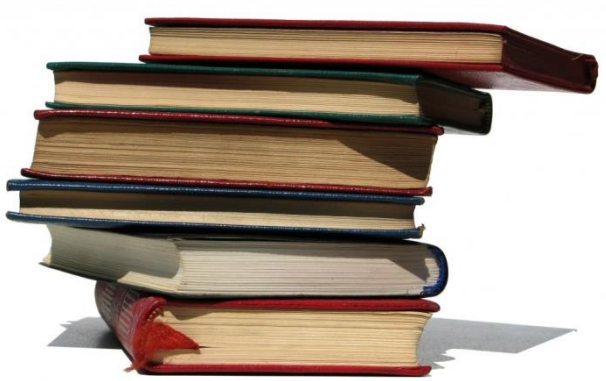
Big Data Analytics

Lecture 1/B

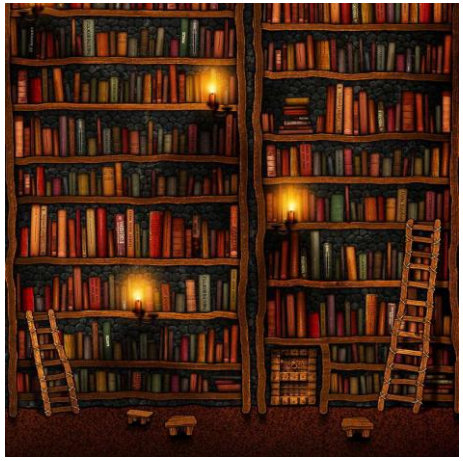
Supervised & Unsupervised scaling algorithms

UNIVERSITÄT LUZERN

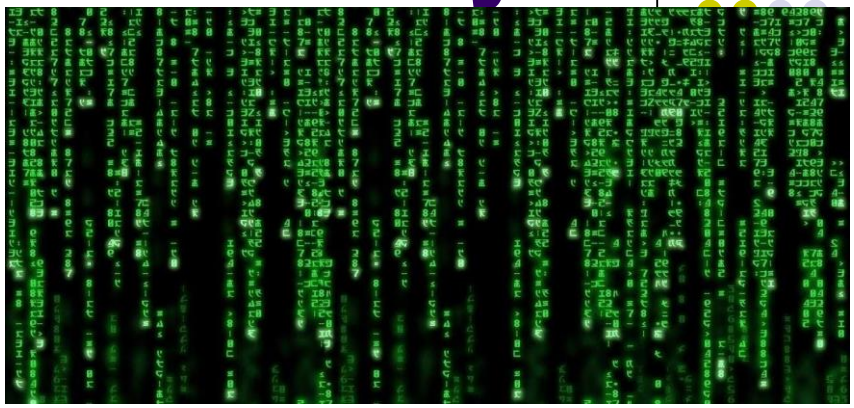
But before that…a summary



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



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



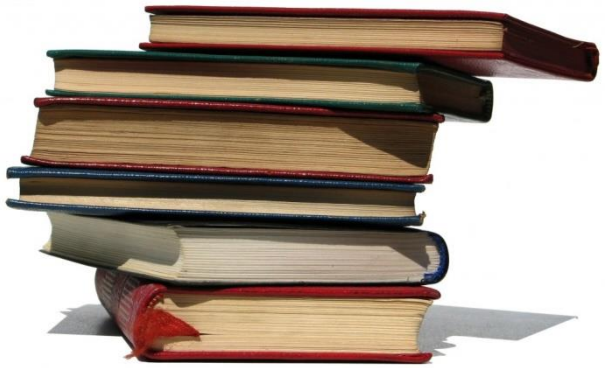
You extract from the corpus the relative document term (or feature) matrix via `dfm`. By doing that you apply the bag of words approach to that corpus of texts

