Week 10: Text Analytics

Slava Mikhaylov

PUBLG088 Advanced Quantitative Methods

Week 10 Outline

Natural Language Processing

- Text analytics pipeline
- Key basic concepts
- Strategies for selecting documents
- Defining features
- Parts of speech
- Filtering features
- "stopwords"

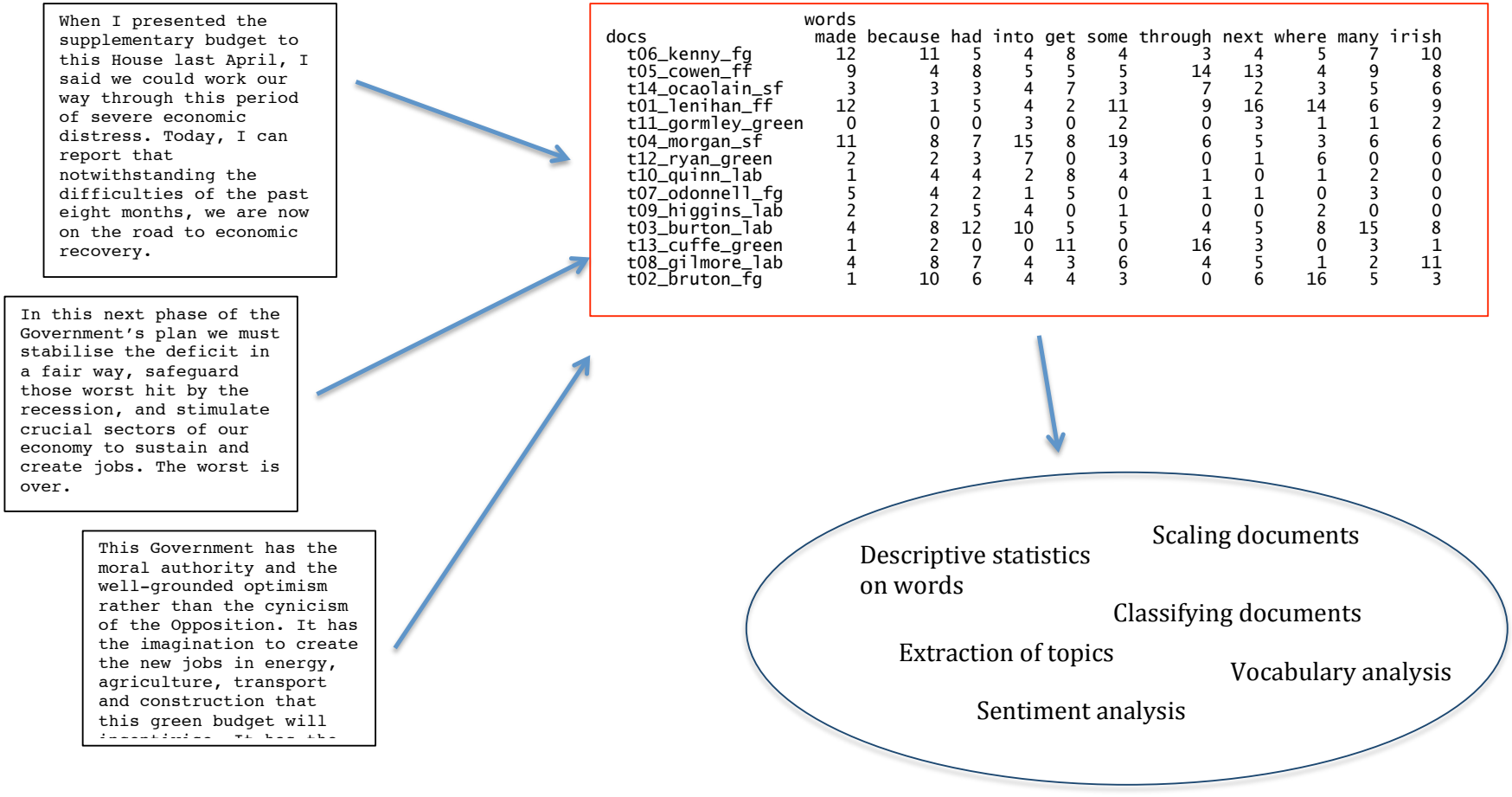
Text Categorization

- Key words in context
- Descriptive text statistics
- Lexical diversity
- Text models

