PPOL 628: Text as Data – Computational Linguistics for Social Scientists

Class 7: Text Reuse

Today

- Lecture: key points from readings.
- Lab: text_reuse.R
- Website: github.com/matthewjdenny/PPOL_628_Text_As_Data

Text Reuse vs. Similarity

- **Text Similarity**: Overall, how similar are these documents?
  - E.g. They focus on the same topic, they use very similar words, they share the same viewpoint, same valence etc.
  - Often what we are interested in when we analyze documents to measure a latent concept.
  - Typically more efficient to calculate than text reuse because order does not matter.
- **Text Reuse**: What character/words sequences are identical?
  - Only makes sense for larger chunks of text (e.g. sentences, paragraphs)
  - Does not necessarily imply same valence, viewpoint (e.g. pull quoting).
  - Does imply a direct connection between documents.

Text Similarity

- **Simplest formulation**: do the rows for two documents in a document term matrix “look similar”?
  - Do they both use roughly the same terms, and in the same proportion/amount?
- **Embeddings approach (preview):** If we were to embed the terms in this document in a k-dimensional space, would they generally be close?
- **Similarity Metrics in Practice**: Standard approach is to define a distance function (usually Cosine or Euclidean distance), then apply to pairs of rows in DTM/contingency table.
  - Interpretation is strongly dependent on average document length.

Distance\Similarity Metrics

- Input: pairs of vectors of term counts/proportions.

- Can apply TF-IDF weighting to focus on differences in rare term use.
  - Shrinks stop-term counts.

- Can be (relatively) efficient to calculate.
  - Apply to sparse term vectors.
  - Do not need to account for term position in documents.

- Use cosine similarity if document length is not important.

Euclidean Distance

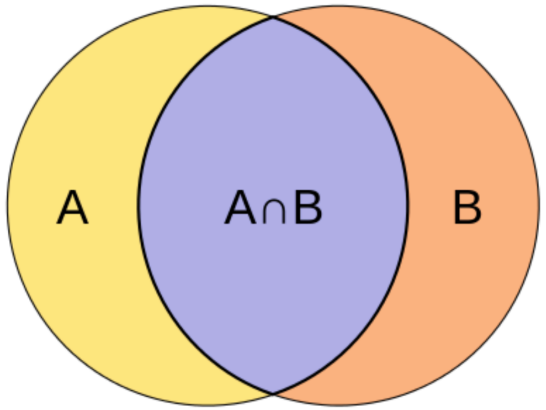
- Input: two term vectors: $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$
- Formula: $$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$
- Increasing in differences between counts of individual elements.
- Must be greater than or equal to zero, unbounded above.
- Can be interpreted as a physical distance between cordinates.

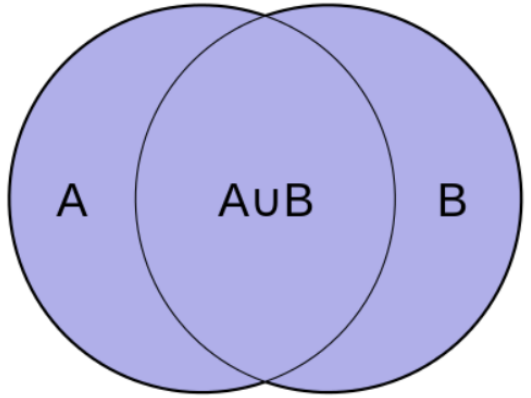
Cosine Similarity

- Input: Let A and B be term vectors as with Euclidean Distance.
- Formula: $$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$
- For text, takes on a value between zero (totally unrelated) to one (meaning exactly terms appear in same proportion).
- Interpreted as angle between term vectors.
- Normalized for document length.

Jaccard Similarity

- Input: Let A and B be term vectors.
- Formula: S = number of **unique** terms in both A and B / number of **unique** terms in the union of A and B





