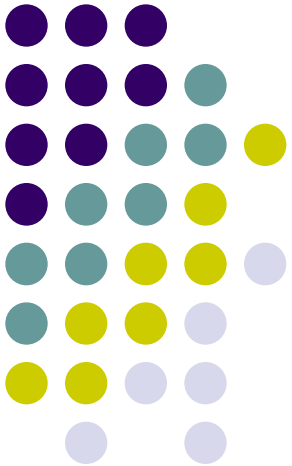
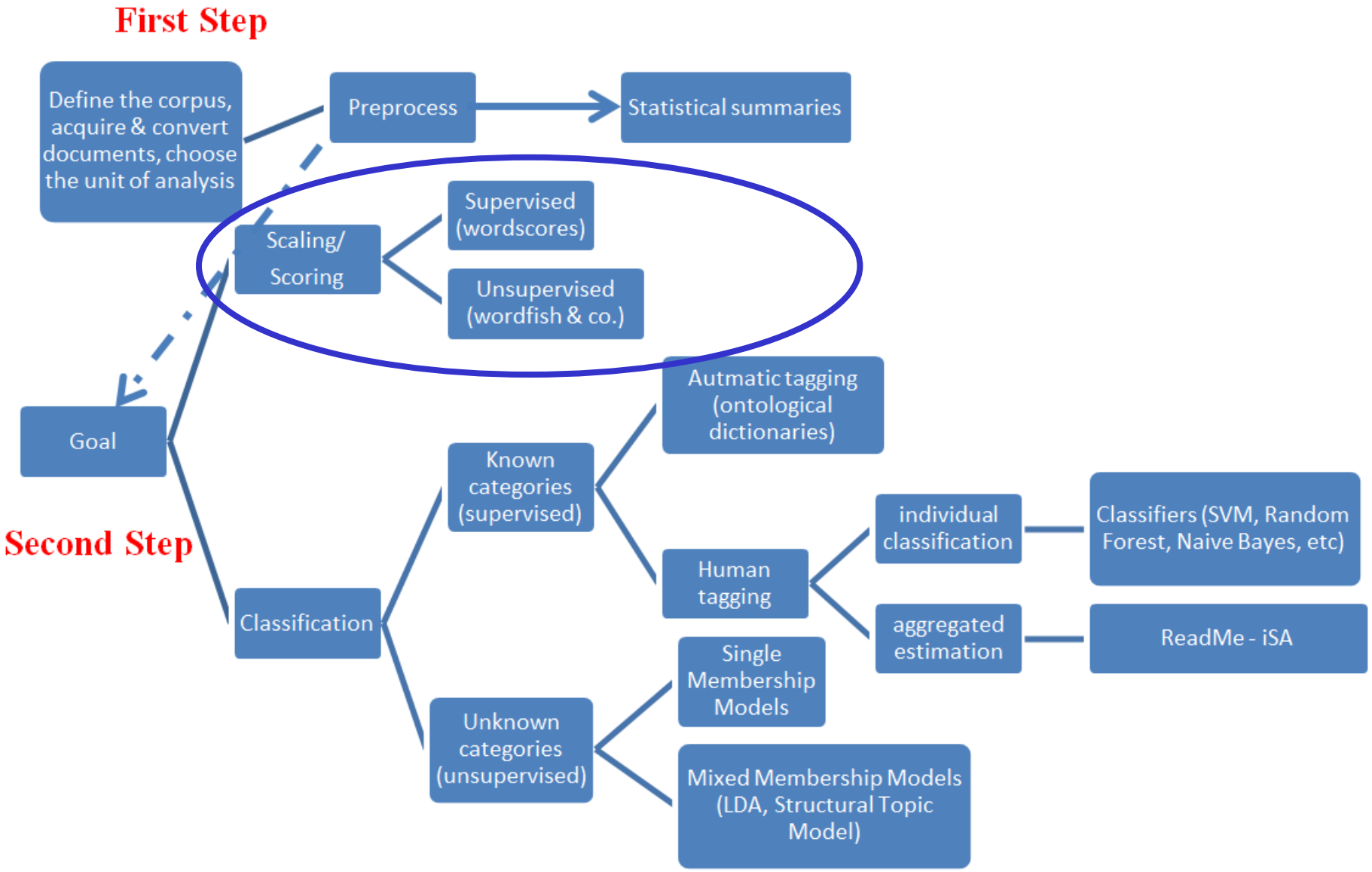
Applied Scaling & Classification Techniques in Political Science

