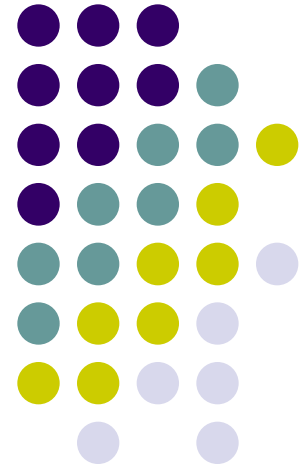
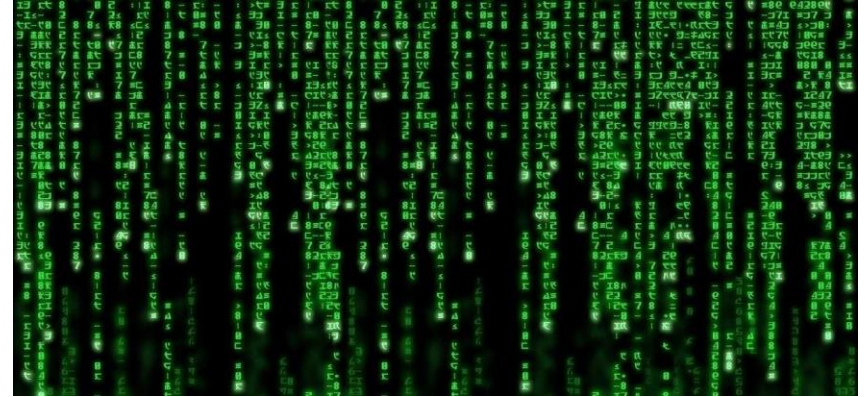
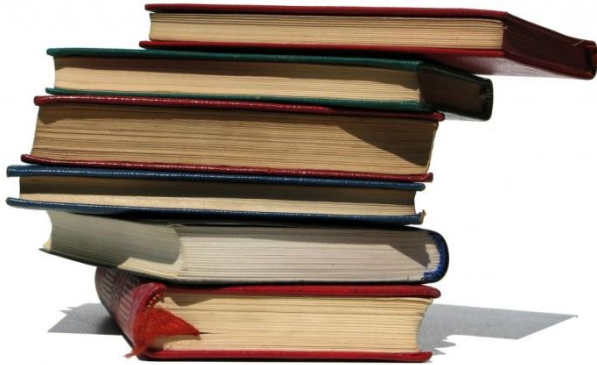


Applied Scaling & Classification Techniques in Political Science

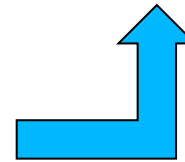
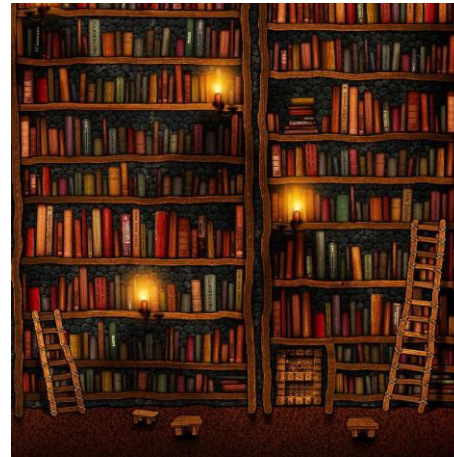
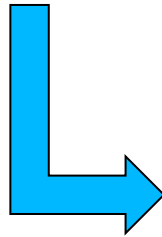
Lecture 2
Wordscore



