Day 12: More text analysis and networks

ME314: Introduction to Data Science and Big Data Analytics

LSE Methods Summer Programme 2018

15 August 2018

Day 12 Outline

Dictionary basics revisited

Hierarchy in dictionaries

Advantages of dictionaries

Keyword analysis and dictionary construction

Scaling and reporting dictionary results

Problems to avoid

Additional useful text skills

Network analysis (of text)

Exam review

Dictionaries: Hybrid between quant and qual

- "Qualitative" since it involves identification of the concepts and associated keys/categories, and the textual features associated with each key/category
- Dictionary construction involves a lot of contextual interpretation and qualitative judgment
- Perfect reliability because there is no human decision making as part of the text analysis procedure

"Dictionary": a misnomer?

- A *dictionary* is really a thesaurus: a canonical term or concept (a "key") associated with a list of equivalent synonyms
- But dictionaries tend to be exclusive: they single out features defined as keys, selecting the terms or patterns linked to each key
- An alternative is a "thesaurus" concept: a tag of key equivalency for an associated set of terms, but non-exclusive
  - WC = `wc, toilet, restroom, bathroom, jack, loo`
  - vote = `poll, suffrage, franchis*, ballot*, ^vot$`

Dictionary basics: A review

- Rather than count words that occur, pre-define words associated with specific meanings
- Two components:

  key the label for the equivalence class for the concept or canonical term

  values (multiple) terms or patterns that are declared equivalent occurences of the key class

- Frequently involves lemmatization: transformation of all inflected word forms to their "dictionary look-up form" — more powerful than stemming

Example: Laver and Garry (2000)

- A *hierarchical* set of categories to distinguish policy domains and policy positions – similar in spirit to the CMP
- Five domains at the top level of hierarchy
  - economy
  - political system
  - social system
  - external relations
  - a " 'general' domain that has to do with the cut and thurst of specific party competition as well as uncodable pap and waffle"
- Looked for word occurences within "word strings with an average length of ten words"
- Built the dictionary on a set of specific UK manifestos

Example: Laver and Garry (2000): Economy

Table 1 Abridged Section of Revised Manifesto Coding Scheme

1 ECONOMY
Role of state in economy

1 1 ECONOMY/+State+
Increase role of state

1 1 1 ECONOMY/+State+/Budget
Budget

1 1 1 1 ECONOMY/+State+/Budget/Spending
Increase public spending

1 1 1 1 1 ECONOMY/+State+/Budget/Spending/Health

1 1 1 1 2 ECONOMY/+State+/Budget/Spending/Educ. and training

1 1 1 1 3 ECONOMY/+State+/Budget/Spending/Housing

1 1 1 1 4 ECONOMY/+State+/Budget/Spending/Transport

1 1 1 1 5 ECONOMY/+State+/Budget/Spending/Infrastructure

1 1 1 1 6 ECONOMY/+State+/Budget/Spending/Welfare

1 1 1 1 7 ECONOMY/+State+/Budget/Spending/Police

1 1 1 1 8 ECONOMY/+State+/Budget/Spending/Defense

1 1 1 1 9 ECONOMY/+State+/Budget/Spending/Culture

1 1 1 2 ECONOMY/+State+/Budget/Taxes
Increase taxes

1 1 1 2 1 ECONOMY/+State+/Budget/Taxes/Income

1 1 1 2 2 ECONOMY/+State+/Budget/Taxes/Payroll

1 1 1 2 3 ECONOMY/+State+/Budget/Taxes/Company

1 1 1 2 4 ECONOMY/+State+/Budget/Taxes/Sales

1 1 1 2 5 ECONOMY/+State+/Budget/Taxes/Capital

1 1 1 2 6 ECONOMY/+State+/Budget/Taxes/Capital gains

1 1 1 3 ECONOMY/+State+/Budget/Deficit
Increase budget deficit

1 1 1 3 1 ECONOMY/+State+/Budget/Deficit/Borrow

1 1 1 3 2 ECONOMY/+State+/Budget/Deficit/Inflation

Example: Laver and Garry (2000)

```
ECONOMY / +STATE
    accommodation
    age
    ambulance
    assist
    ...

ECONOMY / -STATE
    choice*
    compet*
    constrain*
    ...
```

Advantage of dictionaries: Multi-lingual

APPENDIX B
DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS

	NL	UK	GE	IT
Core	elit*	elit*	elit*	elit*
	consensus*	consensus*	konsens*	consens*
	ondemocratisch*	undemocratic*	undemokratisch*	antidemocratic*
	ondemokratisch*			
	referend*	referend*	referend*	referend*
	corrupt*	corrupt*	korrupt*	corrot*
	propagand*	propagand*	propagand*	propagand*
	politici*	politici*	politiker*	politici*
	bedrog	*deceit*	täusch*	ingann*
	bedrieg	*deceiv*	betrüg*	
			betrug*	
	verraa	*betray*	*verrat*	tradi*
	verrad			
	schaam*	shame*	scham*	vergogn*
			schäm*	
	schand*	scandal*	skandal*	scandal*
	waarheid*	truth*	wahrheit*	verità
	oneerlijk*	dishonest*	unfair*	disonest*
			unehrlich*	
Context	establishm*	establishm*	establishm*	partitocrazia
	heersend*	ruling*	*herrsch*	
	capitul*			
	kapitul*			
	kaste*			
	leugen*		lüge*	menzogn*
	lieg*			mentir*