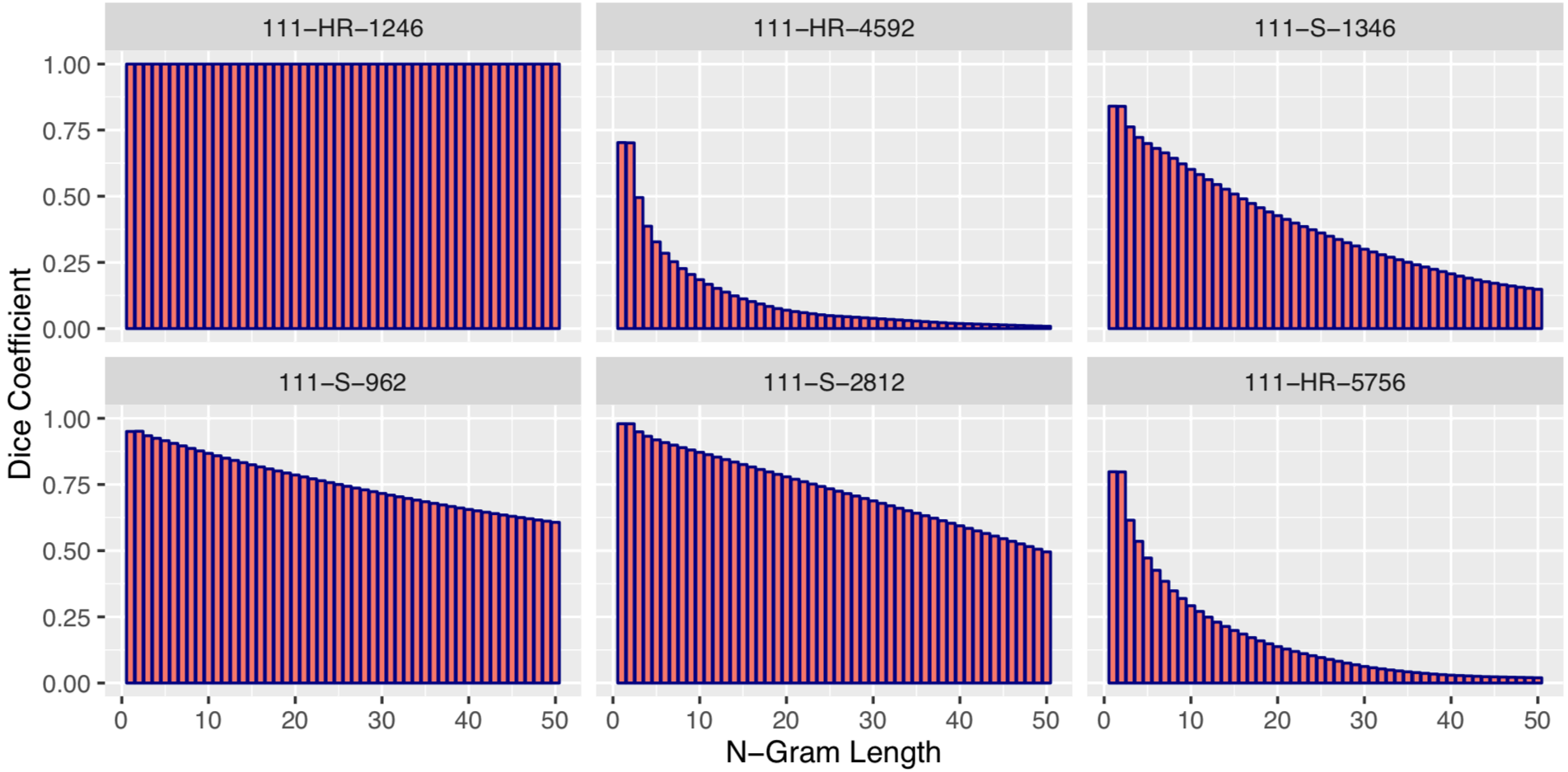
- For text, takes on a value between zero (totally unrelated) to one (meaning exactly the same unique terms appear).

Sørensen–Dice coefficient

- Start by taking documents and for each document, find all unique bigrams.
  - Let n_t be the number of unique bigrams that occur in both documents.
  - Let n_x and n_y be the count of unique bigrams in documents x and y respectively.
- Formula: $$d = \frac{2n_t}{n_x + n_y}$$
- Has a maximum value of 1 when all unique bigrams are the same, zero when none are the same.
- Generalizes to n-grams of any length.
- Potentially more informative because n-gram overlap indicates stronger similarity.

Sørensen–Dice coefficient Ensemble

- Can vary n-gram length to understand similarities at different granularities.



Text Reuse

- When we want to identify chunks of text that are the same between documents.
  - **Plagiarism detection** – want to actually pull out the overlapping parts.
  - **Edit characterization** – want to assess how document changed between versions.
  - **Document subsumption** – want to know if one document is now included in another document.
  - **Document Similarity** – Can serve as the basis for stronger similarity statements.
  - **Citation/Quotation Importance** – some passages that are included in many different documents. May also identify boilerplate passages.

An Example:

- A human reading these two passages could determine that the one on the right is effectively copied from the one on the left.