But before that...a summary

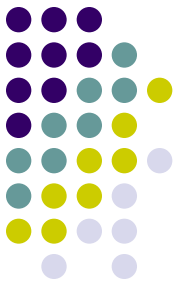


You pass to R the texts you want to analyze via `readtext`



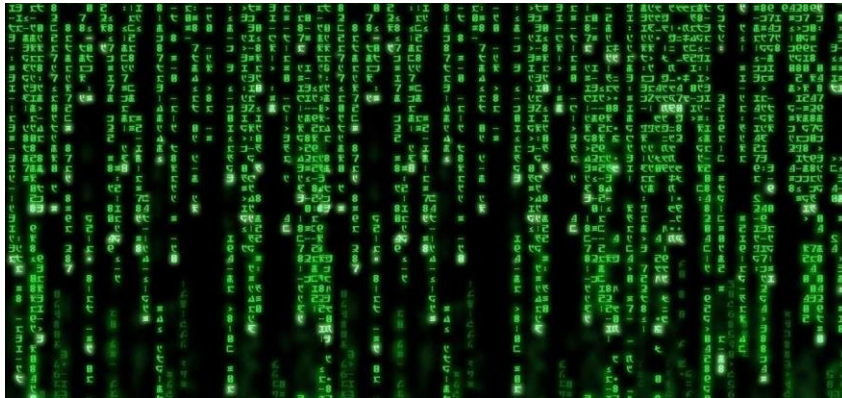
You extract from the corpus the relative document term (or feature) matrix via `dfm`. By doing that you apply the bag of words approach to that corpus of texts

You tell to R that those bunch of texts belong to the *same collection of texts* you want to analyze via `corpus`



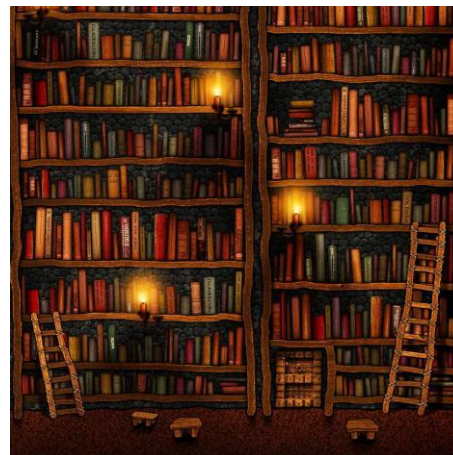
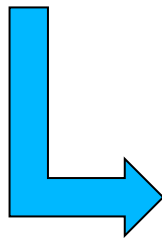
But before that...a summary

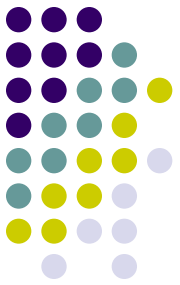
All the statistical models that we will see, **work on this**



| Word | LDP | JSP | KOMEI | DSP |
|------|-----|-----|-------|-----|
| 総理 | 0 | 38 | 29 | 23 |
| 国民 | 11 | 18 | 23 | 22 |
| 基地 | 0 | 3 | 47 | 3 |
| 政府 | 6 | 21 | 8 | 15 |
| 大学 | 7 | 4 | 1 | 38 |
| わが国 | 10 | 2 | 12 | 7 |
| 政治 | 3 | 6 | 6 | 15 |
| 大臣 | 0 | 8 | 1 | 18 |
| 経済 | 8 | 9 | 2 | 6 |
| 安保 | 0 | 4 | 10 | 8 |
| ゴルフ | 0 | 0 | 21 | 0 |
| 社会 | 0 | 10 | 8 | 3 |
| 政策 | 2 | 8 | 1 | 7 |
| 物価 | 2 | 3 | 8 | 5 |
| 学生 | 1 | 2 | 1 | 13 |
| 本土 | 1 | 5 | 3 | 8 |
| 予算 | 1 | 11 | 2 | 3 |
| 考え | 5 | 3 | 5 | 3 |
| 方式 | 0 | 0 | 0 | 16 |
| 公害 | 0 | 0 | 15 | 0 |

NOT on this





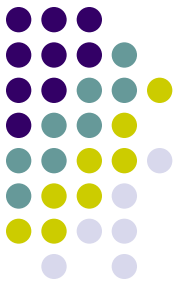
References

- ✓ Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97:311–31

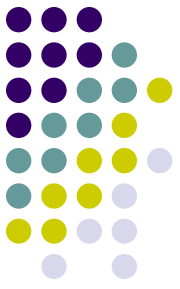
Wordscores

Wordscores is a **supervised method for scaling...**

...that is, it requires **a-priori information** by the researcher to produce estimates



Wordscores



In particular, Wordscores technique estimates policy positions by **comparing two sets of political texts**

On one hand is a set of texts ("**reference**" texts) whose policy positions on well-defined a-priori dimensions are "**known**" to the analyst, in the sense that these can be either estimated with confidence from independent sources or assumed uncontroversial