Examples

Natural language processing

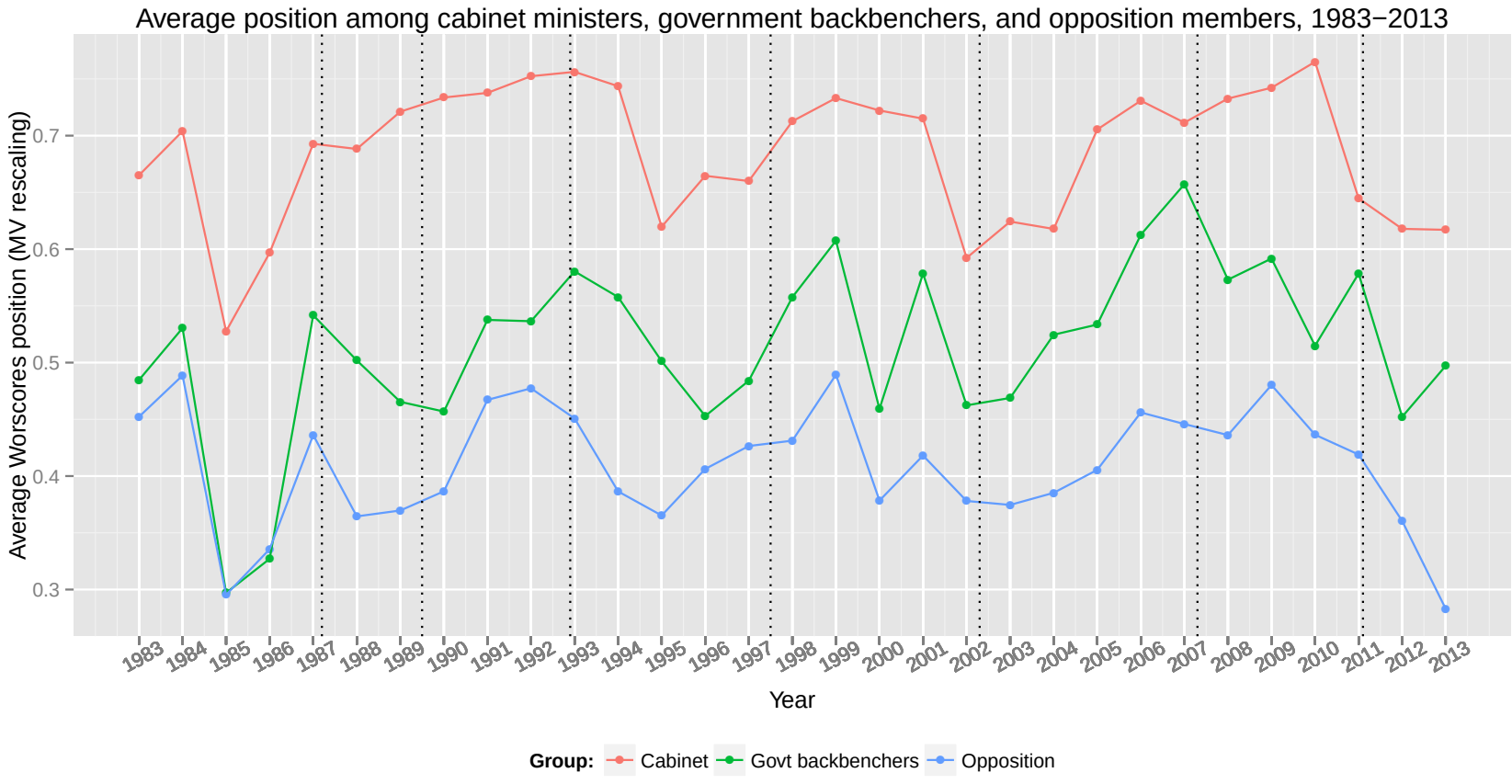
Pipeline: Texts → Feature matrix → Analysis



Pipeline summary

- Conversion of textual features into a quantitative matrix.
- A statistical procedure to extract information from the quantitative matrix.
- Summary and interpretation of the quantitative results.

Example: Text analytics and social science



(from Herzog and Benoit JOP 2015)

This requires assumptions

- That texts represent an observable implication of some underlying characteristic of interest (usually an attribute of the author)
- That texts can be represented through extracting their *features*
  - most common is the bag of words assumption
  - many other possible definitions of "features"
- A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Key features of text analytics

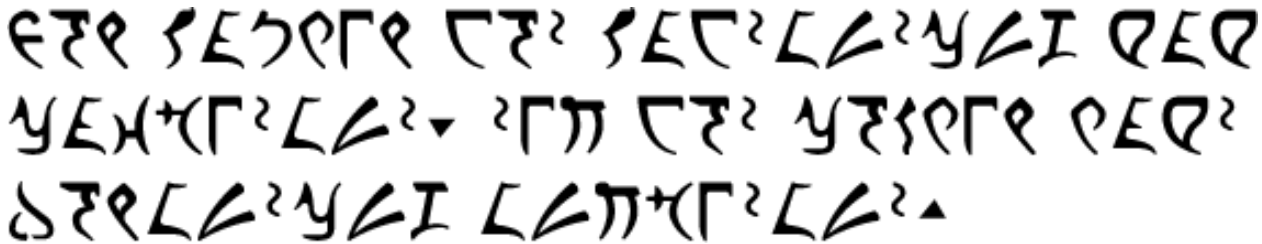
1. Selecting texts: Defining the *corpus*
2. Conversion of texts into a common electronic format
3. Defining documents: deciding what will be the documentary unit of analysis

Key features of text analytics (cont.)