(from Rooduijn and Pauwels 2011)

Disdvantage: Highly specific to context

- Example: Loughran and McDonald used the Harvard-IV-4 TagNeg (H4N) file to classify sentiment for a corpus of 50,115 firm-year 10-K filings from 1994–2008
- found that almost three-fourths of the "negative" words of H4N were typically not negative in a financial context
  e.g. *mine* or *cancer*, or *tax*, *cost*, *capital*, *board*, *liability*, *foreign*, and *vice*
- Problem: polysemes – words that have multiple meanings
- Another problem: dictionary lacked important negative financial words, such as *felony*, *litigation*, *restated*, *misstatement*, and *unanticipated*

How to build a dictionary

- The ideal content analysis dictionary associates all and only the relevant words to each category in a perfectly valid scheme
- Three key issues:

Validity	Is the dictionary's category scheme valid?
Sensitivity	Does this dictionary identify *all* my content?
Specificity	Does it identify *only* my content?

- Imagine two logical extremes of including all words (too sensitive), or just one word (too specific)

Coding scheme fundamentals

1. First key principle: Hierarchy
   - 1.1 First level: Domain
   - 1.2 Second level: subdomain
   - 1.3 (Third+ levels: may be additional sub-domains)
2. Second key principle: Confrontation
   Lowest-level categories should be for/against pairs, or "for/neutral/against"
3. On testing: Not necessary at design stage in the same way as for human coding – this is replaced by sensitivity/specificity testing in dictionary construction

How to build a dictionary

1. Identify "extreme texts" with "known" positions. Examples:
   - Opposition leader and Prime Minister in a no-confidence debate
   - Opposition leader and Finance Minister in a budget debate
   - Five-star review of a product (excellent) and a one-star review (terrible)
2. Search for differentially occuring words using word frequencies
3. Examine these words in context to check their sensitivity and specificity
4. Examine inflected forms to see whether stemming or wildcarding is required
5. Use these words (or their lemmas) for categories

Detecting "keywords"

- Detects words that *discriminate* between partitions of a corpus
- For instance, we could partition the Irish budget speech corpus into "government" and "opposition" speeches, and look for words that occur in one partition with higher relative frequency in opposition than in government speeches
- This is done by constructing a 2×2 table for each word, and testing association between that word and the partition categories

Detecting "keywords": Constructing the association table

	Target	~ Target	
Word 1	n_{11}	n_{12}	$n_{1.}$
~ (Word 1)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

- Once this is constructed, any standard measures of association (similar to those used to detect collocations) can be used to identify keyword associations with a class
- Same association measures are used as with collocation detection

statistical association measures

where m_{ij} represents the cell frequency expected according to independence:

G^2 likelihood ratio statistic, computed as:

$$2 * \sum_i \sum_j (n_{ij} * log \frac{n_{ij}}{m_{ij}}) \quad (1)$$

χ^2 Pearson's χ^2 statistic, computed as:

$$\sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \quad (2)$$

statistical association measures (cont.)

