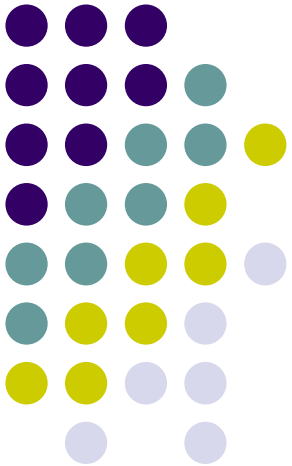
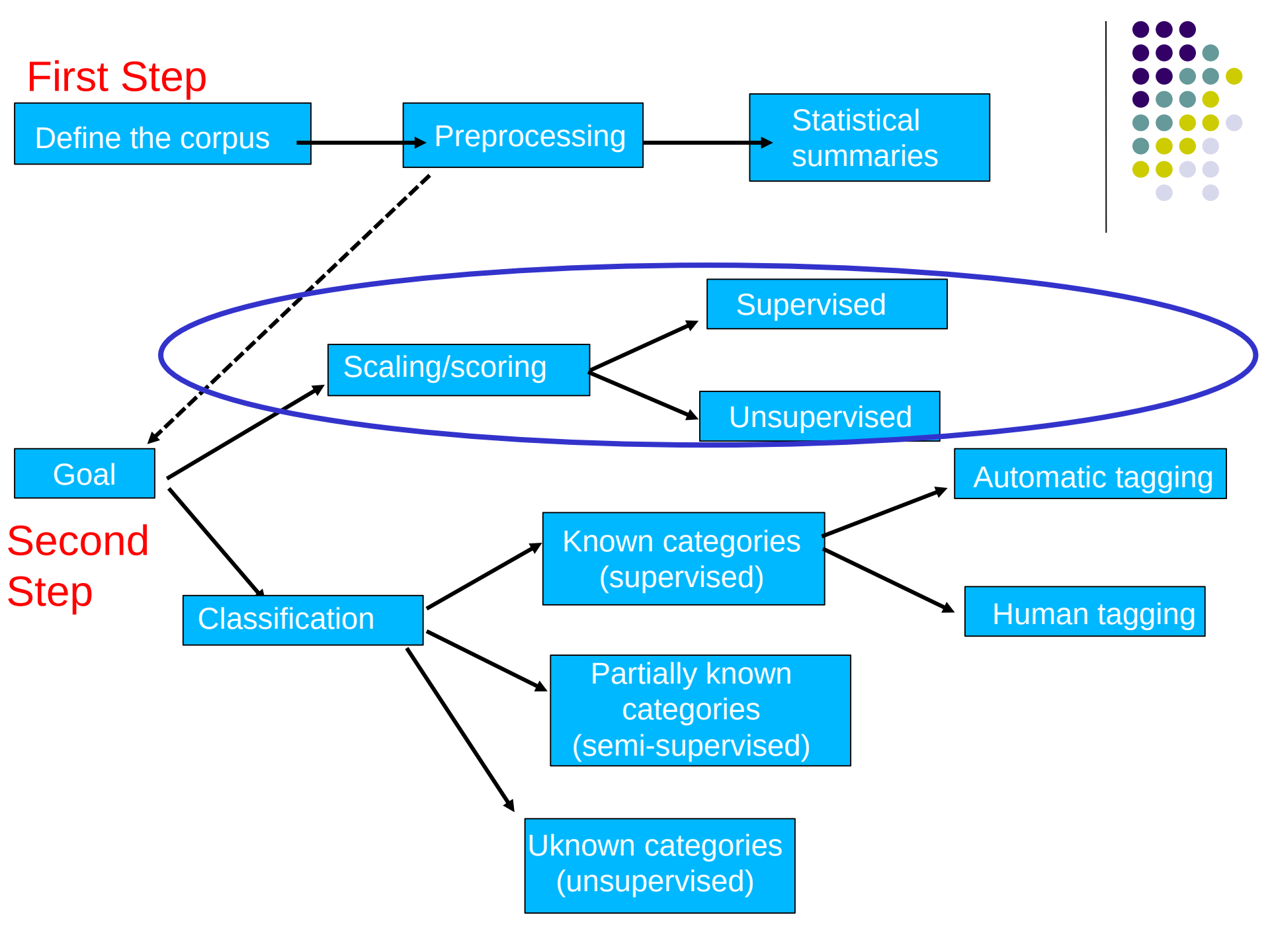
Big Data Analytics

Lecture 2

Supervised scaling algorithms




First Step
Define the corpus
Preprocessing
Statistical summaries
Supervised
Scaling/scoring
Unsupervised
Goal
Automatic tagging
Second Step
Known categories (supervised)
Classification
Human tagging
Partially known categories (semi-supervised)
Uknown categories (unsupervised)

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
- ✓ Martin, Lanny W., and Georg Vanberg. 2008. A robust transformation procedure for interpreting political text. *Political Analysis*, 16: 93-100
- ✓ Bräuninger Thoams and Nathalier Giger. Strategic Ambiguity of Party Positions in Multi-Party Competition, *Political Science Research and Methods*, 6(3), 527-548, 2018

Scaling 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**

This is something you can retrieve for example from your corpus via the `kwic` command

Scaling 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 only ***observable implications***

Is this important? YES!