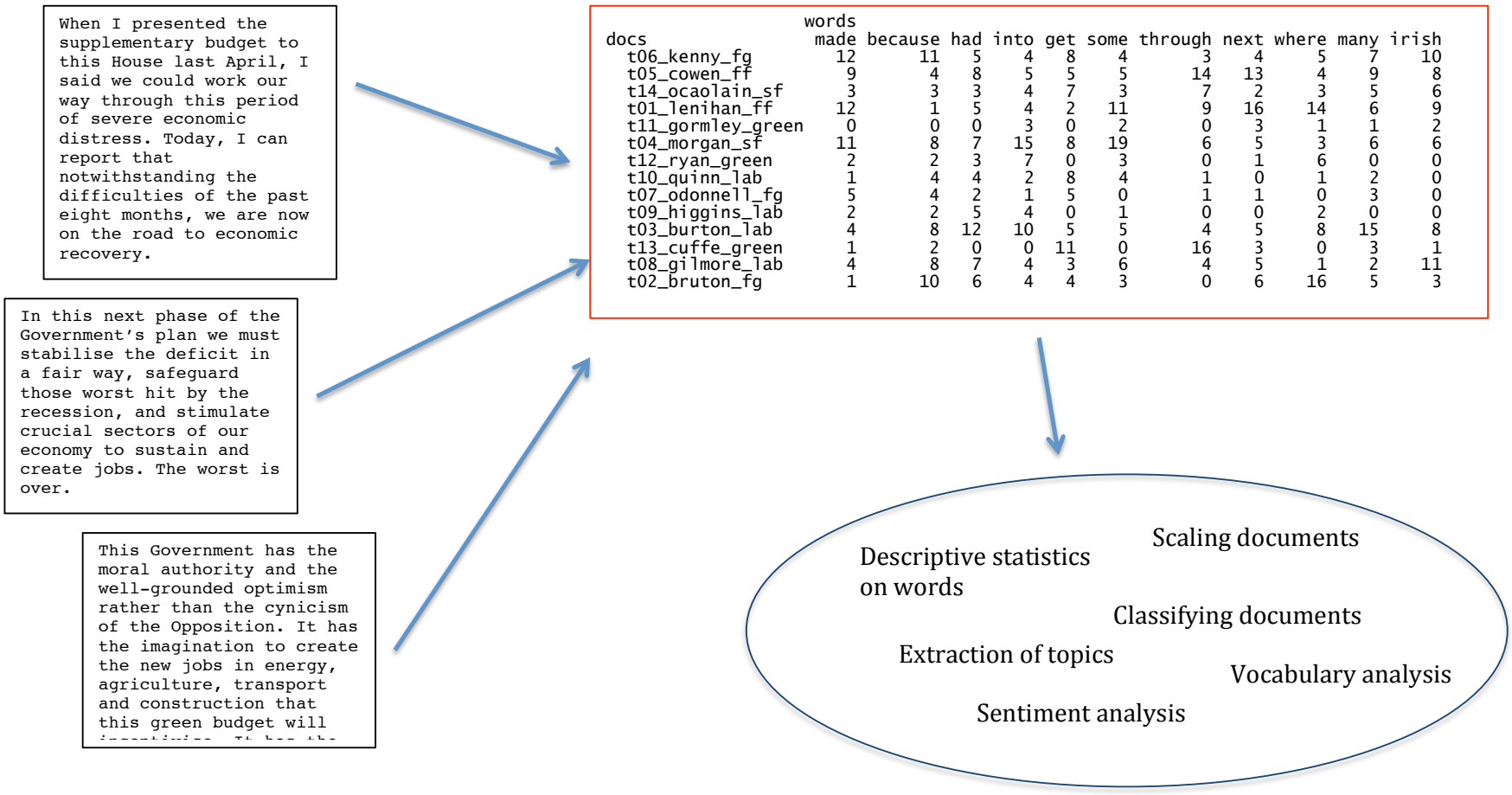
4. Defining features. These can take a variety of forms, including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibily variable length), linguistic features, and more.
5. Conversion of textual features into a data matrix
6. A statistical procedure to extract information from the data matrix
7. Summary and interpretation of the statistical results

Extreme forms of text analytics

- Fully automated technique with minimal human intervention or judgment calls – only with regard to reference text selection
- Methods can "discover" topics with little human supervision
- Language-blind: can analyze anything that occurs with regular patterns (even without knowing what these mean)
- Could potentially work on texts like this:



(See `http://www.kli.org`)


When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.
In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.
This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will
words
docs made because had into get some through next where many irish
t06_kenny_fg 12 11 5 4 8 4 3 4 5 7 10
t05_cowen_ff 9 4 8 5 5 5 14 13 4 9 8
t14_ocaolain_sf 3 3 3 4 7 3 7 2 3 5 6
t01_lenihan_ff 12 1 5 4 2 11 9 16 14 6 9
t11_gormley_green 0 0 0 3 0 2 0 3 1 1 2
t04_morgan_sf 11 8 7 15 8 19 6 5 3 6 6
t12_ryan_green 2 2 3 7 0 3 0 1 6 0 0
t10_quinn_lab 1 4 4 2 8 4 1 0 1 2 0
t07_odonnell_fg 5 4 2 1 5 0 1 1 0 3 0
t09_higgins_lab 2 2 5 4 0 1 0 0 2 0 0
t03_burton_lab 4 8 12 10 5 5 4 5 8 15 8
t13_cuffe_green 1 2 0 0 11 0 16 3 0 3 1
t08_gilmore_lab 4 8 7 4 3 6 4 5 1 2 11
t02_bruton_fg 1 10 6 4 4 3 0 6 16 5 3
Descriptive statistics on words
Scaling documents
Classifying documents
Extraction of topics
Vocabulary analysis
Sentiment analysis

Some key basic concepts

(text) corpus a large and structured set of texts for analysis

types for our purposes, a unique word

tokens any word – so token count is total words

- hapax legomena (or just *hapax*) are types that occur just once

stems words with suffixes removed

lemmas canonical word form (the base form of a word that has the same meaning even when different suffixes (or prefixes) are attached)

keys such as dictionary entries, where the user defines a set of equivalence classes that group different word types

Some more key basic concepts

"key" words	Words selected because of special attributes, meanings, or rates of occurrence
stop words	Words that are designated for exclusion from any analysis of a text
readability	provides estimates of the readability of a text based on word length, syllable length, etc.
complexity	A word is considered "complex" if it contains three syllables or more
diversity	(lexical diversity) A measure of how many types occur per fixed word rate (a normalized vocabulary measure)

Strategies for selecting units of textual analysis

- Words
- *n*-word sequences
- pages
- paragraphs
- Themes
- Natural units (a speech, a poem, a manifesto)
- Key: depends on the research design

Defining Features

- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features
- word segments, especially for languages using compound words, such as German, e.g.
  Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
  (the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef)
  Saunauntensitzer

Defining Features (cont.)

- "word" sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in Chinese

  莎拉波娃现在居住在美国东南部的佛罗里达。今年 4 月 9 日，莎拉波娃在美国第一大城市纽约度过了 1 8 岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

- linguistic features, such as parts of speech
- (if qualitative coding is used) coded or annotated text segments
- linguistic features: parts of speech

Parts of speech

- the Penn "Treebank" is the standard scheme for tagging POS

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Parts of speech (cont.)

