Day 8: Text Scaling Models from Dictionaries to “Word-scoring”

Kenneth Benoit

Essex Summer School 2012

July 18, 2012
Text as Data: Basic Principles

- Data are observed characteristics of underlying tendencies to be estimated – and therefore not *intrinsically* interesting
- Analysis inherit properties of statistics:
  - Precise characterizations of uncertainty (efficiency of estimators)
  - Concerns with reliability (consistency of estimators)
  - Concerns with validity (unbiasedness of estimators)
- We must be concerned with the stochastic processes generating the data
- We must be concerned with functional relationships between characteristics of texts and authors and observed words
Text generation as a stochastic process
Wordscores conceptually

- Two sets of texts
  - Reference texts: texts about which we know something (a scalar dimensional score)
  - Virgin texts: texts about which we know nothing (but whose dimensional score we'd like to know)
- These are analogous to a “training set” and a “test set” in classification
- Basic procedure:
  1. Analyze reference texts to obtain word scores
  2. Use word scores to score virgin texts
The Wordscore Procedure

(Using the UK 1997-2001 Example)

Step 1: Obtain reference texts with a priori known positions ($\text{setref}$)

Step 2: Generate word scores from reference texts ($\text{wordscore}$)

Step 3: Score each virgin text using word scores ($\text{textscore}$)

Step 4: (optional) Transform virgin text scores to original metric

Wordscores Procedure

<table>
<thead>
<tr>
<th>Party</th>
<th>Year</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labour</td>
<td>1992</td>
<td>5.35</td>
</tr>
<tr>
<td>Liberals</td>
<td>1992</td>
<td>8.21</td>
</tr>
<tr>
<td>Conservatives</td>
<td>1992</td>
<td>17.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Party</th>
<th>Year</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labour</td>
<td>1997</td>
<td>9.17</td>
</tr>
<tr>
<td>Liberals</td>
<td>1997</td>
<td>5.00</td>
</tr>
<tr>
<td>Conservatives</td>
<td>1997</td>
<td>17.18</td>
</tr>
</tbody>
</table>

Reference Texts

Scored word list

Table of scored words:

- drugs: 15.66
- corporation: 15.66
- inheritance: 15.48
- successfully: 15.26
- markets: 15.12
- motorway: 14.96
- nation: 12.44
- single: 12.36
- pensionable: 11.59
- management: 11.56
- monetary: 10.84
- secure: 10.44
- minorities: 9.95
- women: 8.65
- cooperation: 8.64
- transform: 7.44
- representation: 7.42
- poverty: 6.87
- waste: 6.83
- unemployment: 6.76
- contributions: 6.68
Wordscores Procedure

The Wordscore Procedure
(Using the UK 1997-2001 Example)

1. Step 1: Obtain reference texts with a priori known positions (setref)
2. Step 2: Generate word scores from reference texts (wordscore)
3. Step 3: Score each virgin text using word scores (textscore)
4. Step 4: (optional) Transform virgin text scores to original metric

Reference Texts

Labour 1992
5.35

Liberals 1992
8.21

Cons. 1992
17.21

Labour 1997
9.17
(.33)

Liberals 1997
5.00 (.36)

Cons. 1997
17.18 (.32)

Scored word list:

- drugs 15.66
- corporation 15.66
- successfully 15.26
- markets 15.12
- motorway 14.96
- nation 12.44
- pensionable 11.59
- management 11.56
- monetary 10.84
- secure 10.44
- minorities 9.95
- women 8.65
- cooperation 8.64
- representation 7.42
- poverty 6.87
- waste 6.83
- unemployment 6.76
- contributions 6.68
The Wordscore Procedure
(Using the UK 1997-2001 Example)

1. Obtain reference texts with a priori known positions (setref).
2. Generate word scores from reference texts (wordscore).
3. Score each virgin text using word scores (textscore).
4. (optional) Transform virgin text scores to original metric.

<table>
<thead>
<tr>
<th>Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>drugs</td>
<td>15.66</td>
</tr>
<tr>
<td>corporation</td>
<td>15.66</td>
</tr>
<tr>
<td>inheritance</td>
<td>15.48</td>
</tr>
<tr>
<td>successfully</td>
<td>15.26</td>
</tr>
<tr>
<td>markets</td>
<td>15.12</td>
</tr>
<tr>
<td>motorway</td>
<td>14.96</td>
</tr>
<tr>
<td>nation</td>
<td>12.44</td>
</tr>
<tr>
<td>single</td>
<td>12.36</td>
</tr>
<tr>
<td>pensionable</td>
<td>11.59</td>
</tr>
<tr>
<td>management</td>
<td>11.56</td>
</tr>
<tr>
<td>monetary</td>
<td>10.84</td>
</tr>
<tr>
<td>secure</td>
<td>10.44</td>
</tr>
<tr>
<td>minorities</td>
<td>9.95</td>
</tr>
<tr>
<td>women</td>
<td>8.65</td>
</tr>
</tbody>
</table>
The Wordscore Procedure

(Using the UK 1997-2001 Example)

1. Labour 1992 5.35
2. Liberals 1992 8.21
3. Conservatives 1992 17.21
4. Labour 1997 9.17 (.33)
5. Liberals 1997 5.00 (.36)
6. Conservatives 1997 17.18 (.32)

Reference Texts

Scored word list

- drugs 15.66
- corporation 15.66
- inheritance 15.48
- successfully 15.26
- markets 15.12
- motorway 14.96
- nation 12.44
- single 12.36
- pensionable 11.59
- management 11.56
- monetary 10.84
- secure 10.44
- minorities 9.95
- women 8.65
- cooperation 8.64
- transform 7.44
- representation 7.42
- poverty 6.87
- waste 6.83
- unemployment 6.76
- contributions 6.68

Step 1: Obtain reference texts with a priori known positions (setref)

Step 2: Generate word scores from reference texts (wordscore)

Step 3: Score each virgin text using word scores (textscore)

Step 4: (optional) Transform virgin text scores to original metric
Wordscores mathematically: Reference texts

- Start with a set of \( I \) reference texts, represented by an \( I \times J \) document-term frequency matrix \( C_{ij} \), where \( i \) indexes the document and \( j \) indexes the \( J \) total word types.