On the other hand is a set of texts whose policy positions we do not know but want to find out ("**virgin**" texts). All we do know about the virgin texts is the words we find in them, which **we compare to the words** we have observed in reference texts with "known" policy positions

Wordscores



More specifically, we use the **relative frequencies** we observe for each of the **different words** in each of the **reference texts** to calculate the **probability** that we are reading a **particular reference text**, given that we are reading a particular word

For a particular a-priori policy dimension, this allows us to generate a **numerical "score" for each word** from the reference texts analysis

This score is the **expected policy position of any possible text**, given only that we are reading the **single word** in question

Wordscores

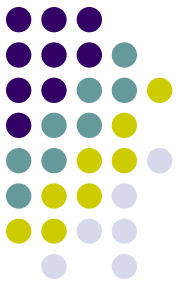


Then, we use the **word scores we generate from the reference texts** to estimate the **positions of virgin texts** on the policy dimensions in which we are interested

Essentially, **each word scored of each virgin text** gives us a small amount of information about which of the reference texts the virgin text **most closely resembles**

This produces a **conditional expectation** of the virgin text's policy position, and **each scored word** in a virgin text adds to this information

Wordscores



Our procedure can thus be thought of as a type of Bayesian reading of the virgin texts, with our estimate of the policy position of any given virgin text being **updated** each time we read a word that is **also found** in one of the reference texts

The more scored words we read, the more confident we become in our estimates

Wordscores: an example

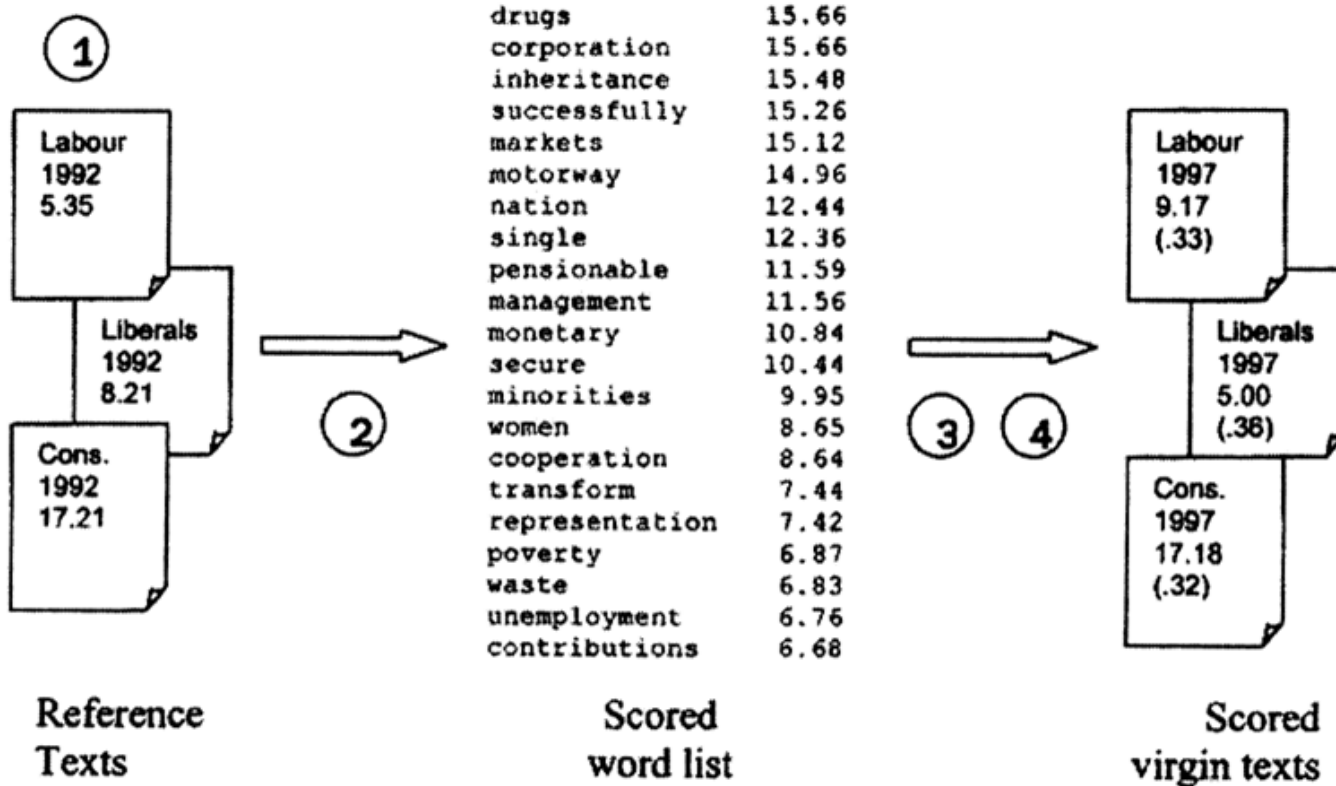
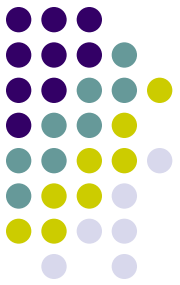