Lecture 3

Supervised scaling algorithms: Wordscores



Our Course Map



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
- ✓ Egerod, Benjamin C.K., and Robert Klemmensen (2020). Scaling Political Positions from text. Assumptions, Methods and Pitfalls. In Luigi Curini and Robert Franzese (eds.), *SAGE Handbook of Research Methods is Political Science & International Relations*, London, Sage, chapter 27
- ✓ Martin, Lanny W., and Georg Vanberg. 2008. A robust transformation procedure for interpreting political text. *Political Analysis*, 16: 93-100

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

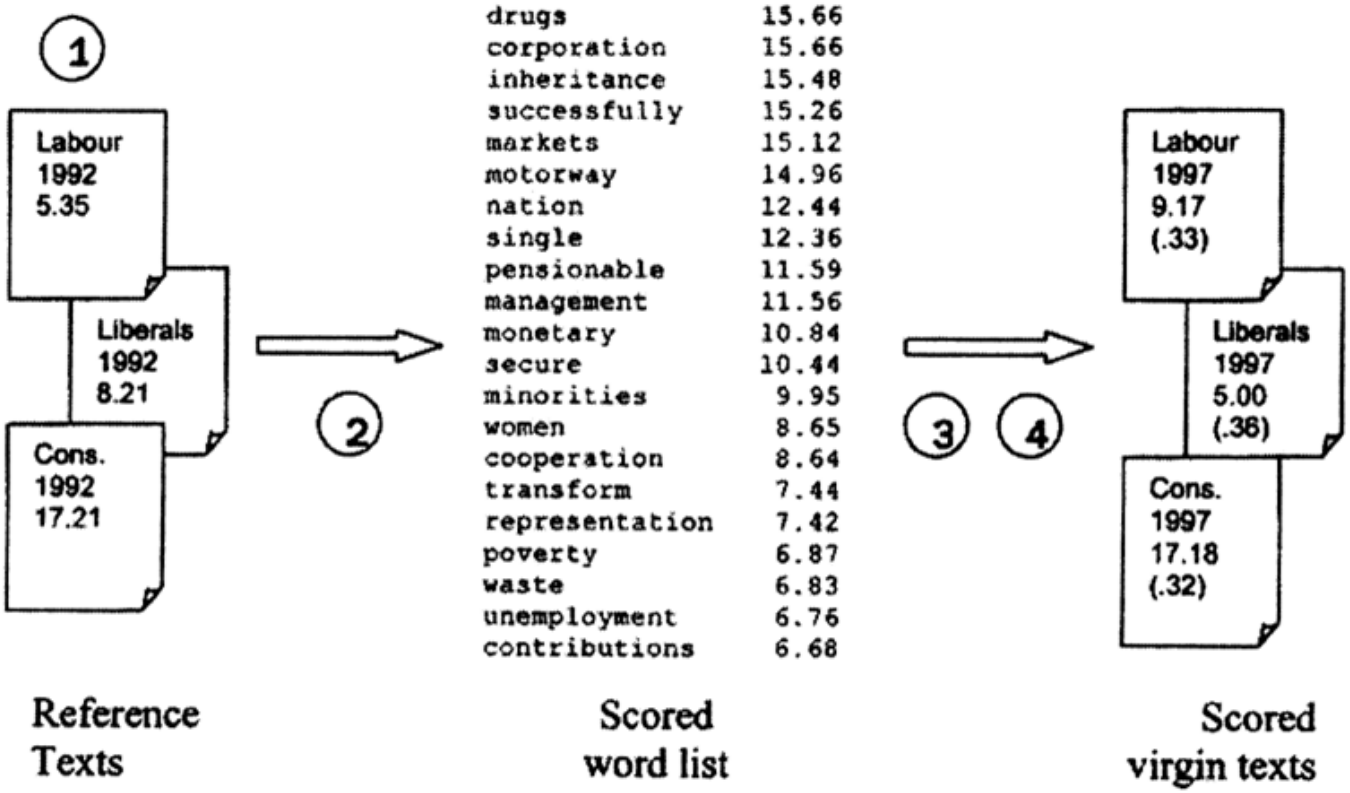
Wordscores: an example

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 along the economic policy dimension

