ACU

جامعة الأهرام الكندية

AHRAM CANADIAN UNIVERSITY

Faculty of computer science

Dr.Ahmed Diaa Hassanein

Associate Professor,

PhD Communication, Engineering Science 2009, University of Oxford

Team Members

- Ahmed Nagy
- Islam Samir
- Ahmed Abooud
- Mohamed Rashad
- Mohamed Elkasaey

Identification of the class of English text among several classes

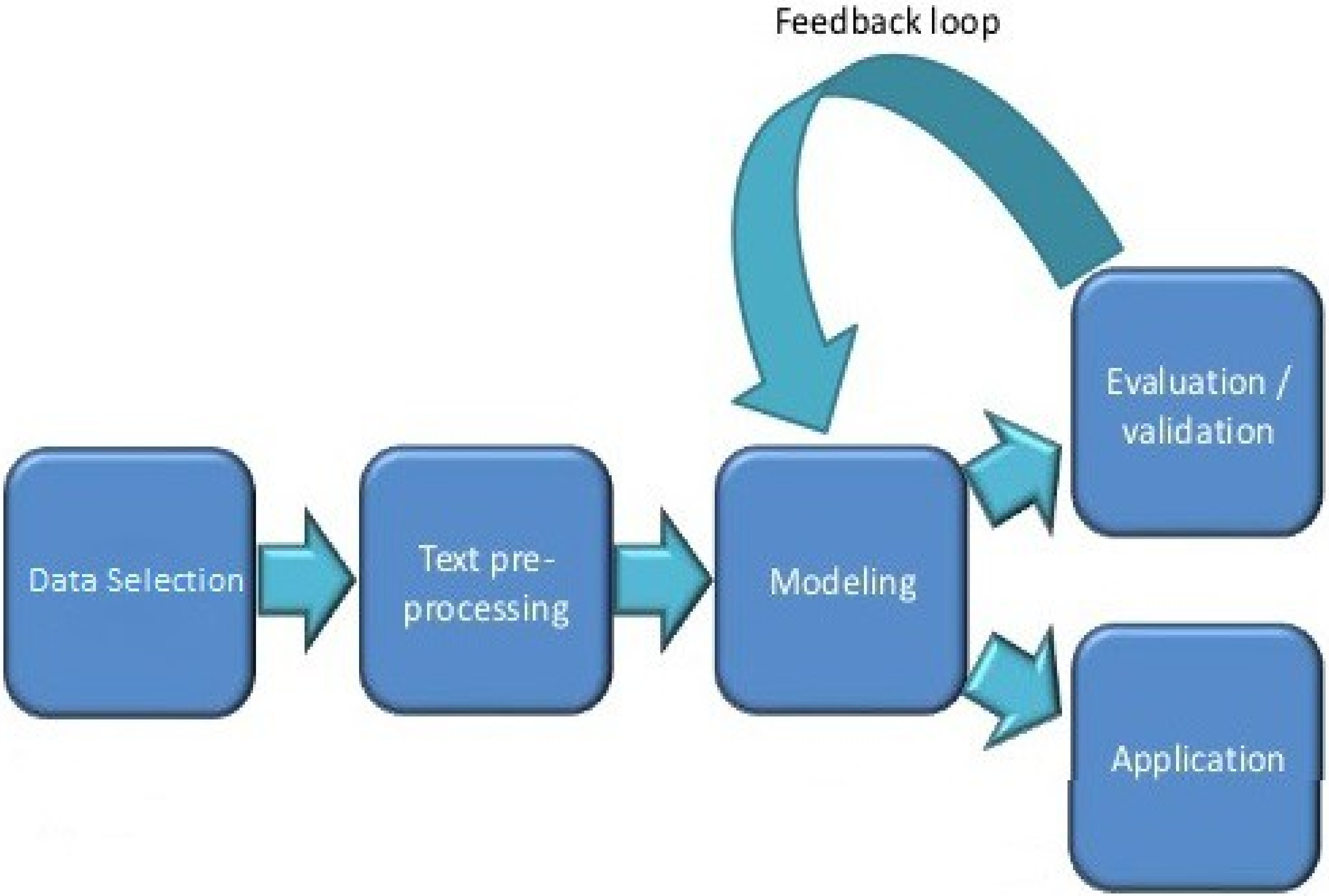
Outline

- Introduction
- Objectives
- General approach
- Steps and Details
- Applications

