Computational Political Science

Session 6

David Broska
Zeppelin University
March 9, 2021

Outline for today

1. **A word about designing research...**
    - Research questions and research design
    - How are we going to evaluate research papers in session 10
    - A recommendation on how to present your research ideas in session 11
2. **Wordscores model**
    - How it relates to Bayes theorem
    - How it is implemented
3. **Regularization** (revisit slides of session 5)
    - Ridge regression
    - Lasso regression
4. **Coding exercise** (revisit exercise of session 5)
    - Which letters are most predictive of female and male names?

Course schedule

Session	Date	Topic	Assignment	Due date
1	Feb 02	Overview and key concepts	-	-
2	Feb 09	Preprocessing and descriptive statistics	Formative	Feb 22 23:59:59
3	Feb 16	Dictionary methods	-	-
4	Feb 23	Machine learning for texts: Classification I	Summative 1	Mar 08 23:59:59
5	Mar 02	Machine learning for texts: Classification II	-	-
6	Mar 09	*Supervised and unsupervised scaling*	Summative 2	Mar 15 23:59:59
7	Mar 16	Similarity and clustering	-	-
8	Mar 23	Topic models	Summative 3	Apr 12 23:59:59
-	-	*Break*	-	-
9	Apr 13	Retrieving data from the web	-	-
10	Apr 20	Published applications	-	-
11	Apr 27	Project Presentations	-	-

A word about designing research...

Quality criteria

While teachers are paid to read student reports, readers outside the university have no such incentive.

Writers must find other ways to convince their audience to read their work.

A good research design begins with a research question:

- for which the answer/s matter
- that builds on an identifiable body of knowledge
- that is feasible to (at least partially) answer

Why? Because good research is *all* of the below:

- consequential: tells us something important
- novel: tells us something new
- valid: tells us something true

The consequentiality criterion

There are lots of practicable yet trivial research questions but good research should be consequential

Explain how your research problem is also your reader's problem by showing the consequence or the costs of not solving it.

Costs of not solving a conceptual problem

Ignorance about a conceptual problem is a lack of understanding that keeps us from understanding something else even more significant.

- If we cannot answer how our depictions of romantic love have changed $_{\text{second question}}$, then we cannot answer how our culture shapes the expectations of young people about marriage and families?$_{\text{consequence/larger, more important question}}$

Costs of not solving a practical problem

The cost of a practical problem is a tangible thing or situation we would like to avoid

- If we do not know the extent of the losses due to the economic crisis, we cannot plan the budget for the next year.

The novelty criterion

There are a lot of people doing social research

- Someone has already tried to answer a question similar to your question.
- Reviewing the research literature demonstrates that the exact question has not been satisfactorily answered.
- One goal for research is novelty, but you cannot assess novelty without an honest assessment of what has already been done. Research should make a specific contribution to an identifiable literature (this should primarily be the scholarly literature but may also include 'grey literature' [see Robson and McCartan 2016: 50- 56])

The feasibility criterion

Research design is about understanding feasibility tradeoffs

- If you deploy unlimited budgets or godlike powers, then you are not really engaging with the difficulty of the problem.
- A research design should include the best arguments in favor of your research design decisions
- Real research must be feasible in order to be completed and generate valid conclusions.
- For the purposes of this course, feasibility means in the context of your degree program.

Expectations on graduate research

Your research proposal is not going to be the most consequential or the most novel, given the fact that supervisors demand that it be feasible for you to complete in your current degree.

What supervisors typically want to see is that you can explain

- what you are proposing to do,
- why it is worth doing,
- how it will work,
- what we will learn from it,
- and what are its limitations.

If you have done this, you will have no difficulty with...

Dinner party test

Scene: You and a stranger at a dinner party / pub

Stranger: "So what do you do?"

You (reluctantly): "I'm working on my student research project"

Stranger (inevitably): "So what's it about?"

You (ramble on for 10 minutes secretly thinking): "this stranger can't possibly understand the complexity and nuance of what I'm doing ..."

Stranger (desperately seeks escape and/or stiff drink, vowing never to ask that question again)

> If you cannot give a synoptic, ordinary language explanation in two or three minutes of what you are focusing on and what you hope to achieve, the **chances are very high that in a very fundamental way you do not yet understand your thesis topic**. (Dunleavy 2003: 22)

Template for a Research Design

1. Background/literature review
2. Research question(s)
3. Data collection strategies
4. Data analysis strategies
5. Potential impact and relevance of the study
6. Limitations and further research
7. References / bibliography

This is an important template that we keep in mind as we go through the reading exercises!

Implicit Questions

Implicitly, a research design asks these questions

1. What do we know already?
2. What are you going to try to learn?
3. What kind of evidence are you going to collect and how will you collect it?
4. How does that evidence enable us to draw conclusions?
5. What might those conclusions be and why do they matter?
6. What are the limitations of what you are going to do? What have you done to mitigate these limitations? What more could be done with extra time and/or resources?
7. References / bibliography

Research questions

[R]esearch questions can provide the key to planning and carrying out a successful research project (Robson and McCartan 2016: 59)

They help to:

Define your project (summarize its focus)

- Set boundaries (demarcate the parameters of your project and so enhance feasibility)
- Give direction (signal what literature to search [relevance vs comprehensiveness], what data to collect, what methods to employ...)
- Define success (answerable research questions enable you to show that you have done what you set out do to)

Setting-up research questions

We need to consider not only what questions enable us to do, but also where we get them from - how can we set-up/motivate our research question/s

- ‘Consequentiality’ and ‘novelty’ criteria is key to this
- Importance of showing that our research matters and that it contributes somehow to our existing understanding of social phenomena