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

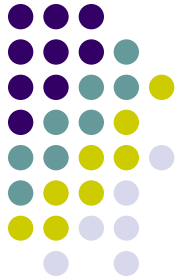
More formally…

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

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:

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

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 $S_{\mathrm{d|w}}$ 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:

$$S_{vd} = \sum_{w} \left(F_{wv} * S_{d|w} \right)$$

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

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:

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

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

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

Wordscores

The LBG (Laver-Benoit-Garry) transformation just shown **can be however problematic** everytime the number of virgin texts change in your analysis. Why?

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

To adjust the dispersion of the raw scores, the transformation relies in fact 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!!!

Wordscores

For example, suppose you use reference texts A and B to score virgin texts C and D

Suppose that the scores for A and B are 3 and 8, and the estimated raw scores for C and D are 5.2 and 5.5

If you want directly compare the raw scores for C and D to the original scores of A and B on the same metric, you need to rescale the raw scores using the previous formula. Let's call the rescaled scores for C and R, C* and D* respectively

Now, let's suppose that you add the virgin text E in the analysis

The raw scores for C and D will not be changed by adding the virgin text E

However, their rescaled scores YES, given that the number of virgin texts is changed, and therefore their standard deviation that affects the way you rescale the raw scores!!!

Wordscores

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

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

Alternatively, you can apply the transformation proposed in Marty and Vanberg (2008) – also implented in Quanteda

Wordscores

So what to do?

Possibile 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. And therefore **just scale** the virgin-texts!
2) Otherwise, compare raw scores to one another. In this case, it is a good idea to scale both the **virgin as well as the reference-texts**!

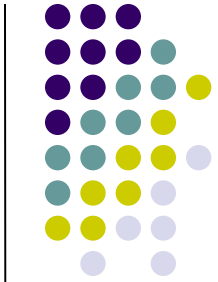
Wordscore: summing up

Wordscores does not make any assumption about words usage (contrary to unsupervised methods for scaling such as Wordfish!)

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!

Wordscore: summing up



Moreover, supervised scaling is robust to **irrelevant text in the virgin documents**

Reference texts that contain language about two extremes of environmental policy, for instance, are unlikely to contain words about health car

Scaling an unknown text using a model fitted to these environmental texts will therefore scale only the terms related to (and hence only the dimension of) environmental policy, even if the document being scaled contained out-of-domain text related to health care

Wordscore: summing up

Supervised scaling approaches have been shown capable of producing valid and robust scale estimates even with relatively small training corpora

The key in scaling applications is more one of the **quality of reference texts** than of their **quantity**

Which reference texts?

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

Three general guidelines in the selection of reference texts

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

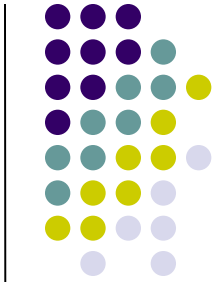
Which reference texts?

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

Which reference texts?



Third: two main conditions should be applied to the **features** included in the reference texts

1) the set of reference texts should contain as **many different words as possible** (i.e., they should include a sufficient range of potential word positions in the virgin texts)

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!

Which reference texts?

2) there should be **sufficient overlap** between distributions of words in the reference texts

Why?

Which reference texts?

It has been shown that Wordscores risks to generate biased word score estimates when there is insufficient overlap of word distributions across reference documents

This happens because **rare words** have always a huge influence in the word scores!

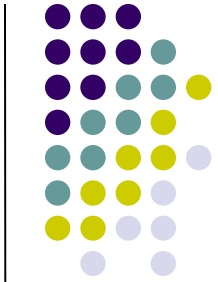
And when such **rare words** are not meaningful discriminators on substantive grounds, but they show up as influential because they only appear **once in the reference speeches**, the estimated probabilities for these words becomes unreliable while their (huge) influence is determined purely by estimation variability

Which reference texts?

