Practical Text Analytics: Text Preparation

Fan Dai

Iowa State University

October 8, 2019

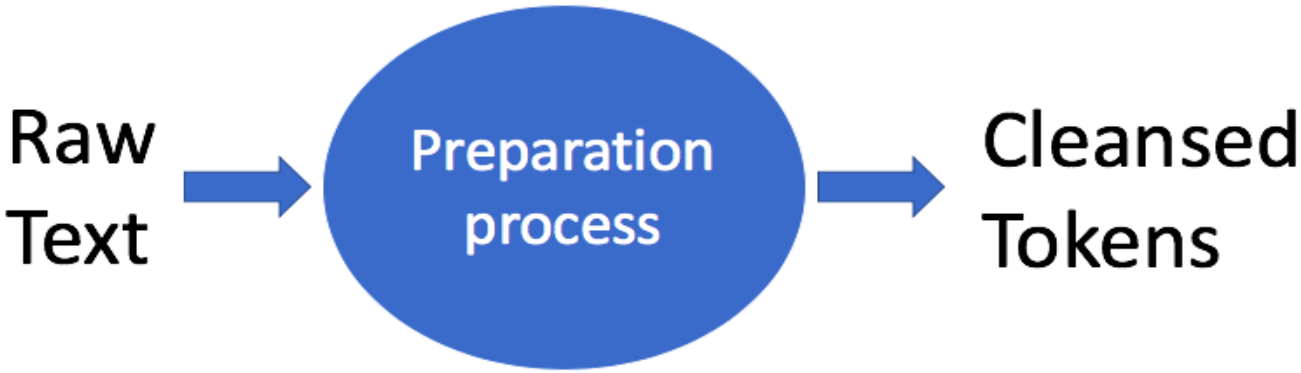
Recall: Text analytics process

Text analytics process

- Planning the text analytics projects
- Preparing and preprocessing text

  (*Chapter 4 & 13, Anandarajan et al.(2019)*)
- Analyzing data
- Interpreting results

Text Preparation



- Raw text data: document collection or corpus
  - Document: each record of the collected text data (unique identifier)
  - Movie reviews:
    Document 1: Excellent cast, story line, performances. ♡
    Document 2: It was so BORING!
- Tokens: single words or groups of words
  - Focus and features of the text analysis
  - *Excellent, cast, story, line, performance,* ♡*, it, was, so, BORING*
  - *Excellent cast, story line, performance,. . .*

Text Preparation: Tokenization

Tokenization: separate the text into a more usable form, known as tokens.

- Single words
  [Excellent] [cast] [,] [story] [line] [,] [performances] [.] [♡]

- N-grams, e.g., n = 2 for document 1,

  The first token is:
  [Excellent cast], story line, performances. ♡

  The second token is:
  Excellent [cast,] story line, performances. ♡

  The third token is:
  Excellent cast [, story] line, performances. ♡

  The last token is:
  [Excellent cast], story line, performances [. ♡]

Text Preparation: Standardization, Cleaning and Stop Word Removal

- Standardization: convert to lower case and check misspelling
- Cleaning: remove numbers, punctuation, and special characters
- Stop word removal: drop frequently used filler words, or stop words
  - *and, the, be, to, of*
- Movie reviews:

  Document 1: Excellent cast, story line, performances. ♡
  - [excellent] [cast] [story] [line] [performances]

  Document 2: It was so BORING!
  - [boring]

Text Preparation: Stemming and Lemmatization

Breaking words down to their root word

- Stemming: remove a word's suffix to reduce the size of the vocabulary
  - Train, Trains, Trained, Training, Trainer ⇒ train
  - A single word could have multiple meanings...
- Lemmatization: group word roots by the part of speech.
  - Group tokens at the cost of added complexity

Word	Stemming	Part of speech	Lemmatization
Fluffiest	Fluffy	Adjective	Fluffy
Fluffiness	Fluffy	Noun	Fluffiness
Fluffily	Fluffy	Adverb	Fluffily

Text Preparation: Synonymy and Polysemy

- Synonymy: two different words having the same meaning.
  - “My cap is brown”, “My hat is white”

    "cap","hat"

- Polysemy: a single word having multiple meanings.
  - “My fluffy dog has a white coat and hat,”

    "coat":

    1. the dog is wearing a white coat
    2. the dog has white fur

Text Preparation: Summaries

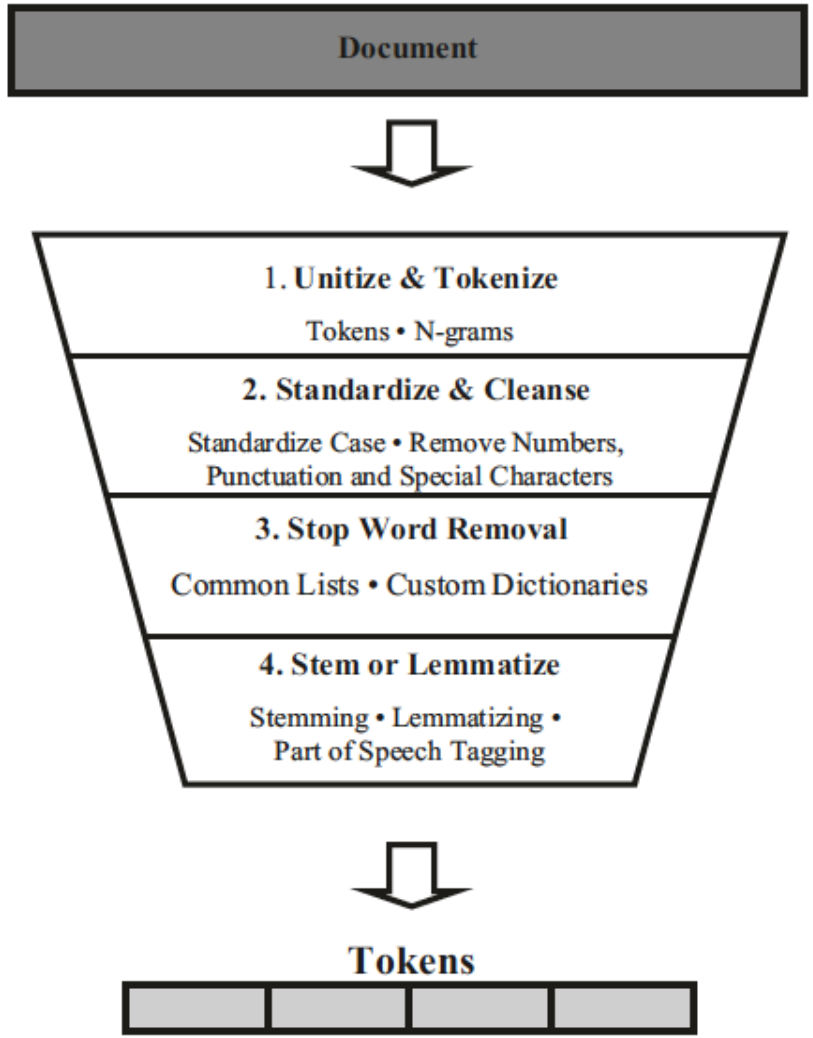


Figure: The text data pre-processing process

Source: **Practical text analytics: maximizing the value of text data**