‘Gap-spotting’ in the academic literature a conventional way of motivating research

- Substantiate our contribution by reviewing what we (collectively) know already
- Multiple modes of ‘gap-spotting’, e.g. confusion spotting, neglect spotting… (Sandberg and Alvesson 2011); also methodological gap spotting

Supplementary sources

Can also signal the importance of our work in other ways. These include framing our research in relation to:

- contemporary social problems/puzzles (e.g. political debates or policy conundrums)
- why does this research matter beyond the ivory towers?
- apparent 'gaps' between official discourse and social practice
- socio-technical developments and trends
- interdisciplinary ('spending time in the next village' – e.g. how have commensurate processes been studied in other disciplines?)
- 'problematisation' (Sandberg and Alvesson 2011)
- personal experience ('starting where you are' [Robson and McCartan 2016: 49-50])

Identifying research questions

2016 *Ertug, Yogev, Lee, and Hedström* 129

Practical Implications

Our study provides practical implications not only for contemporary artists, but also those whose careers involve interacting with clients (or audiences for their work) with diverse concerns and interests. Even after achieving a positive reputation with a particular group of clients, those in such careers need to be aware that this same reputation may not be relevant for a different group of clients. Hence, such people would need to manage their careers to focus on one audience over others, depending on the attributes for which their reputation(s) is (are) relevant. This would also apply to, for example, managers who work as brokers between external clients and internal service providers (e.g., in an R&D department), whose key objectives and concerns differ. Given that a good reputation among external clients may not directly translate into an equally beneficial reputation among internal service providers, these managers should be aware of the differential effect of reputation and might need to separately build a relevant reputation for each audience group. Furthermore, as our arguments and findings regarding the contingent effects suggest, to further enhance one's positive reputation (or to fully benefit from it) with a given audience in a set, it would also be important to consider the accountability of that audience. In the example above, for instance, if the external client (rather than the internal service providers) were more accountable for their decisions vis-à-vis the output of the R&D department, the managers would also need to be mindful that it would become even more important to have other consistent signals of the quality of their work in their dealings with these external clients.

Limitations and Future Research

Our study is subject to certain limitations. First, we find no statistical support for the differential effect of appearing on a magazine cover on success with museums and success with galleries (Hypothesis 1b). We believe that the lack of support is due to the data available to measure a reputation for commercial viability in this setting. Specifically, magazines might also feature artists on their covers who possess artistic qualities in addition to their commercial viability. This is relevant because such occurrences would add to measurement error, making our indicator of a reputation for commercial viability noisier than we would like (as it might also capture some information about artistic quality). While this is true for appearing on a magazine cover, the reverse is not true for awards. Awards are commended and defended as championing artistic quality, therefore making our indicator of a reputation for artistic quality more informative and less noisy. We suggest that this is one reason that we find support for the differential effect of winning an award (Hypothesis 1a) but not for the differential effect of appearing on a magazine cover (Hypothesis 1b). The literature and mechanisms we use for our framework, and the consistent results we find for the differences in contingent effects on status and interaction with other audiences, suggest that the lack of support for Hypothesis 1b is due to the noisy measurement issue noted above. In theory, one might improve on our measure of magazine covers by using data on auctions, for example, with the implication that artists whose works have appeared in auctions more often, or have a higher sales ratio (the proportion of lots sold among those made available), have a reputation for commercial viability. However, for the artists in our estimation sample, we were unable to find a database that offers anywhere close to systematic and comprehensive coverage. We note this, again, to suggest that, based on the support we have for the other predictions and on the foundations of our framework, we expect our predictions to be broadly applicable, despite the lack of support for Hypothesis 1b with the measure we use.

Second, we examine only two signals, among possibly multiple types of other signals. Accordingly, future research can consider signals or intangible assets other than status or interaction with other audiences (Pfarrer et al., 2010; Pollock & Gulati, 2007) and determine whether they, in conjunction with reputation, would be informative for audiences, and how this contingent relationship might again vary on the basis of accountability or other broadly applicable constructs that constrain the decision making of audiences.

Third, the research context of the contemporary art field was used to develop our hypotheses and assess audience-specific reputations. The unique characteristics of this setting reduce the generalizability of our findings, and future studies in other contexts are needed to further establish the generalizability of the framework underlying our hypotheses. However, we view the contextual specificity of our study as a strength rather than a weakness because the nuances of audience-specific reputations require an in-depth understanding of specific audiences and the sources of their concerns and uncertainty in a particular setting. Our framework can be applied to

- Unlike much funded or commissioned research, for independent projects (BA and MA theses) you are expected to come up with research questions yourselves
- Identifying and reading around topics of interest (being sure to include cutting edge studies) should help with this
- You might identify the gap yourself but often empirical papers also include further research directions in their conclusions. Could you feasibly address any of these?
- Journal editorials/review papers can often serve as a source of inspiration

Topic ≠ research question/s

- Once you have a topic, clarifying the purpose/s of your research can be crucial to developing research questions
- Beyond contributing to knowledge, typical broad purposes include exploration, description and explanation (and possibly impact for more applied research)
- Don't be afraid of coming up with multiple questions – this is normal, and a set of (often nested) questions can indeed be advantageous/more readily answerable (as long as the questions are feasible)
- Use the feasibility criterion to help you prioritize your research questions

Ultimately, you should be able to fill in these blanks:

1. Topic: I am studying ...
2. Question: because I want to find out what/why/how ...,
3. Significance: in order to help my reader understand ...

Example immigration-related questions

What is the type of answer you expect from doing research?