Scaling models

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

Ideology, for example, 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

When we deal with a corpus of texts, we can view the **choice of words as the observed outcome**

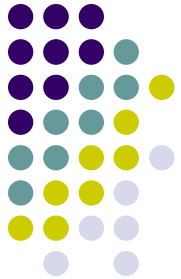
As long as **certain statements are associated with particular preferences**, we can therefore use them to discriminate between positions as expressed in different texts

Accordingly, the goal of **scaling methods** is to use **some observed set of outcomes** (words in our case) to draw inferences about an actor's unobservable position on a **latent dimension(s)** *relative* to other actors

Position is here to be understood as a **preference on that dimension**

Such preferences, as we will discuss, could be related to ideology, or to some other policy (or non-policy) space

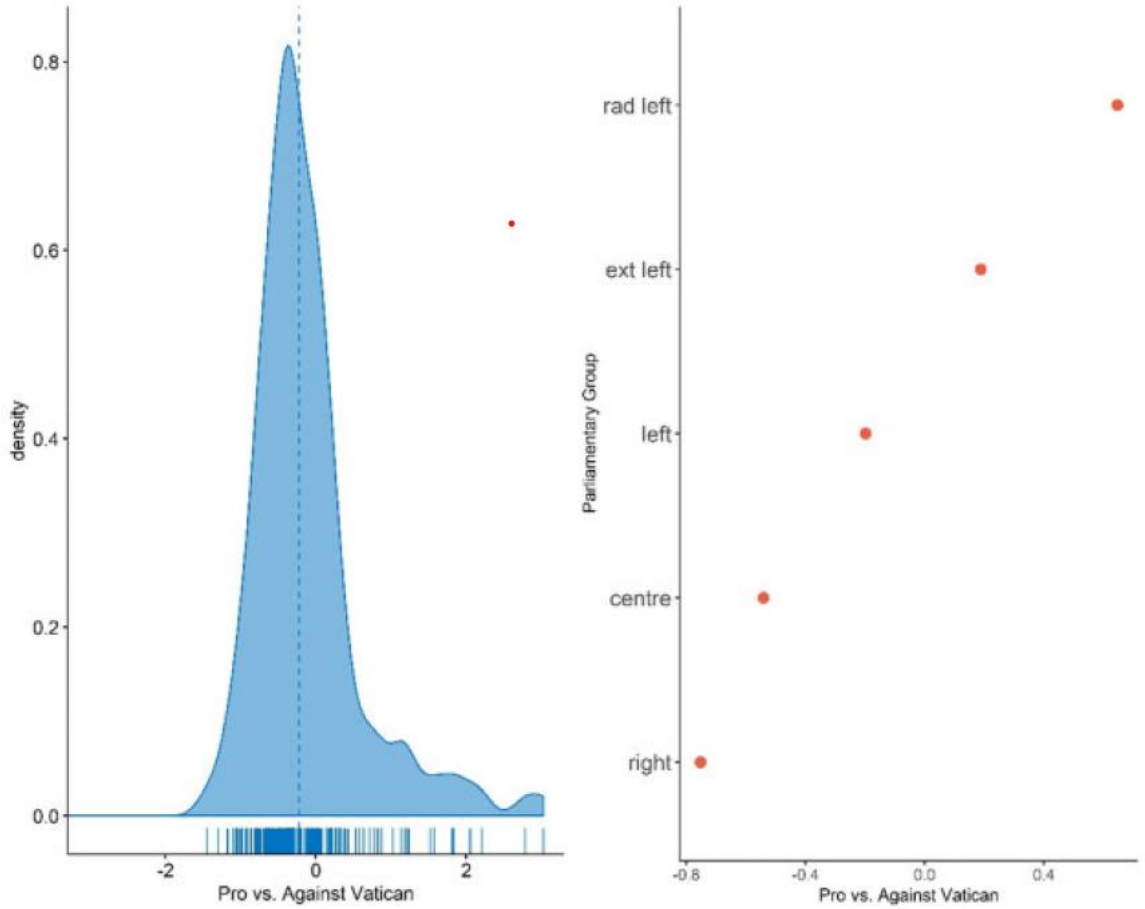
Scaling methods



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

Scaling methods

An example: Italian MPs positions between 1861 and 1870 along a pro vs against Vatican latent dimension (Casiraghi and Curini 2021)



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 some kind of **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 document positions by **comparing two sets of texts**

On one hand we a set of texts ("**reference**" texts) whose 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!)