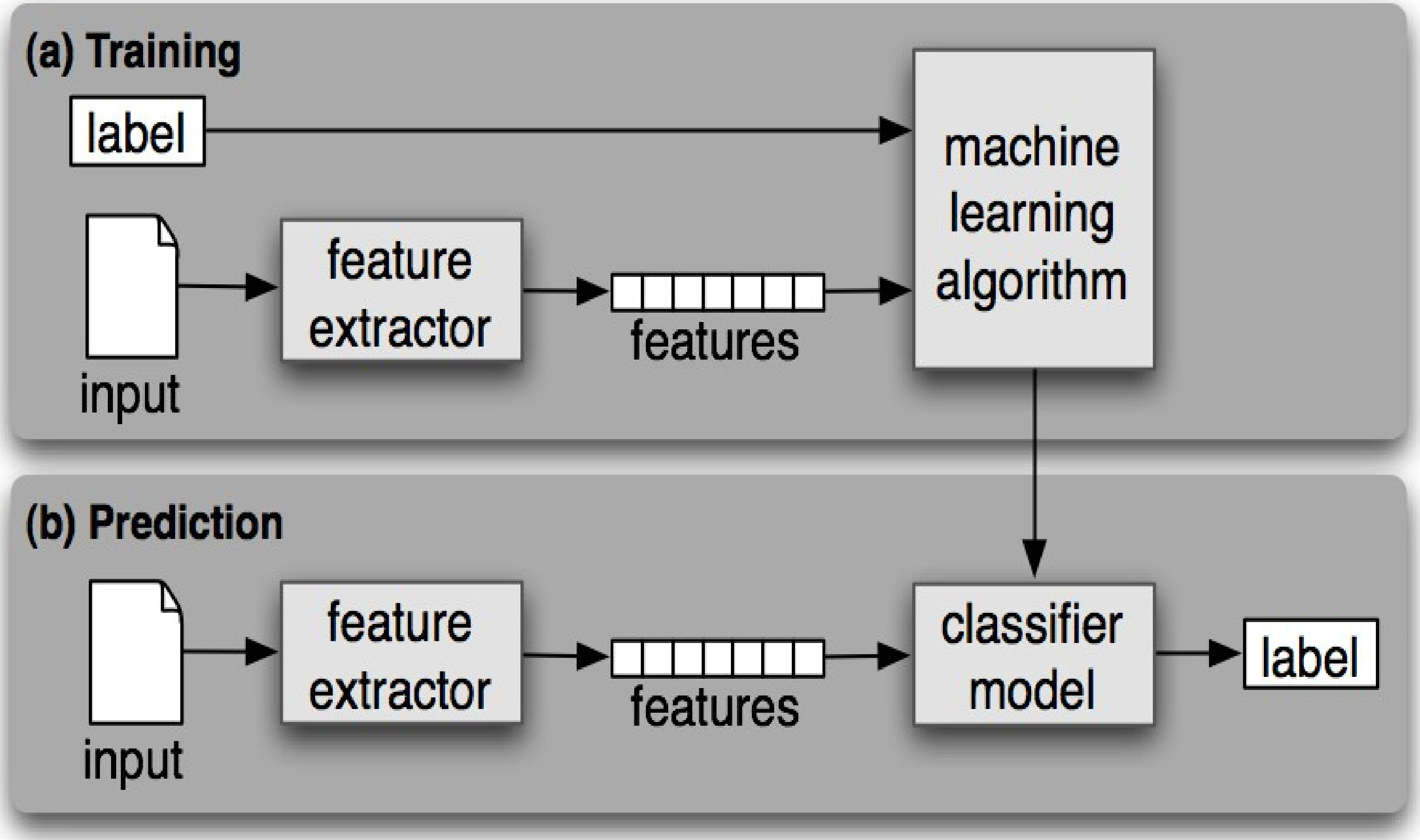
1 - Introduction

- The project is a research-based
- To study the steps to classify the class of English text among several classes
- Using Text processing and Statistical learning techniques
- The Dataset should have two or more classes
- The Dataset should be divided into two groups training group and testing group
- 80% training one , 20% testing one .

Phases


Feedback loop
Data Selection
Text pre-processing
Modeling
Evaluation / validation
Application

Classification process



2 - Objectives

- The main aim of the project is to study and understand a convenient approach that can classify English text among several classes

- With high accuracy

- given a training Dataset

3 - General approach (Theory)

- Dataset Selection
- Preparing The DataSet (Text Processing)
- Building a classifier model
- Accuracy measurement
- Testing / Validation

Each of these will be translated into steps ..

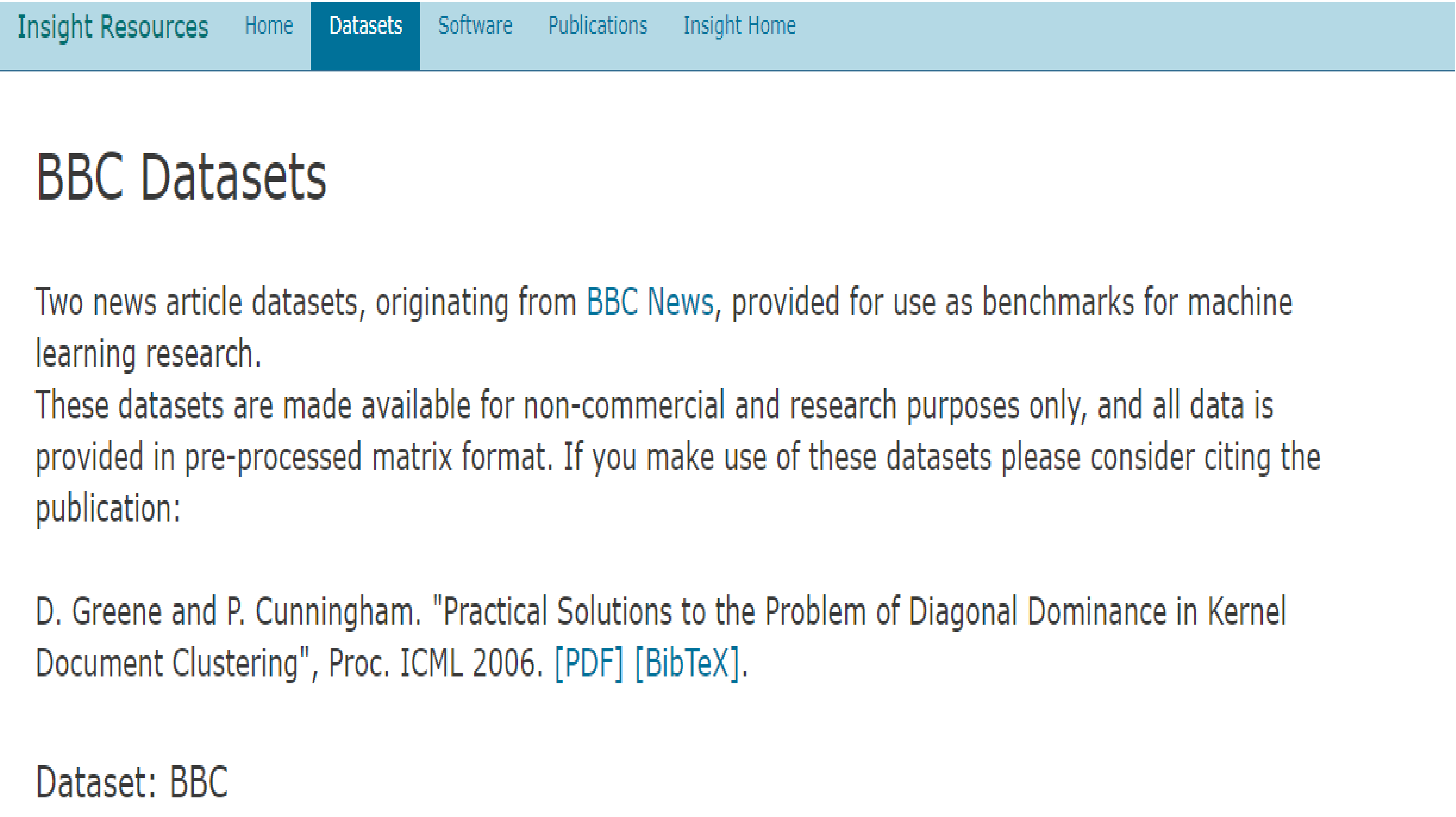
4 – Steps and Details

First Step : Dataset Selection

UCI Dataset VS BBC Dataset

BBC wins !!