1. How many people in my dataset of German residents said they thought there should be less immigration?
2. How do people living in post-industrial towns in Germany perceive immigration in their local areas?
3. What fraction of people in Germany think there should be less immigration?
4. What kinds of people in Germany tend to say there should be less immigration?
5. Do people in Germany become more or less favorable towards immigration if they work with immigrants?
6. Why do some people in Germany say there should be less immigration?

1. A question about particular data
2. An exploratory question about a population
3. A quantifiable question about a population
4. A question about a relationship in a broader population
5. A question about causal relationship size
6. A question about causal relationship mechanisms

Topically Related Research Questions

What is the type of answer you expect from doing research?

1. How many people in my dataset of German residents said they thought there should be less immigration?
2. How do people living in post-industrial towns in Germany perceive immigration in their local areas?
3. What fraction of people in Germany think there should be less immigration?
4. What kinds of people in Germany tend to say there should be less immigration?
5. Do people in Germany become more or less favorable towards immigration if they work with immigrants?
6. Why do some people in Germany say there should be less immigration?

1. Description
2. Exploration
3. Population Inference (description)
4. Population Inference (description)
5. Causal Inference (cause-effect) [explanation]
6. Causal Inference (causal mechanisms) [exploration/explanation]

Tell me about your research

In individual meetings we will discuss your Zeppelin project while the others are encouraged to work on the reading assignments

Preparation for dinner-party test

1. Topic: I am studying ...
2. Question: because I want to find out what/why/how ...,
3. Significance: in order to help my reader understand ...

Projecting outcomes of the study

- What kind of answer do expect from your study?
- How might your results change our view on the social phenomenon that you are studying?

Methods

- Which methods might be useful for collecting and analyzing data?

Supervised and unsupervised learning

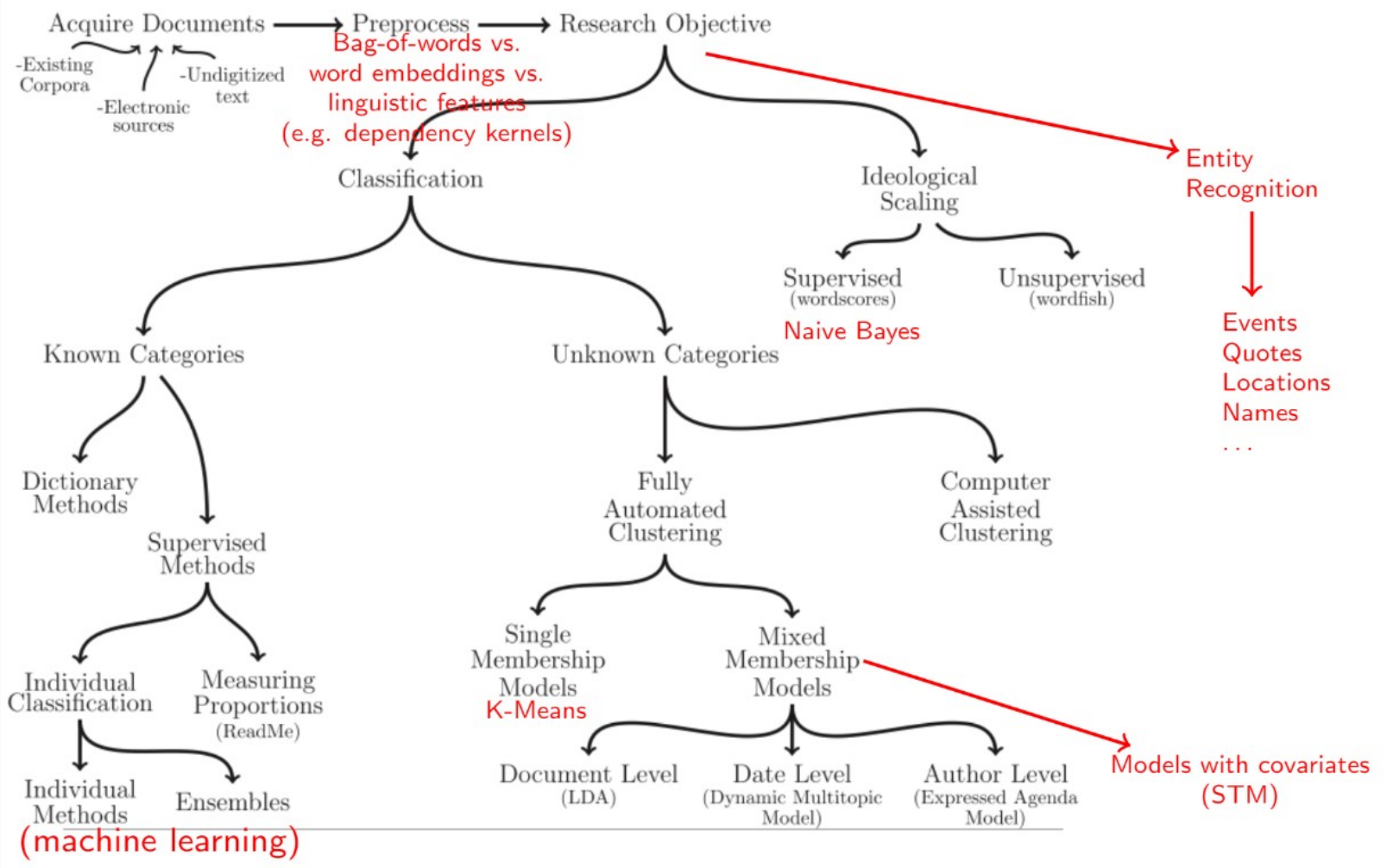


Fig. 1 in Grimmer and Stuart (2013)