- several open-source projects make it possible to tag POS in text, namely Apache's OpenNLP (and R package openNLP wrapper)

```
> s
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
> sprintf("%s/%s", s[a3w], tags)
 [1] "Pierre/NNP"      "Vinken/NNP"      ",/,"             "61/CD"
 [5] "years/NNS"       "old/JJ"          ",/,"             "will/MD"
 [9] "join/VB"         "the/DT"          "board/NN"        "as/IN"
[13] "a/DT"            "nonexecutive/JJ" "director/NN"     "Nov./NNP"
[17] "29/CD"           "./."             "Mr./NNP"         "Vinken/NNP"
[21] "is/VBZ"          "chairman/NN"     "of/IN"           "Elsevier/NNP"
[25] "N.V./NNP"        ",/,"             "the/DT"          "Dutch/JJ"
[29] "publishing/NN"   "group/NN"        "./."
```

Strategies for feature selection

- document frequency How many documents in which a term appears
- term frequency How many times does the term appear in the corpus
- deliberate disregard Use of "stop words": words excluded because they represent linguistic connectors of no substantive content
- purposive selection Use of a *dictionary* of words or phrases
- declared equivalency classes Non-exclusive synonyms, what I call a *thesaurus*

Common English stop words

```
a, able, about, across, after, all, almost, also, am, among,
an, and, any, are, as, at, be, because, been, but, by, can,
cannot, could, dear, did, do, does, either, else, ever,
every, for, from, get, got, had, has, have, he, her, hers,
him, his, how, however, I, if, in, into, is, it, its, just,
least, let, like, likely, may, me, might, most, must, my,
neither, no, nor, not, of, off, often, on, only, or, other,
our, own, rather, said, say, says, she, should, since, so,
some, than, that, the, their, them, then, there, these,
they, this, tis, to, too, twas, us, wants, was, we, were,
what, when, where, which, while, who, whom, why, will, with,
would, yet, you, your
```

- But no list should be considered universal

A more comprehensive list of stop words

as, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, aint, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, arent, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, cmon, cs, came, can, cant, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldnt, course, currently, definitely, described, despite, did, didnt, different, do, does, doesnt, doing, dont, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadnt, happens, hardly, has, hasnt, have, havent, having, he, hes, hello, help, hence, her, here, heres, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, id, ill, im, ive, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isnt, it, itd, itll, its, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, lets, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldnt, since, six, so, some, somebody,

Stemming words

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

example: `produc` from
`production`, `producer`, `produce`, `produces`, `produced`

Text Categorization

Exploring Texts: Key Words in Context

KWIC *Key words in context* Refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.

lime (14)

79[C.10]	4	/Which was builded of **lime** and sand;/Until they came to
247A.6	4	/That was well biggit with **lime** and stane.
303A.1	2	bower,/Well built wi **lime** and stane,/And Willie came
247A.9	2	/That was well biggit wi **lime** and stane,/Nor has he stoln
305A.2	1	a castell biggit with **lime** and stane,/O gin it stands not
305A.71	2	is my awin,/I biggit it wi **lime** and stane;/The Tinnies and
79[C.10]	6	/Which was builded with **lime** and stone.
305A.30	1	a prittie castell of **lime** and stone,/O gif it stands not
108.15	2	/W*hi*ch was made both of **lime** and stone,/Shee tooke him by
175A.33	2	castle then,/Was made of **lime** and stone;/The vttermost
178[H.2]	2	near by,/Well built with **lime** and stone;/There is a lady
178F.18	2	built with stone and **lime**!/But far mair pittie on Lady
178G.35	2	was biggit wi stane and **lime**!/But far mair pity o Lady
2D.16	1	big a cart o stane and **lime**,/Gar Robin Redbreast trail it

Irish Budget Speeches KIWC in quanteda

```
library(quanteda)

##
## Attaching package:  'quanteda'
##
## The following object is masked from 'package:base':
##
##    sample

data("ie2010Corpus")
iebudgets2010 <- subset(ie2010Corpus, year==2010)
kwic(iebudgets2010, "christmas")

##                                                                  prewo
##      [2010_BUDGET_02_Richard_Bruton_FG, 628]                and to see out th
##        [2010_BUDGET_03_Joan_Burton_LAB, 371]             to suggest titles for
##        [2010_BUDGET_03_Joan_Burton_LAB, 379]       Fianna Fils hit single for
##        [2010_BUDGET_03_Joan_Burton_LAB, 922]       women will say goodbye aft
##       [2010_BUDGET_03_Joan_Burton_LAB, 1518]          in single golf clubs th
##       [2010_BUDGET_03_Joan_Burton_LAB, 1726] Community faking its message th
##       [2010_BUDGET_03_Joan_Burton_LAB, 3159]         bags. In previous years
##       [2010_BUDGET_04_Arthur_Morgan_SF, 346]               204 per week or th
##      [2010_BUDGET_04_Arthur_Morgan_SF, 3239]   to social welfare payments th
##      [2010_BUDGET_04_Arthur_Morgan_SF, 3244]         Christmas. The loss of t
##      [2010_BUDGET_04_Arthur_Morgan_SF, 3272]     streets on Santa presents a
```

Basic descriptive summaries of text

Readability statistics Use a combination of syllables and sentence length to indicate "readability" in terms of complexity

Vocabulary diversity (At its simplest) involves measuring a *type-to-token ratio* (TTR) where unique words are types and the total words are tokens

Word (relative) frequency

Theme (relative) frequency

Length in characters, words, lines, sentences, paragraphs, pages, sections, chapters, etc.

Simple descriptive table about texts: Describe your data!

Speaker	Party	Tokens	Types
Brian Cowen	FF	5,842	1,466
Brian Lenihan	FF	7,737	1,644
Ciaran Cuffe	Green	1,141	421
John Gormley (Edited)	Green	919	361
John Gormley (Full)	Green	2,998	868
Eamon Ryan	Green	1,513	481
Richard Bruton	FG	4,043	947
Enda Kenny	FG	3,863	1,055
Kieran ODonnell	FG	2,054	609
Joan Burton	LAB	5,728	1,471
Eamon Gilmore	LAB	3,780	1,082
Michael Higgins	LAB	1,139	437
Ruairi Quinn	LAB	1,182	413
Arthur Morgan	SF	6,448	1,452
Caoimhghin O'Caolain	SF	3,629	1,035
All Texts		49,019	4,840
Min		919	361
Max		7,737	1,644
Median		3,704	991
Hapaxes with Gormley Edited		67	
Hapaxes with Gormley Full Speech		69	

