・1 文の構造

単語＋半角スペース＋単語＋半角スペース＋・・・＋単語＋ピリオド

という形式になっています

・置換

次の記号は簡単のため、ピリオドに置換または消去する。

(テキストデータはすでに処理済みになっています。)

置換前				置換後
,	?	;	:	.
(	)	″	―	(消去)

・アポストロフィの扱い

アポストロフィは以下のように扱う(テキストデータはすでに処理済みになっています。)

〇所有格('s または s')

・・・単なる名詞へ

(例 1)Beauty's→Beauty

(例 2)merchants'→merchants

〇省略表現

・・・省略を戻して元の文へ

(例)don't→do not

〇特別な例(古文的省略 'tis)

・・・省略を戻して元の文へ

(例) 'tis→it is

・その他の扱い

〇2 つの文の間にはスペースを開けない(テキストデータはすでに処理済みになっています。)

(例) there␣was␣once␣a␣very␣rich␣merchant.who␣had␣six␣children.three----

〇大文字、小文字は区別しない(テキストデータはすでに処理済みになっています。)

(例)There と there は同じ単語とする

〇複数形は単数形とは別単語

(例)sons と son は別単語

〇その他の形が違う単語も別単語とする

(例)a と an なども別扱い

〇ハイフン(-)で繋がった言葉は 1 単語として扱う

(例)country-house で 1 単語

引用元:

「Beauty and the Beast - Project Gutenberg」

http://www.gutenberg.org/ebooks/7074