HR-408-IH	HR-146-ENR
[Congressional Bills 111th Congress] [From the U.S. Government Printing Office] [H.R. 408 Introduced in House (IH)] 111th CONGRESS 1st Session H. R. 408 To direct the Secretary of the Interior to convey to the City of Henderson, Nevada, certain Federal land located in the City, and for other purposes. IN THE HOUSE OF REPRESENTATIVES January 9, 2009 Mr. Heller introduced the following bill; which was referred to the Committee on Natural Resources A BILL To direct the Secretary of the Interior to convey to the City of Henderson, Nevada, certain Federal land located in the City, and for other purposes. Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, SECTION 1. SHORT TITLE. This Act may be cited as the "Southern Nevada Limited Transition Area Act". SEC. 2. DEFINITIONS. In this Act: (1) City.–The term "City" means the City of Henderson, Nevada. (2) Secretary.–The term "Secretary" means the Secretary of the Interior. (3) State.–The term "State" means the State of Nevada. (4) Transition area.–The term "Transition Area" means the approximately 502 acres of Federal land located in Henderson, Nevada, and identified as "Limited Transition Area" on the map entitled "Southern Nevada Limited Transition Area Act" and dated March 20, 2006. SEC. 3. SOUTHERN NEVADA LIMITED TRANSITION AREA. (a) Conveyance.–Notwithstanding the Federal Land Policy and Management Act of 1976 (43 U.S.C. 1701 et seq.), on request of the City, the Secretary shall, without consideration and subject to all valid existing rights, convey to the City all right, title, and interest of the United States in and to the Transition Area. (b) Use of Land for Nonresidential Development.– (1) In general.–After the conveyance to the City under subsection (a), the City may sell, lease, or otherwise convey any portion or portions of the Transition Area for purposes of nonresidential development. (2) Method of sale.– (A) In general.–The sale, lease, or conveyance of land under paragraph (1) shall be through a competitive bidding process. (B) Fair market value.–Any land sold, leased, or otherwise conveyed under paragraph (1) shall be for not less than fair market value.	SEC. 2602. SOUTHERN NEVADA LIMITED TRANSITION AREA CONVEYANCE. (a) Definitions.–In this section: (1) City.–The term "City" means the City of Henderson, Nevada. (2) Secretary.–The term "Secretary" means the Secretary of the Interior. (3) State.–The term "State" means the State of Nevada. (4) Transition area.–The term "Transition Area" means the approximately 502 acres of Federal land located in Henderson, Nevada, and identified as "Limited Transition Area" on the map entitled "Southern Nevada Limited Transition Area Act" and dated March 20, 2006. (b) Southern Nevada Limited Transition Area.– (1) Conveyance.–Notwithstanding the Federal Land Policy and Management Act of 1976 (43 U.S.C. 1701 et seq.), on request of the City, the Secretary shall, without consideration and subject to all valid existing rights, convey to the City all right, title, and interest of the United States in and to the Transition Area. (2) Use of land for nonresidential development.– (A) In general.–After the conveyance to the City under paragraph (1), the City may sell, lease, or otherwise convey any portion or portions of the Transition Area for purposes of nonresidential development. (B) Method of sale.– (i) In general.–The sale, lease, or conveyance of land under subparagraph (A) shall be through a competitive bidding process. (ii) Fair market value.–Any land sold, leased, or otherwise conveyed under subparagraph (A) shall be for not less than fair market value. (C) Compliance with charter.–Except as provided in subparagraphs (B) and (D), the City may sell, lease, or otherwise convey parcels within the Transition Area only in accordance with the procedures for conveyances established in the City Charter. (D) Disposition of proceeds.–The gross proceeds from the sale of land under subparagraph (A) shall be distributed in accordance with section 4(e) of the Southern Nevada Public Land Management Act of 1998 (112 Stat. 2345). (3) Use of land for recreation or other public purposes.– The City may elect to retain parcels in the Transition Area for public recreation or other public purposes consistent with the Act of June 14, 1926 (commonly known as the "Recreation and Public Purposes Act") (43 U.S.C. 869 et seq.) by providing to the Secretary written notice of the election.

Table 1: HR-146 bill insertion example. Matches highlighted in red.

Text Reuse from a Technical Perspective

- Primary challenge is computational.
  - **With infinite time**: compare all 1 to number-of-tokens-in-document—grams from each document with those from all others (order preserving). Find longest matches and induce graph structure.
- Also difficult to pay a human to do since it requires exceptional memory.
- In practice, we often want to filter on some similarity metric, topic scores, etc., then only check for exact matches if there is a reasonably high similarity.
- Trade-off between speed and overlap detection quality.

Sequence Alignment Algorithms

- **Smith Waterman**: Algorithm originally designed to align DNA sequences, handles gaps in match very well.

- Very computationally intensive, provides high quality, easy to human understand results (with some fiddling).

- More oriented towards providing an alignment between documents/DNA sequences. Characterization of match quality is up to end user.