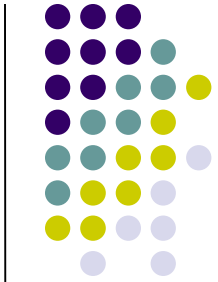
The source of such document positions can derive for example from either expert or mass-surveys if we are dealing with party manifesto (or from your own knowledge if we are dealing with some other texts: i.e., texts that appreciate or not K-pop)

Wordscores

On the other hand we have a set of texts whose 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" positions

Wordscores



More formally…

R = set of reference texts

We assume that we know with confidence the 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 100 words in Text A and 30 times per 100 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 (0.1/(0.1+0.3)); 0.75 probability that we are reading Text B (0.3/(0.1+0.3))

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 $S_{d|w}$ 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

We call “choice” in this case a “non-discriminating word” (a word that does not allow to *discriminate* among reference texts)

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

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

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

Wordscores

Scoring Virgin Texts

First, we must compute the relative frequency of each **virgin text word** that appears **also** in the reference texts, as a proportion of the total number of words in the virgin text that appears *both* in the virgin text and in the reference texts. 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 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 position, and **each scored word** in a virgin text adds to this information

Wordscores

The main **assumption** behind Wordscores is therefore that the **relative frequencies of word usage** in the virgin texts are linked to 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

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

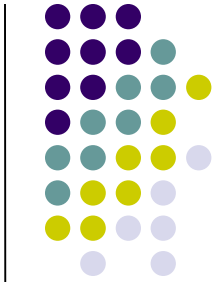
Note that by definition, you do not have any issue in **interpreting the final results**: the virgin texts are going to be scaled along the **a-priori dimension** defined by the researcher (and to which the scores for the reference texts are referring to)

Wordscores

Understanding the logic employed by Wordscores is very important because basically is the same one that is used in other algorithms (such as Machine Learning ones, semi-supervised algorithms, etc.)

If you want to better understand how Wordscores works, take a look at the EXTRA script provided in the home-page of the course

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 position

When these predictions are tightly clustered, we are of course **more confident** in their consensus than when they are scattered more widely

Wordscores

As with any variance, we can use the square root of V_{vd} to produce a standard deviation

This standard deviation can then 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

Note that this measure of uncertainty can reveal you a lot of substantial information per-se!

For example about…**strategic ambiguity**!

Bräuninger and Giger (2016) for example show how extreme parties have incentives to use blurred messages toward different audiences and therefore presenting ambiguous rather than clear-cut policy platforms to attract votes from different groups

How to measure such “blurred messages”? By focusing on the dispersion of wordscores within a party manifesto!

More formally, they use the standard deviation of the word positions as a measure for the level of positional ambiguity

Wordscores

Remember that 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, is:

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

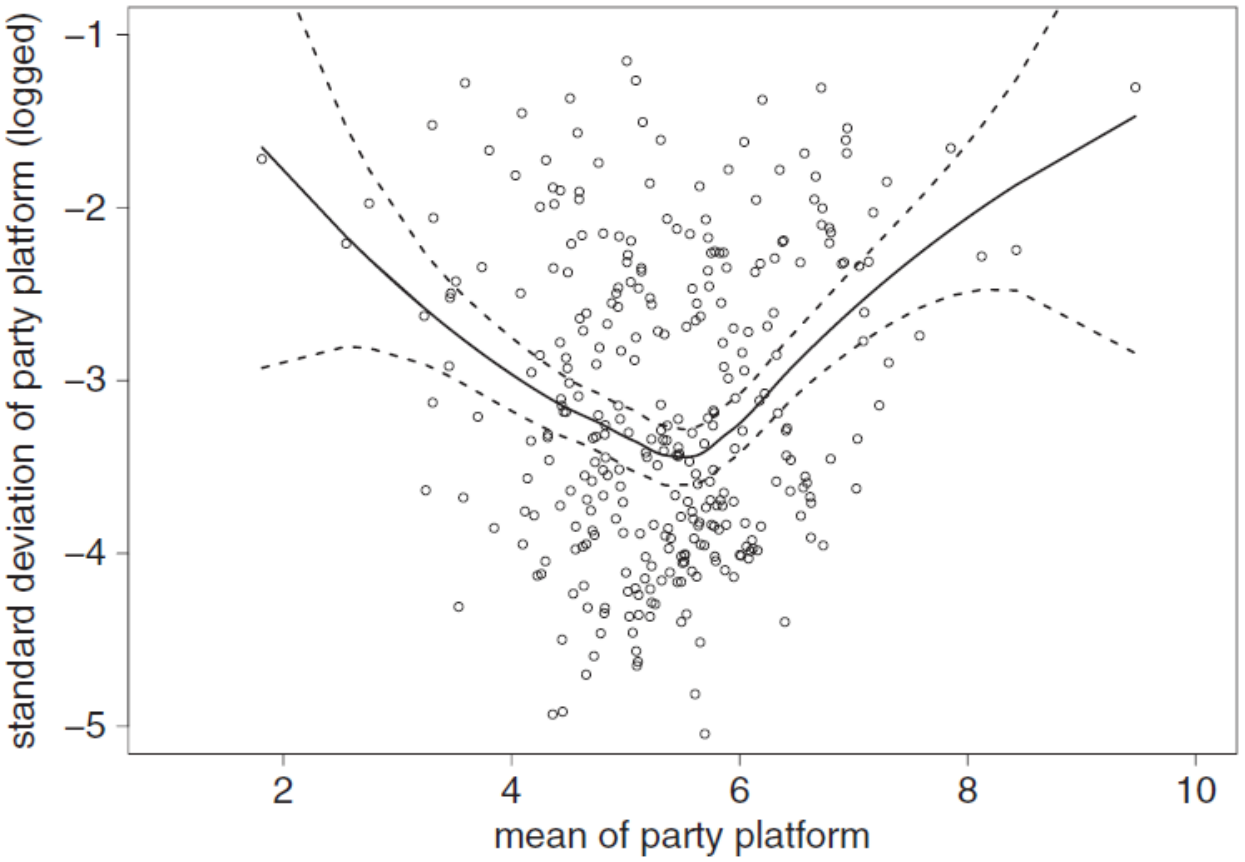
Therefore, σ_{amb} (i.e., the "ambiguity" expressed in a given text) is simply:

$$\sigma_{amb} = \sqrt{V_{vd}}$$

If a text simultaneously contains "rightist" words, that is words associated with a right-wing ideology, but also "leftist" words, that is words used signaling a left ideology, this pattern can be interpreted as an ambiguous position taking. This makes sense!

Wordscores

Bräuninger and Giger (2016) show that extreme parties present more ambiguity than centrist ones



Wordscores

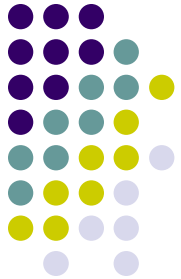
Interpreting Virgin Text Scores

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

Problem: many words are (usually) 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 reference 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), the first one with 205 words; the second with 175 words

There are 4 words included in the corpus of A+B, wherein *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 10 times in A and 30 times in B
2. *Crisis*. It appears 35 times in A and 5 times in B
3. *Government*. It appears 50 times in A and 50 times in B
4. *Britain*. It appears 110 times in A and 90 times in B
5. As a result (where for example 0.22=(10/205)/(10/205+30/175)):

$S_{choices,d}$ = 0.22*(3) + 0.78*(8) = 6.89; $S_{crisis,d}$ = 0.85*(3) + 0.15*(8) = 3.72

$S_{government,d}$ = 0.46*(3) + 0.54*(8) = 5.69; $S_{Britan,d}$ = 0.51*(3) + 0.49*(8) = 5.45

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)*6.89+(10/25)*5.69+(12/25)*5.45=5.71

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)*3.72+(8/24)*5.69+(10/24)*5.45=5.09

The estimated scores for C and D are much clustered to each other than the original scores for A and B (3 & 8)!

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}=(5.71-5.41)*(3.54/0.44) +5.41=7.8

The rescaled score for virgin text D becomes:

S^*_{Dd}=(5.09-5.41)*(3.54/0.44) +5.41=2.8

Therefore: A=3, B=8, C=7.8, D=2.8

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 5.71 and 5.09

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 C* and D* **will be now 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 still 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**

Further advantage: you get a measure of uncertainty also for the reference-text scores!

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)*6.89+(35/205)*3.72+(50/205)*5.69+(110/205)*5.45=5.3

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)*6.89+(5/175)*3.72+(50/175)*5.69+(90/175)*5.45=5.7

Your final raw scores would be: A=5.3, B=5.7, C=5.53, D=4.92

Our rescaled scores were: A=3, B=8, C=7.8, D=2.8

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**!