Our **reference texts** are the 1992 manifestos of the British Labour, Liberal Democrat, and Conservative parties

From some external source, **we know (or we assume to know)** the policy position expressed in each of such party manifestos

The research task is to estimate the **unknown policy positions** revealed by the 1997 manifestos of the same parties, which are thus treated as **virgin texts**, by treating the 1992 manifestos as the **reference texts**

Wordscores: a resume



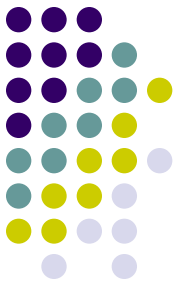
Step 1: Obtain reference texts with a priori known positions

Step 2: Generate word scores from reference texts

Step 3: Score each virgin text using word scores

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

Wordscores



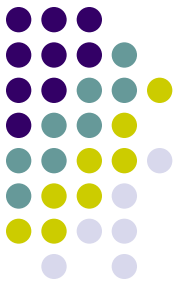
More formally...

R = set of reference texts

We assume that we know with confidence the policy position on dimension d of each reference text r (A_{rd})

F_{wr} = the relative observed frequency of each different word w used in reference text r

Wordscores



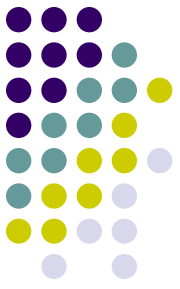
Once we have observed F_{wr} for each of the reference texts, we have a matrix of relative word frequencies that allows us to calculate a matrix of **conditional probabilities**

Each element in this matrix tells us the **probability** that we are reading reference text r , given that we are reading word w

This quantity **is the key** to the Wordscores a-priori approach

More in details...

Wordscores



Given a **set of reference texts**, the probability that an occurrence of word w implies that we are reading text r is:

$$P_{r|w} = \frac{F_{wr}}{\sum_r F_{wr}}$$

As an **example** consider two reference texts, A and B. We observe that the word "*choice*" is used 10 times per 10,000 words in Text A and 30 times per 10,000 words in Text B. If we know simply that we are reading the word "*choice*" in one of the two reference texts, then which is the probability of reading Text A (and Text B?)

0.25 probability that we are reading Text A (10/40); 0.75 probability that we are reading Text B (30/40)

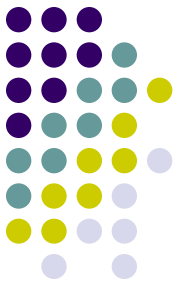
Wordscores



We can then use this matrix $P_{r|w}$ to produce a **score** for each word w on dimension d . This is the expected position on dimension d of any text we are reading, given **only** that we are reading word w , and is defined as:

$$S_{d|w} = \sum_r (P_{r|w} * A_{rd})$$

Wordscores



To continue with our simple example, imagine that Reference Text A is assumed to have a position of 3 on dimension d , and Reference Text B is assumed to have a position of 8 on the same dimension d

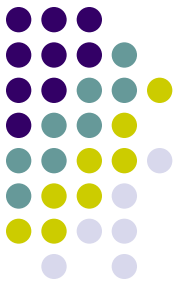
The score of the word "*choice*" is then...what?

$$0.25*(3) + 0.75*(8) = 0.75 + 6 = 6.75$$

Given the pattern of word usage in the reference texts, if we knew only that the word "*choice*" occurs in some text, then this implies that the text's expected position on the dimension under investigation is +6.75

Of course we will **update this expectation** as we gather more information about the text under investigation by reading more words

Wordscores



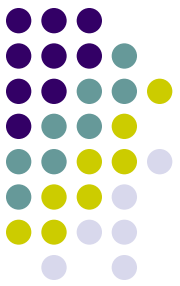
Note that if reference text r contains occurrences of word w and no other text contains word w , then $P_{r|w}$ is equal to what?