S-W Example (Wilkerson et al.,2015)

TABLE 1 A Local Alignment Example

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

Smith-Waterman Algorithm

Fill the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$, $W_1 = 2$

Traceback

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

3

```
G T T - A C
| | |   | |
G T T G A C
```

Result

Sequences

Sequence 1: T G T T A C G G

Sequence 2: G G T T G A C T A

Parameters:

Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$, $W_1 = 2$

Result:

```
Sequence 1    G T T - A C
                | | |   | |
Sequence 2    G T T G A C
```

Shingled n-grams Similarity/Text Reuse Methods

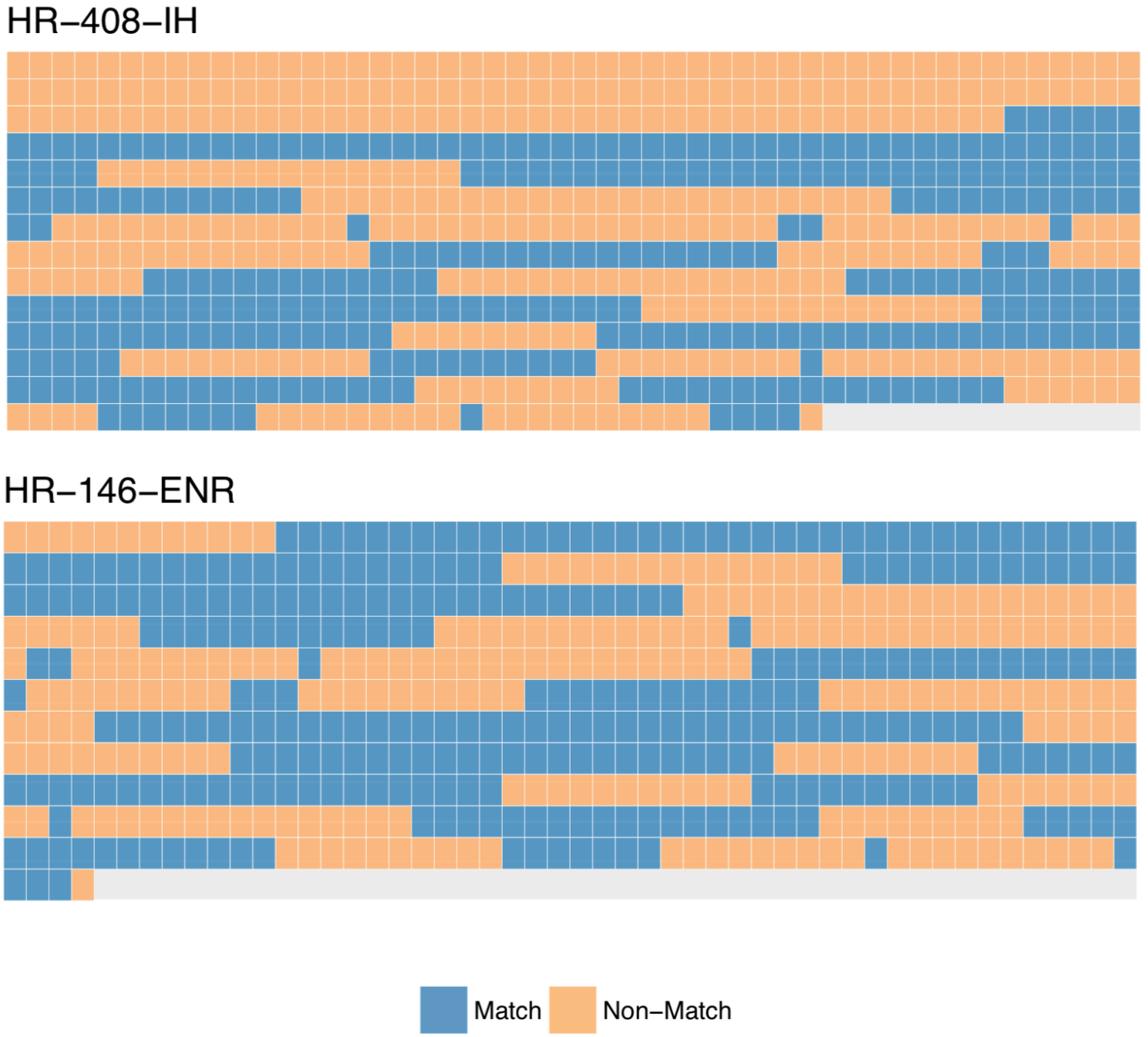
- Designed to identify matching sequences of terms/characters in text (like alignment algorithms), while having computational complexity on par with similarity metrics.

- **Shingled N-grams**: n-grams that overlap with each one starting one token after the next.
  - 3-gram example: “My fast brown cat runs” → [“my fast brown”, “fast brown cat”, “brown cat runs”]

- Using longer n-grams reduces false positive match rate, decreases resolution (ability to identify valid shorter matches).

