Introduction to the Text Mining and Data Mining

Week 07

The basic concept of Text Mining

The corpus is a large, organised collection of texts for analysis.

The document represents each of the corpus's units.

Types mean a unique word.

Tokens are any word, so the token count is the number of words.

e.g. A corpus is a set of documents. This is the second document in the corpus.

It consists of two documents, each of which is a sentence. The first document contains **‘six types’** and **‘seven tokens’**. The second has **‘seven types’** and **‘eight tokens.’**

The basic concept of Text Mining

stems

words with suffixes removed (using a set of rules)

lemmas

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

WORD	win	winning	wins	won	winner
STEM	win	win	win	won	winner
LEMMNA	win	win	win	win	win

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

"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

Document Term Matrices

D1 = "I like databases"
D2 = "I dislike databases",

	I	like	dislike	databases
D1	1	1	0	1
D2	1	0	1	1

- A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents.
- A vector of words represents each document
- Which shows which documents contain which terms and how many times they appear. Note that, unlike representing a document as just a token-count list, the document-term matrix includes all terms in the corpus (i.e. the corpus vocabulary), which is why there are zero counts for terms in the corpus which do not also occur in a specific document.

The process of Text Mining

1. Remove capitalization, punctuation
2. Segment into words, characters, morphemes
3. Discard Order (”Bag of Words” Assumption)
4. Discard stop words
5. Create Equivalence Class: stem, lemmatize, or synonym
6. Discard less useful features
7. Other reduction, specialization

High Frequency does not mean good.

1. High Frequency is useful; if sugar appears near an apricot, that's useful information.
2. But overly frequent words like "the", "it", or "they" are not very informative about the content
3. Some terms carry more information about the contents
4. Need a function that resolves this frequency paradox!

How to weight text data

term frequency Some approaches trim very low-frequency words.

Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

document frequency Could eliminate words appearing in few documents

inverse document frequency Conversely, could weight words more that appear in the most documents

tf-idf a combination of term frequency and inverse document frequency, common method for feature weighting

How to weight text data

term frequency Some approaches trim very low-frequency words.

Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

document frequency Could eliminate words appearing in few documents

inverse document frequency Conversely, could weight words more that appear in the most documents

tf-idf a combination of term frequency and inverse document frequency, common method for feature weighting

How to weight text data

term frequency Some approaches trim very low-frequency words.

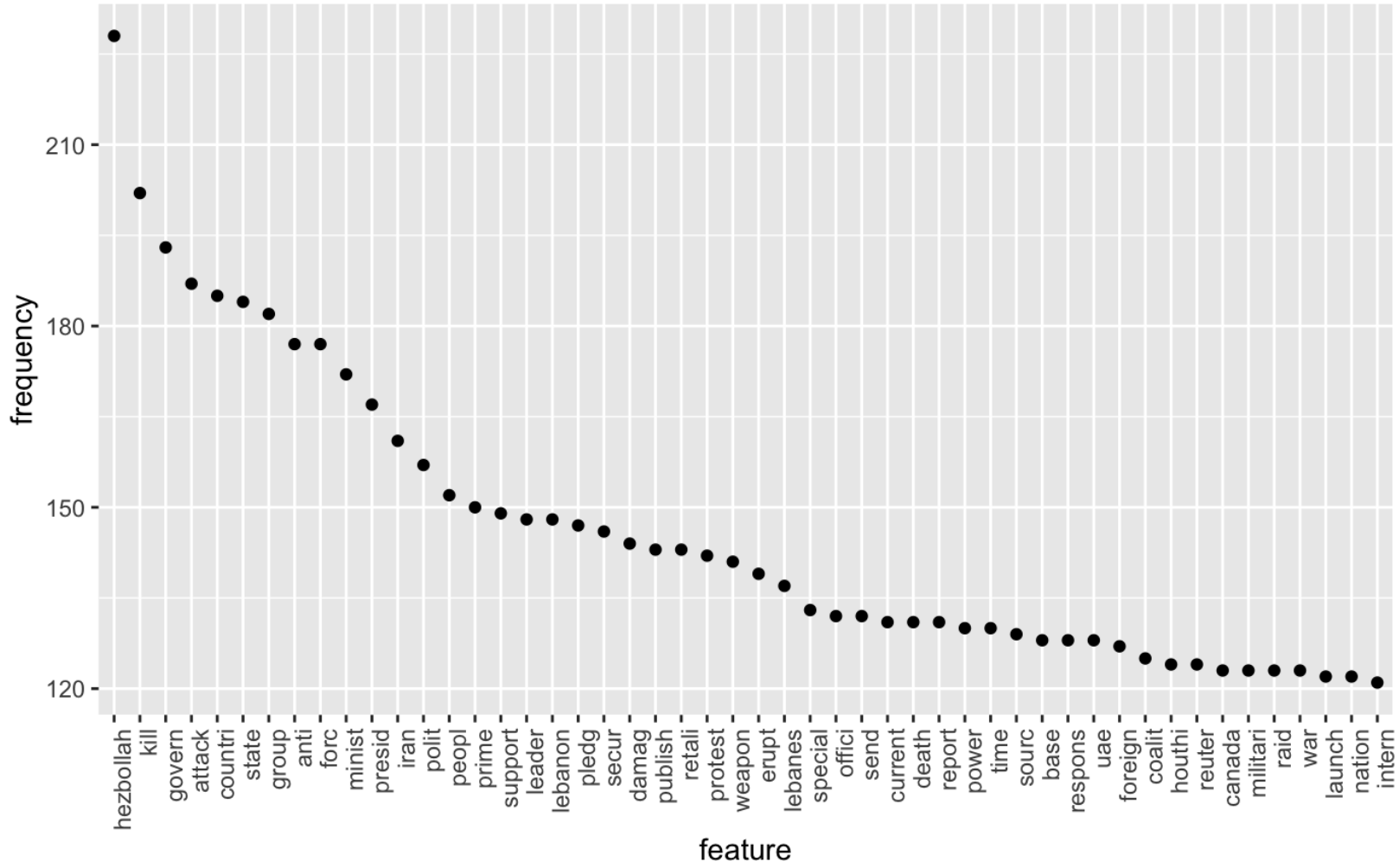
Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

