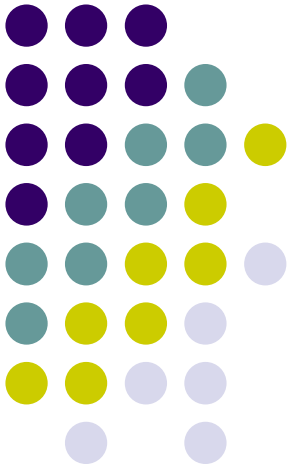
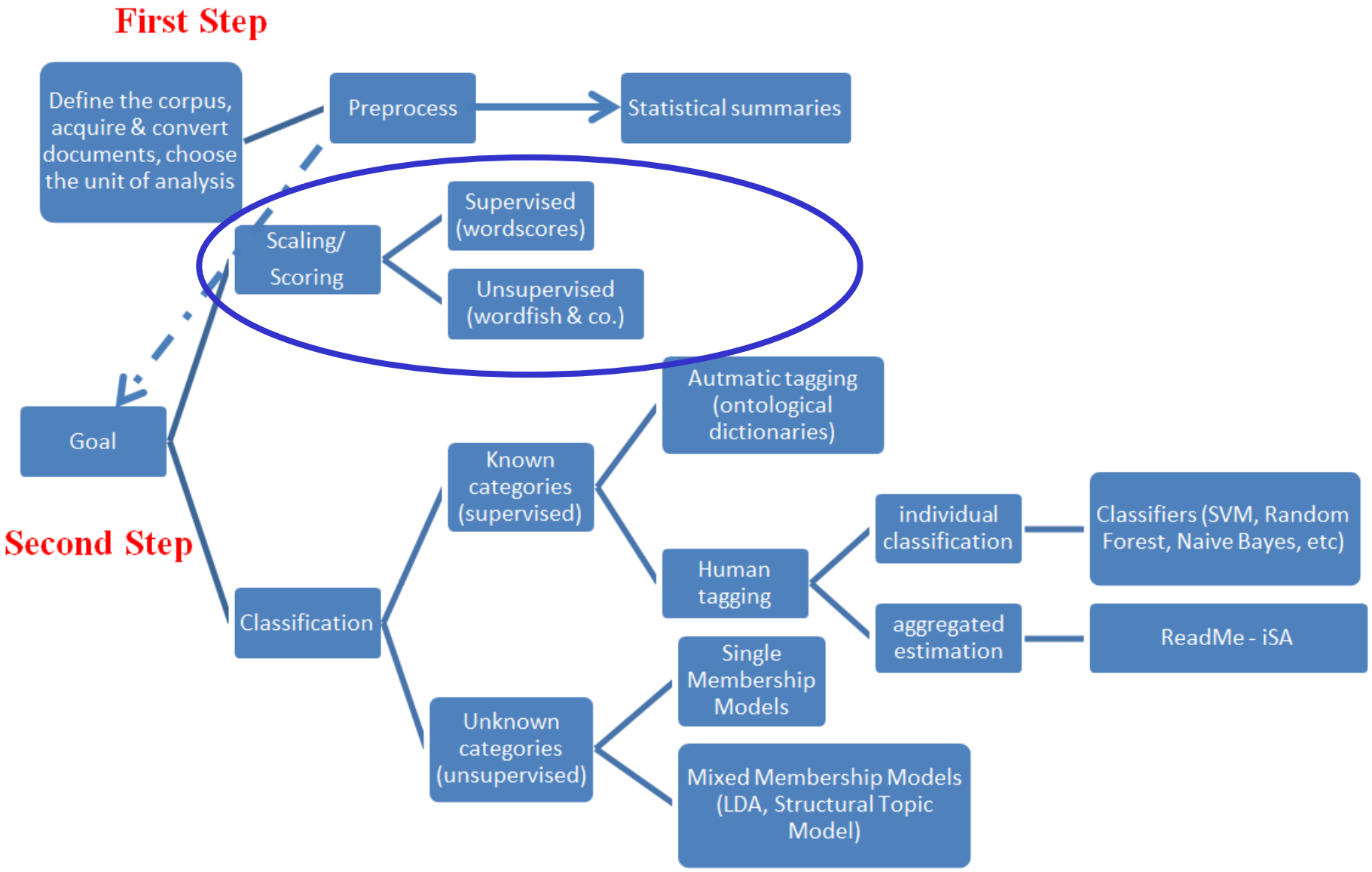
Big Data Analytics

Lecture 2

Supervised scaling algorithms



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

Types of scaling

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

What's the main difference? Remember!

