

Quantitative Text Analysis

Meeting 2

Petro Tolochko

Text Preprocessing

- Texts are *highly* dimensional
- When possible, it is nice to reduce this dimensionality
- Ideally, without losing too much information

Danny & Spirling, 2018

Text Preprocessing For Unsupervised Learning

- Punctuation
- Numbers
- Lowercasing
- Stemming
- Stop-words
- N-grams
- Removal of words by frequency

Punctuation / Numbers / Lowercasing

- Fairly straightforward
- Often we don't care about punctuation and/or numbers – so, might be better to remove them
- We probably do care about the letter case
 - To what extent?
 - Reduction in dimensions might be worth the reduction in accuracy
 - When would letter case be (un)important?

Stemming / Lemmatization

- A stem is the part of the word responsible for lexical meaning
- A stem is invariable part of the word under inflection
- “wait” is a stem of:
 - “Waiting”
 - “Waited”
 - “Waits”
- A lemma is the base / “original” part of the word
- Both are useful for dimension reduction and often produce similar results

Stop Words

- Words that are filtered out before the analysis begins
- Could be any type of words that you do not want in the analysis
- Usually, function words are used as stop words (FORESHADOWING...)
 - “The”
 - “Is”
 - “I”
 - “That”
 - etc.
- Domain-specific words are also often excluded from the analysis
- E.g., “Global Warming” in the corpus of texts about Global Warming

N-grams

- So far, we've only looked at "unigrams" – individual words
- Texts can be broken down into any n-gram sequences
- "I love ice-cream and bananas"
 - "I" "love" "ice-cream" "and" "bananas"
 - "I love" "love ice-cream" "ice-cream and" "and bananas"
 - 3-grams?

Removal of terms by frequency

- Further removal of dimensionality can be achieved by removing either very frequent or very infrequent terms
- If they are very frequent, they probably don't carry much discriminating information for our analysis (think stopwords)
- If they are very infrequent, they probably carry a lot of discriminating information, but very low statistical power

Danny & Spirling, 2018

Text Preprocessing For Unsupervised Learning

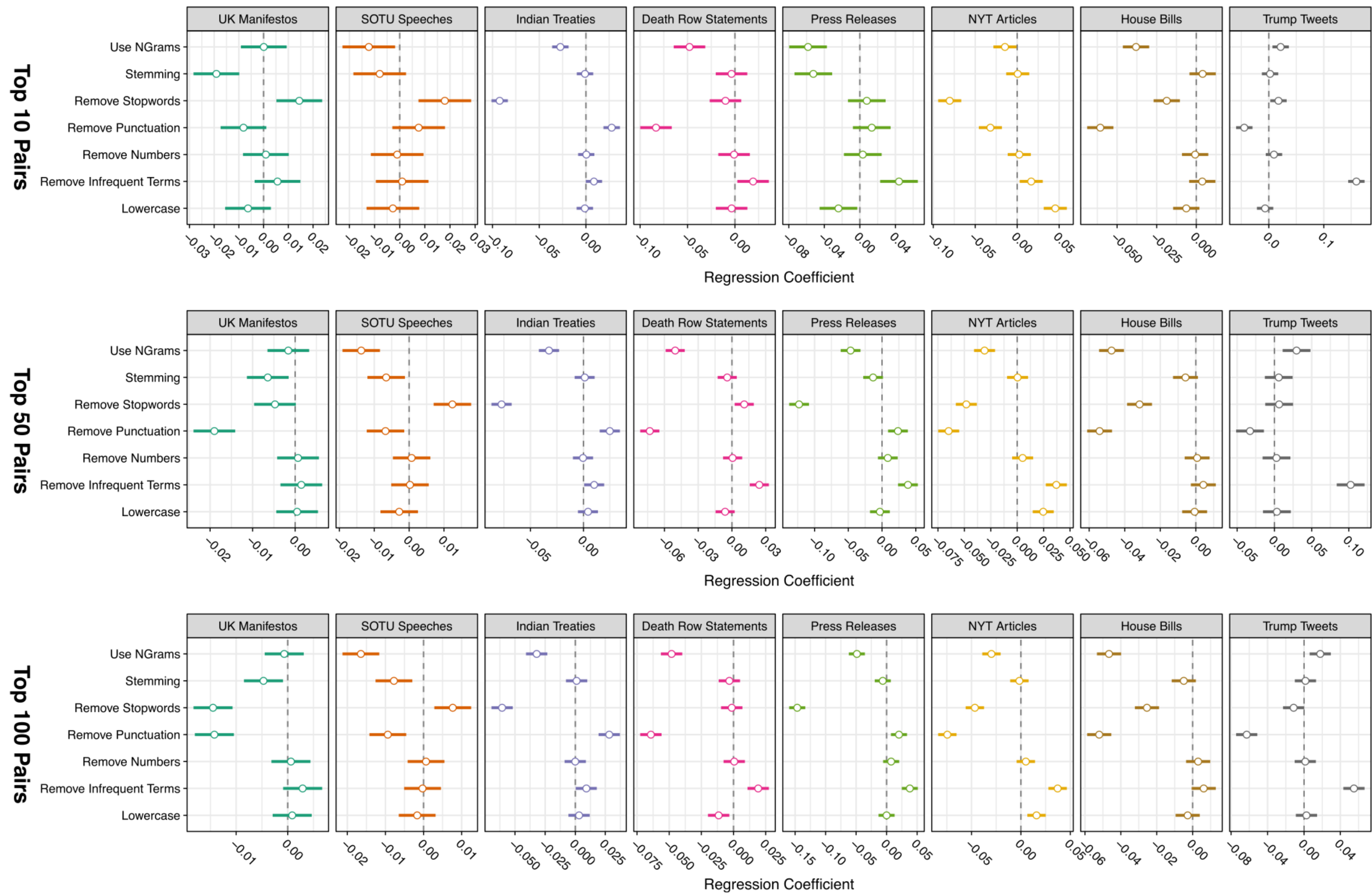


Figure 5. Regression results depicting the effects of each of the seven preprocessing steps on the preText score for that preprocessing combination.