But before that…a summary

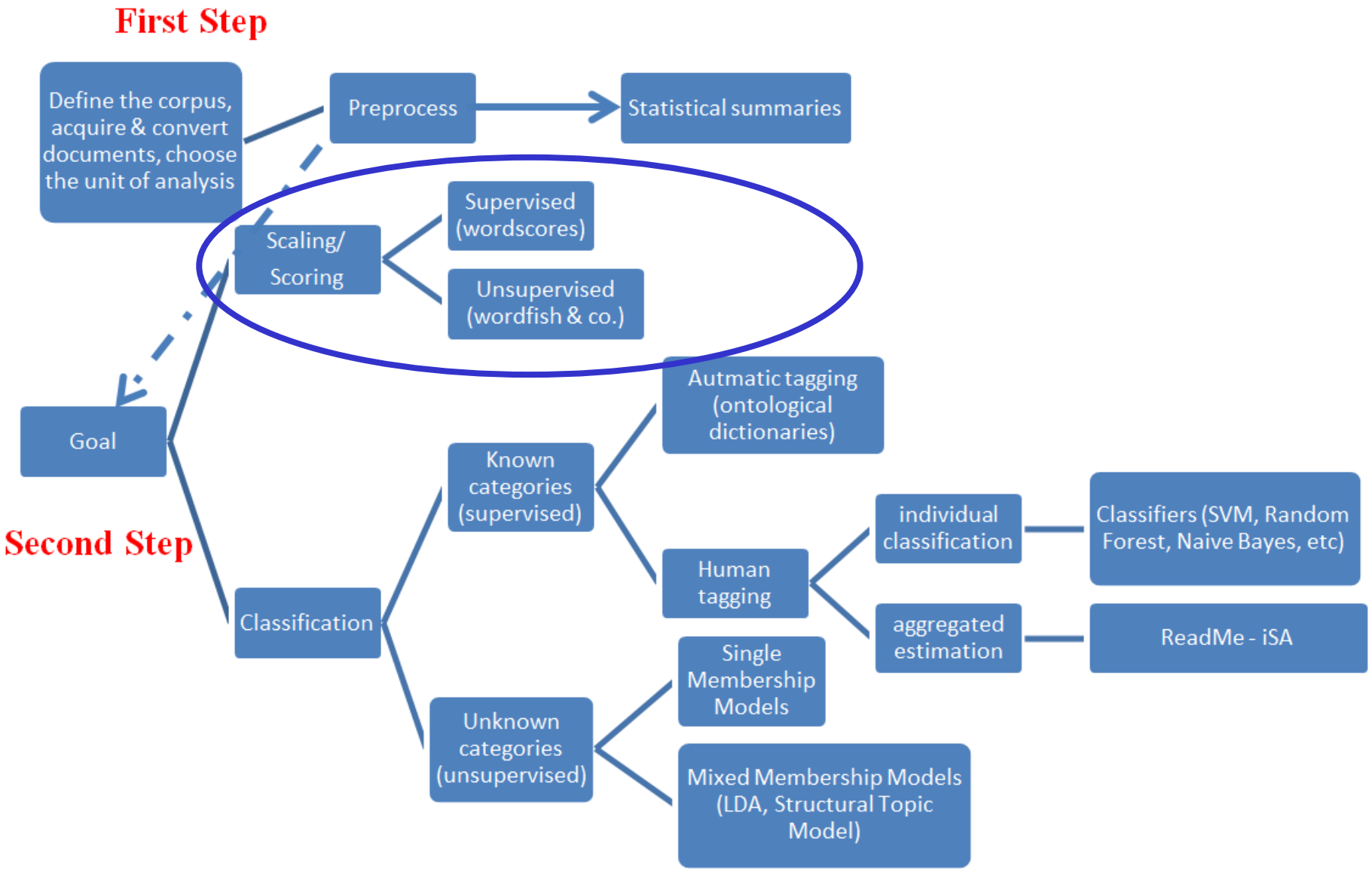
All the statistical models that we will see, **work on this**

docs	voto	programma	priorità	punti	piano	sostegno	famigli	natalità
FDI	1	5	4	5	7	11	4	1
FI	0	2	0	1	11	7	3	1
LEGA	1	9	5	6	18	32	7	3
LEU	0	2	1	0	15	7	3	0
M5S	6	18	13	12	45	20	22	0
PD	1	11	12	8	38	23	25	3

NOT on this



Our Course Map



References (supervised)

- ✓ 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

References (unsupervised)

- ✓ Proksch, Sven-Oliver, and Slapin, Jonathan B. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3): 705-722.
- ✓ Proksch, Sven-Oliver, and Slapin, Jonathan B. 2009. How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany. *German Politics*, 18(3): 323-344
- ✓ Curini, Luigi, Airo Hino, and Atsushi Osaki. 2020. Intensity of government–opposition divide as measured through legislative speeches and what we can learn from it. Analyses of Japanese parliamentary debates, 1953–2013. *Government and Opposition*, 55(2), 184-201

Latent models

Textual data might focus on **manifest characteristics** whose significance lies primarily in **how they were communicated** in the text

To take an example, if we were interested in whether a political speaker used **racist language**, this language **would be manifest directly in the text itself** in the form of racist terms or references, and what would matter is **whether they were used**, not so much **what they might represent**

Latent models

However, sometimes the target of concern is not so much **what the text contains**, but what **its contents reveal as data about the latent characteristics** for which the text provides ***observable implications***

Is this important? YES!

Latent models

Important theories about political and social actors concern qualities that are **unobservable through direct means**

Ideology, in particular, is fundamental to the study of political competition, but we have **no direct measurement instrument** for recording an individual or party's relative preference for liberal policies versus conservative ones

That is, ideology is not something that the researcher can **directly observe**…rather it must be indirectly estimated based (also) upon **observable actions** taken by actors

Scaling methods

The goal of methods **for scaling positions** is to use **some observed set of outcomes** to draw inferences about an actor's (in the widest sense of the word) unobservable position on a **latent dimension** *relative* to other actors

Position is here to be understood as a **preference on that dimension**. To get at such a position, the **observed outcomes** must reveal some kind of preference on the part of the actor

Observable set of outcomes…such as?

TEXTS of course!

Scaling methods

When scaling the political positions of a corpus of texts, we can view the **choice of words as the observed outcome**

Whenever **certain statements are associated with particular preferences**, we can use them to discriminate between positions as expressed in different documents along a certain continuous space (uni- or multi-dimensional)

In other words, the use of a particular (set of) word(s) provides us with **revealed preferences** that could be related to ideology, or to some other policy (or non-policy) space

A big advantage: nearly all political actors speak (or write)! And they used to speak (or write) also **long time ago**…

Types of scaling

Scaling methods can be differentiated between **Supervised & Unsupervised Methods**

What's the main difference? **Supervised Scaling Methods** (but this is true for all supervised methods!) require **a-priori information** by the researcher to produce estimates

Unsupervised Scaling Methods (but this is true for all unsupervised methods!) do not require that!

Let's start with the former ones, and let's discuss about **Wordscores**

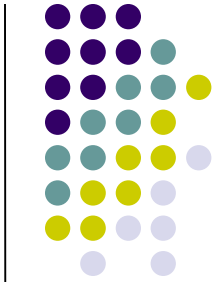
Wordscores

Wordscores technique estimates policy positions by **comparing two sets of texts**

On one hand we a set of texts ("**reference**" texts) whose policy positions on a well-defined *a-priori dimension* are "**known**" to the analyst, in the sense that these can be either estimated with confidence from independent sources or assumed uncontroversial (this is the human input required by the supervised algorithm!)

