Data Mining Pastes

Finding a needle in a haystack
How to analyze unstructured data sets?

Alexandre Dulaunoy

February 27, 2015

An Introduction to the Dataset

- One day of pastes (31MB of compressed unstructured documents)
- Format (Paste-id.gz)
- Various users pasted information with different objectives
- Some of the pastes include sensitive information (e.g. password leaks, vulnerability)

Strategies for Analysis

- Sampling and human-analysis
- File-type detection

```
ls -1 | parallel --gnu 'zcat {1} | file -'
```

- Terms searching
- What else?

Processing Textual Data

- Python TextBlob (using Python NLTK) is a simple library for processing textual data. Extracting nouns, sentences or even sentiment analysis, translation....

```
pip2 install -U textblob
python2 -m textblob.download_corpora
```

- The corpora is installed in your home directory /nltk_data to support the natural language processing functionalities.

Processing Textual Data - A Minimal Example

```
from textblob import TextBlob
w = TextBlob("This is an interesting project but there
    is still a lot of work")
w.noun_phrases
w.sentiment
```