Copyright:

- This Slide deck is licensed under creative commons 3.0. share alike attribution license. It was created by Rene Pickhardt. You can use share and modify this slide deck as long as you attribute the author and keep the same license. All graphics have been self made (unless otherwise stated)
Lesson 2: Modelling the Web with Simple Statistical Descriptive Text Models

Unit 5: Compare the sentence lengths and word lengths of Simple and English Wikipedia

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties
Completing this unit you should

• Get a feeling for interdisciplinary research

• Know the Automated Readability Index

• Have a strong sense of support for our research hypothesis

• Be able to critically discuss the limits of our models
How would linguists tackle this problem?

• Flesch-Kincaid readability test

\[ fkt = 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right) \]

• Wherever the weights and coefficients drop from the idea is clear:
  – first term is low if sentences are shorter
  – second term is low if words have fewer syllables

• Knowing syllables is a non trivial problem for a computer

• Hard to automatically calculate
Interpreting the results of the FKRT

\[ fkt = 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right) \]

<table>
<thead>
<tr>
<th>Score</th>
<th>School Level</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 - 100</td>
<td>5th grade</td>
<td>Very easy to read for average 11 year old</td>
</tr>
<tr>
<td>80-90</td>
<td>6th grade</td>
<td>Easy to read. Conversational English for consumers</td>
</tr>
<tr>
<td>70-80</td>
<td>7th grade</td>
<td>Fairly easy to read</td>
</tr>
<tr>
<td>60-70</td>
<td>8th &amp; 9th grade</td>
<td>Plain English. Easily understood by 13 – 15 year old students</td>
</tr>
<tr>
<td>50-60</td>
<td>10th to 12th grade</td>
<td>Fairly difficult to read</td>
</tr>
<tr>
<td>30-50</td>
<td>college</td>
<td>Difficult to read</td>
</tr>
<tr>
<td>0-30</td>
<td>College graduate</td>
<td>Very difficult to read.</td>
</tr>
</tbody>
</table>
Automated Readability Index

\[
ari = 4.71 \left( \frac{\text{total characters}}{\text{total words}} \right) + 0.5 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 21.43
\]

- Wherever the weights and coefficients drop from the idea is clear:
  - first term is low if words have fewer characters
  - second term is low if sentences are shorter

- Counting words, sentences and characters is easy for a computer

- Formula corresponds to our testable prediction
Interpreting the results of the ARI

\[
ari = 4.71 \left( \frac{\text{total characters}}{\text{total words}} \right) + 0.5 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 21.43
\]

<table>
<thead>
<tr>
<th>Score</th>
<th>Age</th>
<th>Grade Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5-6</td>
<td>Kindergarten</td>
</tr>
<tr>
<td>2</td>
<td>6-7</td>
<td>First grade</td>
</tr>
<tr>
<td>3</td>
<td>7-8</td>
<td>Second grade</td>
</tr>
<tr>
<td>4</td>
<td>8-9</td>
<td>Third grade</td>
</tr>
<tr>
<td>5</td>
<td>9-10</td>
<td>Fourth grade</td>
</tr>
<tr>
<td>6</td>
<td>10-11</td>
<td>Fifth grade</td>
</tr>
<tr>
<td>7</td>
<td>11-12</td>
<td>Sixth grade</td>
</tr>
<tr>
<td>8</td>
<td>12-13</td>
<td>Seventh grade</td>
</tr>
<tr>
<td>9</td>
<td>13-14</td>
<td>Eighth grade</td>
</tr>
<tr>
<td>10-13</td>
<td>15-18</td>
<td>High school</td>
</tr>
<tr>
<td>&gt; 14</td>
<td>18-22</td>
<td>College</td>
</tr>
</tbody>
</table>
What does the ARI for Wikipedia look like?

```python
In [83]:
def ari(fp):
    numSentences = 0
    numWords = 0
    numChars = 0
    for sentence in f:
        words = sentence.split(" ")
        numSentences = numSentences + 1
        numWords = numWords + len(words)
        for word in words:
            numChars = numChars + len(word)
    return 4.71*(float(numChars)/numWords) + 0.5*float(numWords)/numSentences - 21.43

f = open("../datasets/simpleWikiAbstractsOneScentencePerLine")
print "SimpleEnglish ari: ", ari(f)
print "English Wikipedia ari", ari(f)
```

SimpleEnglish ari: 7.16189918182
English Wikipedia ari 9.6751455226

• Can we depend on the result?
Lots of fluctuation for the readability index
Remember! Logarithmic scales are tricky
The full cycle of research... Making new observations asking questions

More words are needed to understand 50% of sentences in English Wikipedia than in Simple Wikipedia

The ARI in Simple English is lower than in English Wikipedia

Could the distribution of sentence lengths be the reason?

Research starts over again with new question
Thank you for your attention!

Contact:
Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de
Copyright:

- This Slide deck is licensed under creative commons 3.0. share alike attribution license. It was created by Rene Pickhardt. You can use share and modify this slide deck as long as you attribute the author and keep the same license. All graphics have been self made (unless otherwise stated)
The instantiated model reflects a particular situation in the world

• When we take a collection of web pages in order to build a text model

• Model characterizes how the world might work in general

• But the models we study only have a special snapshot of a special situation

• also das Modell charakterisiert wie ein Ausschnitt der Welt im Allgemeinen funktioniert und das spezielle Modell instantiiert eine spezielle Situation in der Welt