On the other hand we have 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 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 in Text A and 30 times 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?

S_{wd} = 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:

S_{wd} = 0.5*(3) + 0.5*(8) = 5.5

Wordscores

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

For a given 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

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 **estimated 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

In words: we use the **word scores we generated from the reference texts** to estimate the **positions of virgin texts** on the a-priori policy dimension 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

This procedure can thus be thought of as a type of **Bayesian reading of the virgin texts**, with the 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

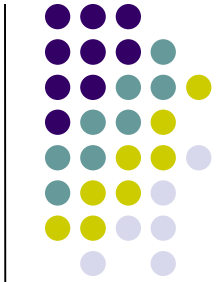
Wordscores

Of course, **only** the words included both in the reference texts **as well as** in the virgin texts are useful to compute S_{vd}!

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

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 (necessarily) shared frequently 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

Why is this important?

Cause 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

An example: you have two reference texts (A=3; and B=8)

There are 4 words included in the corpus of A+B, where *Government* and *Britain* appear much more often, and frequently, in both reference texts, compared to the *Choice* and *Crisis* (i.e., they are non-discriminating words)

1. *Choice*. It appears overall 40 times. 10 times in A and 30 times in B. As a result: $S_{choice,d}$ = 0.25*(3) + 0.75*(8) = 6.75
2. *Crisis*. It appears overall 40 times. 35 times in A and 5 times in B. As a result: $S_{crisis,d}$ = 0.875*(3) + 0.125*(8) = 3.616
3. *Government*. It appears overall 100 times. 50 times in A and 50 times in B. As a result: $S_{government,d}$ = 0.5*(3) + 0.5*(8) = 5.5
4. *Britain*. It appears overall 200 times. 110 times in A and 90 times in B. As a result: $S_{Britain,d}$ = 0.55*(3) + 0.45*(8) = 5.25

Wordscores

Then you have two virgin texts (C and D)

In text C, word *choice* appear 3 times, *government* 10 times and *Britain* 12 times. The total frequency of the words included in both text C *as well as* in the reference texts is therefore (3+10+12)=25. As a result:

S_{Cd} = (3/25)*6.75+(10/25)*5.5+(12/25)*5.25=5.53

In text D, word *crisis* appear 6 times, *government* 8 times and Britain 10 times. The total frequency of the words included in both text D *as well as* in the reference texts is 24. As a result:

S_{Dd} = (6/24)*3.616+(8/24)*5.5+(10/24)*5.25=4.92

The estimated scores for C and D are much clustered to each other than the original scores for A and B!

Wordscores

Because raw scores are dispersed **on a much smaller scale**, they cannot therefore 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/standardize** 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

Back to our example, the rescaled score for virgin text C becomes:

S^*_{Cd}=(5.53-5.23)*(3.54/0.43) +5.23=7.73

The rescaled score for virgin text D becomes:

S^*_{Dd}=(4.92-5.23)*(3.54/0.43) +5.23=2.73

Therefore: A=3, B=8, C=7.73, D=2.73

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 D, 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 **will be changed**, 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?

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)

Wordscores

Moreover, if we wish 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

Back to our example! Let's estimate the raw scores for reference texts A and B employing the $S_{d|w}$ (the dictionary) extracted from them!

In text A, word *choice* appear 10 times, *crisis* 35 times, *government* 50 times and *Britain* 110 times. The total frequency of the words included in both text A and reference texts=205. As a result:

$S_{Ad} =$ (10/205)*6.75+(35/205)*3.6161+(50/205)*5.5+(110/205)*5.25=5.11

In text B, word *choice* appear 30 times, *crisis* 5 times, *government* 50 times and *Britain* 90 times. The total frequency of the words included in both text B and reference texts=175. As a result:

$S_{Bd} =$ (30/175)*6.75+(5/175)*3.6161+(50/175)*5.5+(90/175)*5.25=5.54

Your final raw scores would be: A=5.11, B=5.54, C=5.53, D=4.92

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

Wordfish

Unsupervised methods for scaling texts produce estimates using **only the information available** in the textual data itself

How to do that?

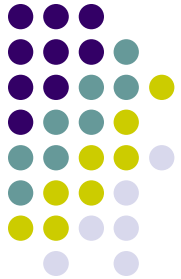
Let's introduce **Wordfish**!

Wordfish

Wordfish assumes that **relative word usage** within documents conveys information about their positions in some policy space

To give an example, this algorithm assumes that if one party uses the word '**freedom**' more frequently than the word '**equality**' in a document on economic policy while another party uses 'equality' more often than 'freedom' in a similar document, these **two words** – 'equality' and 'freedom' – **provide information** about party preferences with regard to an underlying policy dimension, and **discriminate** between the parties