$P_{r|w} = 1!$ If we are reading word w , then we conclude from this that we are certainly reading text r

And what about $S_{d|w}$ in this case?

In this event, the score of word w on dimension d is the position of reference text r on dimension d : thus $S_{d|w} = A_{rd}$

Wordscores



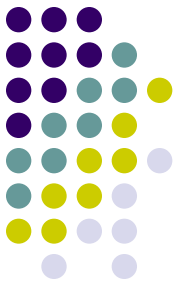
On the contrary, if all reference texts contain occurrences of word w at precisely **equal frequencies**, then reading word w leaves us **none the wiser** about which text we are reading

In this case S_{wd} is the **mean position** of all reference texts

Back to previous example, if the word “choice” is found with the same frequencies in Reference Text A and Reference Text B, then the score of the word "choice" is simply the mean position of Reference Texts A (i.e., 3) and B (i.e., 8), that is:

$$0.5*(3) + 0.5*(8) = +5.5$$

Wordscores



Scoring Virgin Texts

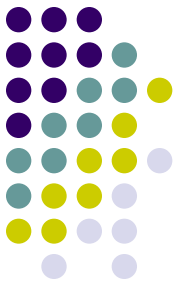
Having calculated scores for all **words in the word universe of the reference texts**, the analysis of any set of virgin texts V of any size is straightforward

First, we must compute the relative frequency of each **virgin text word**, as a proportion of the total number of words in the virgin text. We call this frequency F_{wv}

The **score** of any virgin text v on dimension d , S_{vd} , is then the **mean dimension score** of all of the scored words that it contains, **weighted** by the frequency of the scored words:

$$S_{vd} = \sum_w (F_{wv} * S_{d|w})$$

Wordscores



This single numerical score represents the **expected position of the virgin text** on the a-priori dimension under investigation

This inference is based on the assumption that the **relative frequencies of word usage** in the virgin texts are linked to policy positions **in the same way** as the relative frequencies of word usage in the reference texts

This is why the selection of **appropriate reference** texts is such an important matter (more on this below)

Wordscores



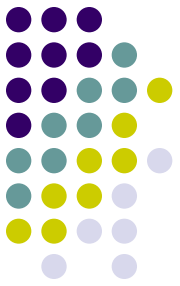
Interpreting Virgin Text Scores

Once raw estimates have been calculated for each virgin text, we need to interpret these in **substantive terms**

Problem: many words are shared across reference texts

As a result of that, such words receive a centrist score, i.e., they take as their scores **the mean overall scores of the reference texts** (given that they do not discriminate among texts)

Wordscores



As a consequence, for any set of virgin texts containing the **same set of non-discriminating words** found in the reference texts, the presence of these overlapping words pulls raw scores toward the interior of the interval defined by the reference scores, that is...

...the raw virgin text scores tend to be much more **clustered** together than the reference text scores

Wordscores

Because raw scores are dispersed **on a much smaller scale**, they cannot be directly **compared to the exogenous scores** attached to the reference texts.

To compare the virgin scores directly with the reference scores, therefore, we need then to **transform/standardized** the scores of the virgin texts so that they have **same dispersion metric as the reference texts**



Wordscores



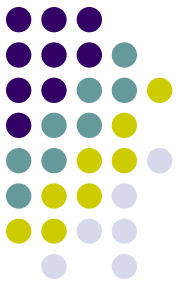
For each virgin text v on a dimension d (where the total number of virgin texts $V > 1$), this is done as follows:

$$S^*_{vd} = (S_{vd} - S_{\bar{v}d}) \left(\frac{SD_{rd}}{SD_{vd}} \right) + S_{\bar{v}d}$$

where $S_{\bar{v}d}$ is the average score of the virgin texts, and the SD_{rd} and SD_{vd} are the sample standard deviations of the reference and virgin text scores, respectively