document frequency Could eliminate words appearing in few documents

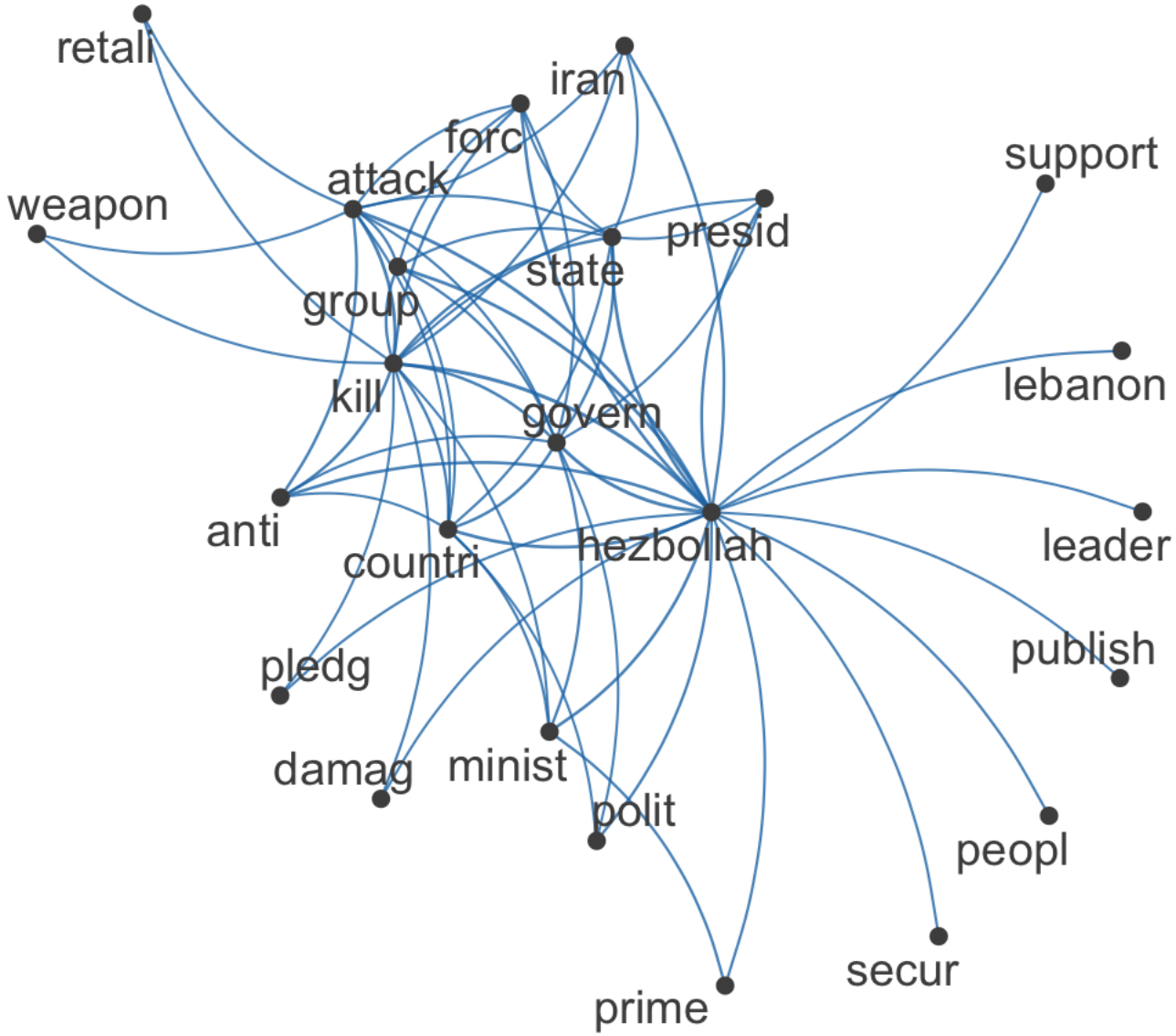
inverse document frequency Conversely, could weight words more that appear in the most documents