Wordfish Estimation Process

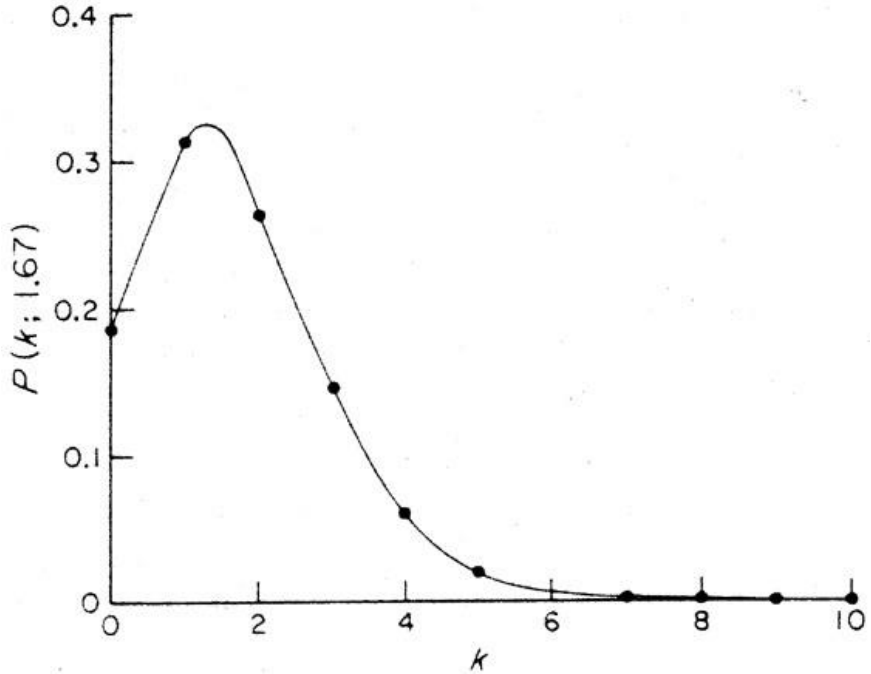


The *discover* of words that distinguish locations on a political spectrum is made possible by adopting some statistical assumptions on the **distribution of words** employed in texts

Wordfish Estimation Process

But which is the **statistical distribution** which most **accurately approximate word usage**?

Wordfish assumes that word frequencies (the number of times an actor *i* mentions word *j)* are generated by/drawn from a **Poisson process**, a distribution that is **heavily skewed**, as is the case of **word usage**



More formally

Formally, the functional form of the model is as follows:

$y_{ijt} \approx POISSON(\lambda_{ijt})$ where y_{ijt} is the **count** of word *j* in document *i*'s (i.e., party manifesto; speech; etc.) at time t

The lambda parameter has the following systematic component:

$$\lambda_{ijt} = exp(\alpha_{it} + \psi_j + \beta_j * \theta_{it})$$

Wordfish Estimation Process

The **systematic component** of this process contains 4 parameters: 1) *document fixed effects at time t* α; 2) *word fixed effects* Ψ (psi); 3) *word weights* β; 4) *document positions* θ at time *t* (theta)

The **document fixed effect** parameters control for the possibility that some documents in the analysis may be **significantly longer** than others

When using manifestos to estimate party positions, for example, this can happen when some parties in some years write much longer manifestos

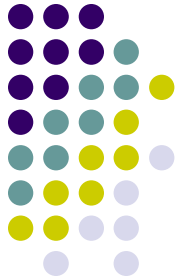
Wordfish Estimation Process

The **systematic component** of this process contains 4 parameters: 1) *document fixed effects at time t* α; 2) *word fixed effects* Ψ (psi); 3) *word weights* β; 4) *document positions* θ at time *t* (theta)

Word fixed effects are included to capture the fact that some words need to be used **much more often** in a language

Such words may serve a grammatical purpose but they have no substantive or ideological meaning, such as conjunctions or definite and indefinite articles

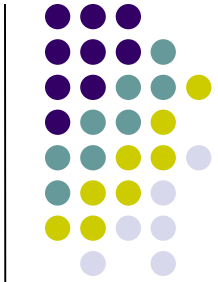
Wordfish Estimation Process



The **systematic component** of this process contains 4 parameters: 1) *document fixed effects at time t* α; 2) *word fixed effects* Ψ (psi); 3) *word weights* β; 4) *document positions* θ at time *t* (theta)

The **word discrimination parameters** allow the researcher to analyze **which words differentiate documents (party) positions**

Wordfish Estimation Process

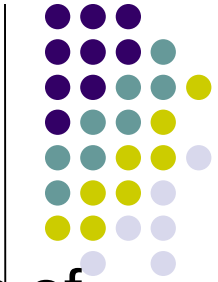


The **systematic component** of this process contains 4 parameters: 1) *document fixed effects at time t* α; 2) *word fixed effects* Ψ (psi); 3) *word weights* β; 4) *document positions* θ at time *t* (theta)

Finally, and *crucially*, the **document positions parameters** tells us the positions of each document relative to the other documents in the recovered latent space

This allows the researcher to estimate party positions and uncover the variations in political language that are responsible for placing parties on this latent dimension

Wordfish

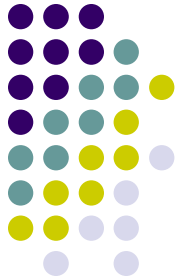


Note one important aspect: the substantial interpretation of the **estimated latent dimension** in Wordfish is completely left to the researcher

In the previous example, Wordfish **does not tell the researcher** whether ‘equality’ is a ‘left-wing word’ while ‘freedom’ is a ‘right-wing word’

The algorithm will simply use the relative frequencies of these words as data to locate the documents on a latent continuous scale, and it is up to the researcher to make an assessment about what constitutes ‘left’ and ‘right’ based upon her **knowledge of politics** (*a-posteriori* method!)