BBC Dataset



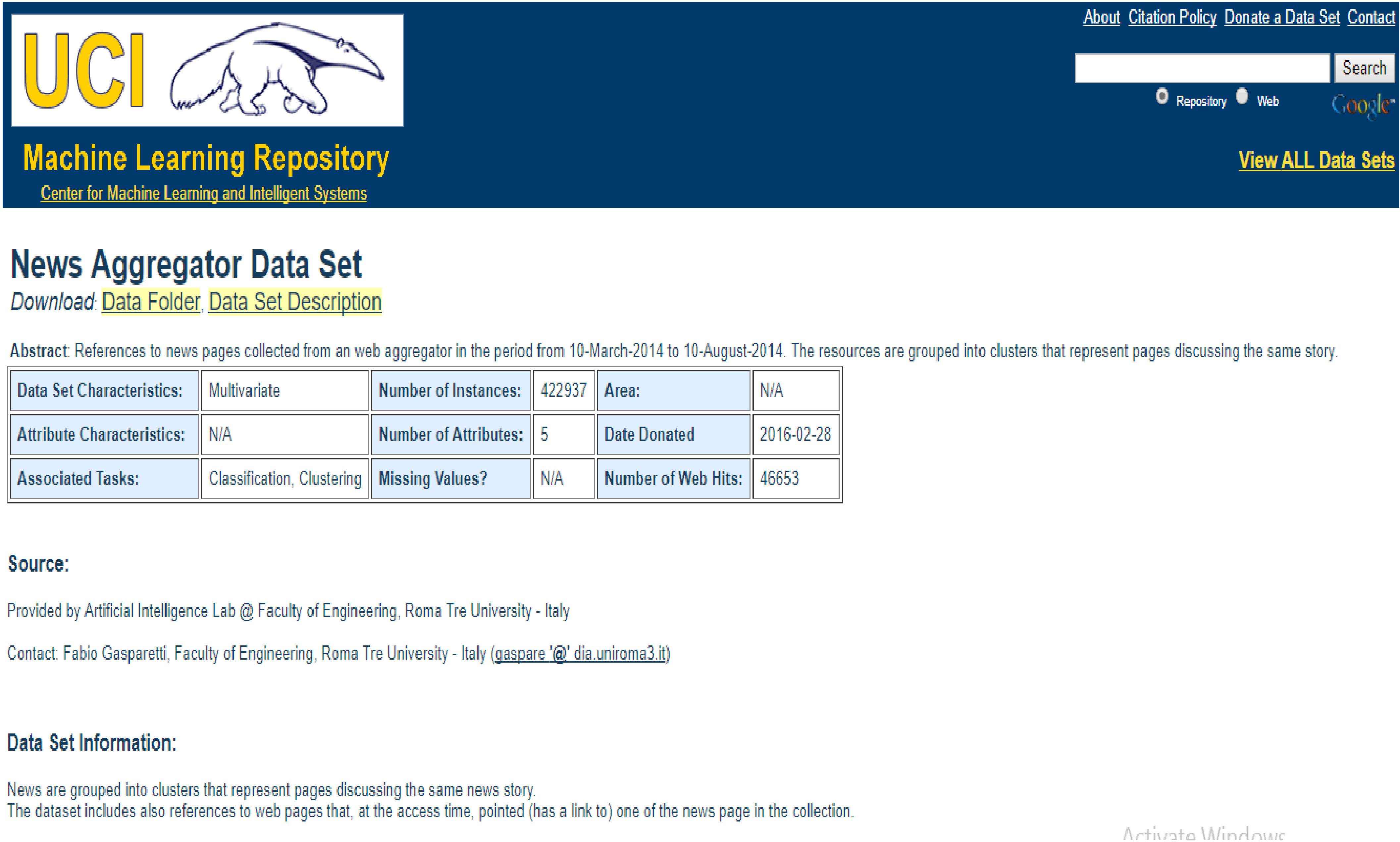
BBC Datasets

Two news article datasets, originating from BBC News, provided for use as benchmarks for machine learning research.
These datasets are made available for non-commercial and research purposes only, and all data is provided in pre-processed matrix format. If you make use of these datasets please consider citing the publication:

D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006. [PDF] [BibTeX].

Dataset: BBC

UCI Dataset



Machine Learning Repository

Center for Machine Learning and Intelligent Systems

News Aggregator Data Set

Download: Data Folder, Data Set Description

Abstract: References to news pages collected from an web aggregator in the period from 10-March-2014 to 10-August-2014. The resources are grouped into clusters that represent pages discussing the same story.

Data Set Characteristics:	Multivariate	**Number of Instances:**	422937	**Area:**	N/A
Attribute Characteristics:	N/A	**Number of Attributes:**	5	**Date Donated**	2016-02-28
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	46653

Source:

Provided by Artificial Intelligence Lab @ Faculty of Engineering, Roma Tre University - Italy

Contact: Fabio Gasparetti, Faculty of Engineering, Roma Tre University - Italy (gaspare '@' dia.uniroma3.it)

Data Set Information:

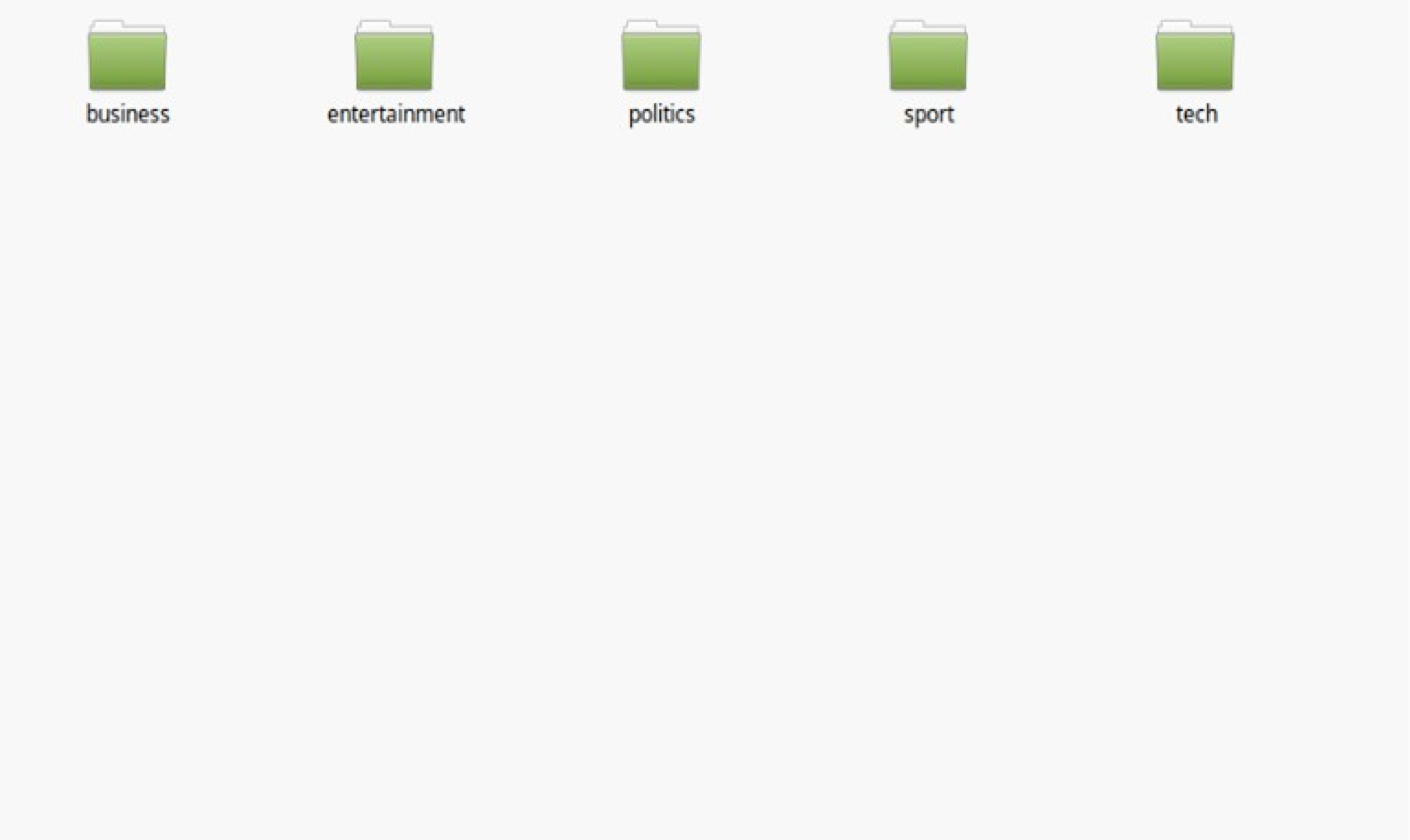
News are grouped into clusters that represent pages discussing the same news story.
The dataset includes also references to web pages that, at the access time, pointed (has a link to) one of the news page in the collection.

Why BBC's Dataset is selected?

- Certificated
- Scientifically collected
- Clean
- Highly recommended

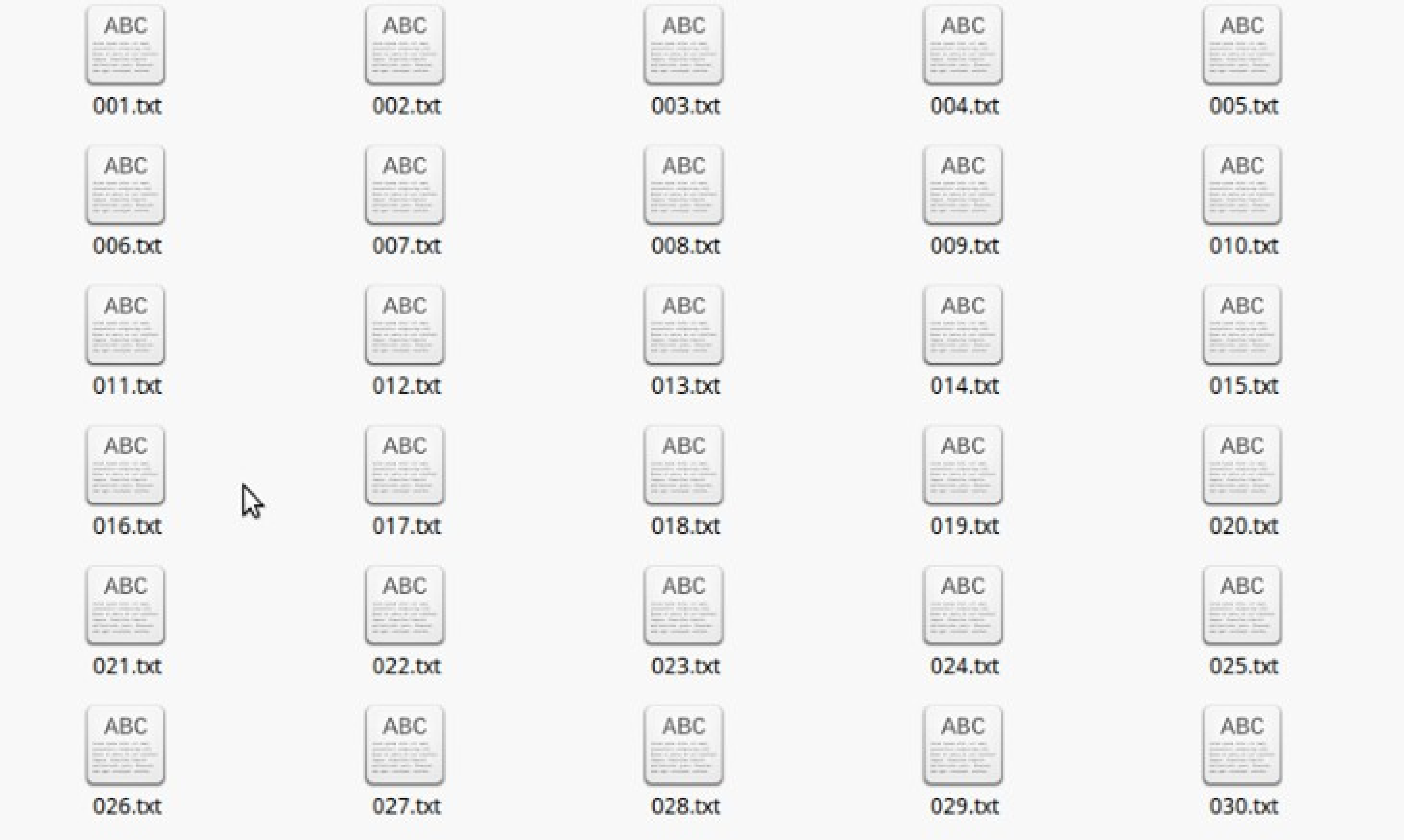
The Selected Dataset description (1)

