# Exploring Linguistic Features for Web Spam Detection
# A Preliminary Study

Jakub Piskorski[1] Marcin Sydow[2] Dawid Weiss[3]

[1] Joint Research Centre of the European Commission, Ispra, Italy

[2] Web Mining Lab, Polish-Japanese Institute of Information Technology, Warsaw, Poland

[3] Institute of Computing Science, Poznan University of Technology, Poland

# Background

There is a recent interest in machine-learning approach to Web spam detection.

The main motivations are:

- complexity: too many factors to consider
- scale: too much data to analyse by humans
- need for adaptivity: a dynamic problem (arms race)

# Previous work on content analysis, etc.

Various content-based factors have been already studied:

- statistic-based approach (Fetterly et al. ’04)
- checksums, term weighting (Drost et al. ’05, Ntoulas et al. ’06)
- blog spam detection by language model disagreement (Mishne et al. ’05)
- auto-generated content (Fetterly et al. ’05)
- HTML structure (Urvoy et al. ’06)
- commercial attractiveness of keywords (Benczur et al. ’07)

# Previous work on content analysis, etc.

Various content-based factors have been already studied:

- statistic-based approach (Fetterly et al. ’04)
- checksums, term weighting
  (Drost et al. ’05, Ntoulas et al. ’06)
- blog spam detection by language model disagreement
  (Mishne et al. ’05)
- auto-generated content (Fetterly et al. ’05)
- HTML structure (Urvoy et al. ’06)
- commercial attractiveness of keywords (Benczur et al. ’07)

Also other dimensions of data were explored: link-based, query-log based, combined, etc.

# Previous work on content analysis, etc.

Various content-based factors have been already studied:

- statistic-based approach (Fetterly et al. ’04)
- checksums, term weighting
  (Drost et al. ’05, Ntoulas et al. ’06)
- blog spam detection by language model disagreement
  (Mishne et al. ’05)
- auto-generated content (Fetterly et al. ’05)
- HTML structure (Urvoy et al. ’06)
- commercial attractiveness of keywords (Benczur et al. ’07)

Also other dimensions of data were explored: link-based, query-log based, combined, etc.

What about linguistic analysis of Web documents?

# Motivation

Linguistic analysis:

- have not been used before in the Web spam detection problem (except some corpus-based statistics)
- proved successful in **deception** detection in textual human-to-human communication
  (Zhou et al. “Automating Linguistics-based Cues for detecting deception of text-based Asynchronous Computer-Mediated Communication”)

# Linguistic Analysis

We applied light-weight linguistic analysis to compute new attributes for Web spam detection problem.

Two different NLP software tools were used:

- Corleone (developed at JRC, Ispra)
- General Inquirer (`www.wjh.harvard.edu/~inquirer`)

Why only a *light-weight* analysis?

- computationally cheap
- more immune in the context of the open-domain nature of the Web documents

General linguistic, document-level analysis without any prior knowledge about the corpus.

# Contributions

1. the two Yahoo! Web Spam Corpora of human-labelled hosts were taken
2. the two different NLP software tools were applied to them
3. over 200 linguistic-based attributes were computed and made publicly available for further research. Info:
   `http://www.pjwstk.edu.pl/~msyd/linguisticSpamFeatures.html`
4. over 1200 histograms were generated and analysed (also available)
5. the most promising attributes were preliminarily selected with the use of 2 different distribution-distance metrics

# Corleone-based attributes, examples

- **Type:**

$$\textit{Lexical validity} = \frac{\text{\# of valid word forms}}{\text{\# of all tokens}}$$

$$\textit{Text-like fraction} = \frac{\text{\# of potential word forms}}{\text{\# of all tokens}}$$