pmi point-wise mutual information score, computed as $\log n_{11}/m_{11}$

dice the Dice coefficient, computed as

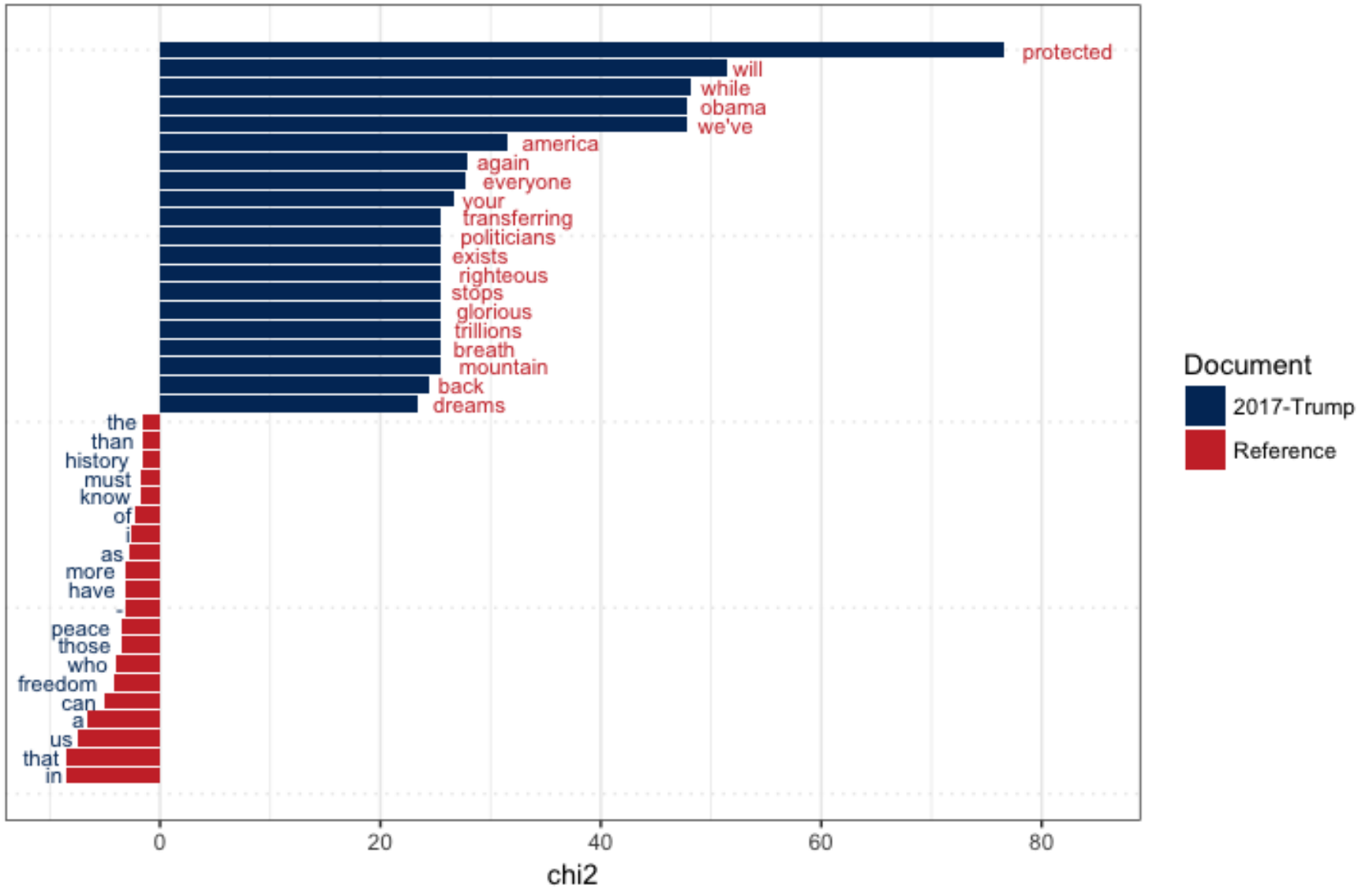
$$\frac{n_{11}}{n_{1.} + n_{.1}} \tag{3}$$

Examples

```
# compare Trump 2017 to other post-war preseidents
period <- ifelse(docvars(data_corpus_inaugural, "Year") < 1945,
                 "pre-war", "post-war")
pwdfm <- dfm(corpus_subset(data_corpus_inaugural, period == "post-war"))

textstat_keyness(pwdfm, target = "2017-Trump") %>%
    head(n = 7)
#     feature     chi2            p n_target n_reference
# 1 protected 76.64466 0.000000e+00        5           1
# 2      will 51.44795 7.351897e-13       40         299
# 3     while 48.23022 3.790079e-12        6           7
# 4     obama 47.85727 4.584000e-12        3           0
# 5     we've 47.85727 4.584000e-12        3           0
# 6   america 31.45537 2.040775e-08       18         112
# 7     again 27.81145 1.337322e-07        9          33
```

Examples

```
# using the likelihood ratio method
textstat_keyness(dfm_smooth(pwdfm), measure = "lr", target = "2017-Trump") %>%
    head()
#    feature        G2            p n_target n_reference
# 1     will 24.604106 7.040156e-07       41         317
# 2  america 14.040255 1.789387e-04       19         130
# 3     your 10.435140 1.236402e-03       12          68
# 4    again  9.758516 1.784939e-03       10          51
# 5    while  9.504990 2.049139e-03        7          25
# 6 american  8.877690 2.886766e-03       12          76


textstat_keyness(pwdfm, target = "2017-Trump") %>%
    textplot_keyness()
```


protected
will
while
obama
we've
america
again
everyone
your
transferring
politicians
exists
righteous
stops
glorious
trillions
breath
mountain
back
dreams
the
than
history
must
know
of
i
as
more
have
-
peace
those
who
freedom
can
a
us
that
in
0
20
40
60
80
chi2
Document
2017-Trump
Reference

Examples

Table 5
Keywords by gender in interview text: Selected categories[a]

Prostate	Breast
Treatment	
Catheter, brachytherapy, hormone, Zoladex, treatment, seeds, prostatectomy, Casodex, injection, radiation, injections, operation, Viagra, beam, radical, bag, Spes, Flutamide, tubes, capsule, Prazosin, tablets, watchful [waiting], cryosurgery, cryotherapy, Muse, probes, [watchful] waiting, therapy, strapped	Chemotherapy, Tamoxifen, mastectomy, prosthesis, chemo, lumpectomy, needle, HRT, scar, drains
Support	
NO KEYWORDS	Help, supportive, support, helped
Feelings	
Concerned, embarrassment	Feel, felt, want, need, cope, scared, crying, ups [and downs], wanted, depressed, scary, brave, cried, angry, coping, coped, feelings, fight, hard, upset
People	
Wife, he, men, man, chap, male, his, chaps, guy	I, she, husband, her, you, women, my, people, mum, sister, everybody, me, children, mother, friends, woman, lady, dad, she'd, daughter, she's, yourself, myself, sisters, I'd, auntie, ladies, who've, someone, somebody, your
Superlatives	
NO KEYWORDS	Wonderful, lovely, lots, amazing, marvellous

[a]Each section lists words in descending order of 'keyness'; 'split' words are excluded.