- Each text will have an associated “score” \( a_i \), which is a single number locating this text on a single dimension of difference.
  - This can be on a scale metric, such as 1–20.
  - Can use arbitrary endpoints, such as -1, 1.

- We normalize the document-term frequency matrix within each document by converting \( C_{ij} \) into a relative document-term frequency matrix (within document), by dividing \( C_{ij} \) by its word total marginals:

\[
F_{ij} = \frac{C_{ij}}{C_i}.
\]  

where \( C_i = \sum_{j=1}^{J} C_{ij} \).
Compute an \( I \times J \) matrix of relative document probabilities \( P_{ij} \) for each word in each reference text, as

\[
P_{ij} = \frac{F_{ij}}{\sum_{j=1}^{J} F_{ij}}
\]

This tells us the probability that given the observation of a specific word \( j \), that we are reading a text of a certain reference document \( i \)
Assume we have two reference texts, A and B

The word “choice” is used 10 times per 1,000 words in Text A and 30 times per 1,000 words in Text B

So $F_i \text{ "choice"} = \{.1, .3\}$

If we know only that we are reading the word choice in one of the two reference texts, then probability is 0.25 that we are reading Text A, and 0.75 that we are reading Text B

(3)
Compute a $J$-length “score” vector $S$ for each word $j$ as the average of each document $i$’s scores $a_i$, weighted by each word’s $P_{ij}$:

$$S_j = \sum_{i=1}^{l} a_i P_{ij}$$  \hspace{1cm} (4)

In matrix algebra, $S = a \cdot P$

This procedure will yield a single “score” for every word that reflects the balance of the scores of the reference documents, weighted by the relative document frequency of its normalized term frequency.
Continuing with our example:

- We “know” (from independent sources) that Reference Text A has a position of 1.0, and Reference Text B has a position of +1.0
- The score of the word choice is then
  \[0.25(-1.0) + 0.75(1.0) = -0.25 + 0.75 = +0.50\]
Wordscores mathematically: Scoring “virgin” texts

- Here the objective is to obtain a single score for any new text, relative to the reference texts.
- We do this by taking the mean of the scores of its words, weighted by their term frequency.
- So the score $v_k$ of a virgin document $k$ consisting of the $j$ word types is:

$$v_k = \sum_j (F_{kj} \cdot s_j)$$

(5)

where $F_{kj} = \frac{C_{kj}}{C_k}$ as in the reference document relative word frequencies.

- Note that new words outside of the set $J$ may appear in the $K$ virgin documents — these are simply ignored (because we have no information on their scores).

- Note also that nothing prohibits reference documents from also being scored as virgin documents.
Wordscores mathematically: Rescaling raw text scores

- Because of overlapping or non-discriminating words, the raw text scores will be dragged to the interior of the reference scores (we will see this shortly in the results)
- Some procedures can be applied to rescale them, either to a unit normal metric or to a more “natural” metric
- Martin and Vanberg (2008) have proposed alternatives to the LBG (2003) rescaling
Computing confidence intervals

- The score $v_k$ of any text represents a weighted mean.
- LBG (2003) used this logic to develop a standard error of this mean using a weighted variance of the scores in the virgin text.
- Given some assumptions about the scores being fixed (and the words being conditionally independent), this yields approximately normally distributed errors for each $v_k$.
- An alternative would be to bootstrap the textual data prior to constructing $C_{ij}$ and $C_{kj}$ — see Lowe and Benoit (2012).
Pros and Cons of the Wordscores approach

- Fully automated technique with minimal human intervention or judgment calls – only with regard to reference text selection
- Language-blind: all we need to know are reference scores
- Could potentially work on texts like this:

(See http://www.kli.org)
Pros and Cons of the Wordscores approach

- Estimates unknown positions on a priori scales – hence no inductive scaling with a posteriori interpretation of unknown policy space
- Very dependent on correct identification of:
  - appropriate reference texts
  - appropriate reference scores
Suggestions for choosing reference texts

- Texts need to contain information representing a clearly dimensional position
- Dimension must be known a priori. Sources might include:
  - Survey scores or manifesto scores
  - Arbitrarily defined scales (e.g. -1.0 and 1.0)
- Should be as discriminating as possible: extreme texts on the dimension of interest, to provide reference anchors
- Need to be from the same lexical universe as virgin texts
- Should contain lots of words
Suggestions for choosing reference values

- Must be “known” through some trusted external source
- For any pair of reference values, all scores are simply linear rescalings, so might as well use (-1, 1)
- The “middle point” will not be the midpoint, however, since this will depend on the relative word frequency of the reference documents
- Reference texts if scored as virgin texts will have document scores more extreme than other virgin texts
- With three or more reference values, the mid-point is mapped onto a multi-dimensional simplex. The values now matter but only in relative terms (we are still investigating this fully)