Language Trees and Zipping
Dario Benedetto, Emanuele Caglioti and Vittorio Loreto

Bernhard Reinke

September 25, 2015
published 2002 in Physical Review Letters

introduces a measure for remoteness of general sequences using zipping

Applications in language recognition, authorship attribution and language classification
General Idea

- Texts from the same author have little relative entropy.
- Data compression is a good tool for approximating entropy.
- Approximate relative entropy by comparing lengths of compressed files.
Relative entropy

- Given large sources $\mathcal{A}$ and $\mathcal{B}$, we want to approximate the relative entropy.
- Take large samples $A, B$ and small samples $a, b$ from $\mathcal{A}, \mathcal{B}$.
- Let $C(x)$ be the length of $x$ compressed.
- The relative entropy is

$$S_{\mathcal{A}\mathcal{B}} \approx \frac{(C(Ab) - C(A)) - (C(Bb) - C(B))}{|b|}$$
Suppose we have texts $A_i$ with known authors and an unknown text $x$.

Take samples $A_i$ and $x$.

Minimize $C(A_i, x) - C(A_i)$ over all texts.

Our output is the author of the minimizing text.
gzip, zlib and DEFLATE

- The authors used gzip for compressing files.
- gzip is implemented by zlib, which uses the DEFLATE algorithm.
- DEFLATE is a combination of Lempel-Ziv-’77 and Huffman encoding.
I used python3.4.3 and the zlib library in the standard python library.

Large samples have length 48 KB, small samples 8 KB.

The output were very noisy (based on the random choice of the samples), so I perform multiple runs and take the most common answer.

zlib offers compression with a pre-set dictionary, compressing $x$ with dictionaries $A_i$ gives similar results and is much faster.
Original dataset

- Italian novels from www.liberliber.it
- 90 texts from 11 authors
- each text is tested against all others
- 84 texts are correctly attributed, 93.3% rate of success
- No comparison to other methods is given.
Substitute dataset

- English novels from PAN12 competition
- $I$ corpus for training
- $I$ test cases and closed $J$ test cases for testing
- 28 texts from 14 authors
- file sizes range from 100 KB to 1.1 MB
### Results of reimplementation

<table>
<thead>
<tr>
<th>Compression level</th>
<th>single run</th>
<th>10 runs</th>
<th>20 runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>17</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>