Pythonではじめる機械学習

7章　テキストデータの処理

7.4　ストップワード

15T4061L　南濱　篤

7.4 ストップワード

頻出頻度があまりに多く役に立たない単語

2つの手法

- 言語固有のストップワードリスト作成
- 頻度の高い単語を捨てる

7.4 ストップワード

ストップリストは

小さいデータセットに関して有効

→ ストップワードを決めるだけの情報がないから

大きいデータセットには性能にも解釈の容易さにそれほど影響を与えない

→ あまり意味がない