What to do with dictionary results

- Describe the results
- Scale quantities: pro- v. anti-, left v. right, etc. Example: Laver and Garry (see Lowe et al 2011 for alternatives)
- Could use these as features to measure similarity using (e.g.) cosine similarity
- Treat as other features and use machine learning or data mining methods

Scaling Issues

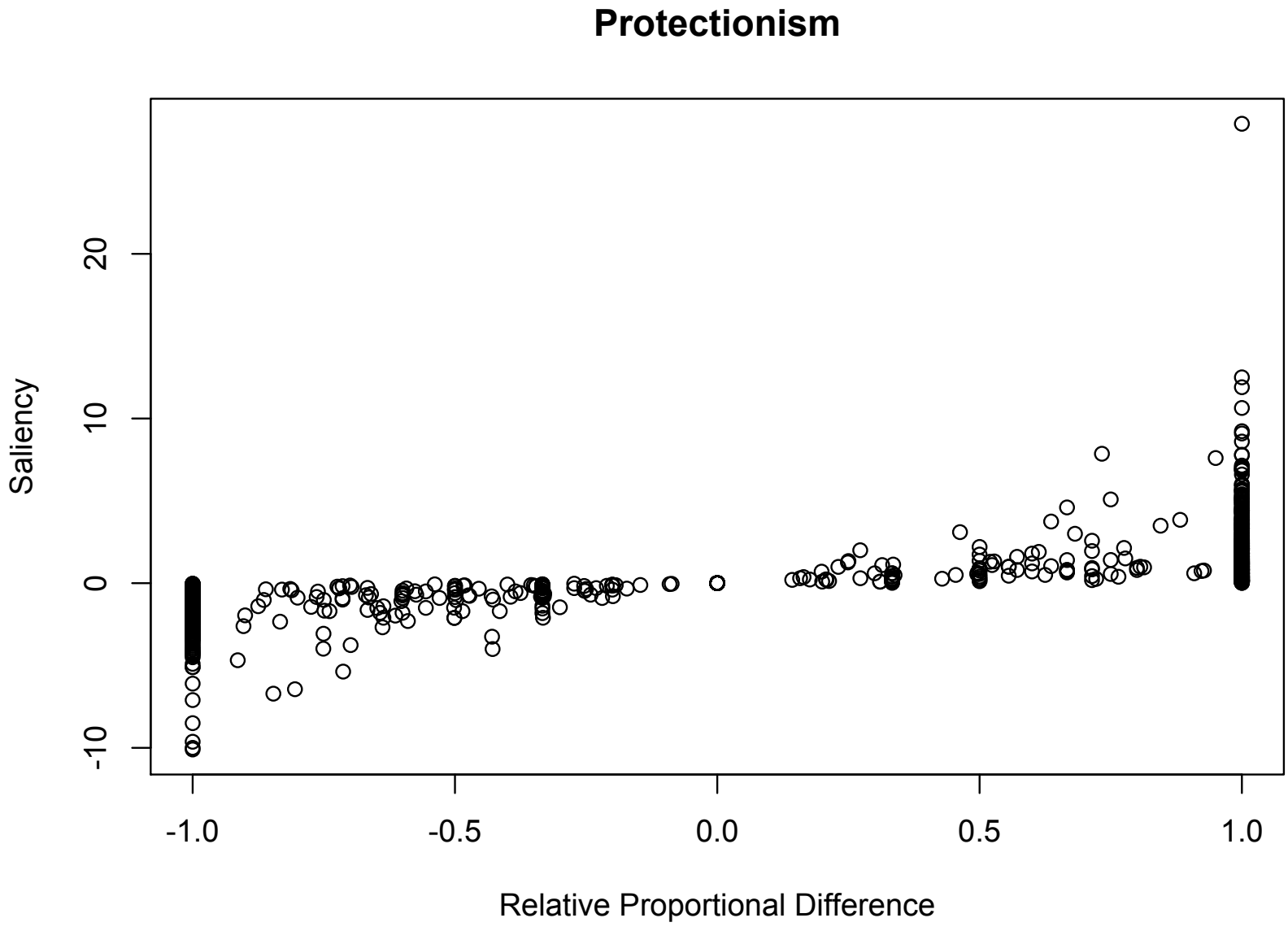
- Scaling becomes a major issue when we wish to construct quantities of interest from quantitative content analyses
- Simple example: Proportion of content of a given type (e.g. anti-Lisbon treaty)
- Complex example: Left-right policy positions (e.g. CMP "Rile")
- Are the metrics "natural"?
- Does the output metric resemble the input metric (if any)?
- What properties should the scale have, such as boundaries, type of increase, etc?
- How can uncertainty be characterized for the given scale?