Lexical Diversity

- Basic measure is the TTR: Type-to-Token ratio
- Problem: This is very sensitive to overall document length, as shorter texts may exhibit fewer word repetitions
- Special problem: length may relate to the introdution of additional subjects, which will also increase richness

Vocabulary diversity and corpus length

- In natural language text, the rate at which new types appear is very high at first, but diminishes with added tokens

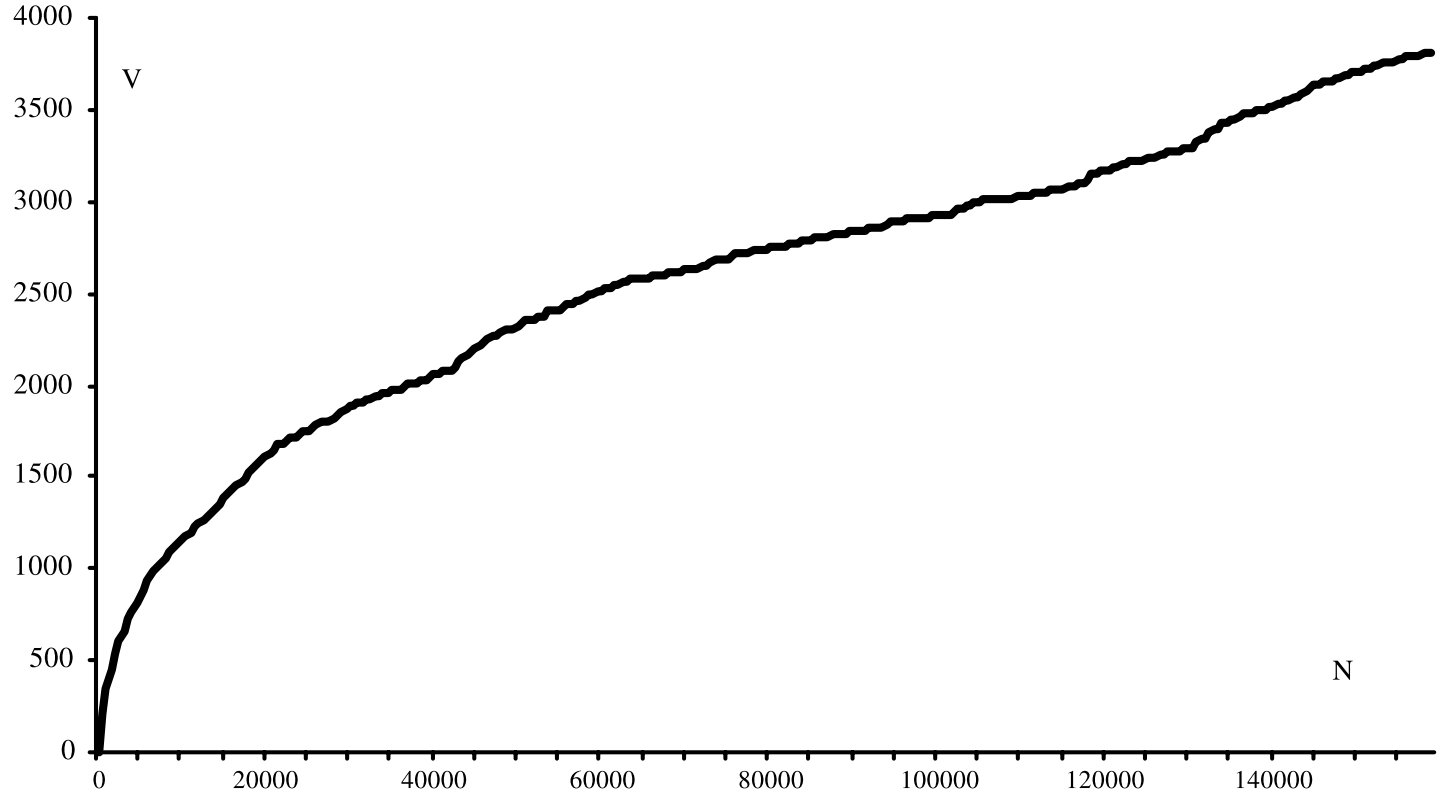


Fig. 1. Chart of vocabulary growth in the tragedies of Racine (chronological order, 500 token intervals).

Vocabulary Diversity Example

- Variations use automated segmentation – here approximately 500 words in a corpus of serialized, concatenated weekly addresses by de Gaulle (from Labbé et. al. 2004)
- While most were written, during the period of December 1965 these were more spontaneous press conferences