This **preserves** the mean and relative positions of the virgin scores but **sets their variance equal to that of the reference texts**

Wordscores



Estimating the Uncertainty of Text Scores

Recall that each virgin text score S_{vd} is the **weighted mean score** of the words in text v on dimension d

If we can compute a mean for any set of quantities, then we can also compute a variance...and from here a measure of uncertainty

In this context our interest is in how, for a given text, the scores $S_{d|w}$ of the words in the text vary around this mean

Wordscores



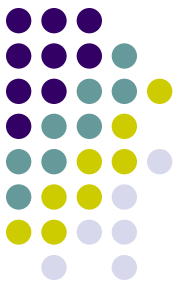
The variance of $S_{d|w}$ for a given text **measures how dispersed the individual word scores** are around the text's mean score. The less this variance, the more the words in the text all correspond to the final score

Because the text's score S_{vd} is a weighted average, the variance we compute also needs to be weighted

We therefore compute V_{vd} , the variance of each word's score around the text's total score, weighted by the frequency of the scored word in the virgin text:

$$V_{vd} = \sum_w F_{wv} (S_{d|w} - S_{vd})^2$$

Wordscores

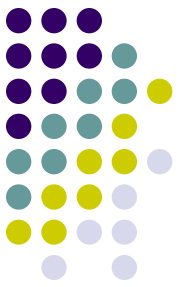


This measure produces a familiar quantity directly analogous to the unweighted variance, summarizing the "consensus" of the scores of each word in the virgin text

Intuitively, we can think of each scored word in a virgin text as generating an independent prediction of the text's overall policy position. When these predictions are tightly clustered, we are **more confident** in their consensus than when they are scattered more widely

As with any variance, we can use the square root of V_{vd} to produce a standard deviation. This standard deviation can be used in turn, along with the total number of scored virgin words N^v , to generate a standard error $\sqrt{V_{vd}}/\sqrt{N^v}$ for each virgin text's score S_{vd}

Wordscores: which reference texts?

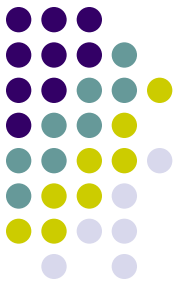


Wordscores does not make any assumption about words usage (contrary to unsupervised methods for scaling: more on this later!)

But to produce an answer (i.e., a score for unknown texts), it requires the **information present in some reference texts**

A big advantages of using Wordscores is that by changing the scores of the dimension d (i.e., first a score for the economic dimension; then a score for the foreign-policy dimension, etc.), we can use the **same reference texts** to score the position of the same virgin texts **on different dimensions** as we will see in the Lab class!

Wordscores: which reference texts?



Selection of Reference Texts

The selection of an appropriate set of reference texts is clearly a crucial aspect of the research design of the type of a-priori analysis

Three general guidelines in the selection of reference texts

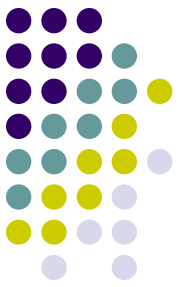
Wordscores: which reference texts?



First: the reference texts should use the **same lexicon**, in the same context, as the virgin texts being analyzed

For example, if you analyze party manifestos, use as reference texts other party manifestos, if you analyze speeches in a legislature, use as reference texts other speeches, and so on

Wordscores: which reference texts?

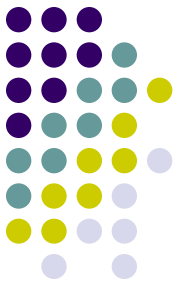


Second: policy positions of the reference texts should "**span**" the dimensions in which we are interested. Trivially, if all reference texts have the **same policy position** on some dimension under investigation, then their content contains no information that can be used to distinguish between other texts on the same policy dimension

An ideal selection of reference texts will contain texts that **occupy extreme positions, as well as positions at the center**, of the dimensions under investigation

This allows differences in the content of the reference texts to form the basis of inferences about differences in the content of virgin texts.

Wordscores: which reference texts?



