Workshop in Methods (WIM)

Introduction to Text Mining for Social Scientists

Helge Marahrens
Indiana University Bloomington
Department of Sociology
Email correspondence to: hmarahre@iu.edu

The New York Times

twitter

PRO PUBLICA

Text Mining

- Deriving quantitative information from texts
  - from text to numbers
  - unstructured to structured

Male	Height	Religion
1	5.8	Protestant
0	6.0	Jewish
0	5.5	None
…	….	…

Statistical analyses

Text Mining

- Deriving quantitative information from texts
  - from text to numbers
  - unstructured to structured

# of Words	Male	Height	Religion
20	1	5.8	Protestant
1200	0	6.0	Jewish
350	0	5.5	None
…	…	….	…

Statistical analyses

- Python 3.X
  - data types
    - lists and dictionaries
  - function vs. method
  - indentation is key (4 spaces)
  - counting begins at zero

Example Codes

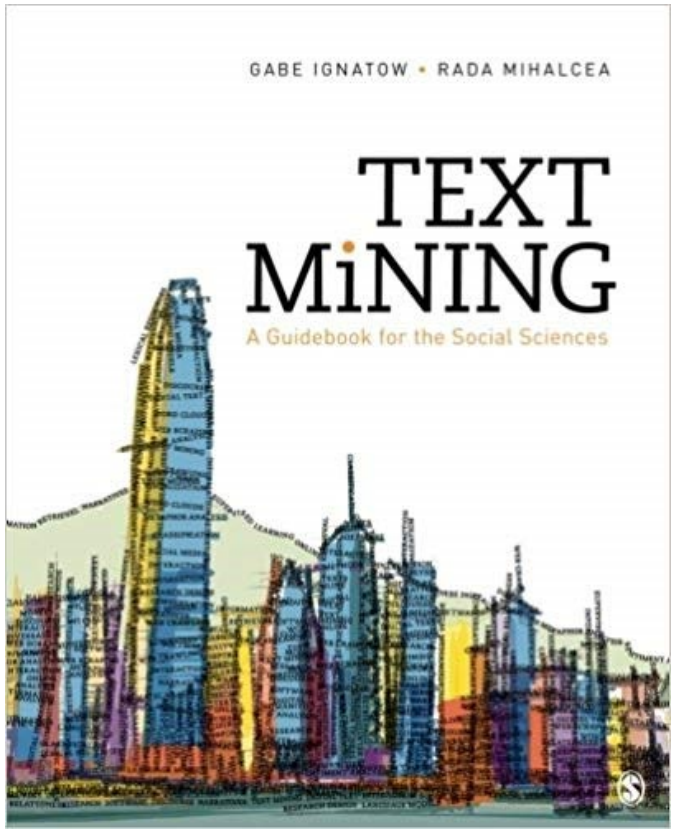
recoding strings

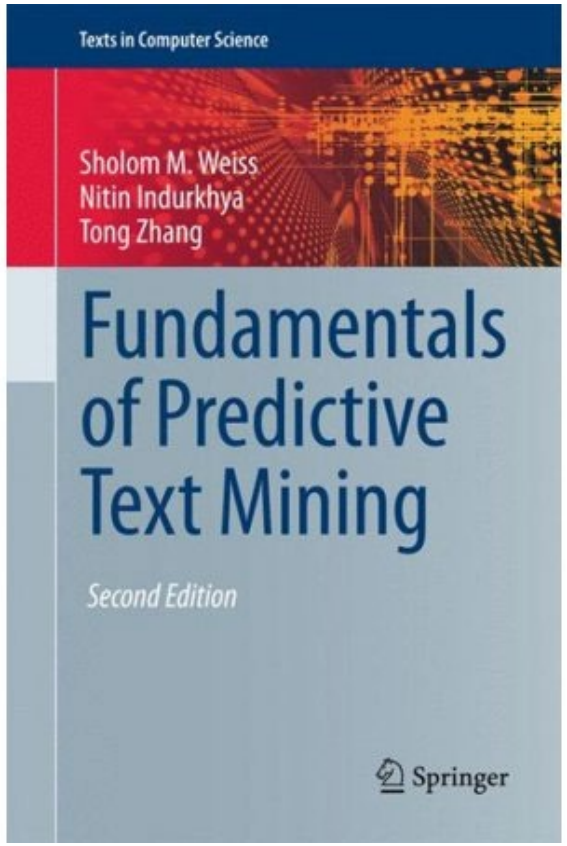
regular expressions

sentiment analysis

dimension reduction

advanced NLP

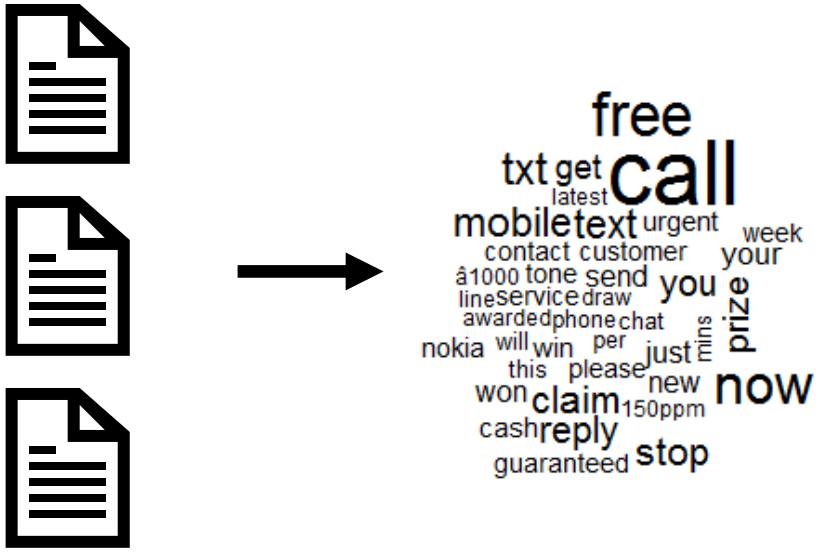




Ignatow, G., & Mihalcea, R. (2017). *Text Mining. A Guidebook for the Social Sciences*. Thousand Oaks, CA: Sage Publications.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of predictive text mining*. London, UK: Springer.

bag of words



	Word 1	Word 2	Word 2	Word 4
Document 1	3	0	0	0
Document 2	0	0	0	0
Document 3	0	0	0	1
	…	…	….	…

basic text operations

- tokenization – breaking the stream of characters into “words” (tokens)

  “How are you?” → ['How', 'are', 'you', '?']

basic text operations

- tokenization – breaking the stream of characters into “words” (tokens)

- lower case – “Hello” → “hello”
- stemming – find root “computer”, ”computing” → “comput”
- stopword deletion

[‘the’, ‘a’, 'i', 'me', ‘my’, …]

basic_string_2020-02-14_hmarahre.py