Unsupervised Scaling Methods (but this is true for all unsupervised methods!) do not require **a-priori information** by the researcher to produce estimates

Supervised Scaling Methods (but this is true for all supervised methods!) do require such **a-priori information**

Let's move to the latter methods, and let's discuss about **Wordscores**

Wordscores

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

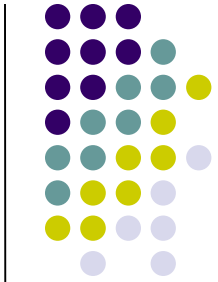
On one hand we a set of texts ("**reference**" texts) whose policy positions on a well-defined *a-priori dimension* are "**known**" to the analyst, in the sense that these can be either estimated with confidence from independent sources or assumed uncontroversial (this is the human input required by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

Wordscores

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

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

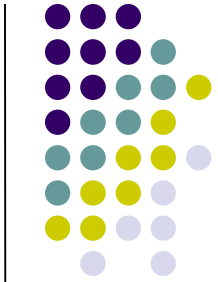
Wordscores

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

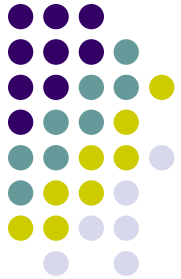
Wordscores

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**!

Some challenges of doing supervied scaling



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

A-priori assumptions

Once again, we do not want that authors **censor their statements for any reason**

Plus, the documents **should be informative about the differences** we seek to estimate (as we already discussed with Wordfish)

Document Selection

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**

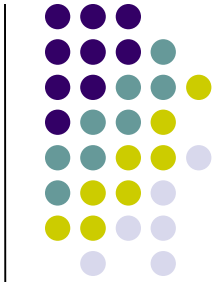
Reference texts that contain language about two extremes of environmental policy, for instance, are unlikely to contain words about health-care

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

Moreover, and by definition, you do not have any issue in interpreting the final results (no a-posteriori judgements!)

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

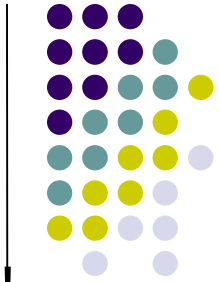
Third, the set of reference texts should contain as **many different words as possible** (i.e., they should include a sufficient range of potential word positions in the virgin texts). Why that?

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



Note that this point is also important (very) for Wordfish!

However, for Wordscores corpora consisting of relatively short texts (below 400 words on average) can still be scaled, if the reference documents provide good coverage of the virgin texts

Document Selection

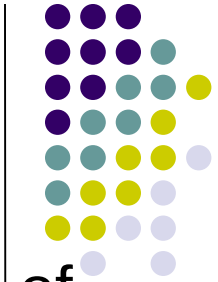
Finally, there should be **sufficient overlap** between distributions of words in the reference texts

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)

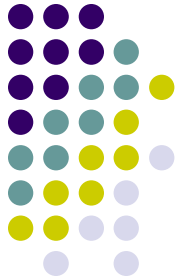
Therefore…

(a) Employ not short-reference texts

(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)!

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

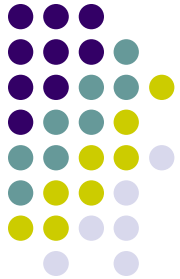
In this case, however, all the information related to “black” and “Afro-Americans” will be lost

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 is an 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

Further unsupervised scaling algorithms

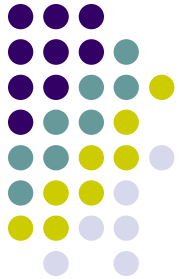


Correspondence Analysis (old stuff, good stuff!)