Basic Intuition

- Each square represents an (overlapping) 10-gram in each document.
- Blue squares are matches orange squares are mismatches.
- Read from left to right then top to bottom.
- Output as a vector of zeros and ones.



Shingled n-grams Implementation

- Tokenize each document into shingled n-grams, **preserve order**!
- For each shingle in each document to be compared, check to see if it appears anywhere in the other document, if so record a 1, else 0.
  - Note that for any one n-gram this could represent a false positive match, but with longer sequences of matched shingles, probability of true match increases.

- End up with two vectors of zeros and ones (one for each document).
  - Can use these to highlight what parts of text have a match in other document.
  - Can calculate statistics one these match vectors like longest match sequence.

Effects of Preprocessing on n-gram matches

HR-408-IH	HR-146-ENR
[Congressional Bills 111th Congress] [From the U.S. Government Printing Office] [H.R. 408 Introduced in House (IH)] 111th CONGRESS 1st Session H. R. 408 To direct the Secretary of the Interior to convey to the City of Henderson, Nevada, certain Federal land located in the City, and for other purposes. IN THE HOUSE OF REPRESENTATIVES January 9, 2009 Mr. Heller introduced the following bill; which was referred to the Committee on Natural Resources A BILL To direct the Secretary of the Interior to convey to the City of Henderson, Nevada, certain Federal land located in the City, and for other purposes. Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, SECTION 1. SHORT TITLE. This Act may be cited as the "Southern Nevada Limited Transition Area Act". SEC. 2. DEFINITIONS. In this Act: (1) City.–The term "City" means the City of Henderson, Nevada. (2) Secretary.–The term "Secretary" means the Secretary of the Interior. (3) State.–The term "State" means the State of Nevada. (4) Transition area.–The term "Transition Area" means the approximately 502 acres of Federal land located in Henderson, Nevada, and identified as "Limited Transition Area" on the map entitled "Southern Nevada Limited Transition Area Act" and dated March 20, 2006. SEC. 3. SOUTHERN NEVADA LIMITED TRANSITION AREA. (a) Conveyance.–Notwithstanding the Federal Land Policy and Management Act of 1976 (43 U.S.C. 1701 et seq.), on request of the City, the Secretary shall, without consideration and subject to all valid existing rights, convey to the City all right, title, and interest of the United States in and to the Transition Area. (b) Use of Land for Nonresidential Development.– (1) In general.–After the conveyance to the City under subsection (a), the City may sell, lease, or otherwise convey any portion or portions of the Transition Area for purposes of nonresidential development. (2) Method of sale.– (A) In general.–The sale, lease, or conveyance of land under paragraph (1) shall be through a competitive bidding process. (B) Fair market value.–Any land sold, leased, or otherwise conveyed under paragraph (1) shall be for not less than fair market value.	SEC. 2602. SOUTHERN NEVADA LIMITED TRANSITION AREA CONVEYANCE. (a) Definitions.–In this section: (1) City.–The term "City" means the City of Henderson, Nevada. (2) Secretary.–The term "Secretary" means the Secretary of the Interior. (3) State.–The term "State" means the State of Nevada. (4) Transition area.–The term "Transition Area" means the approximately 502 acres of Federal land located in Henderson, Nevada, and identified as "Limited Transition Area" on the map entitled "Southern Nevada Limited Transition Area Act" and dated March 20, 2006. (b) Southern Nevada Limited Transition Area.– (1) Conveyance.–Notwithstanding the Federal Land Policy and Management Act of 1976 (43 U.S.C. 1701 et seq.), on request of the City, the Secretary shall, without consideration and subject to all valid existing rights, convey to the City all right, title, and interest of the United States in and to the Transition Area. (2) Use of land for nonresidential development.– (A) In general.–After the conveyance to the City under paragraph (1), the City may sell, lease, or otherwise convey any portion or portions of the Transition Area for purposes of nonresidential development. (B) Method of sale.– (i) In general.–The sale, lease, or conveyance of land under subparagraph (A) shall be through a competitive bidding process. (ii) Fair market value.–Any land sold, leased, or otherwise conveyed under subparagraph (A) shall be for not less than fair market value. (C) Compliance with charter.–Except as provided in subparagraphs (B) and (D), the City may sell, lease, or otherwise convey parcels within the Transition Area only in accordance with the procedures for conveyances established in the City Charter. (D) Disposition of proceeds.–The gross proceeds from the sale of land under subparagraph (A) shall be distributed in accordance with section 4(e) of the Southern Nevada Public Land Management Act of 1998 (112 Stat. 2345). (3) Use of land for recreation or other public purposes.– The City may elect to retain parcels in the Transition Area for public recreation or other public purposes consistent with the Act of June 14, 1926 (commonly known as the "Recreation and Public Purposes Act") (43 U.S.C. 869 et seq.) by providing to the Secretary written notice of the election.