Classes



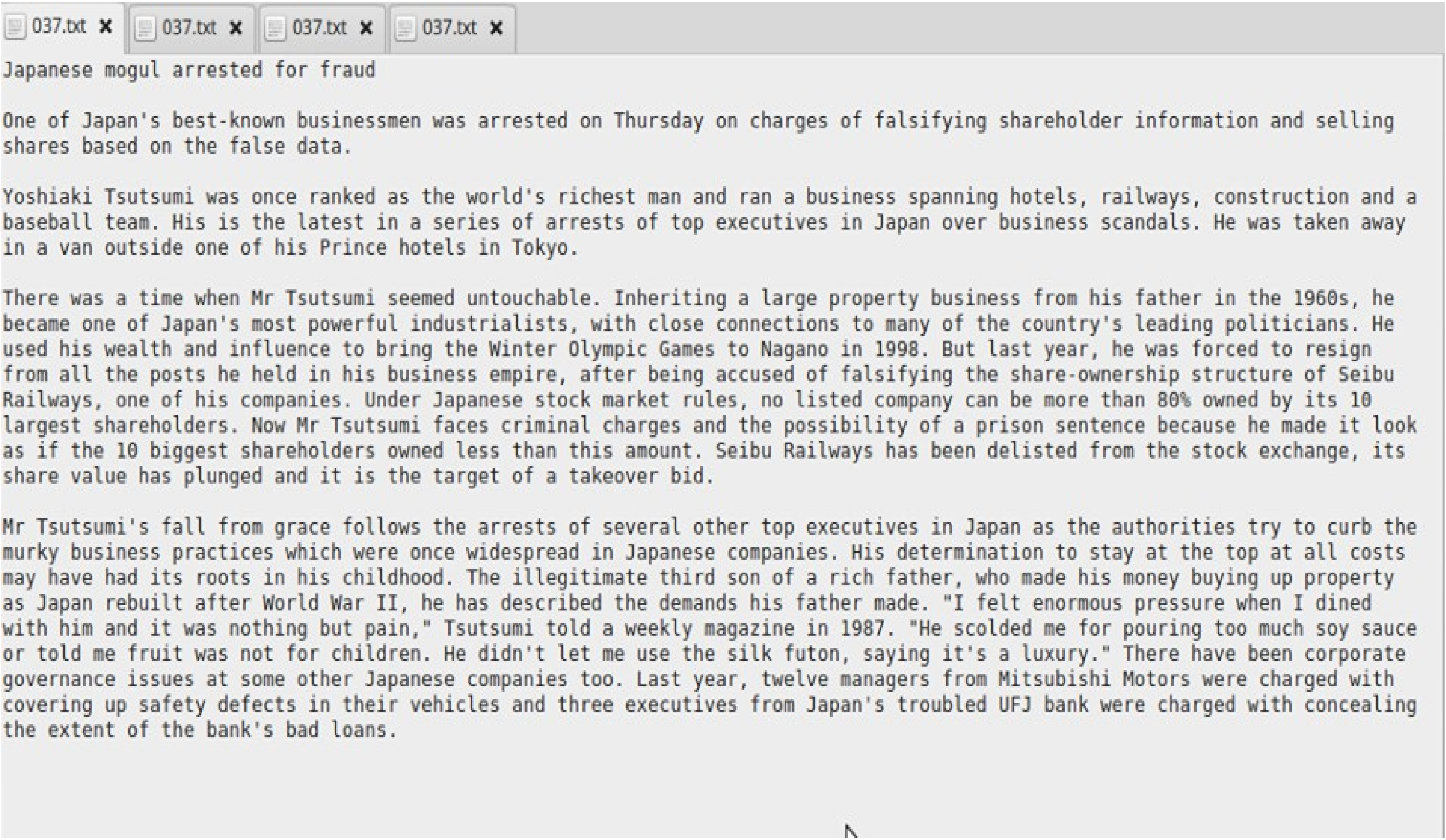
The Selected Dataset description (2)

Files within a class



The Selected Dataset description (3)

