

Bloat-Ectomy: A Method for the Identification and Removal of Duplicate Text in the Bloated Notes of Electronic Health Records and Other Documents

Roselie A. Bright, ScD, MS (FDA/OC/OCS/OHI)
Summer Rankin, PhD (Booz Allen Hamilton)

The authors are part of a team that is using the text
notes in electronic healthcare records (EHRs).



FDA/OC/OCS/OHI (OHI@FDA.HHS.gov): Targeting solutions where health, science, informatics, and technology converge...



DATA

"Multiparameter Intelligent Monitoring in Intensive Care (MIMIC III)"

MIT database of 60,000 Beth Israel Deaconess Hospital
admissions with critical care, 2001-2012

<http://mimic.physionet.org/about/mimic/>

Privacy protection method:

- Preserved:
 - Text notes
 - Timelines for each patient
- Not preserved:
 - Calendar time
 - Ability to simulate real time analysis
 - Proper names

Bloat-Ectomy: Example notes from one admission show duplications*



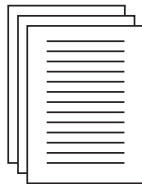
Progressively longer nurse's notes
over 1 shift

Example of 2 physicians' notes from
same time period



VISION

Identify the duplicate sections of text.



Mark them for manual record review.



Delete them for statistical analyses.



Bloat-Ectomy: Our needs

LITERATURE re “NOTE BLOAT”

Current methods don’t meet our need for:

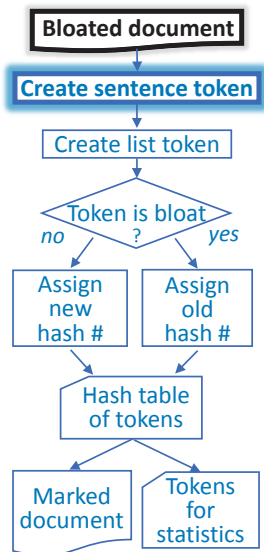
- Exact detection of all pasted material
- Sentence-level (minimal level of meaning) algorithm
- Eliminate less, rather than more, text
- Simple implementation

METHOD

- The LZW compression algorithm* works by hashing **words** by assigning sequential numeric addresses to them.
- We are using **sentences**.

*Welch modification of compression technique developed by Lempel and Ziv, <https://ieeexplore.ieee.org/document/1659158?arnumber=1659158&tag=1> (1984); suggested by Walter Bright WalterBright.com.

Bloat-Ectomy: Create sentence tokens



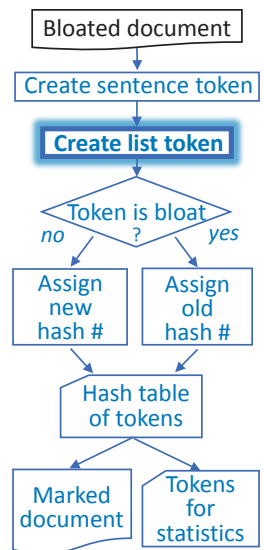
No CP. Became tachycardic to 160s on dopa. No CP.

A sentence is:

- Any character(s) + a period
- followed by 1 of:
 - a space
 - a tab
 - a newline character (\n)

Token #	Token
1	No CP.
2	Became tachycardic to 160s on dopa.
3	No CP.

Bloat-Ectomy: Create list tokens

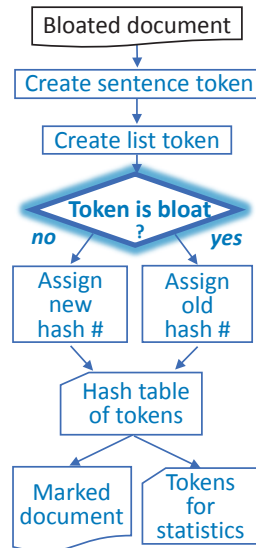


Tmax: 36.6
C (97.8
HR: 100 (97 - 166) bpm
Tmax: 36.6
C (97.8

List of $\mathcal{X}$ (5 in this example) items defined by newline characters.

