# Text Similarity (Plagiarism Detection)

Alexa Federice and Linh Phan

# Definition: Text Reuse

- Text similarity detection: used to find the similarity between a source text and a possibly reused text
    - Examples of things you can do with it:
        - Detect plagiarism
        - Compare the language of bills in the U.S. Congress
        - Study the diffusion of memes or other online content
        - Trace historical influence of important books
        - Compare musical scores

# Purpose:

- In political science, scholars use text reuse algorithms primarily to identify similarities between policy proposals. Manually identifying similarities is costly and time consuming, making text reuse algorithms extremely useful. Scholars are able to process and analyze the similarity of thousands of bills or other political documents based on different sets of criteria (content, structure, style etc.) much more easily than manually comparing documents.

# Process:

1. Input: give the algorithm a set of documents that are relatively short
2. The text reuse algorithm then takes each of the documents and compares it to every other document in the set.
3. If the algorithm finds text that is shared between two documents (criteria for “reused text” can be different depending on ) then it sets the documents aside
4. The text reuse algorithm compares each document pair using the Smith-Waterman algorithm. This algorithm goes through each document character by character and decides whether each character in the one document matches characters in the other. When characters do not match, a space is inserted.

# Process:

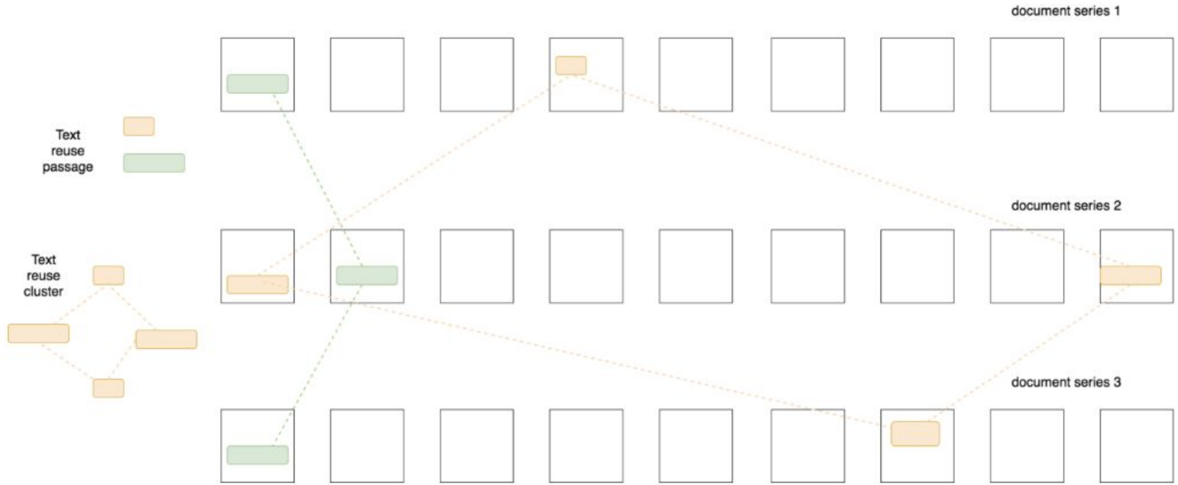
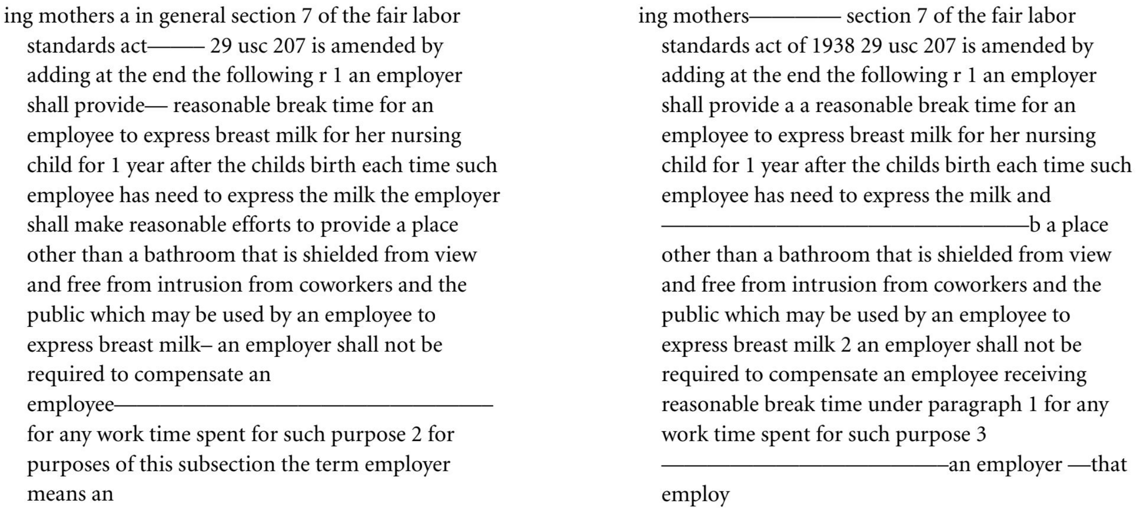


Figure 1. Schematic representation of text reuse clusters; each cluster consists of similar passages found in several series of documents.