Third: the set of reference texts should contain as **many different words as possible**

The content of the virgin texts is analyzed in the context of the **word universe of the reference texts**

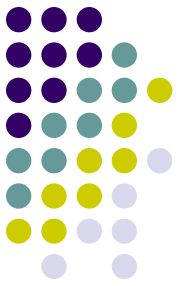
The more comprehensive this word universe, and thus the **less often we find words in virgin texts that do not appear in any reference text**, the better

In the extreme scenario where no word in virgin texts appears in any reference text, Wordscores become completely useless!

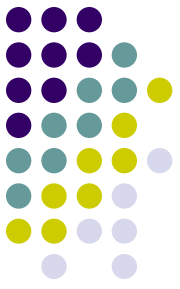
Wordscores challenges (1)

In using particular reference texts, we are assuming that party manifestos in country c at election t are valid points of reference for the analysis of party manifestos at election $t + 1$ in the same country...

...we have here a **temporal dynamic challenge**



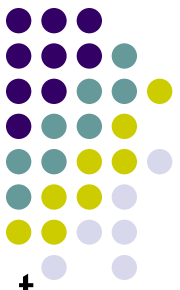
Wordscores challenges (1)



That is, computing word scoring runs into significant problems when it comes to generating **long time series** of the policy positions of particular texts authors

Essentially this is because **words change their political associations over time**, which makes it difficult for us to know, if we estimate the positions of the **same author** of different texts issued at different time points, whether any **movement** we observe can be attributed to a changing meaning of the words, or to a changing underlying policy position of the author

Wordscores challenges (1)



Example: imagine that you want to use reference texts at time $t-1$, to estimate texts at time t and time $t+1$

Imagine that in times $t-1$, the text uses the word “nigger” to identify Afro-Americans, while in time t the text uses the word “black” and at time $t+1$ the word “Afro-Americans”. Different words that refer to the same concept

In this case, however, all the information related to “black” and “Afro-Americans” will be lost

Two possible answer in this respect: 1) you modify the words “nigger” and “black” in your texts with the words “Afro-Americans”. Through that you avoid the problem of word-comparability; 2) you select reference texts from time $t-1$, t and $t+1$, so that through that you increase the “universe of words” used in both the reference and the virgin texts

Wordscores challenges (2)



The transformation that LBG apply to raw scores to have the same dispersion in virgin texts as the reference text scores can be problematic:

$$S_{vd}^* = (S_{vd} - S_{\bar{v}d}) \left(\frac{SD_{rd}}{SD_{vd}} \right) + S_{\bar{v}d}$$

To adjust the dispersion of the raw scores, the transformation relies on the standard deviation of the virgin text raw scores. But this **standard deviation depends** on the particular set of virgin texts that are analyzed.

Put simply, the LBG-transformed scores are inherently non-robust to the selection of virgin texts

How to develop a transformation that makes scores independent of such aspect?

Wordscores challenges (2)



Possible answer: why bothering in transforming the raw scores?

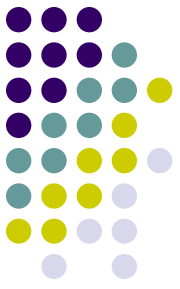
The most direct way to use Wordscores output is to interpret the **virgin text scores directly** since these scores contain substantive information on an interval scale (as well as the relative ordering of parties in a policy space)

If we wish moreover to compare estimated virgin text positions to reference texts, **we can simply score reference texts too as if they were “virgin” texts**

Because they are all generated by a single dictionary, these scores tell us *now* how the word usage across texts (*both* virgin and reference) differs **as evaluated by the same dictionary.**

The resulting raw estimates are robust, in the sense of being the same regardless of **the set of virgin texts chosen**

Wordscores challenges (2)



Alternatively, you can apply the transformation proposed in Marty and Vanberg (2008)

- ✓ Martin, Lanny W., and Georg Vanberg. 2008. A robust transformation procedure for interpreting political text. *Political Analysis*, 16: 93-100

Wordscores challenges (2)



So what to do?

Possible suggestions:

- 1) If transformation is motivated by a desire to compare like-for-like reference and virgin texts on the same absolute metric, use the LBG transformation
- 2) Otherwise, compare raw scores to one another

Hint: personally, I tend to adopt the LBG transformation