Wordfish Estimation Process



Let's see an example

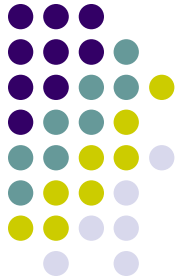
In Curini et al. (2018), we have selected all the speeches in which Japanese Prime Ministers make a general policy speech (*shoshin hyoumei enzetsu*) in the following situations:

i) after being nominated in the Special session

ii) after having succeeded a predecessor during a parliamentary session

iii) and in the beginning of the Extraordinary session

Wordfish Estimation Process



Overall 439 speeches over 82 sessions, and almost 20,000 words/kanji

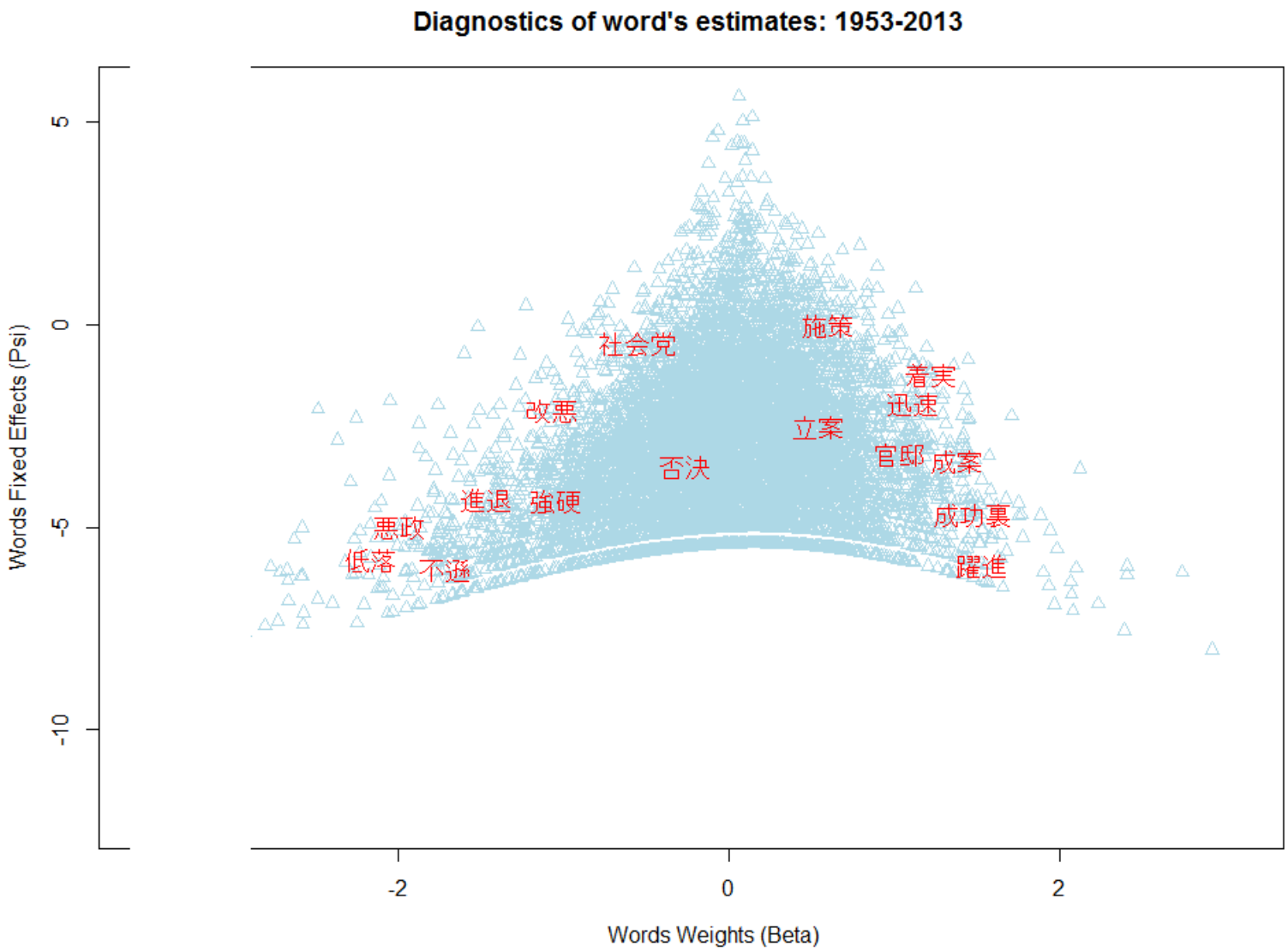
URL to get access to Japanese legislative speeches:

http://kokkai.ndl.go.jp/

Of course, we **tokenized** all the texts!!!

Our time range: 1953/2013 (pretty long period…more on this below…)

The discriminating words



The discriminating words

Positive betas: *breakthrough, successfully, bills passed, steady, prompt, policy measure, policy making*

Negative betas: *decline, misgovernment, arrogance, decision to leave from a position, deterioration, by force, rejecting bills*

What we have to do is therefore **linking** the discriminating words parameters β with the documents' position θ parameters to infer the substantial content of the latent dimension along which the documents are going to be scaled

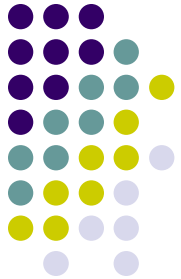
The discriminating words

In the example just saw, *bills passed* has a high absolute positive value for its discrimination value. Therefore, party's documents using that words with high frequency will receive a positive score along the latent dimension (cabinet parties?)

The word *rejecting bills* would also have a high absolute value **but with the opposite (negative) sign**. Therefore, party's documents using that words with high frequency will receive a negative score along the latent dimension (opposition parties?)

Therefore the latent dimension is a *opposition-cabinet one*?