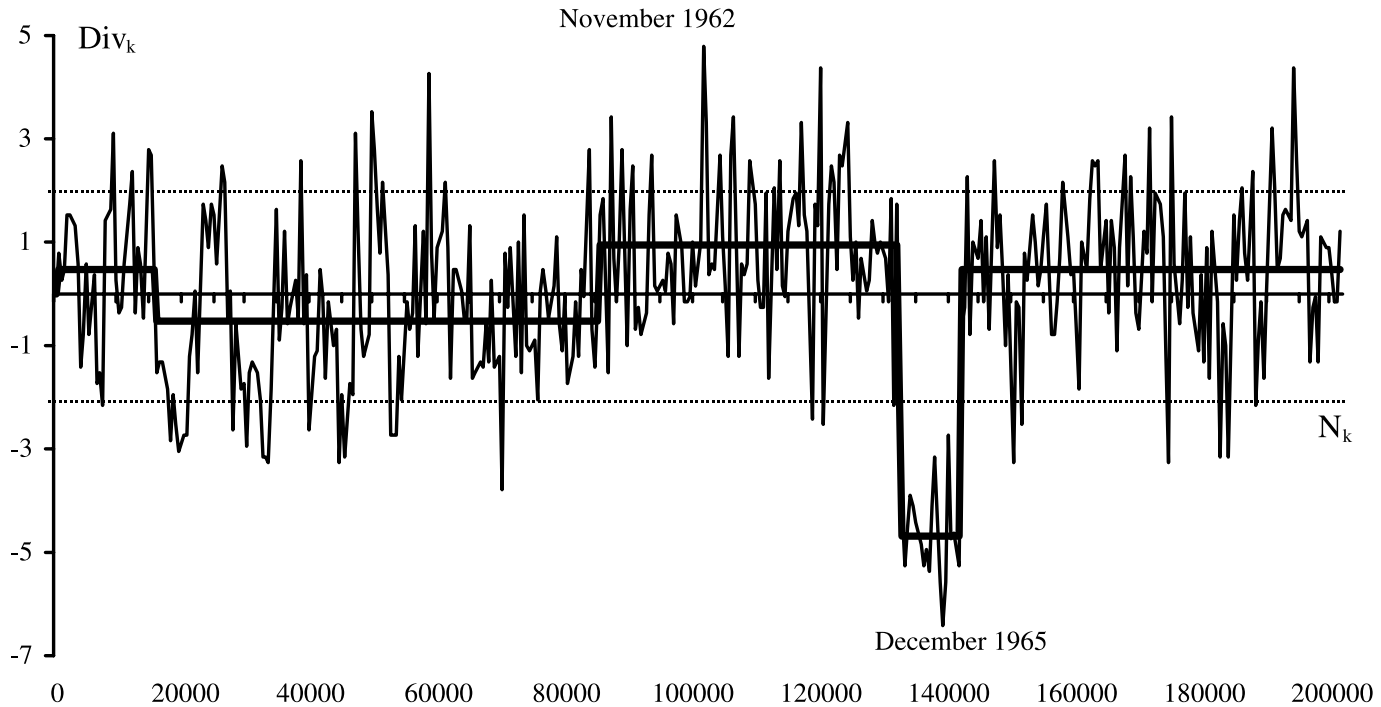


Fig. 8. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958–April 1969).

Text models

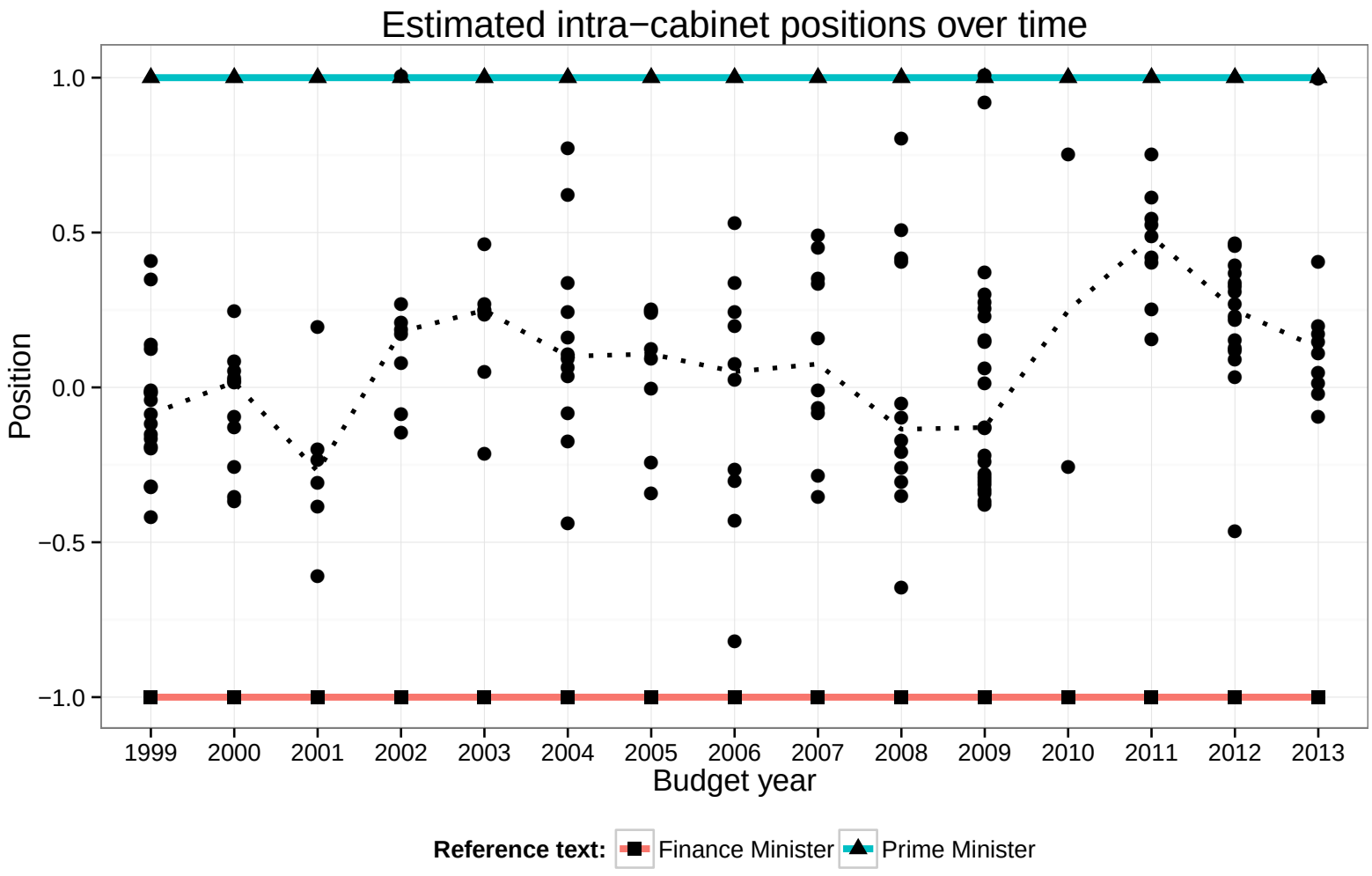
- You can apply most of the methods we covered in the course so far to model text or retrieve necessary information from text. E.g.,
  - Naive Bayes & kNN
  - Logistic regression
  - Support vector machines
  - k-Means
  - Clustering
  - Correspondence analysis
  - Probabalistic topic modeling