Wordscores

From classification to scaling

Machine learning focuses on identifying classes (*classification*), while social science is typically interested in locating things on latent traits (*scaling*), for example:

- Policy positions on economic vs social dimension
- Inter- and intra-party differences
- Soft news vs hard news
- ...and any other continuous scale

But the two methods overlap and can be adapted - will demonstrate later using the Naive Bayes classifier

In fact, the class predictions for a collection of words from Naive Bayes can be adapted to scaling

Wordscores

Analogous to a "training set" and a "test set" in classification, the Wordscores method by Laver, Benoit, and Garry (2003) uses two sets of texts:

Reference texts

- texts about which we know something (a scalar dimensional score)

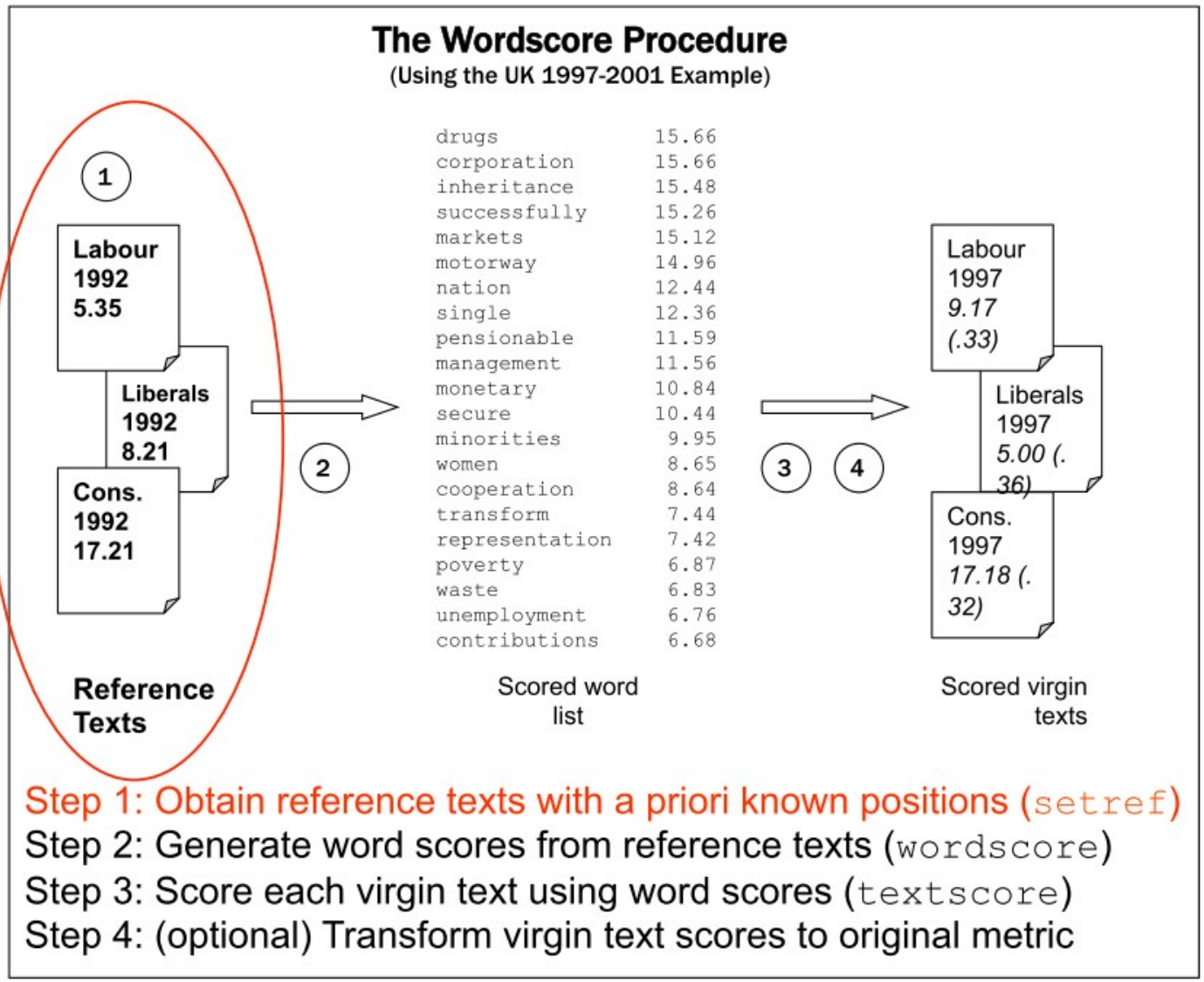
Virgin texts

- texts about which we know nothing (but whose dimensional score we'd like to know)