More formally



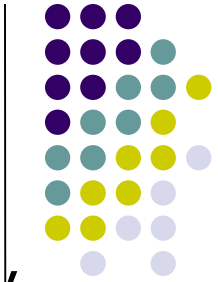
WORDFISH uses an **expectation maximization (EM) algorithm** to retrieve maximum likelihood estimates for all parameters

The implementation of this algorithm entails an **iterative process**:

first *document parameters* are held fixed at a certain value while *word parameters* are estimated, **then** *word parameters* are held fixed at their new values while the *document parameters* are estimated

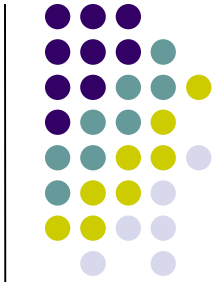
This process is **repeated until the parameter estimates** reach an acceptable level of convergence

Some challenges of scaling



1. A-priori assumptions (to be satisfied) to meaningfully scale a corpus
2. Document selection
3. Dynamic estimation

A-priori assumptions



First, if the cost to articulate a position is low, authors' might engage in *cheap talk*

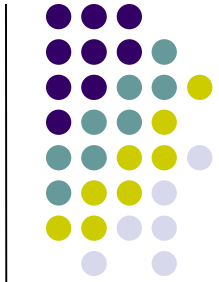
Conversely, if costs are high, they might choose not to articulate the position for *strategic reasons*

All the scaling techniques we focus on, assume on the contrary that authors do **not censor their statements for political reasons**

This assumption, in some given circumstances, could however cause significant measurement error

A-priori assumptions

A-priori assumptions

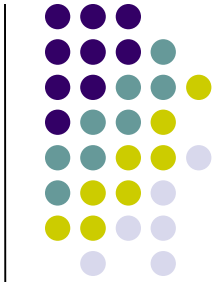


Second, the documents **should be informative about the differences** we seek to estimate

Particularly in contexts where there are **strong common norms about how to phrase a documen**t (as with highly technical legislative or legal documents) or the texts do not communicate any preference at all, it can be difficult to scale documents

If authors of different preferences use similar choices of words, we cannot in fact use the texts to discriminate between their positions

Document Selection



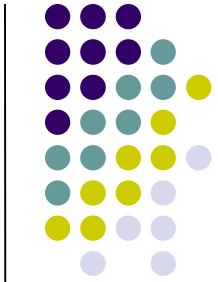
Document selection is essential and possibly the most tricky task in the estimation process both in Wordscores as well as in Wordfish

Let's start with Wordfish

Wordfish estimates a **single dimension**, and the information contained in this **dimension depends only upon the texts** that the researcher chooses to analyze (w/o any a-priori human contribution contrary to Wordscores)

Therefore, the **selection of texts should depend** on the particular dimension the researcher would wish to examine

Document Selection



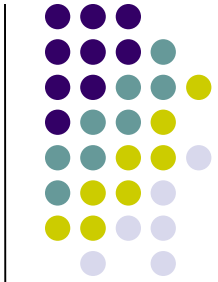
Quite often analyses dealing with political texts rely on the strong assumption of **ideological dominance in speech** (i.e., that actors' ideological leanings determine what is discussed in texts)…sometimes this makes sense, other times no!

This is **not a shortcoming** of Wordfish!

This simply suggests that one **should not blindly assume** that Wordfish output measures an ideological location of political actors without careful validation

In the previous example about Japan, we actually capture an opposition-cabinet latent dimension!

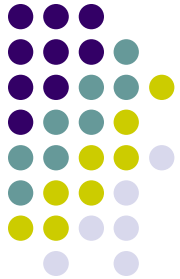
Document Selection



For instance, if a researcher is interested in comparing **foreign policy statements** of parties in country X, then only such texts should be included in the analysis

On the other hand, if the research question is to determine a **general ideological position** using all aspects of policy, then the analysis should perhaps be conducted using all parts of an election manifesto, for example, assuming that such documents are encyclopedic statements of policy positions

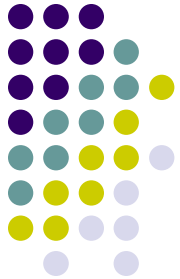
Document Selection



The estimated single latent dimension **will thus be a function** of the selection of the text corpus

This also implies that when the generative model specifies a unidimensional space, when it really is *multidimensional*, we risk miss-specifying the dimension we extract! **Why?**

Document Selection



Wordfish will recognize differences in word use between two texts as indicative of their different positions

These differences could be however also due to the topics addressed by the authors , i.e., situations where texts do not address **similar topics at all**

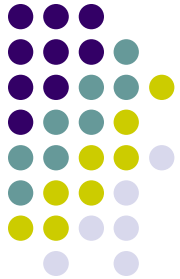
In these situations texts cannot be reasonably scaled together, and if they are, it will often result in the main latent dimension being grossly miss-specified

Document Selection

For example, if you have a set of texts discussing about K-pop and a set of texts discussing about Japanese politics, and you scale them together…

…you will obtain a latent scale that will differentiate between K-pop texts on one extreme of the latent dimension and texts discussing about Japanese politics on the other extreme. What's the utility of that?

Document Selection



WORDFISH does not estimate **multiple dimensions**, but it does allow the estimation of **different dimensions** if you use different text sources

For instance, if your interest is in estimating positions of **presidential candidates on foreign policy and economic policy**, then you could estimate separate positions using foreign policy speeches only on the one hand and economic policy speeches on the other hand and from this creating a **2-dimensional space**

Document Selection

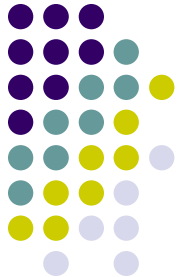
Let's now move to **Wordscores**

Wordscores does not make any assumption about words usage (contrary to Wordfish)

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

A **nice property** of using Wordscores in this sense 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!

Document Selection



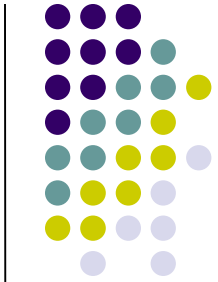
Moreover, supervised scaling is robust to **irrelevant text in the virgin documents**

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

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

This is a big advantage with respect to Wordfish

Document Selection



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

Three general guidelines in the selection of reference texts

Document Selection

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

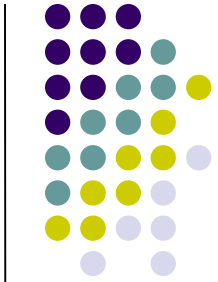
Document Selection

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

Document Selection



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). Why that?

