スパム3分間クッキング

スパムコレクションを使って SpamAssassinのテストルールを 作ってみよう！

日本SpamAssassinユーザ会／サードウェア
滝澤隆史

本題に入る前に

- SpamAssassinとは何か？
- 日本SpamAssassinユーザ会

SpamAssassinって何？

- スパム（迷惑メール）を検出するための
  - フィルタプログラム
  - Perlモジュールライブラリ

SpamAssassinって何？

- スパムメールらしさをスコアで算出し，メールにヘッダを挿入する
  - X-Spam-Flag: YES
  - X-Spam-Level: *******
  - X-Spam-Status: Yes, score=7.3, ...

様々なテストで総合的に判定する

- リアルタイム・ブラックリスト（RBL）
- URIブラックリスト（URIBL）
- 送信ドメイン認証（SPF, DomainKeys/DKIM）
- 協調型フィルタリング
- コンテンツフィルタリング
  - ヘッダの解析
  - 本文の解析
  - 画像ファイル解析
- ベイジアンフィルタ

SpamAssassinの日本語対応

- 最新版のSpamAssassin 3.2系列の日本語対応パッチあります
  - http://spamassassin.jp/download/sa3.2/
- 日本語でテストルールが書けます
  - body　HOGOHOGE　/ほごほげ/
- ベイジアンフィルタも日本語対応

日本SpamAssassinユーザ会

- http://spamassassin.jp/
- ウェブサイト放置気味
- 基本は各メンバーがやりたいことを勝手にやっているだけ
  - SpamAssassinを使いたい人
  - SpamAssassinのルールを整備したい人
  - SpamAssassinを日本語に対応させたい人
  - SpamAssassinを利用するソフトウェアを作りたい人

本題に戻る

作るもの

- SpamAssassinの日本語のテストルール

材料

- スパム（迷惑メール）たくさん
- ハム（正常なメール）たくさん
-
- ハムよりスパムを多めに用意してくださいね

用意するもの

- スクリプト
  - http://spamassassin.jp/download/experimental/taki/
  - sa-tokenizer.pl　　--- トークナイザー
  - sa-ja-testmaker.pl --- テスト生成スクリプト
- SpamAssassin 3.2（日本語パッチ適応済み）
- MeCabとPerlモジュールを少々

下ごしらえ

- スパムとハムをそれぞれ単語毎に分解

スパムメールから単語を抽出

- トークナイザー sa-tokenizer.pl を使って，メール構造を解析し，日本語の分かち書きをする

```
$ ./sa-tokenizer.pl -r ~/Mail/spam > spam.txt

The number of messages      : 39525
The number of uniq messages: 28863
```

こんなのができました

$ cat spam.txt

私
の
名前
は
中野
です
。

ハム(正常なメール)についても同様

```
$ ./sa-tokenizer.pl -r ~/Mail/ham > ham.txt

The number of messages      : 4956
The number of uniq messages: 4778
```

料理本番

- 下ごしらえしたデータをテストルール生成スクリプトsa-ja-testmaker.plに投入

```
$ ./sa-ja-testmaker.pl -s spam.txt -h ham.txt > body-ja.cf

The spam words:
  The number of Japanese words    : 1412565
  The number of uniq words        : 83553
The ham words:
  The number of Japanese words    : 1841092
  The number of uniq words        : 197798
The number of removed words       : 80325
The number of remaining uniq words: 3228
The number of made tests          : 200
```

こんなのができました。

```
# BODY_JA_HITOZUMA: 人妻 spam=2583/1325054, ham=1/1841092, ratio=0.00194
body     BODY_JA_HITOZUMA       /人妻/
describe BODY_JA_HITOZUMA       HITOZUMA
score    BODY_JA_HITOZUMA       0.6

# BODY_JA_ANATA: 貴方 spam=2645/1325054, ham=11/1841092, ratio=0.00193
body     BODY_JA_ANATA          /貴方/
describe BODY_JA_ANATA          ANATA
score    BODY_JA_ANATA          0.6

# BODY_JA_ICHIHACHIMIMAN: 18未満 spam=2446/1325054, ham=0/1841092, ratio=0.00184
body     BODY_JA_ICHIHACHIMIMAN     /18未満/
describe BODY_JA_ICHIHACHIMIMAN     ICHIHACHIMIMAN
score    BODY_JA_ICHIHACHIMIMAN     0.6

# BODY_JA_ADARUTO: アダルト spam=2426/1325054, ham=0/1841092, ratio=0.00183
body     BODY_JA_ADARUTO        /アダルト/
describe BODY_JA_ADARUTO        ADARUTO
score    BODY_JA_ADARUTO        0.5

# BODY_JA_DEAI: 出会い spam=2444/1325054, ham=9/1841092, ratio=0.00179
body     BODY_JA_DEAI           /出会い/
describe BODY_JA_DEAI           DEAI
score    BODY_JA_DEAI           0.5
```

自動生成されたルール

- テスト名称(ローマ字)を自動生成
- 出現頻度によりスコアの割り付け

```
# BODY_JA_DEAI: 出会い spam=2444/1325054,
#    ham=9/1841092, ratio=0.00179
body      BODY_JA_DEAI          /出会い/
describe BODY_JA_DEAI          DEAI
score     BODY_JA_DEAI          0.5
```

人柱募集

- テストルールを評価してくれる人募集！
- SpamAssassin-JPメーリングリストで意見ください

おわり