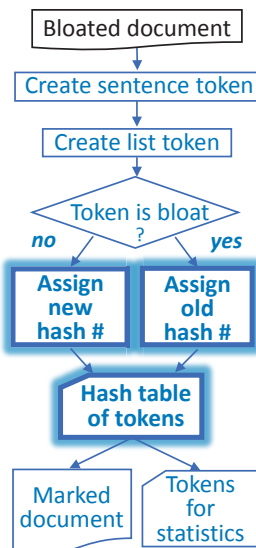
Token #	Token
4	Tmax: 36.6
5	C (97.8
6	HR: 100 (97 - 166) bpm
7	Tmax: 36.6
8	C (97.8

Bloat-Ectomy: Identify bloat



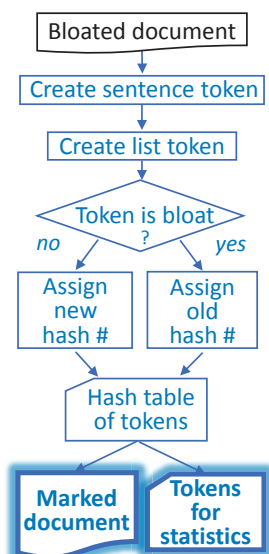
Token #	Token	Bloat ?
1	No CP.	no
2	Became tachycardic to 160s on dopa.	no
3	No CP.	yes
4	Tmax: 36.6	no
5	C (97.8	no
6	HR: 100 (97 - 166) bpm	no
7	Tmax: 36.6	yes
8	C (97.8	yes

Bloat-Ectomy: Add assigned hash numbers to hash table



Token #	Token	Bloat ?	Hash #
1	No CP.	no	1
2	Became tachycardic to 160s on dopa.	no	2
3	No CP.	yes	1
4	Tmax: 36.6	no	3
5	C (97.8	no	4
6	HR: 100 (97 - 166) bpm	no	5
7	Tmax: 36.6	yes	3
8	C (97.8	yes	4

Bloat-Ectomy: Produce outputs



Marked document for manual review

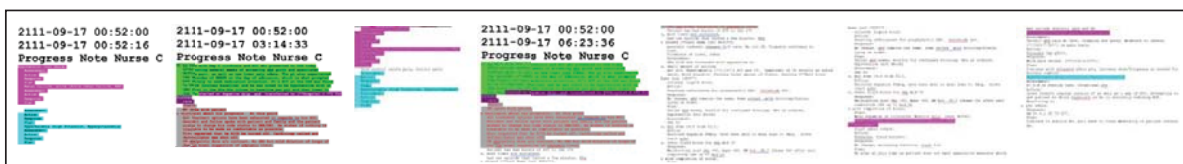
No CP.
 Became tachycardic to 160s on dopa.
 No CP.
 Tmax: 36.6
 C (97.8
 HR: 100 (97 - 166) bpm
 Tmax: 36.6
 C (97.8

Deduplicated tokens for statistics

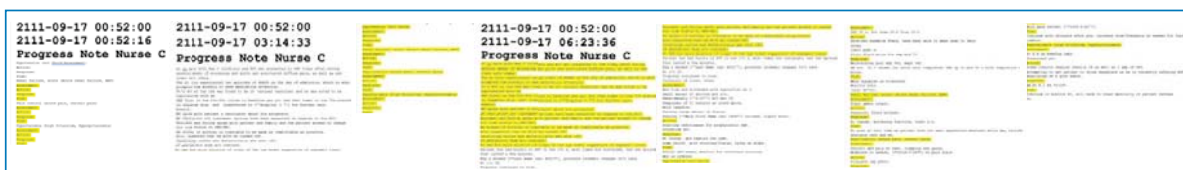
No CP.
 Became tachycardic to 160s on dopa.
 Tmax: 36.6
 C (97.8
 HR: 100 (97 - 166) bpm

Bloat-Ectomy: Results for real example notes Nurse's notes

Manual



Bloat-ectomy



Bloat-Ectomy: Results for real example notes Physicians' notes



Bloat-Ectomy: Flexibility



Adaptation to other texts can be user-made and will depend on the characteristics of the text and the analysis purpose(s).

Token definition options:

- Sentence demarcation (default is period)
- List demarcations: symbols that start a new token in a list (i.e. bullet point, star)

What to do with identified bloat:

- Mark for review.
- Delete for statistical analysis.

Output Formats:

- Word: mark bloat (bold, italics, highlight) and output as docx for a word processor.
- HTML option: highlighted bloat in printout.
- Text: output as a string or token strings to be ingested into a database (i.e., PostgreSQL)

Bloat-Ectomy: Conclusions



Bloat-ectomy:

- Works
- Customizable