Logit scale for left-right

- The Comparative Manifesto Project scales policy positions as absolute porportional difference, measured by proportion of "Right" mentions less proportion of "Left" mentions: $\frac{(R-L)}{N}$
- Problems:
  - Addition of irrelevant content shifts the scale toward zero
  - Assumes the additional mentions increase emphasis in a linear scale
- The alternative is to scale $\frac{(R-L)}{(R+L)}$ (Kim and Fording 2002; Laver and Garry 2000), but this too has problems:
  - Still linear shift in position for increase in repetition
  - Quickly maxes out at the extremes
- Lowe, Benoit, Mikhaylov and Laver (2010) propose using a logistic odds-ratio scale $\log \frac{R}{L}$

Comparing scales:

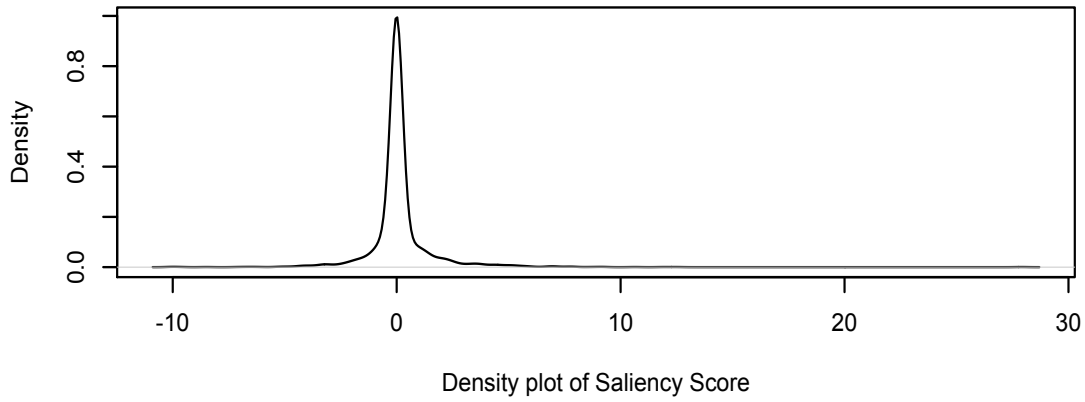
$\hat{\theta}^{(S)}$ v. $\hat{\theta}^{(R)}$

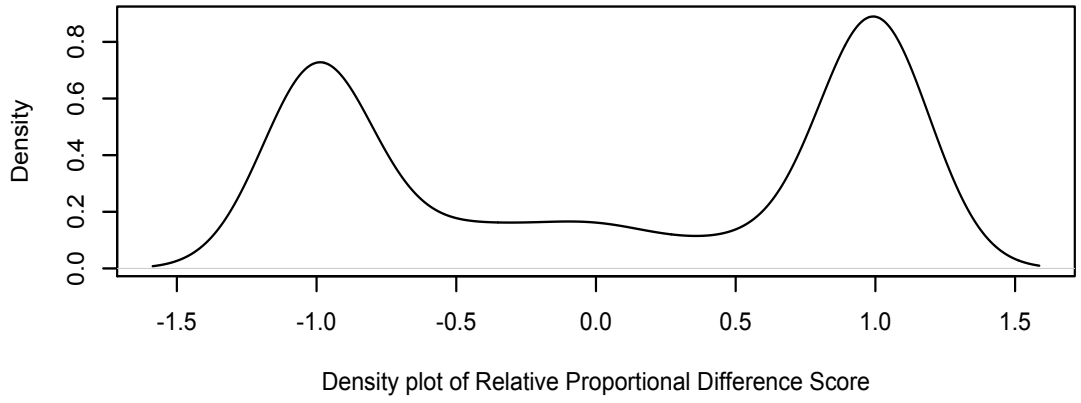


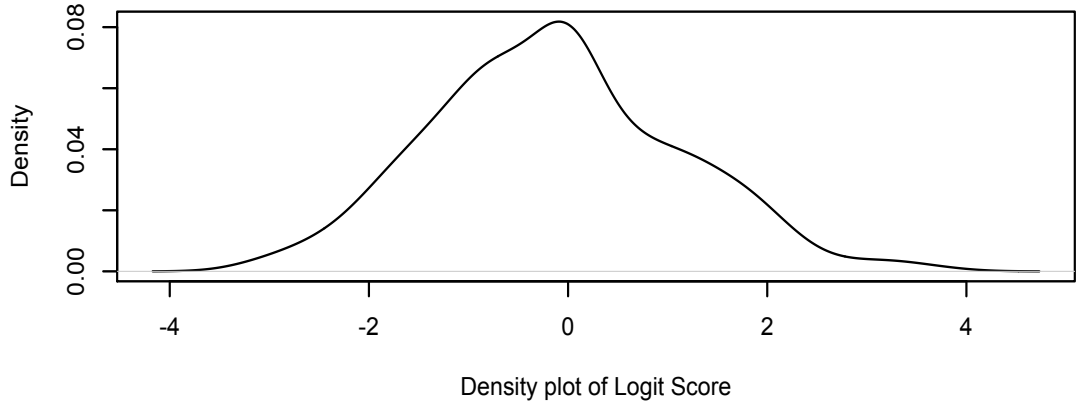
Comparing scales

Protectionism

distributions







Dictionary false positives

```
library("quanteda", quietly = TRUE, warn.conflicts = FALSE, verb

## Package version:  1.3.4
## Parallel computing:  2 of 8 threads used.
## See https://quanteda.io for tutorials and examples.

thedfm <- dfm(data_corpus_irishbudget2010, dictionary = data_dic
              verbose = TRUE)

## Creating a dfm from a corpus input...
##    ...  lowercasing
##    ...  found 14 documents, 5,140 features
##    ...  applying a dictionary consisting of 4 keys
##    ...  created a 14 x 4 sparse dfm
##    ...  complete.
## Elapsed time:  0.413 seconds.
```