Table 1: HR-146 bill insertion example. Matches highlighted in red.

HR-408-IH	HR-146-ENR
cited southern nevada limited transition area conveyance notwithstanding federal land policy management et seq request city without consideration subject valid existing rights convey city right interest united transition area use land nonresidential development general conveyance city city sell lease otherwise convey portion portions transition area purposes nonresidential development method sale general sale lease conveyance land competitive bidding process fair market value land sold leased otherwise conveyed less fair market value compliance charter except paragraphs city sell lease otherwise convey parcels within transition area accordance procedures conveyances established city charter disposition proceeds gross proceeds sale land distributed accordance southern nevada public land management stat use land recreation public purposes city elect retain parcels transition area public recreation public purposes consistent june commonly known recreation public purposes et seq providing written notice election noise compatibility requirements city plan manage transition area accordance united code relating airport noise compatibility planning regulations promulgated accordance agree land transition area sold leased otherwise conveyed city sale lease conveyance contain limitation require uses compatible airport noise compatibility planning reversion general parcel land transition area conveyed nonresidential development reserved recreation public purposes years enactment parcel land discretion revert united inconsistent use city uses parcel land within transition area manner inconsistent uses specified discretion parcel revert united make election city	march southern nevada limited transition area conveyance notwithstanding federal land policy management et seq request city without consideration subject valid existing rights convey city right interest united transition area use land nonresidential development general conveyance city city sell lease otherwise convey portion portions transition area purposes nonresidential development method sale general sale lease conveyance land competitive bidding process fair market value land sold leased otherwise conveyed less fair market value compliance charter except subparagraphs city sell lease otherwise convey parcels within transition area accordance procedures conveyances established city charter disposition proceeds gross proceeds sale land distributed accordance southern nevada public land management stat use land recreation public purposes city elect retain parcels transition area public recreation public purposes consistent june commonly known recreation public purposes et seq providing written notice election noise compatibility requirements city plan manage transition area accordance united code relating airport noise compatibility planning regulations promulgated accordance agree land transition area sold leased otherwise conveyed city sale lease conveyance contain limitation require uses compatible airport noise compatibility planning reversion general parcel land transition area conveyed nonresidential development reserved recreation public purposes years enactment parcel land discretion revert united inconsistent use city uses parcel land within transition area manner inconsistent uses specified discretion parcel revert united make election clause

Table 2: HR-146 bill insertion example after pre-processing. Matches highlighted in red.

Deriving Information from Match Vectors

- Output from shingled N-Grams matching:
  - Doc 1: {1,1,1,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,0,0,0,1,1,1,1,0,0,0,0,0,0,0}
  - Doc 2: {0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1}

- Can now calculate arbitrary statistics on these vectors.

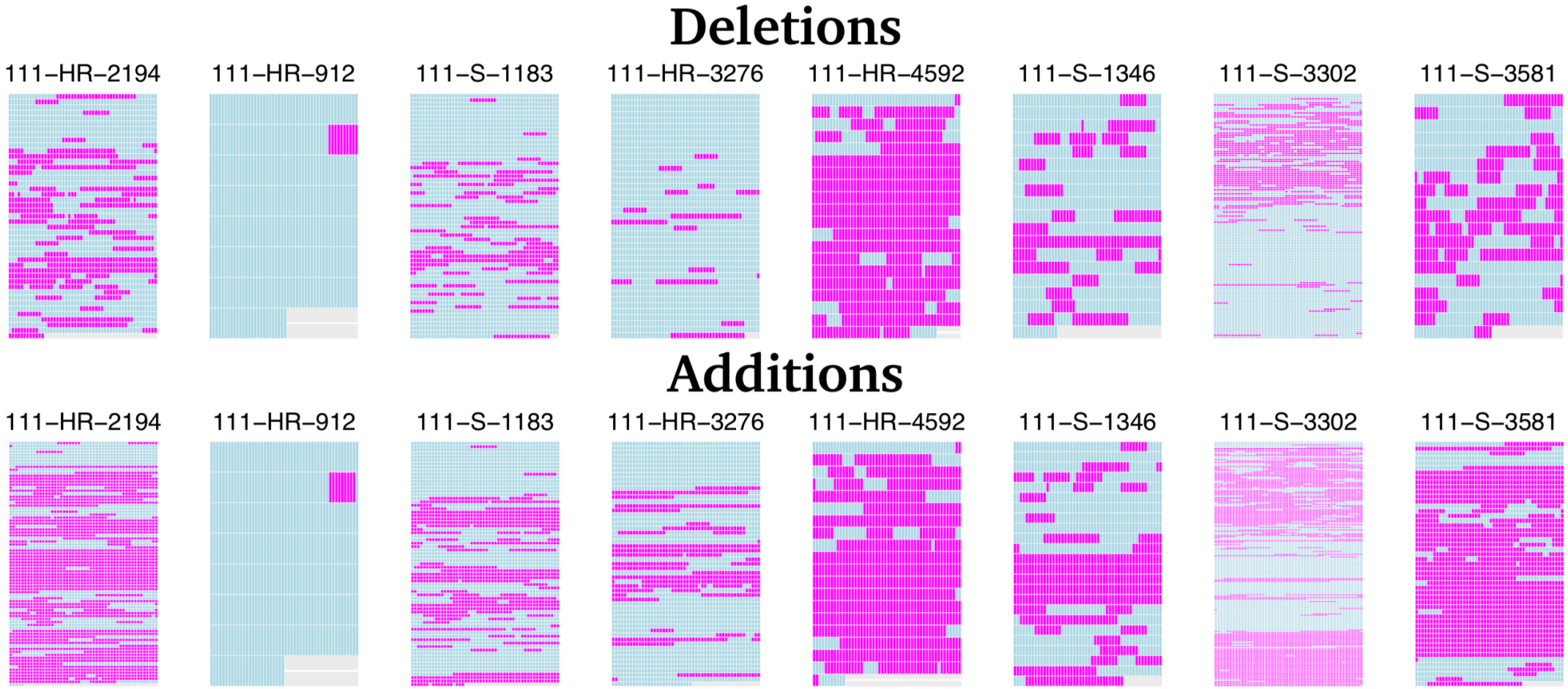
- These data can be fed into a supervised learning model or used with some thresholds to identify plagiarism, document insertions into other documents, identical documents, etc.