Practical side

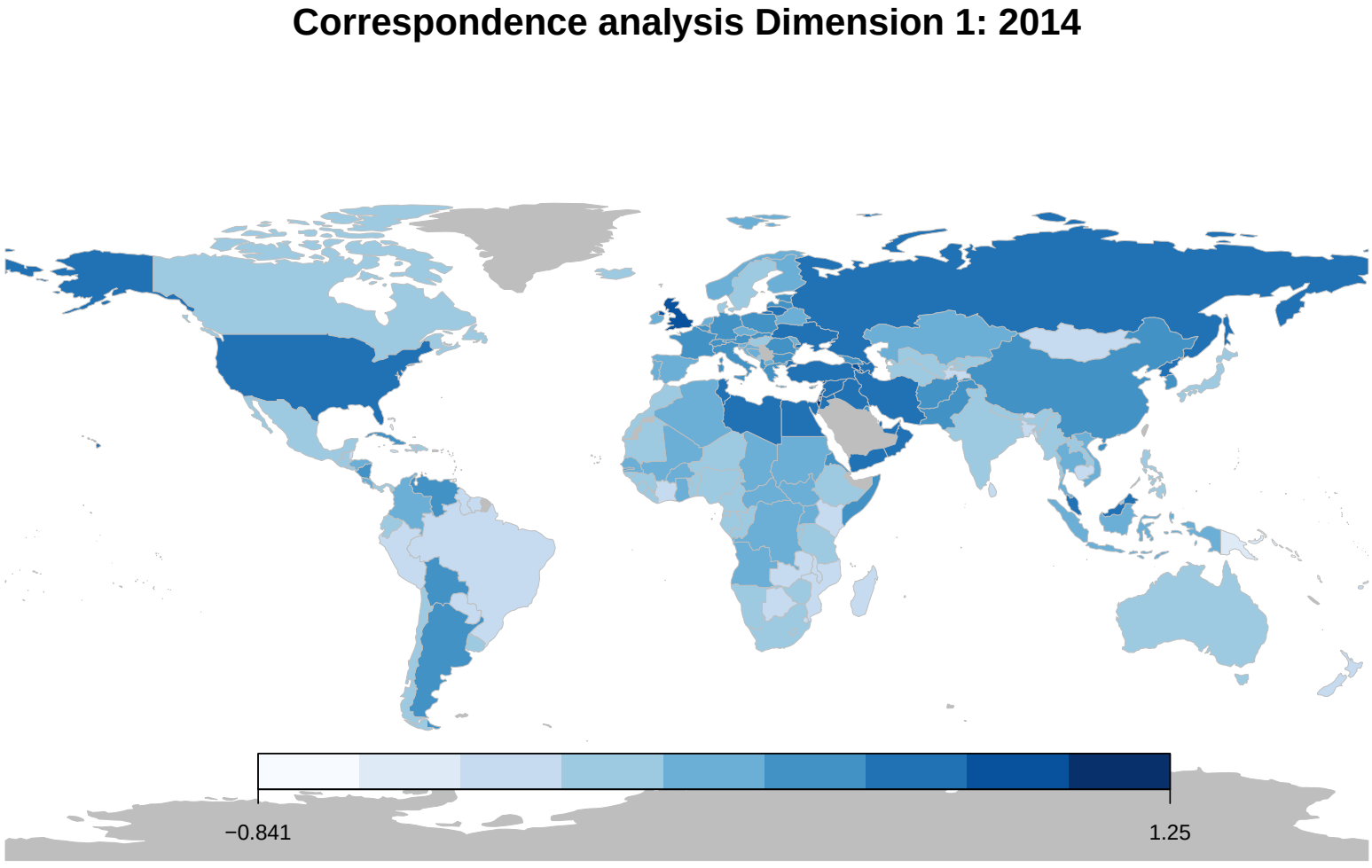
- Most of the things can be done in Quanteda package or taking text transformation output from Quanteda and then using other models we covered earlier.
- `https://github.com/kbenoit/quanteda`
- `https://github.com/kbenoit/ITAUR`