Cause 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!

Document Selection

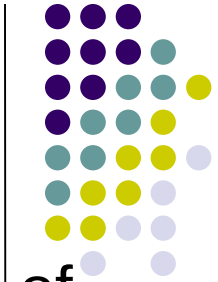
2) there should be **sufficient overlap** between distributions of words in the reference texts

Why?

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

Document Selection



Summing up: use Wordscores alongside a good choice of reference texts (defined by the above conditions)

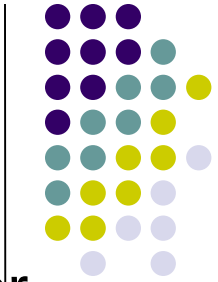
Therefore…

(a) Employ not short-reference texts (i.e., do not use Worscores on tweets, unless…)

(b) …with a reasonable amount of correlation among them (i.e., an *overall average* >.6 is a good value)…

(c) …and drop all the **unique words** from the DfM (to ensure 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)!

Document Selection



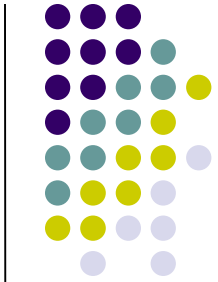
Note that the first point above is also important (very) for Wordfish!

With respect to point (a): Egerod and Klemmensen (2020) found that scaling corpora with fewer than approximately 1,800 words in the average text is infeasible using both Wordfish and Wordscores

So, using Wordfish to scale for example tweets (i.e., very short texts) is not a great idea…

For Wordscores 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

Dynamic Estimation



Using texts to estimate policiy positions **over time** creates an additional challenge

On the one hand, we would like to use as much information in the texts as possible. On the other hand, we would like to estimate position change over time

But there is a **trade-off** here both for unsupervised as well as supervised approaches

Let’s start with the former ones

Dynamic Estimation

For example, if **public debate changes and new vocabulary** enters the public lexicon at time *t*, then this fact per-se (i.e., the change in the vocabulary) will differentiate texts at point *t* from those at point *t-1* irrespective (or above of) any “true” change in in party (or politician) positions along the same latent dimension!

Dynamic Estimation

Take as an example the set of parties' manifestos in Germany since 1970 to 2005. Assume that you want to analyze such documents with Wordfish

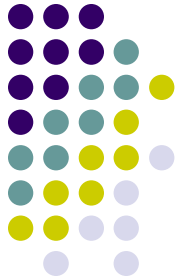
Now assume that the political lexicon in the manifestos at election time *t* contains an issue (and a vocabulary) that is no longer relevant at time *t+1*, e.g. official relations with the GDR (East Germany)

If all parties make a statement with regard to the GDR at point *t* but not at *t+1*, then the words **will not only distinguish** parties at point *t*, but also **distinguish the elections**

As a result, if all words are counted, even the rare ones, the parties are more likely to be **clustered by election**


German Party Position Estimates, 1969-2005
(Dataset A: all words)
Wordfish Position
1
0
-1
-2
FDP
Grüne
Linke
SPD
Union
1970
1975
1980
1985
1990
1995
2000
2005
Year
41,684 unique words, 44 documents.

Dynamic Estimation



Which are the potential route to addressing this issue? We could carefully select the **words** that enter the analysis!!!

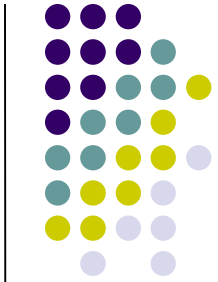
Thus, if there is movement of parties, it can only be due to **different word usage**

This requires that the **word data over time** must be comparable at a minimum level

Which word inclusion criteria then?

Two (main) options

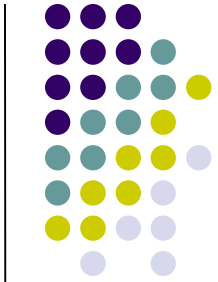
Dynamic Estimation



First alternative (non-informative priors):

- ✓ in the DfM includes words that are **mentioned in a minimum number of documents** (say, in at least 20%), thus essentially keeping words that are deemed important enough to be mentioned either over time by one party or by several parties