As for a FA, the idea in CA is to reduce the dimensionality of a data matrix and visualize it in a subspace of low-dimensionality

The data of interest in simple CA are usually a two-way contingency table for which *relative* (not absolute) values are of primary interest

Further unsupervised scaling algorithms

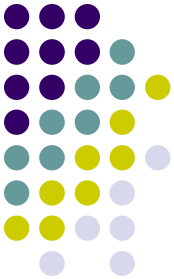


Correspondence Analysis (old stuff, good stuff!)

Lowe (2008; 2016) shows that CA provides an approximation to a Poisson ideal point model for text data (i.e., Wordfish!)

In most applications it does **not make much difference** which model is used; however, it has been found that Wordfish is **more robust** when a single document is **very different** than the others, which happens not infrequently in political documents

Further unsupervised scaling algorithms



Correspondence Analysis (old stuff, good stuff!)

The advantages of CA:

✓ You can easily estimate a 2-dimensional world

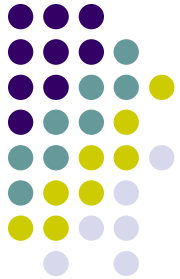
The limits of CA:

✓ no uncertainty estimation

✓ validating the latent space extracted it more tricky than Wordfish

Quanteda command: `textmodel_ca`

Further unsupervised scaling algorithms

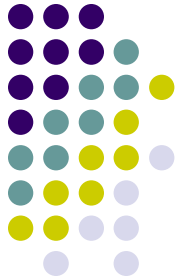


A further possible unsupervised scaling algorithm: the Poisson Reduced Rank Model

- ✓ Poisson Reduced Rank Model (see: C. Jentsch, E. R. Lee and E. Mammen (2020) Time-dependent Poisson reduced rank models for political text data analysis. *Computational Statistics and Data Analysis*, 142, 106813; C. Jentsch, E. Mammen and E. R. Lee (2021) Poisson reduced rank models with an application to political text data. *Biometrika*, 108, 2, 455 – 468)
- ✓ Take a look at here: https://github.com/chroetz/poisrrr
- ✓ To download the package:

```
remotes::install_github("chroetz/poisrrr")
```

Other scaling algorithms



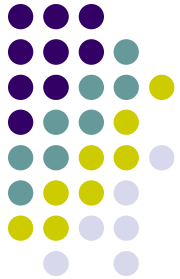
Class affinity (an extension of Wordscores)

Class affinity is attractive every time you have a few examples of documents at the extremes of a hypothesized ideological or stylistic spectrum and you want to estimate the probability (the degree of similarity) of your set of documents/texts to belong to one out of two categories (0/1, Government/Opposition, etc.)

Perry, P.O. & Benoit, K.R. (2017). Scaling Text with the Class Affinity Model. arXiv:1710.08963 [stat.ML]

Quanteda command: `textmodel_affinity`

Other scaling algorithms

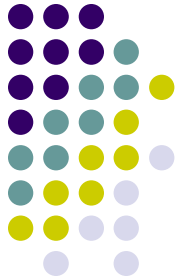


Class affinity (an extension of Wordscores)

The basic conceptual behind a class affinity model:

- ✓ Over the course of a speech, for example, a speaker orientation **switches back and forth** between Government mode and Opposition mode
- ✓ When she is in Government mode, she chooses words in the same manner as the government leadership
- ✓ Likewise, when she is Opposition mode, she chooses words in the same manner as the opposition leadership

Other scaling algorithms



Class affinity (an extension of Wordscores)

- ✓ We should therefore place the speaker on the spectrum between the two extremes of pro-government and pro-opposition according to what **proportion of time she spends in each mode**
- ✓ In the class affinity framework, this learning step requires not large volumes of training data, but rather texts that are clearly polar examples of each reference class (also **more than ones at each extreme**), to form benchmarks for estimating the other texts' affinities to these classes