File sample



Japanese mogul arrested for fraud

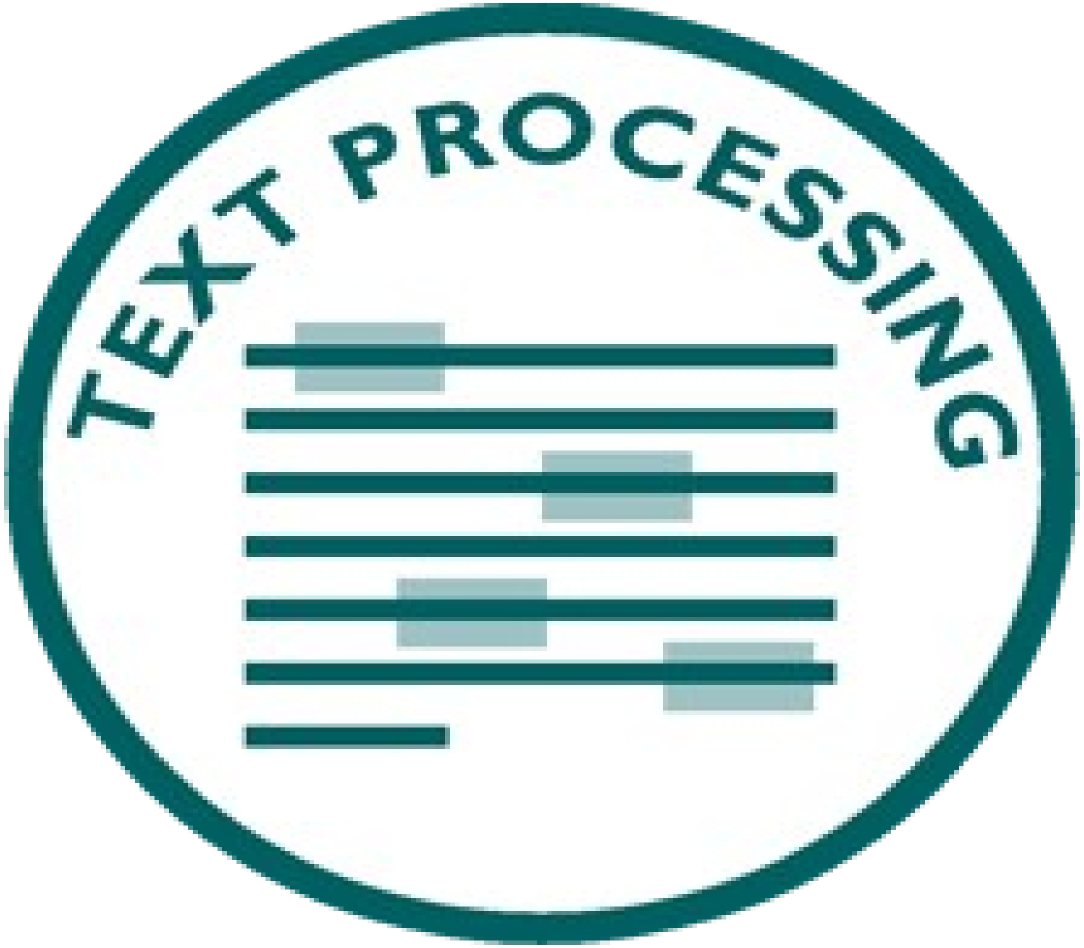
One of Japan's best-known businessmen was arrested on Thursday on charges of falsifying shareholder information and selling shares based on the false data.

Yoshiaki Tsutsumi was once ranked as the world's richest man and ran a business spanning hotels, railways, construction and a baseball team. His is the latest in a series of arrests of top executives in Japan over business scandals. He was taken away in a van outside one of his Prince hotels in Tokyo.

There was a time when Mr Tsutsumi seemed untouchable. Inheriting a large property business from his father in the 1960s, he became one of Japan's most powerful industrialists, with close connections to many of the country's leading politicians. He used his wealth and influence to bring the Winter Olympic Games to Nagano in 1998. But last year, he was forced to resign from all the posts he held in his business empire, after being accused of falsifying the share-ownership structure of Seibu Railways, one of his companies. Under Japanese stock market rules, no listed company can be more than 80% owned by its 10 largest shareholders. Now Mr Tsutsumi faces criminal charges and the possibility of a prison sentence because he made it look as if the 10 biggest shareholders owned less than this amount. Seibu Railways has been delisted from the stock exchange, its share value has plunged and it is the target of a takeover bid.

Mr Tsutsumi's fall from grace follows the arrests of several other top executives in Japan as the authorities try to curb the murky business practices which were once widespread in Japanese companies. His determination to stay at the top at all costs may have had its roots in his childhood. The illegitimate third son of a rich father, who made his money buying up property as Japan rebuilt after World War II, he has described the demands his father made. "I felt enormous pressure when I dined with him and it was nothing but pain," Tsutsumi told a weekly magazine in 1987. "He scolded me for pouring too much soy sauce or told me fruit was not for children. He didn't let me use the silk futon, saying it's a luxury." There have been corporate governance issues at some other Japanese companies too. Last year, twelve managers from Mitsubishi Motors were charged with covering up safety defects in their vehicles and three executives from Japan's troubled UFJ bank were charged with concealing the extent of the bank's bad loans.

Preparing data (Text Processing)