- **Diversity:**

$$\textit{Lexical diversity} = \frac{\text{\# of different tokens}}{\text{\# of all tokens}}$$

$$\textit{Content diversity} = \frac{\text{\# of different nouns \& verbs}}{\text{\# of all nouns \& verbs}}$$

$$\textit{Syntactical diversity} = \frac{\text{\# of different POS n-grams}}{\text{\# of all POS n-grams}}$$

$$\textit{Syntactical entropy} = -\sum_{g \in G} p_g \cdot \log p_g$$

# General Inquirer attribute groups

- ‘Osgood’ semantic dimensions
- pleasure, pain, virtue and vice
- overstatement/understatement
- language of a particular ‘institution’
- roles, collectivities, rituals, and interpersonal relations
- references to people/animals
- processes of communicating
- valuing of status, honour, recognition and prestige
- references to locations
- references to objects
- cognitive orientation
- pronoun types
- negation and interjections
- verb types
- adjective types
- skill categories
- motivation
- adjective types
- power
- rectitude
- affection
- wealth
- well-being
- enlightenment

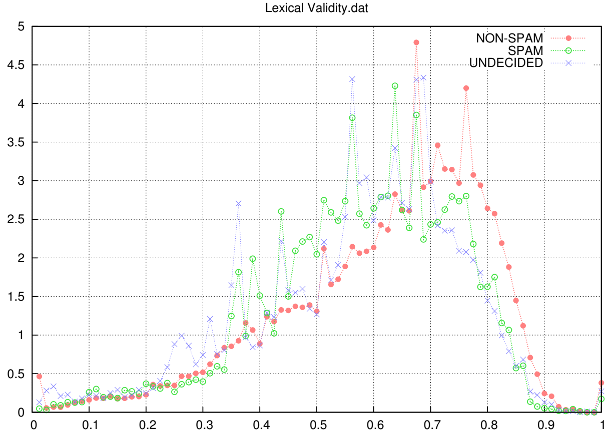
# Computation, input data sets

Map-reduce jobs (Hadoop) for processing (40 CPU cluster).

| | 2006 | 2007 |
|---|---|---|
| pages | 3 396 900 | 12 533 652 |
| pages without content | 65 948 | 1 616 853 |
| pages with HTTP/404 | 281 875 | 230 120 |
| TXT SQF (compressed file, GB) | 2.87 | 8.24 |

# Reducing noise

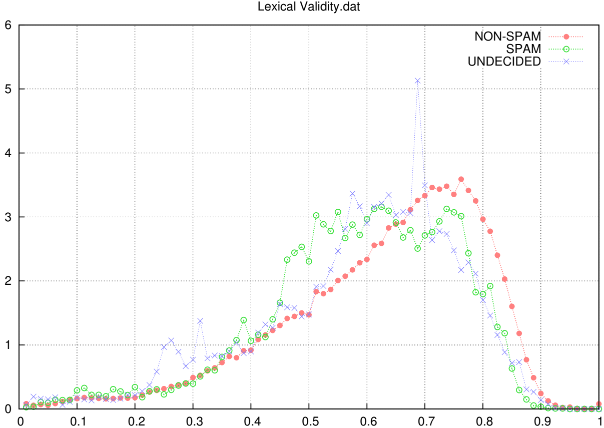
- Removed binary content-type pages.
- Different “modes” of page filtering:
  (0) $<$ 50k tokens, (1) 150–20k tokens, (2) 400–5k tokens.




Lexical validity for **unfiltered input**, **Corleone**, **WebSpam-Uk2007**.

# Reducing noise

- Removed binary content-type pages.
- Different "modes" of page filtering:
  (0) $<$ 50k tokens, (1) 150–20k tokens, (2) 400–5k tokens.




Lexical validity for **mode-1 filtered input**, **Corleone**, **WebSpam-Uk2007**.

# Discriminancy Measures

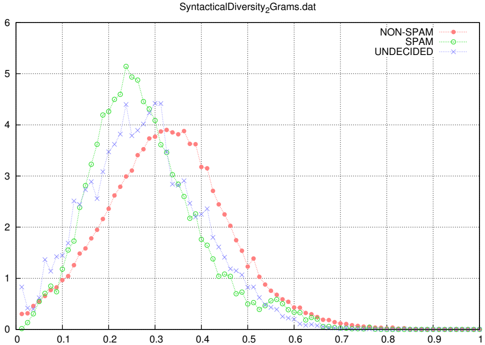
$$\mathrm{absDist}(h) = \sum_{i \in I} |s_i^h - n_i^h|/200 \tag{1}$$