Tf-idf

- We can do more than just count words
- We can transform these counts
- Use some sort of a weight in order to transform
- Term frequency inverse document frequency is one form of weighting

Term Frequency

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Term Frequency

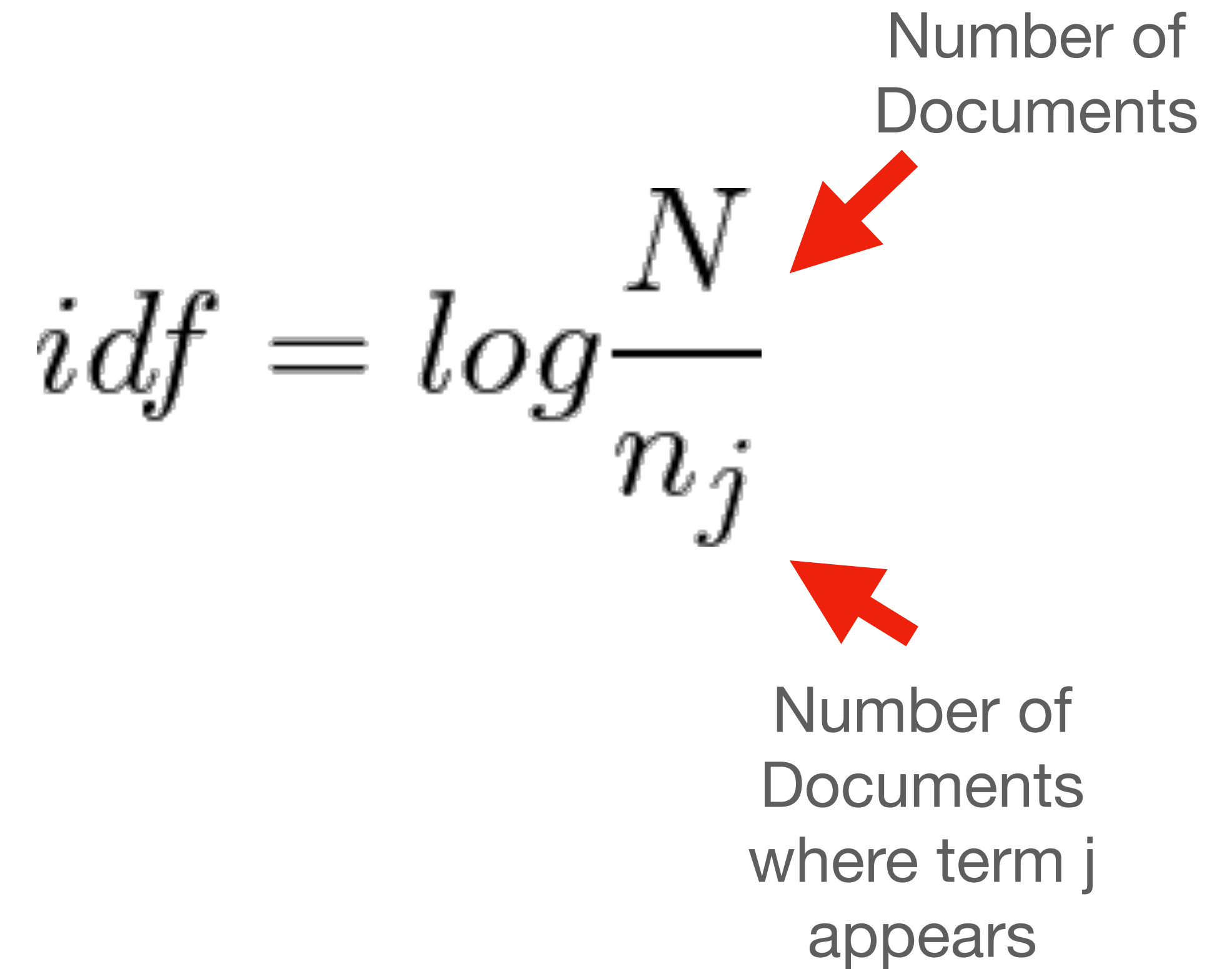
$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Inverse Document Frequency

$$idf = \log \frac{N}{n_j}$$

Number of Documents

Number of Documents where term j appears



tfidf

$$W_{ij} \times \log \frac{N}{n_j}$$

What the hell?

What the hell?

- *Exactly!*
- Introduced in 1972 by a computer scientist (Spärck Jones, 1972)
- No theoretical justification
- Apart from “it seems to work...”

What the hell?

- *Exactly!*
- Introduced in 1972 by a computer scientist (Spärck Jones, 1972)
- No theoretical justification
- Apart from “it seems to work...”
- Sometimes it does...

Log Odds / Log Odds Ratio

$$\log O_w^i = \log \frac{f_w^i}{1 - f_w^i}$$

$$\log \frac{O_w^i}{O_w^j} = \log \frac{f_w^i}{1 - f_w^i} / \frac{f_w^j}{1 - f_w^j} = \log \frac{f_w^i}{1 - f_w^i} - \log \frac{f_w^j}{1 - f_w^j}$$