Examples

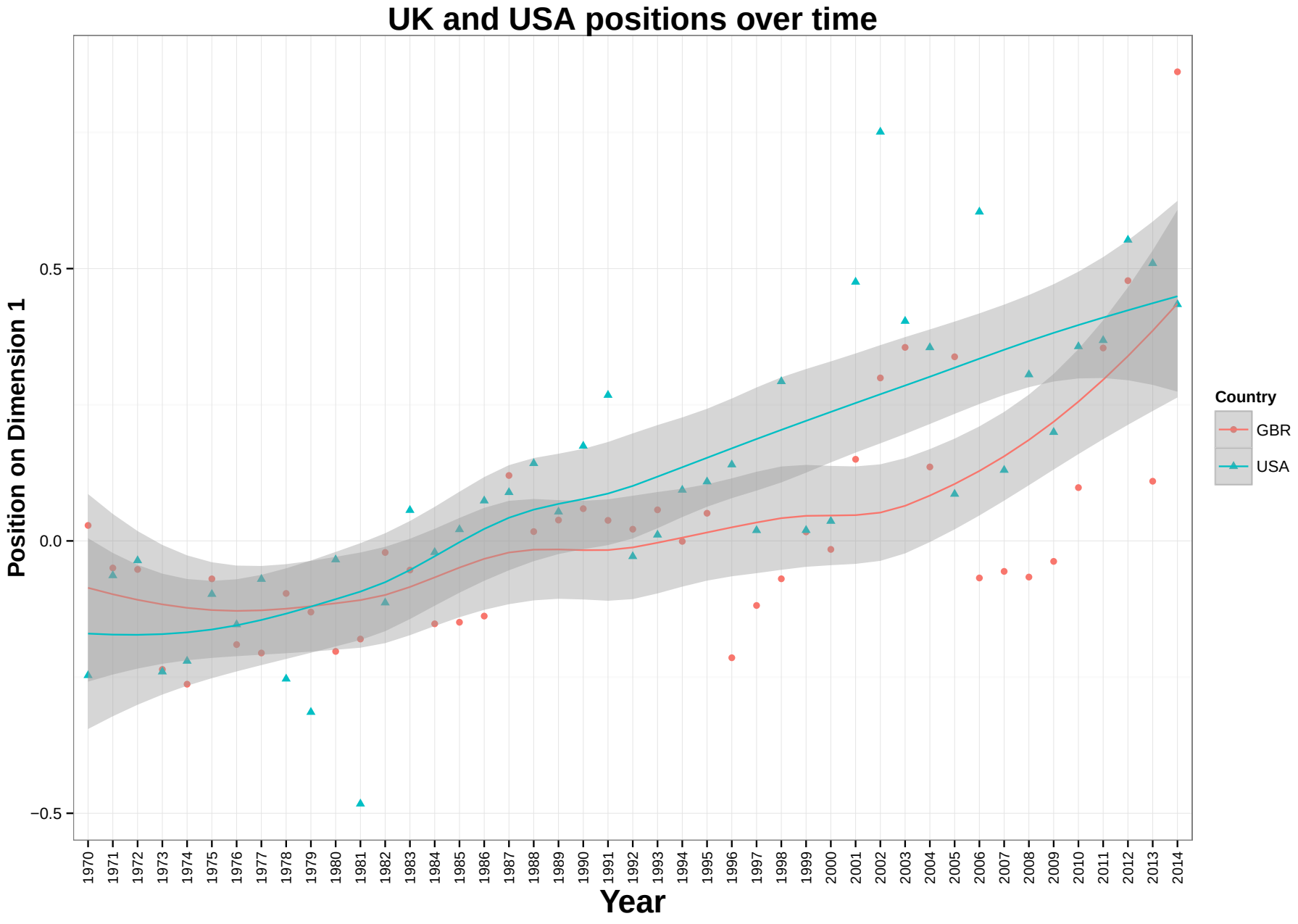
Intra-cabinet politics in Ireland



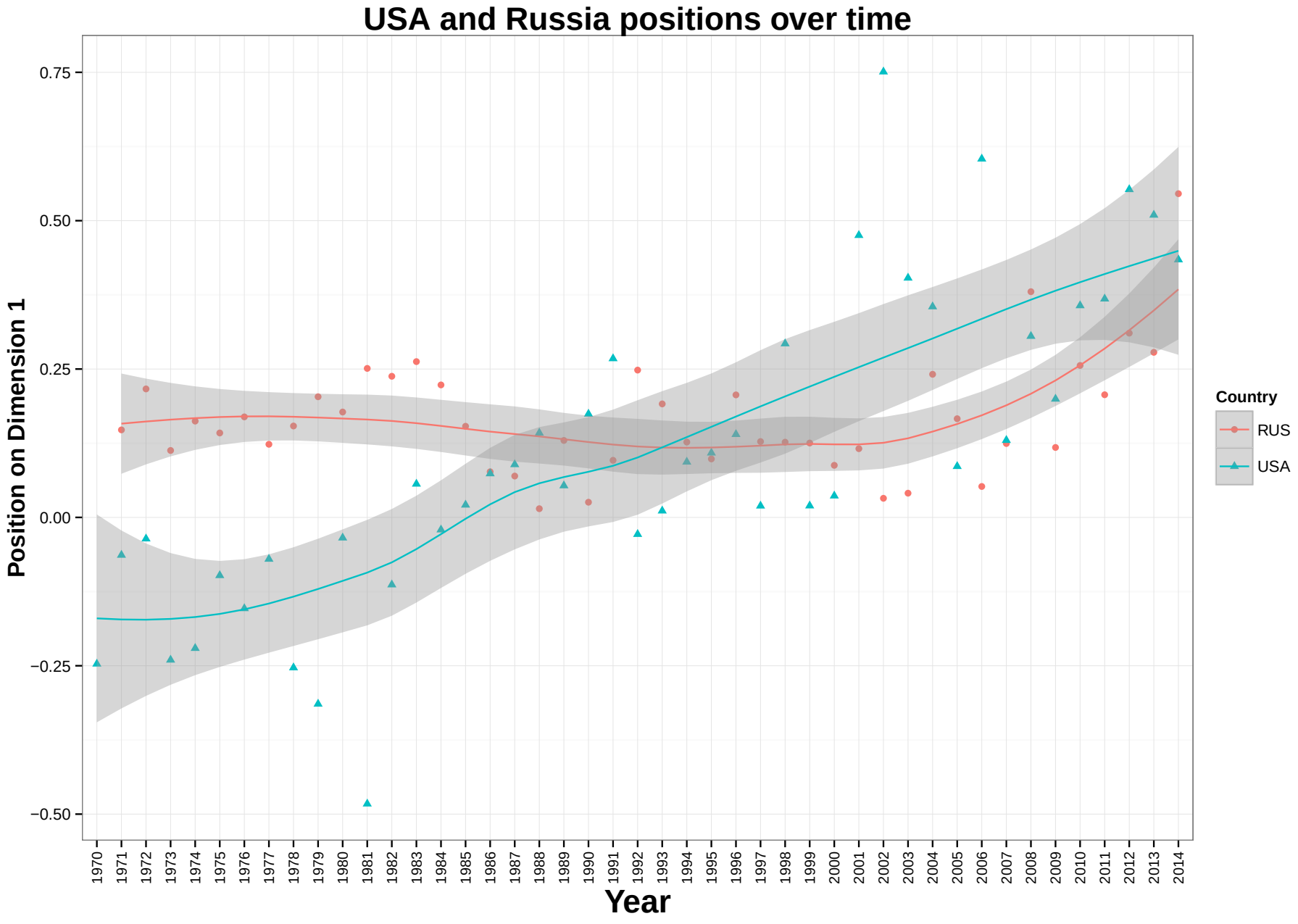
Foreign policy preferences from speeches



Foreign policy preferences from speeches



Foreign policy preferences from speeches



Semantic networks in UK House of Commons