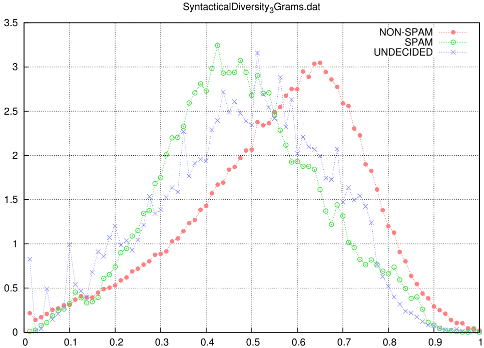
$$\mathrm{sqDist}(h) = \sum_{i \in I} (s_i^h/max_h - n_i^h/max_h)^2/|I| \tag{2}$$

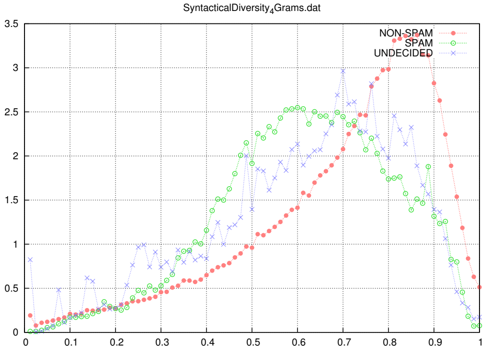
# The Most Promising Features (Corleone)

The most discriminating **Corleone** attributes wrt *absDist* and *sqDist* metric.

| **Corleone (absDist)** | 2007 | 2006 | **Corleone (sqDist)** | 2007 | 2006 |
|---|---|---|---|---|---|
| Passive Voice | 0.263 | 0.273 | Syn. Diversity (4g) | 0.053 | 0.054 |
| Syn. Diversity (4g) | 0.255 | 0.245 | Syn. Diversity (3g) | 0.050 | 0.067 |
| Content Diversity | 0.234 | 0.331 | Syn. Diversity (2g) | 0.037 | 0.036 |
| Syn. Diversity (3g) | 0.230 | 0.253 | Content Diversity | 0.032 | 0.065 |
| Pronoun Fraction | 0.224 | 0.261 | Syn. Entropy (2g) | 0.029 | 0.026 |
| Syn. Diversity (2g) | 0.221 | 0.232 | Lexical Diversity | 0.026 | 0.043 |
| Lexical Diversity | 0.213 | 0.262 | Lexical Validity | 0.024 | 0.033 |
| Syn. Entropy (2g) | 0.208 | 0.179 | Pronoun Fraction | 0.024 | 0.031 |
| Text-Like Fraction | 0.188 | 0.184 | Text-Like Fraction | 0.023 | 0.017 |



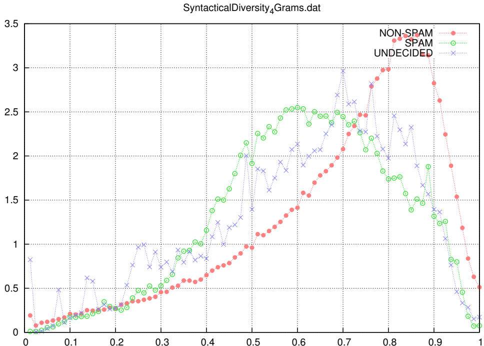

## Corleone, Syntactical diversity mode-1 filtered, 2006 data set

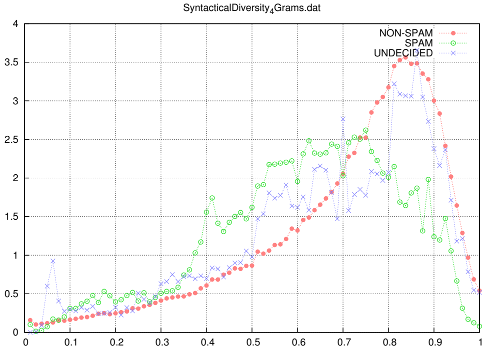
- 2, 3 and 4-grams
- different Y scale to illustrate shape
- increasing skewness of NON-SPAM