Summing up: when using Wordscores alongside a good choice of reference texts (defined by the above conditions) estimates are generally less sensitive to differences in the meanings and uses of words

How to increase the probability of reaching such outcome given what just underlined?

Which reference texts?

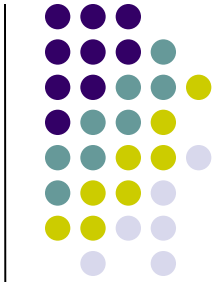


(a) Employ rather long reference texts…

(b) …with a reasonable amount of correlation among them (i.e., >.6)…

(c) …and drop all the **unique words** from the DfM (to ensure through that, that the words included in the reference texts are also included in the virgin texts - only the unique words in the reference texts of course matter, given that the unique words in the virgin texts are NOT scored by definition)!

Which reference texts?



Finally, note that with Wordscores, **dropping stopwords** is (very) relevant! Why?

Averaging word scores to estimate a document's score implies that each word adds the same amount of information about the document, that is, Wordscores **treats all words as equally informative**

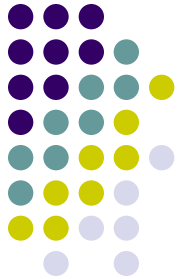
To give an example…

Which reference texts?

Words like ‘‘*taxes*’’ are informative about economic policy in a way that words like ‘‘*the*’’ are not. However, Wordscores has no way to represent the difference between a **genuinely informative politically centrist word** - one that is used preferentially by center parties to denote centrist policy positions - and a word that all documents contain in roughly equal numbers for functional linguistic rather than political reasons

The problem is that if document scores are spread evenly across a policy dimension, then centrist words and politically uninformative words will both have word scores close to the overall scoring mean

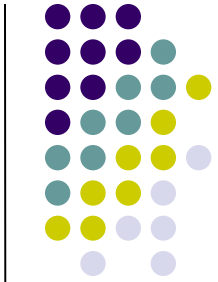
Wordscores and Wordfish



Note that the above suggestions are also important also for the reliability of Wordfish estimates!

As word use in documents becomes **more dissimilar**, any automatic scaling becomes less feasible!!!

Wordscores and Wordfish



Wordfish relies heavily on documents addressing the same concerns using the same words. If they do not, the algorithm is likely to pick up differences in the topics the authors address, not in their political positions. So **correlation in word use** is important!

For Wordscores, performance relies more on the reference texts being representative of the broader universe of texts in the corpus. As long as that is the case, differences in word frequencies matter less (although they are not irrelevant), but as they become less representative (e.g. because the number of unique words increase), performance of Wordscores decreases markedly

Wordscores challenges (2)

Regarding the length of the included documents Egerod and Klemmensen (2020) found that scaling corpora with fewer than approximately 1,800 words in the average text is infeasible using both algorithms

For Wordscores, however, corpora consisting of very short texts (below 400 words on average) can be scaled, if the reference documents provide good coverage of the virgin texts

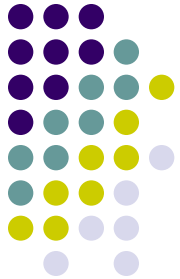
Which reference texts?

Finally…

In using particular reference texts, we are sometimes assuming that, for example, 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 (once again!) a **temporal dynamic challenge**

Which reference texts?



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

Which reference texts?

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

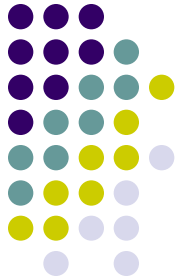
Imagine that in times t-1, the text uses the word “nigger” to identify Afro-Americans, while in time t the text uses the word “black” and at time t+1 the word “Afro-Americans”. Different words that refer to the same concept

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

Which reference texts?

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



IMPORTANT!!!

Before using rtweet

Open an R session and type the following commands. Plz let me know if you are able (or not) to download some tweets or not:

```
library(rtweet)

library(httpuv)


dt <- stream_tweets( "trump", timeout = 30)

print(dt$text[1:10])
```