Additions and Deletions

- **Scope of Deletions**: the scope of deletions is just the proportion of shingled n-grams in the first version of the document that do not match an n-gram in the second version. This gives us a measure of how much of the original text survived to the second version of the text.

- **Scope of Additions**: the scope of additions is similar to the scope of deletions, but is instead defined as the proportion of n-grams in the second version of the document that do not match an n-gram in the first version. This gives us a measure of how much of the edited version of the text was not present in the original version of the text.

Additions and Deletions



Edit Size

- **Average Edit Size:** the average edit size is defined as the average length of sequences of non-matching n-grams across both versions of a document. The shorter the average edits size, the shorter the length of edits made to a document. We do not normalize this measure by document length because this length has a natural interpretation.

- **Proportion of Possible Changes:** this metric captures the number of unique sequences of mismatches in both versions of the text relative to the maximum number of sequences of mismatches possible given the n-gram length used for comparison. Thus is can range between zero, indicating that no changes were made at all, to one, indicating that a maximal number of unigram changes were made to the document (relative to the resolution limit implied by the choice of n-gram size).

Other Match Sequence Statistics

- **Edit granularity:** related to the average edits size. Ranges between 0 and 1, and maximized when all edits are just a single term, and minimized when the document is just wholesale replaced by another document with unrelated text.