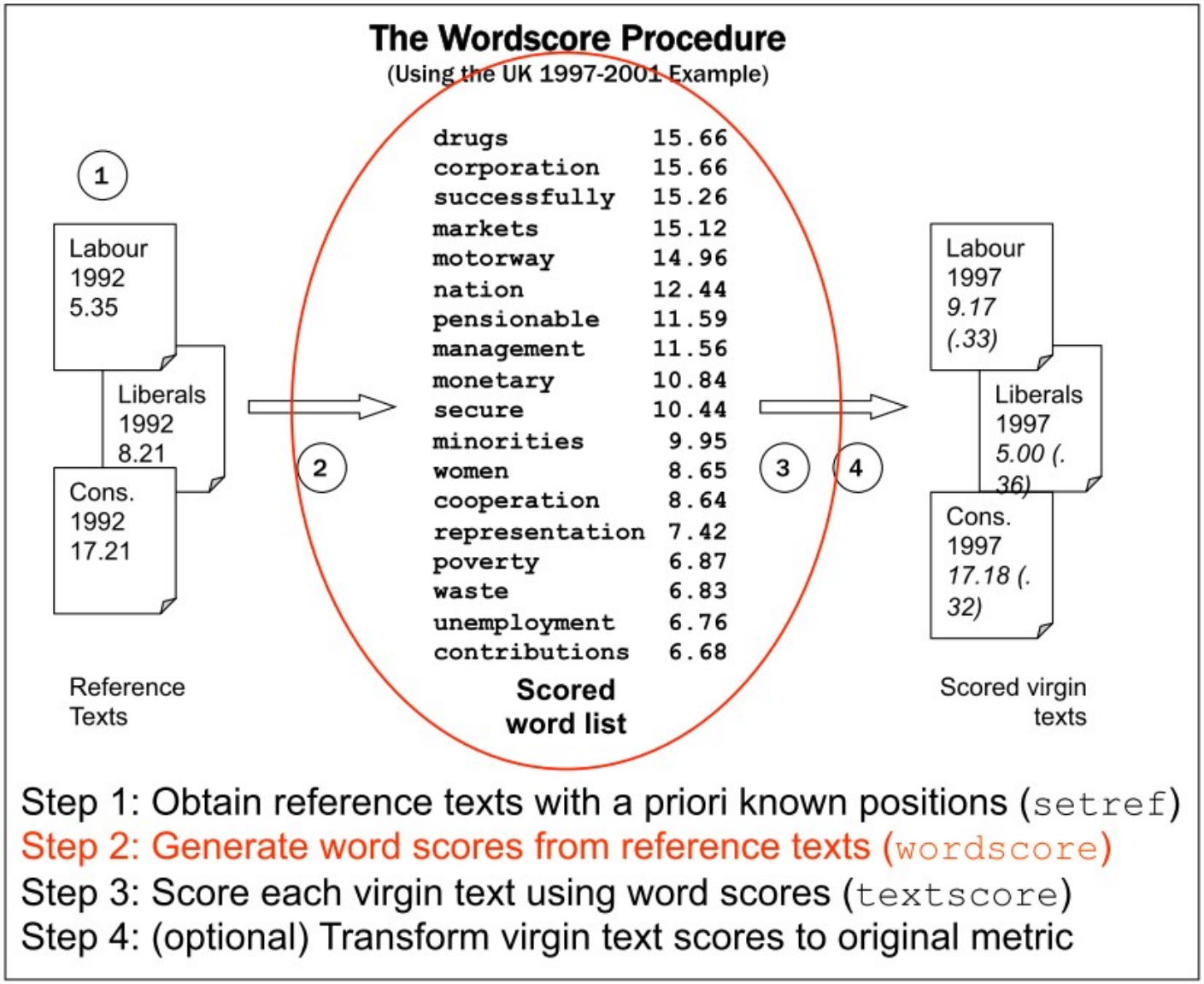
Basic procedure

1. Analyze reference texts to obtain a single "score" for every word
2. Use word scores to score virgin texts

Wordscores procedure (I)



Wordscores procedure (II)



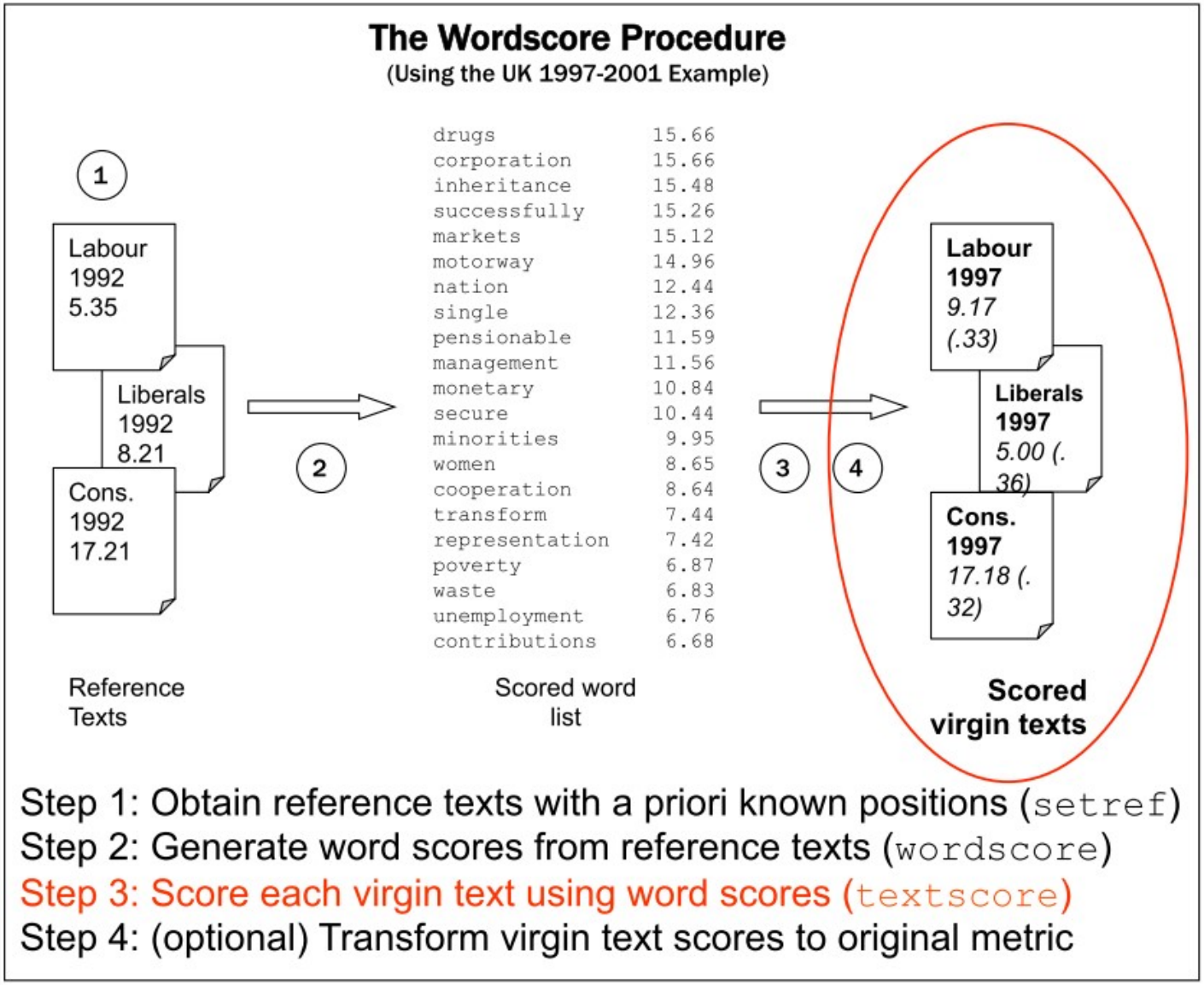
Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