Looks ok...

```
head(thedfm)

## Document-feature matrix of: 6 documents, 4 features (0% spars
## 6 x 4 sparse Matrix of class "dfm"
##                                      features
## docs                                  negative positive neg_posi
##   2010_BUDGET_01_Brian_Lenihan_FF          188      397
##   2010_BUDGET_02_Richard_Bruton_FG         163      147
##   2010_BUDGET_03_Joan_Burton_LAB           225      266
##   2010_BUDGET_04_Arthur_Morgan_SF          260      249
##   2010_BUDGET_05_Brian_Cowen_FF            150      368
##   2010_BUDGET_06_Enda_Kenny_FG             104      146
##                                      features
## docs                                  neg_negative
##   2010_BUDGET_01_Brian_Lenihan_FF                1
##   2010_BUDGET_02_Richard_Bruton_FG               2
##   2010_BUDGET_03_Joan_Burton_LAB                 3
##   2010_BUDGET_04_Arthur_Morgan_SF                2
##   2010_BUDGET_05_Brian_Cowen_FF                  2
##   2010_BUDGET_06_Enda_Kenny_FG                   3
```

...but isn't

```
data_dictionary_LSD2015$negative[1500:1510]

##  [1] "invas*"         "invisib*"       "involuntary*"  "irascib
##  [5] "irascibility*" "irat*"          "ire*"          "irk*"
##  [9] "ironic*"        "irony*"         "irrational*"

kwic(data_corpus_irishbudget2010, "ire*", window = 2) %>% head()

##
##  [2010_BUDGET_01_Brian_Lenihan_FF, 1164] people reaffirmed |
##  [2010_BUDGET_01_Brian_Lenihan_FF, 1199]        support to |
##  [2010_BUDGET_01_Brian_Lenihan_FF, 1266]         prices in |
##  [2010_BUDGET_01_Brian_Lenihan_FF, 2594]     from Northern |
##  [2010_BUDGET_01_Brian_Lenihan_FF, 2697]         regard to |
##  [2010_BUDGET_01_Brian_Lenihan_FF, 4265]   fiscal problems |
##
##  place at
##  in the
##  are now
##  where lower
```

Regular expressions

- an expanded version of the "glob" matching implemented in most command line interpreters, i.e.
  - `*` matches zero or more characters
  - `?` matches any one character (and in some environments, zero trailing characters)
  - `[]` may match any characters within a range inside the brackets
- a much more powerful version are *regular expressions*, which also exist in several (slightly) different versions
- R has both the POSIX 1003.2 and the Perl Compatible Regular Expressions implemented, see `?regex`
- Additional materials:
  - great cheat sheet
  - useful tutorial and reference

Incorporating parts of speech

```
library("spacyr")

txt <- c(d1 = "The democratic United Nations called its members
         d2 = "The vote was to sanction the Democratic People's

spacy_initialize()

## Found 'spacy_condaenv'.  spacyr will use this
environment
## successfully initialized (spaCy Version:  2.0.10,
language model:  en)
## (python options:  type = "condaenv", value =
"spacy_condaenv")
```

Incorporating parts of speech

```
txtparsed <- spacy_parse(txt)
head(txtparsed)
```

```
##   doc_id sentence_id token_id      token      lemma   pos ent
## 1     d1           1        1        The        the   DET
## 2     d1           1        2 democratic democratic   ADJ
## 3     d1           1        3     United     united PROPN  OR
## 4     d1           1        4    Nations    nations PROPN  OR
## 5     d1           1        5     called       call  VERB
## 6     d1           1        6        its     -PRON-   ADJ
```

Incorporating parts of speech

```
entity_consolidate(txtparsed)

##    doc_id sentence_id token_id
## 1      d1           1        1
## 2      d1           1        2
## 3      d1           1        3
## 4      d1           1        4
## 5      d1           1        5
## 6      d1           1        6
## 7      d1           1        7
## 8      d1           1        8
## 9      d1           1        9
## 10     d1           1       10
## 11     d1           1       11
## 12     d1           1       12
## 13     d2           1        1
## 14     d2           1        2
## 15     d2           1        3
## 16     d2           1        4
## 17     d2           1        5
```

