集合知プログラミング
ー特徴の検出の改良

06t4071f

林華

不具合点

- 非アルファベットを境界にして単語を分割

  短所：非アルファベットが存在する場合バグとなる

- すべての単語を小文字に変換

  短所：大文字単語の使いすぎを検出できない

改善策

- 大文字と小文字を区別して扱う
- テキストの代わりにエントリを丸ごと利用
- タイトル、要約などの単語の集合も利用
- メタ情報をもっと捕捉する
- URLと数を完全な状態で保存する

Akismetを利用する

Akismetは人々が自身のブログに投稿されたスパムコメントを報告できるWordPressのプラグインである。新しいコメントは他の人々がスパムとして報告済みのコメントたちとの類似度を基にフィルタされる。

- 文字列を受け取り、判断する
- スパムコメント以外のが動作しないかも
- 回答の根拠となる計算を見ることできない

その他の手法

１．ニューラルネットワークを適用する

短所：複雑しすぎてネットワーク中のニューロン間のコネクション強度は簡単に解釈できない。

長所：ある特徴の確率を他の特徴が存在かどうかで変更ができ、単純ベイズ分類器では捉えない相互依存関係を捉えることができる

２．サポートベクトルマシンを適用する

(第9章参照）