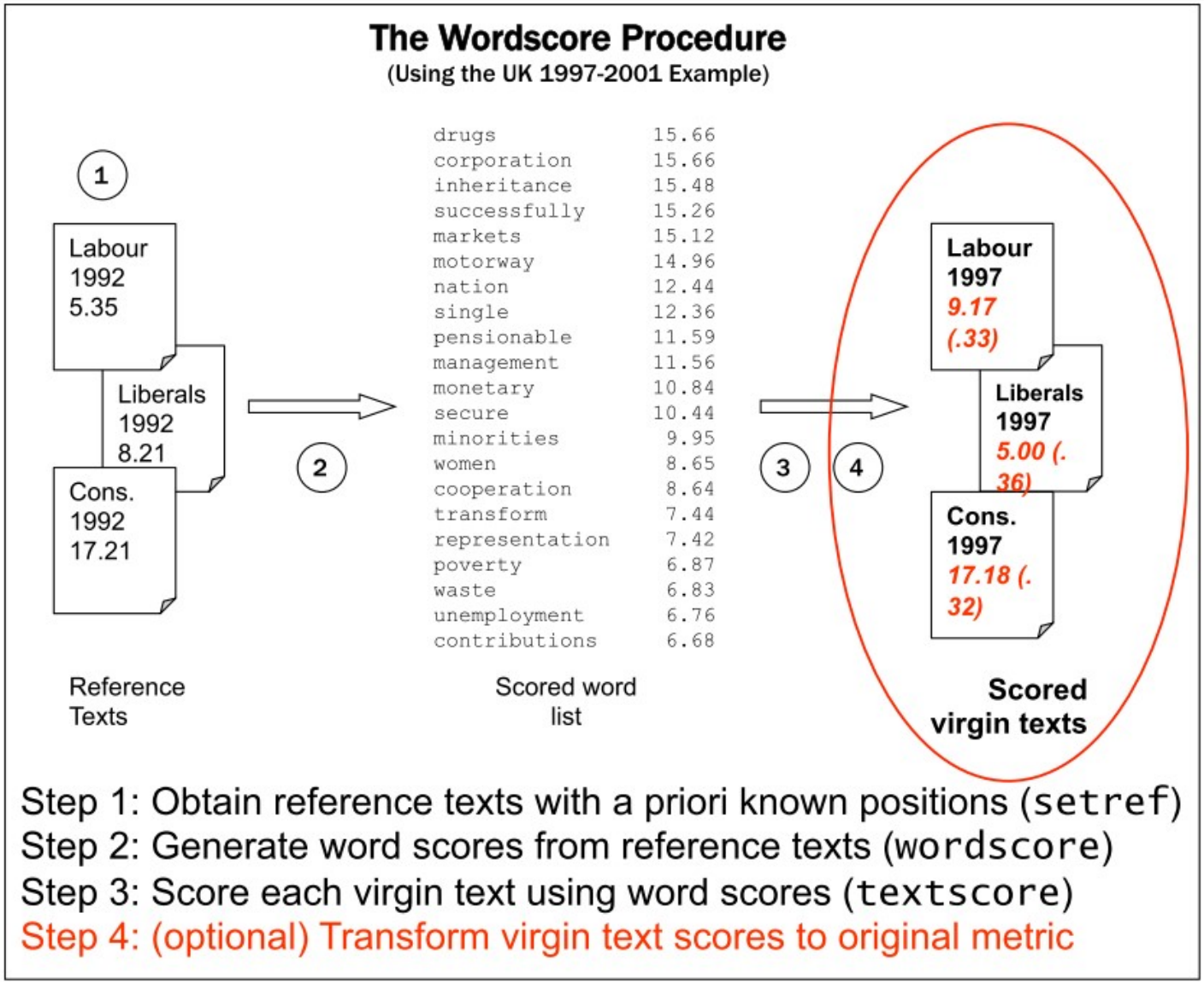
Step 4: (optional) Transform virgin text scores to original metric

Wordscores procedure (III)



Step 1: Obtain reference texts with a priori known positions (`setref`)
Step 2: Generate word scores from reference texts (`wordscore`)
Step 3: Score each virgin text using word scores (`textscore`)
Step 4: (optional) Transform virgin text scores to original metric