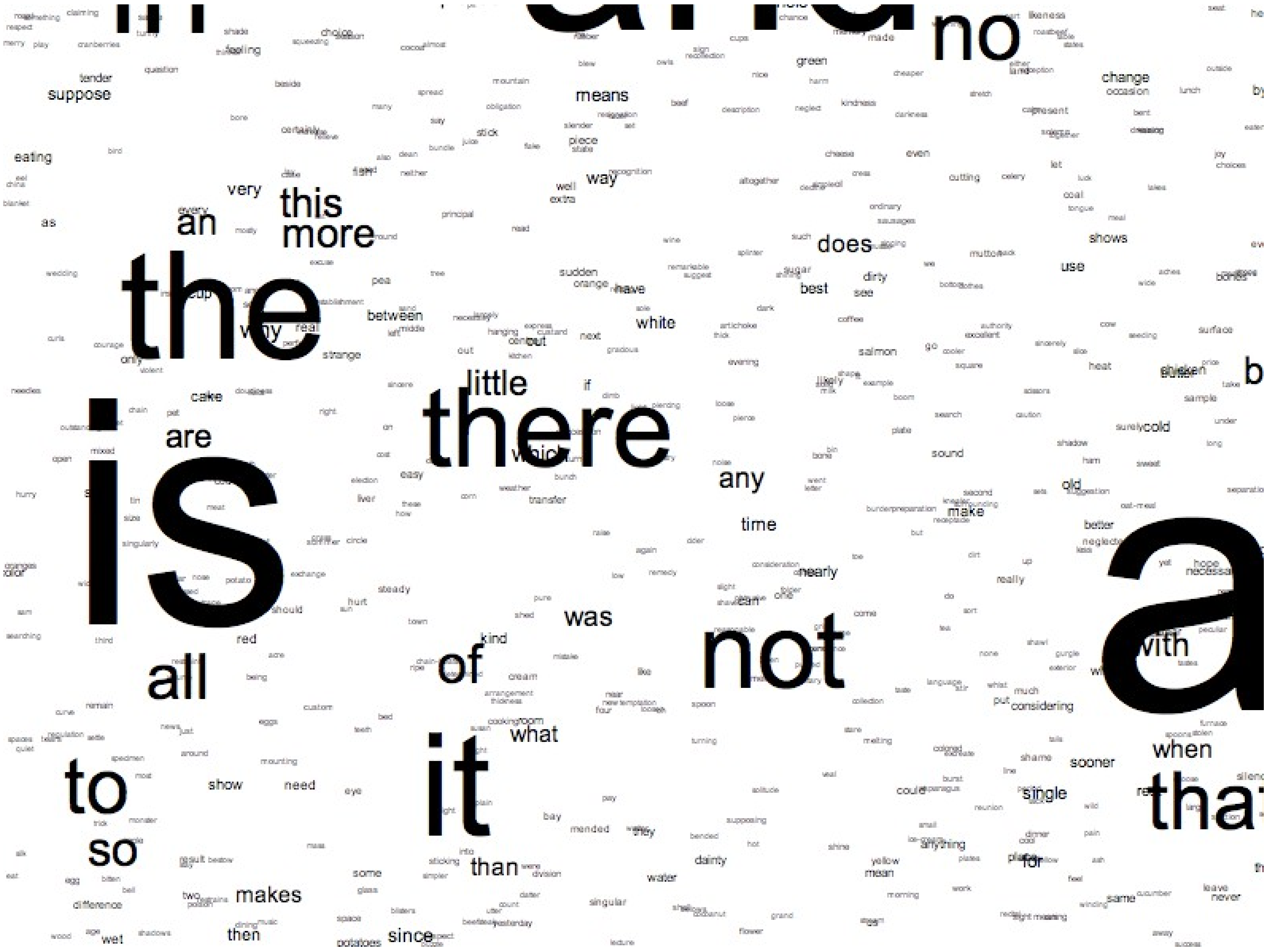
Preparing the dataset (1)

Steps

- Removing Stopwords
- Removing punctuation marks & symbols
- Getting the root form for each word (stemming)
- Getting each class file as a vector of words
- Calculating the frequency of each remaining word within each class file
- Getting only the words with highest frequency to represent the file
- Merging the vectors for each class in one vector without repetition to represent each class

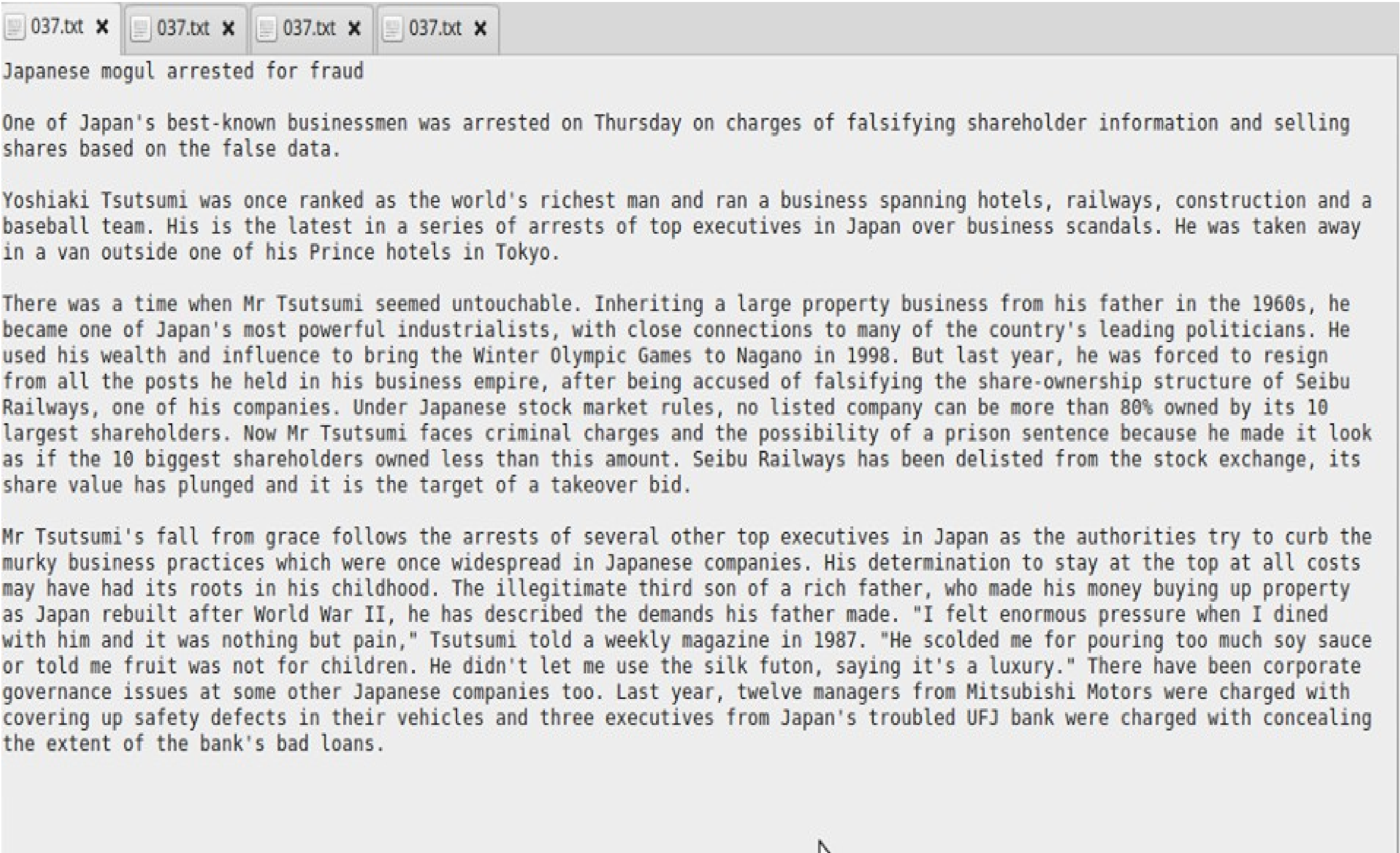
The steps should be done in the same order !!