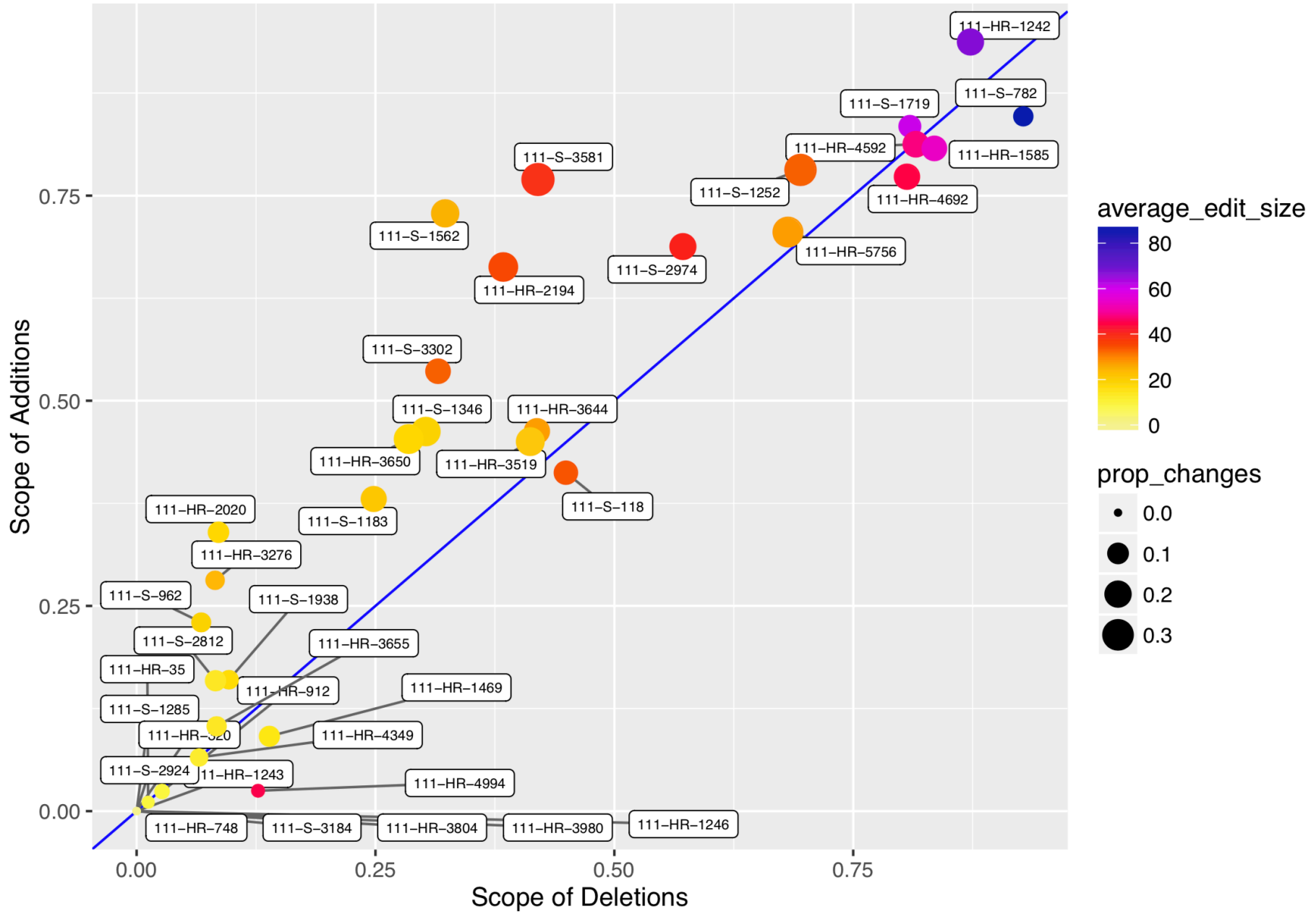
- **Match/Mismatch Sequence Variance, etc.:** These statistics can tell us more about the variation in the sizes of matching bits of text and edits, which can be useful both as a substantive quantity of interest, and as input to downstream a classifier.

Algorithm Output as Input for Supervised Learning

- Use these features to train a classifier to identify identical documents, plaigarism, etc.

addition_granularity	deletion_granularity	addition_scope	deletion_scope	average_addition_size	average_deletion_size	scope	average_edit_size
0.971711684	0.9563314981	0.876937810	0.8733700377	945.2258	150.70000	0.875153924	633.64706
0.959554986	0.9590318886	0.930235312	0.8603303390	1366.4348	141.38095	0.895282825	781.75000
1.000000000	1.0000000000	0.000000000	0.0000000000	0.0000	0.00000	0.000000000	0.00000
0.153861079	0.3103448276	0.846138921	0.6896551724	4361.0000	2380.00000	0.767897047	3370.50000
0.809267631	0.8112865836	0.762929476	0.7548536656	1054.7500	651.25000	0.758891571	853.00000
0.994327391	0.9878296146	0.005672609	0.0121703854	28.0000	42.00000	0.008921497	35.00000
0.995247148	0.9820341930	0.004752852	0.0179658070	20.0000	62.00000	0.011359329	41.00000
0.954728950	0.9895682411	0.045271050	0.0104317589	157.0000	36.00000	0.027851404	96.50000
0.992824874	0.5514343669	0.007175126	0.4485656331	27.0000	1548.00000	0.227870380	787.50000
0.919730010	0.9910170965	0.160539979	0.0179658070	386.5000	31.00000	0.089252893	208.75000
0.119492934	0.7758620690	0.880507066	0.4482758621	4237.0000	773.50000	0.664391464	1928.00000
0.990543735	0.9886989278	0.009456265	0.0113010722	36.0000	39.00000	0.010378668	37.50000
0.688821498	0.9854148556	0.933535507	0.0437554332	1231.3333	50.33333	0.488645470	640.83333

Characterizing Document Editing



General Points

- **Text similarity metrics** can be useful for document clustering, and as a first step in assessing **text reuse**.
  - Some limitations as compared to other document clustering methods like topic models.

- How useful text reuse methods end up being will greatly depend on how your corpus was generated.
  - Same authors, copying, etc. vs. incidental overlap.

- Studying document editing is very new --> lots of theoretical and computational challenges.

- **There is no one globally optimal approach**.