tf-idf a combination of term frequency and inverse document frequency, common method for feature weighting

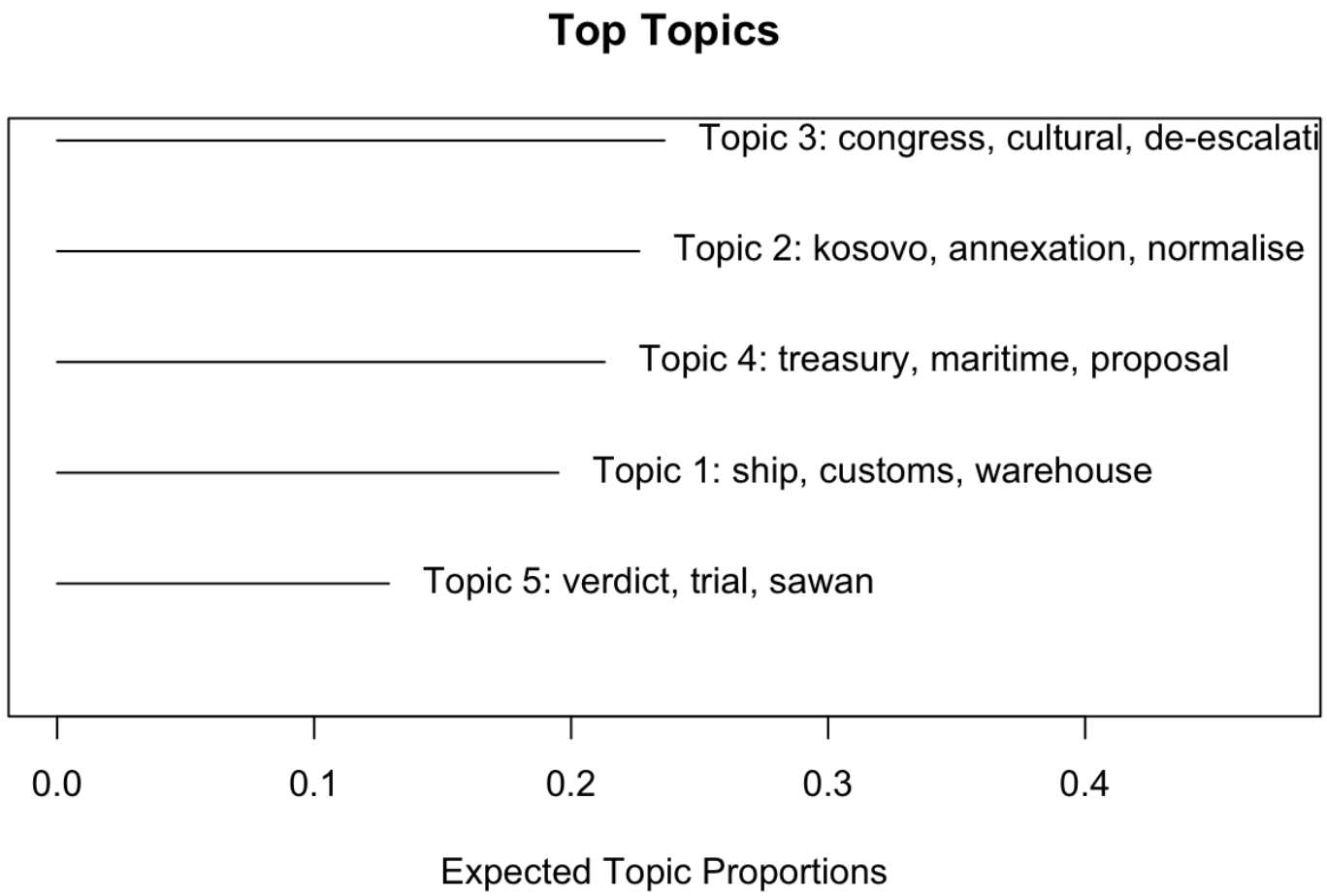
Frequency


frequency
210
180
150
120
hezbollah
kill
govern
attack
countri
state
group
anti
forc
minist
presid
iran
polit
peopl
prime
support
leader
lebanon
pledg
secur
damag
publish
retali
protest
weapon
erupt
lebanes
special
offici
send
current
death
report
power
time
sourc
base
respons
uae
foreign
coalit
houthi
reuter
canada
militari
raid
war
launch
nation
intern
feature