Romanello and Hengchen (2021)

# Simple Example:

Passages below share  all content that is typed out. The spaces indicate parts of the passages that do not match up:



ing mothers a in general section 7 of the fair labor
standards act——— 29 usc 207 is amended by
adding at the end the following r 1 an employer
shall provide— reasonable break time for an
employee to express breast milk for her nursing
child for 1 year after the childs birth each time such
employee has need to express the milk the employer
shall make reasonable efforts to provide a place
other than a bathroom that is shielded from view
and free from intrusion from coworkers and the
public which may be used by an employee to
express breast milk– an employer shall not be
required to compensate an
employee————————————————————
for any work time spent for such purpose 2 for
purposes of this subsection the term employer
means an

ing mothers——— section 7 of the fair labor
standards act of 1938 29 usc 207 is amended by
adding at the end the following r 1 an employer
shall provide a a reasonable break time for an
employee to express breast milk for her nursing
child for 1 year after the childs birth each time such
employee has need to express the milk and
————————————————————b a place
other than a bathroom that is shielded from view
and free from intrusion from coworkers and the
public which may be used by an employee to
express breast milk 2 an employer shall not be
required to compensate an employee receiving
reasonable break time under paragraph 1 for any
work time spent for such purpose 3
———————————————an employer —that
employ

Wilkerson et al. 2015

# Cool Example 1:

Wilkerson, John, David Smith, and Nicholas Stramp. 2015. “Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach.” American Journal of Political Science 59 (4): 943–56

- They use an algorithm (Smith-Waterman) that was developed for genetic sequencing and use it to develop an alignment algorithm which looks for similar words aligned in similar sequence
- “Alignment algorithms stand in contrast to bag-of-words methods, which compare documents in terms of the use of similar words, but do not account for the ordering/sequence in which the words appear. (Linder et al. 2020).

# Cool Example 1 (continued)

Findings:

- The Affordable Care Act shared ideas with 232 bills that had been introduced earlier.
- “We were also able to show how many ideas in the law and the main markup bills could be traced to provisions of earlier bills, and that many of these antecedent provisions were sponsored by Republican lawmakers (who ultimately voted against the law). (945).”

# Cool Example 2

Stefano Pagliari, Kevin L. Young, Exploring information exchange among interest groups: a text-reuse approach, Journal of European Public Policy, 10.1080/13501763.2020.1817132, 27, 11, (1698-1717), (2020).

- The authors also utilized the Smith-Waterman algorithm.
  - They note that an advantage of this algorithm is that the approach does not limit any analysis to one single language. The algorithm captures instances where responses written in any language share text with others written in the same language.
- “We find that there are significant differences between the structure of information exchange networks and more formal lobbying coalitions in the EU, as well as between the groups that engage in these forms of coordination.”

# Common Assumptions/Mistakes

- Assumption: Text reuse detection examines similarities between words
    - Text reuse detection can examine more than just similarities between words. These algorithms can detect word sequences, stylistic features (vocabulary richness), and structural similarities (length of sentences).
- Mistake: Using text reuse algorithms for big data only.
    - Text reuse algorithms have been designed specifically for small and medium size data.

# Suggested Readings

- Hoad, Timothy C., and Justin Zobel. 2003. “Methods for Identifying Versioned and Plagiarized Documents.” Journal of the American Society for Information Science and Technology 54(3): 203–15.
- Linder, Fridolin, Bruce A. Desmarais, Matthew Burgess, and Eugenia Giraudy. 2020. “Text as Policy: Measuring Policy Similarity through Bill Text Reuse.” Policy Studies Journal 48: 546–74.
- Burgess, Matthew, Eugenia Giraudy, Julian Katz-Samuels, Joe Walsh, Derek Willis, Lauren Haynes, and Rayid Ghani. 2016. “The Legislative Influence Detector: Finding Text Reuse in State Legislation.” KDD 2016, San Francisco, CA.

# Software

The R textreuse package (R) written by Lincoln Mullen

- Tutorial and information https://lincolnmullen.com/blog/an-introduction-to-the-textreuse-package-with-suggested-applications/
- Package created to deal with “medium data.”

WCopyFind

- Used by Wilkerson et al. 2015

Passim (Scala) developed by David Smith (Northeastern University)

- Tutorial on how to use Passim: https://programminghistorian.org/en/lessons/detecting-text-reuse-with-passim