Un

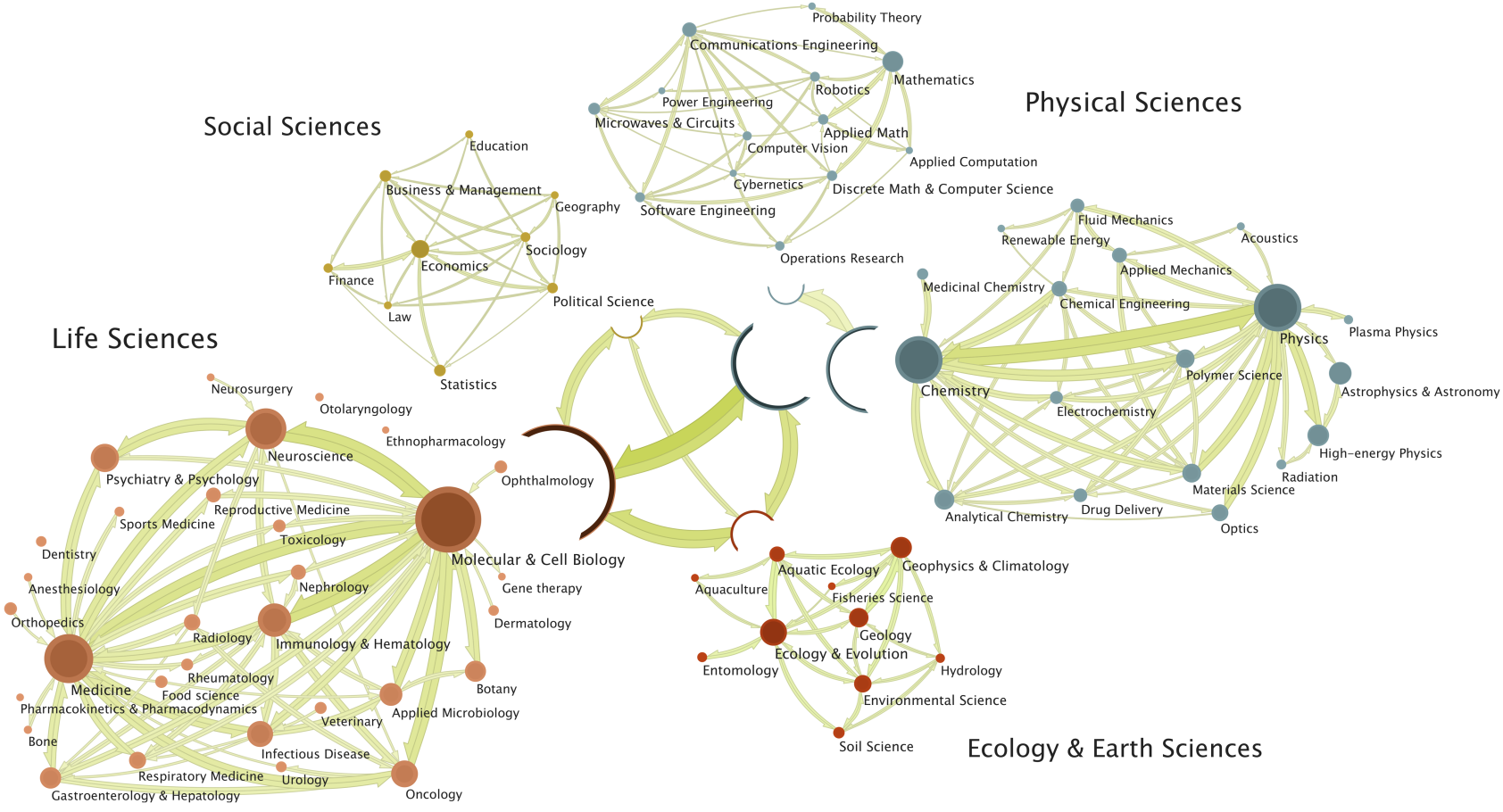
Incorporating parts of speech cont.

```
library("quanteda")
as.tokens(txtparsed) %>%
    dfm()

## Document-feature matrix of: 2 documents, 19 features (34.2% s
## 2 x 19 sparse Matrix of class "dfm"
##     features
## docs the democratic united nations called its members to vote
##   d1   1          1      1       1      1   1       1  1    1
##   d2   2          1      0       0      0   0       0  1    1
##     features
## docs sanction . was people 's republic of korea
##   d1        1 1   0      0  0        0  0     0
##   d2        1 1   1      1  1        1  1     1

entity_consolidate(txtparsed) %>%
    as.tokens(include_pos = "pos") %>%
    dfm()

## Document-feature matrix of: 2 documents, 16 features (40.6% s
## 2 x 16 sparse Matrix of class "dfm"
```

Incorporating parts of speech cont.

