normalized keyword rank

![Graph showing the relationship between frequency and normalized average word frequency.](image)
Frequency v.s. Normalized keyword rank
Support Vector Machine

Iter 517: *(NumConst=136, SU=59, CEps=0.1081, QPEps=0.0521)
Iter 518: ..........*(NumConst=136, SU=60, CEps=0.1266, QPEps=0.0616)
Iter 519: *(NumConst=135, SU=65, CEps=0.1423, QPEps=0.0627)
Iter 520: *(NumConst=135, SU=57, CEps=0.1108, QPEps=0.0531)
Iter 521: *(NumConst=136, SU=68, CEps=0.1503, QPEps=0.0523)
Iter 522: *(NumConst=137, SU=58, CEps=0.1525, QPEps=0.0551)
Iter 523: ..........(NumConst=137, SU=58, CEps=0.0944, QPEps=0.0551)
Final epsilon on KKT-Conditions: 0.09437
Upper bound on duality gap: 484.58858
Dual objective value: dval=396111.24400
Primal objective value: pval=396595.83258
Total number of constraints in final working set: 137 (of 522)
Number of iterations: 523
Number of calls to 'find_most_violated_constraint': 2715
Number of SU: 58
Norm of weight vector: \|w\|=71.22732
Value of slack variable (on working set): \(x_i=78.74293\)
Value of slack variable (global): \(x_i=78.81183\)
Norm of longest difference vector: \(|\psi(x,y) - \psi(x,y)_{\text{bar}}|\) = 21.42347
Runtime in cpu-seconds: 1021.30
Final number of constraints in cache: 905
Compacting linear model... done
Writing learned model... done
Support Vector Machine
Support Vector Machine

Iter 387: *(NumConst=131, SU=63, CEps=0.1860, QPEps=0.0758)
Iter 388: *(NumConst=132, SU=62, CEps=0.1910, QPEps=0.0749)
Iter 389: *(NumConst=133, SU=67, CEps=0.1592, QPEps=0.0753)
Iter 390: *(NumConst=134, SU=73, CEps=0.2068, QPEps=0.0989)
Iter 391: *(NumConst=135, SU=74, CEps=0.1408, QPEps=0.0584)
Iter 392: *(NumConst=135, SU=72, CEps=0.1076, QPEps=0.0521)
Iter 393: ........<NumConst=135, SU=72, CEps=0.0893, QPEps=0.0521>
Final epsilon on KKT-Conditions: 0.08934
Upper bound on duality gap: 523.48954
Dual objective value: dval=436951.87426
Primal objective value: pval=437475.36380
Total number of constraints in final working set: 135 (of 392)
Number of iterations: 393
Number of calls to 'find_most_violated_constraint': 2565
Number of SU: 72
Norm of weight vector: \|w\|=95.93821
Value of slack variable (on working set): \(x_i=86.50360\)
Value of slack variable (global): \(x_i=86.57466\)
Norm of longest difference vector: \(||\Psi(x,y)-\Psi(x,ybar)||=29.56010||\)
Runtime in cpu-seconds: 2475.92
Final number of constraints in cache: 855
Compacting linear model...done
Writing learned model...done
Support Vector Machine

Dual objective value: dual=436951.87426
Primal objective value: pval=437475.36380
Total number of constraints in final working set: 135 (of 392)
Number of iterations: 393
Number of calls to 'find_most_violated_constraint': 2565
Number of SU: 72
Norm of weight vector: |w| = 95.93821
Value of slack variable (on working set): x_i = 86.50360
Value of slack variable (global): x_i = 86.57466
Norm of longest difference vector: ||Psi(x,y) - Psi(x,ybar)|| = 29.56010
Runtime in cpu-seconds: 2475.92
Final number of constraints in cache: 855
Compacting linear model... done
Writing learned model... done

D:\tm>svm_multiclass_classify.exe svm/NB_2_test.txt svm/NB_2_model svm/NB_2_predictions
Reading model... done.
Reading test examples... (43 examples) done.
Classifying test examples... done
Runtime (without IO) in cpu-seconds: 0.00
Average loss on test set: 39.5349
Zero/one-error on test set: 39.53% (26 correct, 17 incorrect, 43 total)

D:\tm>
Conclusion
Questions?
Thank You!