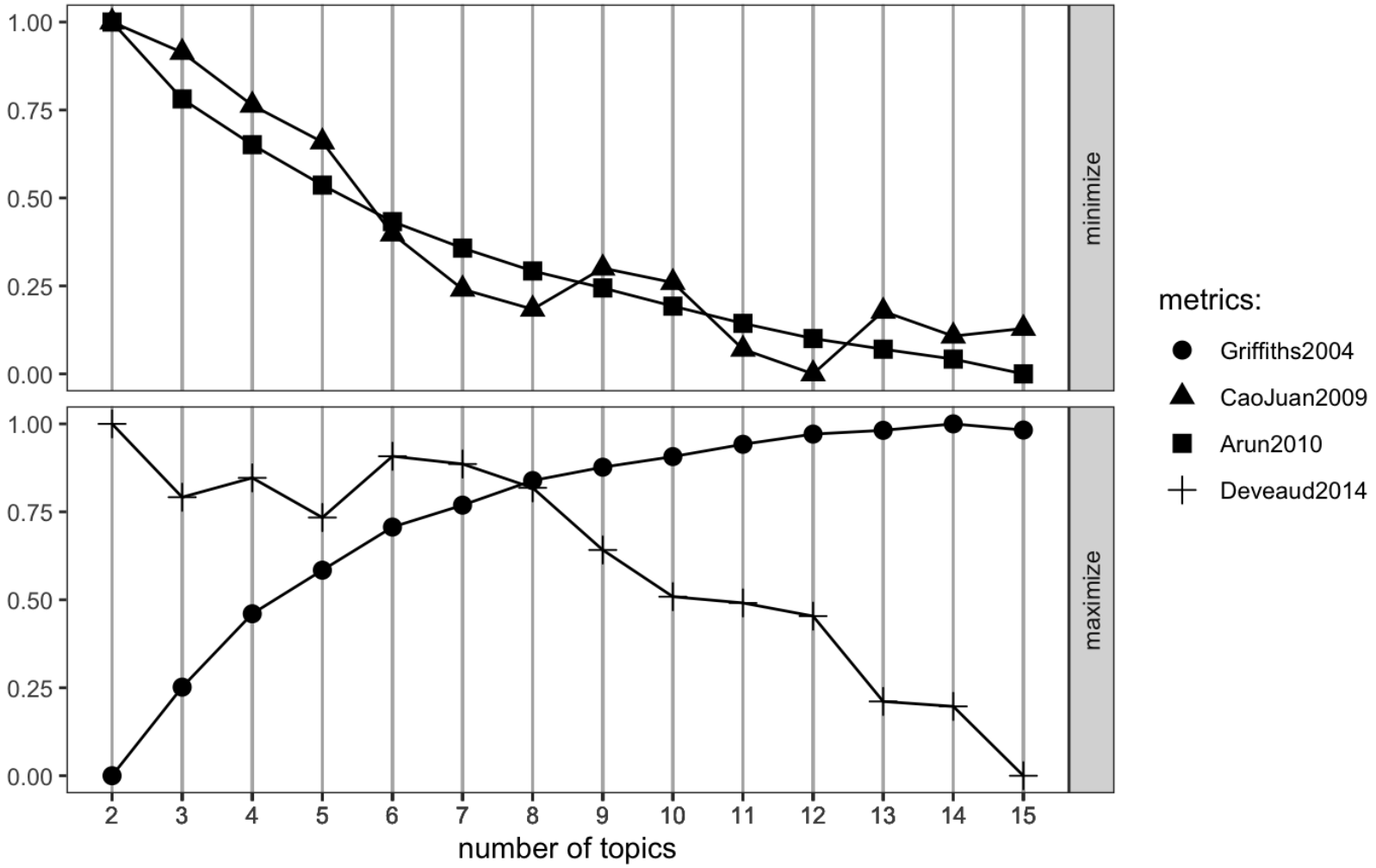
Co-occurrence Network



Topic Analysis



Topic Analysis



Next Week

We don’t have a lecture next week

Week 8: Mid Terms