Wordscores

A **nice property** 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!

Document Selection

Wordscores does not make any assumption about the statistical distribution of words usage (contrary to a unsupervised algorithm, as we will see)

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

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

Let me give you some general guidelines in the selection of reference texts…

Document Selection

First: positions of the reference texts should "**span**" the dimensions in which we are interested. Trivially, if all reference texts have the **same position** on some dimension under investigation (say only texts from the left of the ideological spectrum), then their content contains no information that can be used to distinguish between other texts on the same dimension

An ideal selection of reference texts will contain texts that **occupy extreme positions, as well as intermediate positions,** along 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

Second: 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

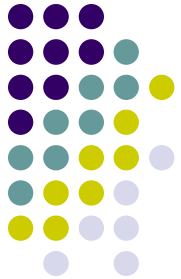
Third, the set of reference texts should contain as **many different words as possible** (i.e., they should include a sufficient range of potential word positions in the virgin texts). Why that?

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

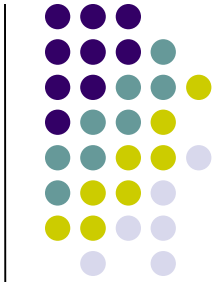


As a result of all this: avoid very-short reference texts!

Egerod and Klemmensen (2020) note that corpora consisting of very short texts (below 400 words on average) can still be reasonably scaled via Wordscores, if the reference documents provide good coverage of the virgin texts

However with shorter texts, problems arise

Document Selection



Fourth, there should be also a **sufficient overlap** between distributions of words in the reference texts

Why?

Because you can better compute the score for each word by taking advantage of such overlap (i.e., w/o overlap, all wordscores coming from the same reference text, will be exactly the same!)

For example, an *average of cosine-similarity* >.6 is a good value

Dynamic Estimation

Note that computing word scoring runs into significant problems when it comes to generating **long time series** of the positions of particular texts authors

In using particular reference texts, we are in fact for example 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

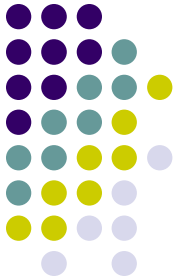
Imagine that in times t-1, the text uses the word “nigger” to identify Afro-Americans, while in time t the text uses the word “black” and at time t+1 the word “Afro-Americans”. Different words that refer to the same concept

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

Dynamic Estimation

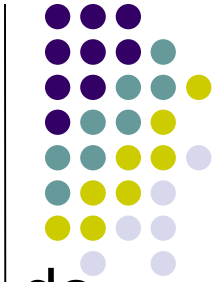
Three possible answers in this respect:

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, this 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 that we will discuss next week when dealing about unsupervised scaling algorithms



IMPORTANT!!!

About social media data



When dealing with social-media data (as we will start to do since tomorrow), you should be always very careful about data-privacy:

1. what kind of information can be ethically gathered about the users (public information)
2. how published data should look like to comply with privacy regulations (like the GDPR)
3. and what consequences violating the social network's terms of service may entail for the researcher

About social media data

For example: Twitter restricts the redistribution of Twitter content to third parties, so what you can share or publish online must be datasets consisting of tweet IDs of relevant tweets (or, *dehydrated* tweets)

You have then to *hydrate* these tweet IDs to obtain the tweet content

Relatively easy via `rtweet` package – as we will see

About social media data

Some good readings about these points:

Computational Research in the Post-API Age

What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data

(Supplementary Material with a great code in R to use with Facebook!)

An interesting paper with a review of nine different free-of-charge and low-cost software tools for studying Twitter: "Free and Low-Cost Twitter Research Software Tools for Social Science"

Before our second Lab

install.packages("cowplot", repos='http://cran.us.r-project.org')

install.packages("psych", repos='http://cran.us.r-project.org')

install.packages("PerformanceAnalytics", repos='http://cran.us.r-project.org')

install.packages("stringr", repos='http://cran.us.r-project.org')

install.packages("dplyr", repos='http://cran.us.r-project.org')

install.packages("gridExtra", repos='http://cran.us.r-project.org')

devtools::install_version("rtweet", version = "0.7.0", repos = "http://cran.us.r-project.org")

install.packages("httpuv", repos='http://cran.us.r-project.org')

devtools::install_github("hadley/emo")

devtools::install_github("gadenbuie/tweetrmd")

devtools::install_github("slowkow/ggrepel")

install.packages("readr", repos='http://cran.us.r-project.org')