Wordscores procedure (IV)



Wordscore implementation

```
# 4 texts with known and 3 texts with unknown category
txt <- c(k1 = "$ Win $",
         k2 = "$ Prize $",
         k3 = "Earn $ Easily",
         k4 = "Paypal 100 $",
         u1 = "$",
         u2 = "$ $",
         u3 = "Paypal 100 $ $")
x <- dfm(txt)
y <- c(1, 1, 1, -1, NA, NA, NA)
```

training dfm from references texts

	$	win	prize	earn	easily	paypal	100
k1	2	1	0	0	0	0	0
k2	2	0	1	0	0	0	0
k3	1	0	0	1	1	0	0
k4	1	0	0	0	0	1	1

training vector with known positions

y
1
1
1
-1

Wordscores

Compute probability of a reading document given a word

Start with a set of D reference texts, represented by an $D \times W$ document-feature matrix C_{dw} , where d indexes the document and w indexes the W total word types.

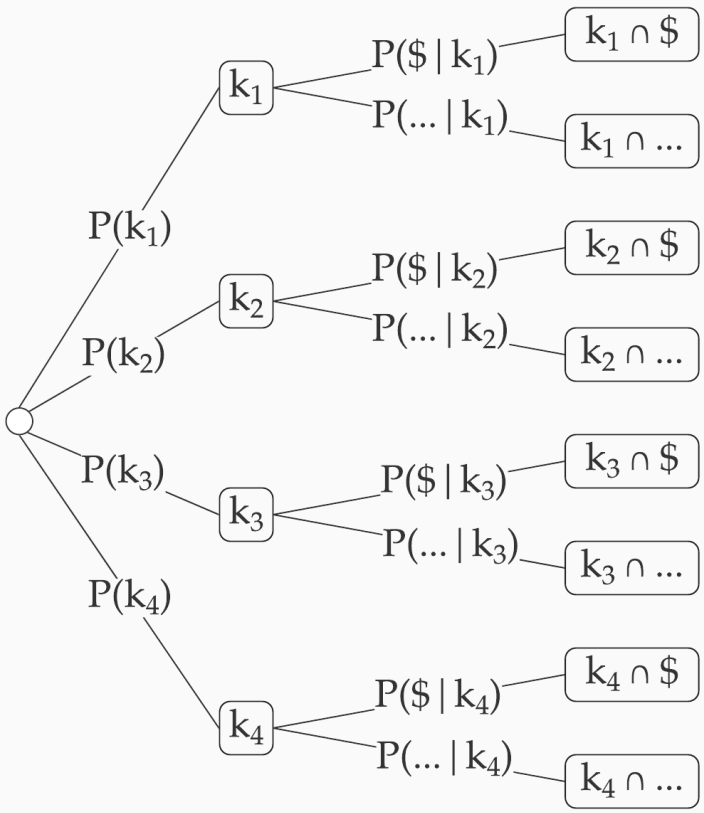
We normalize the document-feature matrix within each document by converting C_{ij} into a relative document-feature matrix (within document), by dividing C_{ij} by its word total marginals

Probability of word given the document

```
( PwGd <- dfm_weight(x[1:4,],scheme="prop") )
```

```
##     features
## docs    $  win prize earn easily paypal  100
##   k1 0.67 0.33  0.00 0.00   0.00   0.00 0.00
##   k2 0.67 0.00  0.33 0.00   0.00   0.00 0.00
##   k3 0.33 0.00  0.00 0.33   0.33   0.00 0.00
##   k4 0.33 0.00  0.00 0.00   0.00   0.33 0.33
```

P(k_1 | $)



Uniform priors: $P(k_1)=...=P(k_4)= ¼$

If we only read "$" the probability of reading the document k_1 is ⅓.

Probability of word given the document:

	$	win	prize	earn	easily	paypal	100
k1	0.67	0.33	0.00	0.00	0.00	0.00	0.00
k2	0.67	0.00	0.33	0.00	0.00	0.00	0.00
k3	0.33	0.00	0.00	0.33	0.33	0.00	0.00
k4	0.33	0.00	0.00	0.00	0.00	0.33	0.33

$$
\begin{aligned}
&P(k_1|\$) \\
&= \frac{P(k_1)P(\$|k_1)}{P(k_1)P(\$|k_1) + ... + P(k_2)P(\$|k_4)} \\
&= \frac{P(\$|k_1)}{P(\$|k_1) + ... + P(\$|k_4)} \\
&= \frac{\frac{2}{3}}{\frac{2}{3} + \frac{2}{3} + \frac{1}{3} + \frac{1}{3}} = \frac{1}{3}
\end{aligned}
$$

P(document | word)

Now let's compute all probabilities of reading a document given a word

```
PwGd # recall our matrix containing all P(word | document)
```

```
##     features
## docs    $  win prize earn easily paypal  100
##   k1 0.67 0.33  0.00 0.00   0.00   0.00 0.00
##   k2 0.67 0.00  0.33 0.00   0.00   0.00 0.00
##   k3 0.33 0.00  0.00 0.33   0.33   0.00 0.00
##   k4 0.33 0.00  0.00 0.00   0.00   0.33 0.33
```

```
# transpose PwGd matrix
( tPwGd <- t(PwGd) )
```

```
##          docs
## features   k1   k2   k3   k4
##   $      0.67 0.67 0.33 0.33
##   win    0.33 0.00 0.00 0.00
##   prize  0.00 0.33 0.00 0.00
##   earn   0.00 0.00 0.33 0.00
##   easily 0.00 0.00 0.33 0.00
##   paypal 0.00 0.00 0.00 0.33
```

```
# P(document | word)
( PdGw <- tPwGd / rowSums(tPwGd) )
```

```
##          docs
## features   k1   k2   k3   k4
##   $      0.33 0.33 0.17 0.17
##   win    1.00 0.00 0.00 0.00
##   prize  0.00 1.00 0.00 0.00
##   earn   0.00 0.00 1.00 0.00
##   easily 0.00 0.00 1.00 0.00
##   paypal 0.00 0.00 0.00 1.00
```