Dynamic Estimation

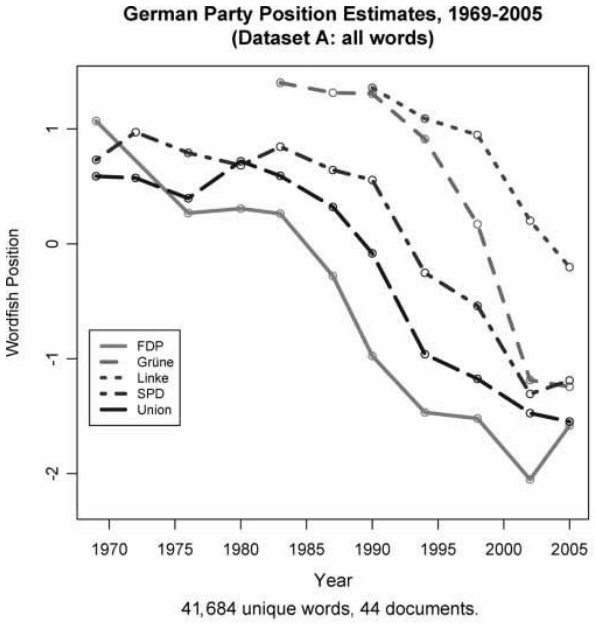


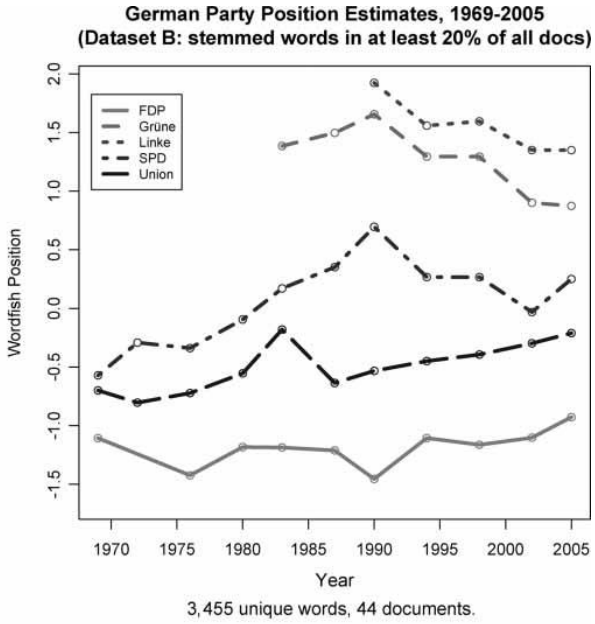
Second alternative (informative priors)

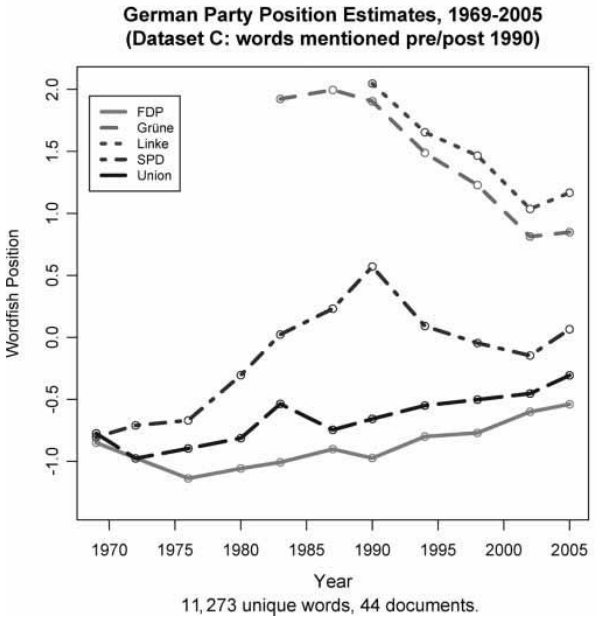
- ✓ in the DfM includes **only those words that appear both pre- and post-1990**, i.e., reunification added words to the German political lexicon that were not in it previously. Likewise, some words that were previously important likely fell out of use

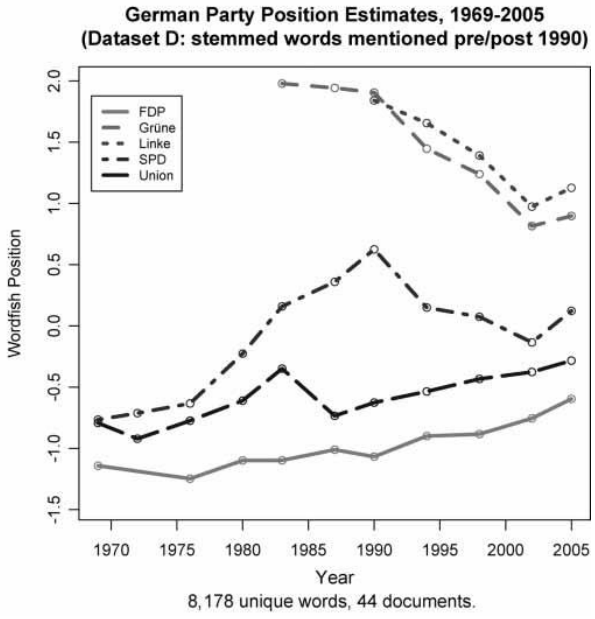
If we do not control for this fact, we would see a **large jump** in all parties around 1990 as they all shift their word usage to account for new political realities

And indeed…

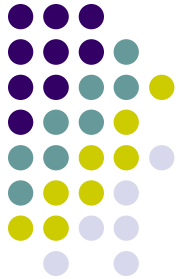

German Party Position Estimates, 1969-2005
(Dataset A: all words)
Wordfish Position
Year
FDP
Grüne
Linke
SPD
Union
41,684 unique words, 44 documents.


German Party Position Estimates, 1969-2005
(Dataset B: stemmed words in at least 20% of all docs)
Wordfish Position
Year
FDP
Grüne
Linke
SPD
Union
3,455 unique words, 44 documents.


German Party Position Estimates, 1969-2005
(Dataset C: words mentioned pre/post 1990)
Wordfish Position
Year
FDP
Grüne
Linke
SPD
Union
11,273 unique words, 44 documents.


German Party Position Estimates, 1969-2005
(Dataset D: stemmed words mentioned pre/post 1990)
Wordfish Position
Year
FDP
Grüne
Linke
SPD
Union
8,178 unique words, 44 documents.

Dynamic Estimation



Time (and the possibility of a vocabulary changes) creates problems also for supervised scaling algorithms

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

In using particular reference texts, we are in fact 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…however this shouldn't be always the case if words change their political/social associations over time

Dynamic Estimation

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

Dynamic Estimation

Three possible answers 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. Of course this makes sense for one feature; for many features isan infeasible task!
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
3) you follow the paths already discussed with Wordfish