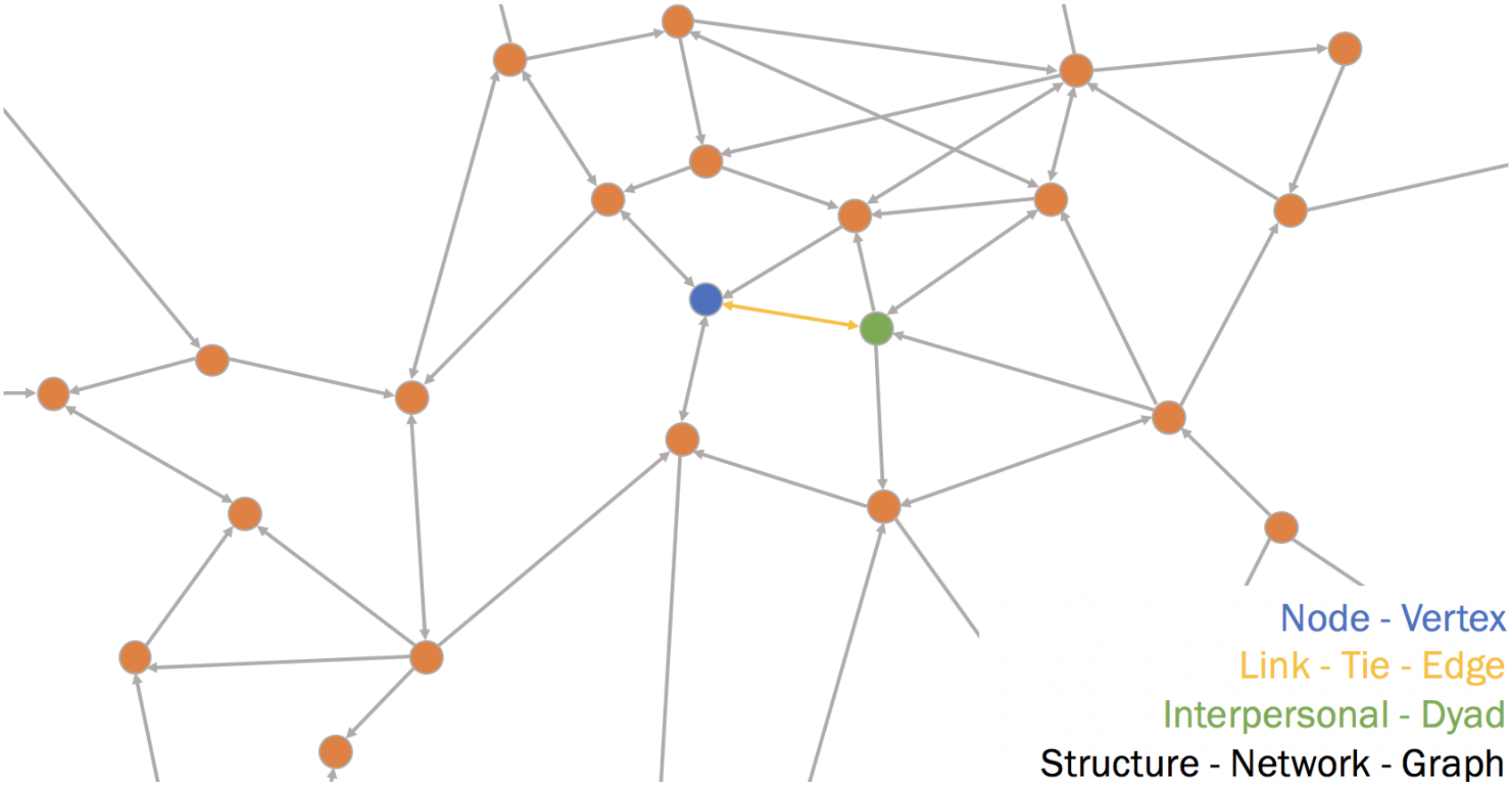
```
entity_consolidate(txtparsed) %>%
    as.tokens(include_pos = "pos") %>%
    dfm()

## Document-feature matrix of: 2 documents, 16 features (40.6% s
## 2 x 16 sparse Matrix of class "dfm"
##     features
## docs the/det democratic/adj united_nations/entity called/verb
##   d1       1              1                       1           1
##   d2       1              0                       0           0
##     features
## docs members/noun to/part vote/verb on/adp a/det sanction/nou
##   d1            1       1         1      1     1
##   d2            0       1         0      0     0
##     features
## docs vote/noun was/verb sanction/verb
##   d1         0        0              0
##   d2         1        1              1
##     features
## docs the_democratic_people_'s_republic_of_korea/entity
```

Networks



Source: Rosvall, M. and Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLOS ONE* 6(4), e18209.

Network basic terminology



Text networks

1. Tokenize the text.
2. Construct a "feature-co-occurrence matrix" (fcm)
3. Use `textplot_network()`

Text networks: Example

```
toks <- corpus_subset(data_corpus_irishbudget2010) %>%
    tokens(remove_punct = TRUE) %>%
    tokens_tolower() %>%
    tokens_remove(stopwords("english"), padding = FALSE)
myfcm <- fcm(toks, context = "window", tri = FALSE)
feat <- names(topfeatures(myfcm, 30))
fcm_select(myfcm, feat, verbose = FALSE) %>%
    textplot_network(min_freq = 0.8)
```

Exam review

- Counts for 75% of the grade
- Released Thursday at 18:30, due Monday 17:00.
- Same technical tools as assignments, and HTML should be submitted via Moodle.
- Content Hints
  - Linear regression
  - (Simple) classification models and computing accuracy-related measures
  - Some bootstrapping
  - Simple text analysis, using dictionaries
  - Clustering