Corleone, Syntactical diversity
mode-1 filtered, 2006 and 2007 data set

- 4-grams
- different Y scale to illustrate shape
- 2006 (left), 2007 (right)

- **results very similar**



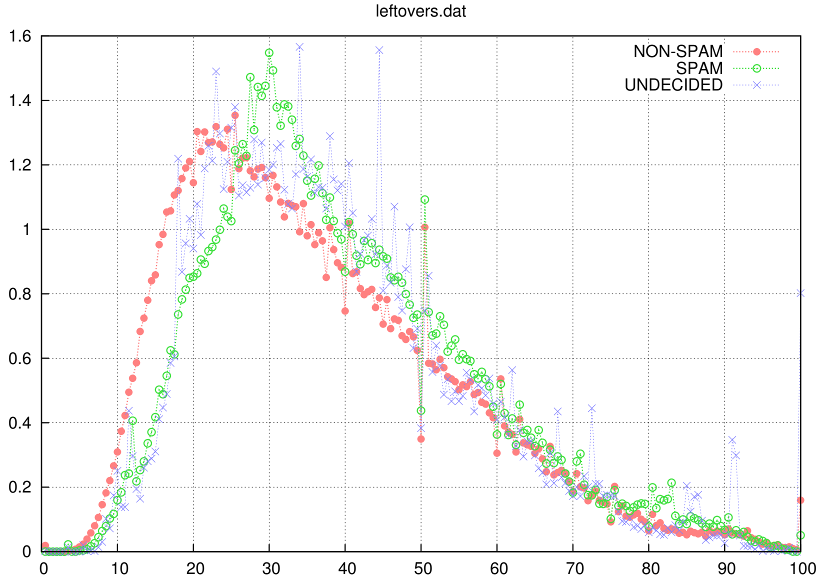



# The Most Promising Features (GI)

The most discriminating **General Inquirer** attributes according to *absDist* and *sqDist* metric.

| **GI (absDst)** | 2007 | 2006 | **GI (sqDist)** | 2007 | 2006 |
|---|---|---|---|---|---|
| WltTot | 0.287 | 0.346 | leftovers | 0.0150 | 0.0128 |
| WltOth | 0.285 | 0.341 | EnlOth | 0.0085 | 0.0072 |
| Academ | 0.270 | 0.263 | EnlTot | 0.0082 | 0.0118 |
| Object | 0.255 | 0.282 | Object | 0.0073 | 0.0086 |
| EnlTot | 0.249 | 0.247 | text-length | 0.0056 | 0.0048 |
| Econ@ | 0.228 | 0.356 | ECON | 0.0038 | 0.0034 |
| SV | 0.206 | 0.260 | Econ@ | 0.0038 | 0.0031 |
| | | | WltTot | 0.0038 | 0.0027 |
| | | | WltOth | 0.0037 | 0.0024 |

Leftovers attribute, **General Inquirer**, mode-1 filtered, 2006 data set:



# Conclusions and Further Work

Positive outcomes:

- Features showing different characteristic between `normal` and `spam` classes: content diversity, lexical diversity, syntactical diversity, ...

Limitations and problems:

- Spam pages generated from legitimate content.
- Graphical spam (images overlaid over legitimate text).
- Multi-lingual pages.

Further steps:

- new attributes should be tested directly in the Web classification task

# The Data sets

There are 4 data sets available ($\{$'06, '07$\} \times \{$Corleone, GI$\}$):

- the data sets are document-level
- the assigned labels are host-level
- for '07 corpus the labels are taken from the training set + merged with '06 labels
- easy, line-record, tab-separated ASCII format
- the histograms are also available

# Availability of the Data

Data sets: $\rightarrow$

http://www.pjwstk.edu.pl/~msyd/lingSpamFeatures.html

Enquiries: $\rightarrow$

msyd@pjwstk.edu.pl
jpiskorski@googlemail.com
dawid.weiss@cs.put.poznan.pl

Thank you for your attention.