Scoring words

Compute a J-length "score" vector S for each word j as the average of each document i's scores a_i, weighted by each word's P_{ij} so that $S_j = \sum_i^I a_i P_{ij}$

```
y[1:4] # the "a" vector with the positions of the document
```

```
## [1]  1  1  1 -1
```

```
t(PdGw) * y[1:4] # transpose matrix so we can multiply PdGw with the doc positions
```

```
##     features
## docs     $ win prize earn easily paypal 100
##   k1  0.33   1     0    0      0      0   0
##   k2  0.33   0     1    0      0      0   0
##   k3  0.17   0     0    1      1      0   0
##   k4 -0.17   0     0    0      0     -1  -1
```

```
( ws <- colSums( t(PdGw) * y[1:4] )) # then, sum up the result column-wise
```

```
##      $    win  prize   earn easily paypal    100
##   0.67   1.00   1.00   1.00   1.00  -1.00  -1.00
```

Scoring words

We obtain the scored words *also* by using matrix multiplication. In matrix algebra,

$$\underset{1\times J}{S} = \underset{1\times I}{a} \cdot \underset{I\times J}{P}$$

```
PdGw # P(document | word)
```

```
##          docs
## features   k1   k2   k3   k4
##   $      0.67 0.67 0.33 0.33
##   win    0.33 0.00 0.00 0.00
##   prize  0.00 0.33 0.00 0.00
##   earn   0.00 0.00 0.33 0.00
##   easily 0.00 0.00 0.33 0.00
##   paypal 0.00 0.00 0.00 0.33
##   100    0.00 0.00 0.00 0.33
```

```
y[1:4] # documents scale
```

```
## [1]  1  1  1 -1
```

```
 # matrix multiplication with P(document|words) and scores
( ws <- PdGw %*% y[1:4] )
```

```
##      $    win  prize   earn easily paypal    100
##   0.67   1.00   1.00   1.00   1.00  -1.00  -1.00
```

Scoring texts

The goal is to obtain a single score for any new text, relative to the reference texts

We do this by taking the mean of the scores of its words, weighted by their term frequency

```
PwGd <- dfm_weight(x, scheme="prop")
t(PwGd) # transpose matrix
```

```
##          docs
## features   k1   k2   k3   k4 u1 u2   u3
##    $      0.67 0.67 0.33 0.33  1  1 0.50
##    win    0.33 0.00 0.00 0.00  0  0 0.00
##    prize  0.00 0.33 0.00 0.00  0  0 0.00
##    earn   0.00 0.00 0.33 0.00  0  0 0.00
##    easily 0.00 0.00 0.33 0.00  0  0 0.00
##    paypal 0.00 0.00 0.00 0.33  0  0 0.25
##    100    0.00 0.00 0.00 0.33  0  0 0.25
```

```
# row-wise PwGd * score
t(PwGd) * ws[,1]
```

```
##          docs
## features   k1   k2   k3    k4   u1   u2
##    $      0.44 0.44 0.22  0.22 0.67 0.67
##    win    0.33 0.00 0.00  0.00 0.00 0.00
##    prize  0.00 0.33 0.00  0.00 0.00 0.00
##    earn   0.00 0.00 0.33  0.00 0.00 0.00
##    easily 0.00 0.00 0.33  0.00 0.00 0.00
##    paypal 0.00 0.00 0.00 -0.33 0.00 0.00
##    100    0.00 0.00 0.00 -0.33 0.00 0.00
```

```
colSums( t(PwGd) * ws[,1] )
```

```
##    k1    k2    k3    k4    u1    u2    u3
##  0.78  0.78  0.89 -0.44  0.67  0.67 -0.17
```

Scoring texts

We obtain the scored words *also* by using matrix multiplication.

```
# matrix multiplication with P(word | document) and obtained wordscores
PwGd %*% ws
```

```
##    k1    k2    k3    k4    u1    u2    u3
##  0.78  0.78  0.89 -0.44  0.67  0.67 -0.17
```

Does this result make sense in the context of the spam example?

k1 (s)	k2 (s)	k3 (s)	k4 (¬s)	u1	u2	u3
$ Win $	$ Prize $	Earn $ Easily	Paypal 100 $	$	$ $	Paypal 100 $ $

Final remarks

- Note that new words outside of the set J may appear in the K virgin documents — these are simply ignored (because we have no information on their scores)
- Note also that nothing prohibits reference documents from also being scored as virgin documents

Scoring texts

Using textmodel_wordscores()

For convenience we can use the quanteda function to obtain the above results

```
ws_mod <- textmodel_wordscores(x,y)
```

Wordscores

```
summary(ws_mod)
```

```
## textmodel_wordscores.dfm(x = x, y = y)

## (showing first 7 elements)
##      $    win  prize   earn easily paypal    100
##   0.67   1.00   1.00   1.00   1.00  -1.00  -1.00
```

Scaled documents

```
predict(ws_mod)
```

```
##    k1    k2    k3    k4    u1    u2    u3
##  0.78  0.78  0.89 -0.44  0.67  0.67 -0.17
```

Wordscore coding exercise