Stopwords



is
the
a
there
not
it
and
no
that
to
so
all
of
this
more
an
are
little
was
with
than
what
any
does
time
makes
then
since
when
means
suppose
eating
very
as
cake
kind
way
sudden
white
salmon
best
use
shows
change
single
sooner
could
water
dainty
mean
cream
difference
wet
show
need
eye
potatoes
some
red
between
strange
easy
liver
transfer
make
sound
better
old
cold
heat
really
nearly
green
piece
stick
tender
choice
feeling
made
cutting
sugar
dirty
surface
hurt
steady
should
mended
singular
considering
much
like
bay
yellow
for
anything
way
well
extra
present
pea
why
real

Before Removing Stopwords



Japanese mogul arrested for fraud

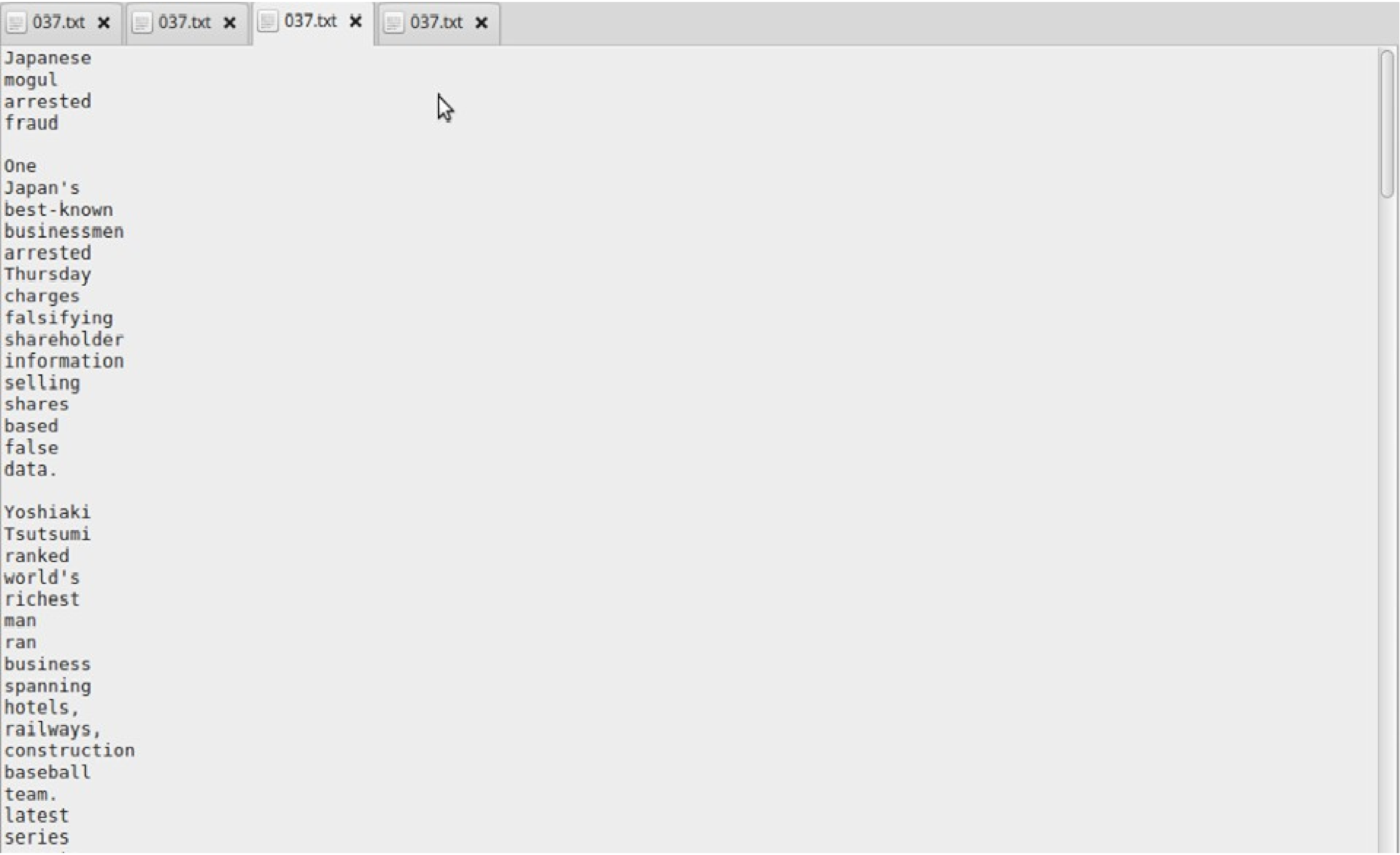
One of Japan's best-known businessmen was arrested on Thursday on charges of falsifying shareholder information and selling shares based on the false data.

Yoshiaki Tsutsumi was once ranked as the world's richest man and ran a business spanning hotels, railways, construction and a baseball team. His is the latest in a series of arrests of top executives in Japan over business scandals. He was taken away in a van outside one of his Prince hotels in Tokyo.

There was a time when Mr Tsutsumi seemed untouchable. Inheriting a large property business from his father in the 1960s, he became one of Japan's most powerful industrialists, with close connections to many of the country's leading politicians. He used his wealth and influence to bring the Winter Olympic Games to Nagano in 1998. But last year, he was forced to resign from all the posts he held in his business empire, after being accused of falsifying the share-ownership structure of Seibu Railways, one of his companies. Under Japanese stock market rules, no listed company can be more than 80% owned by its 10 largest shareholders. Now Mr Tsutsumi faces criminal charges and the possibility of a prison sentence because he made it look as if the 10 biggest shareholders owned less than this amount. Seibu Railways has been delisted from the stock exchange, its share value has plunged and it is the target of a takeover bid.

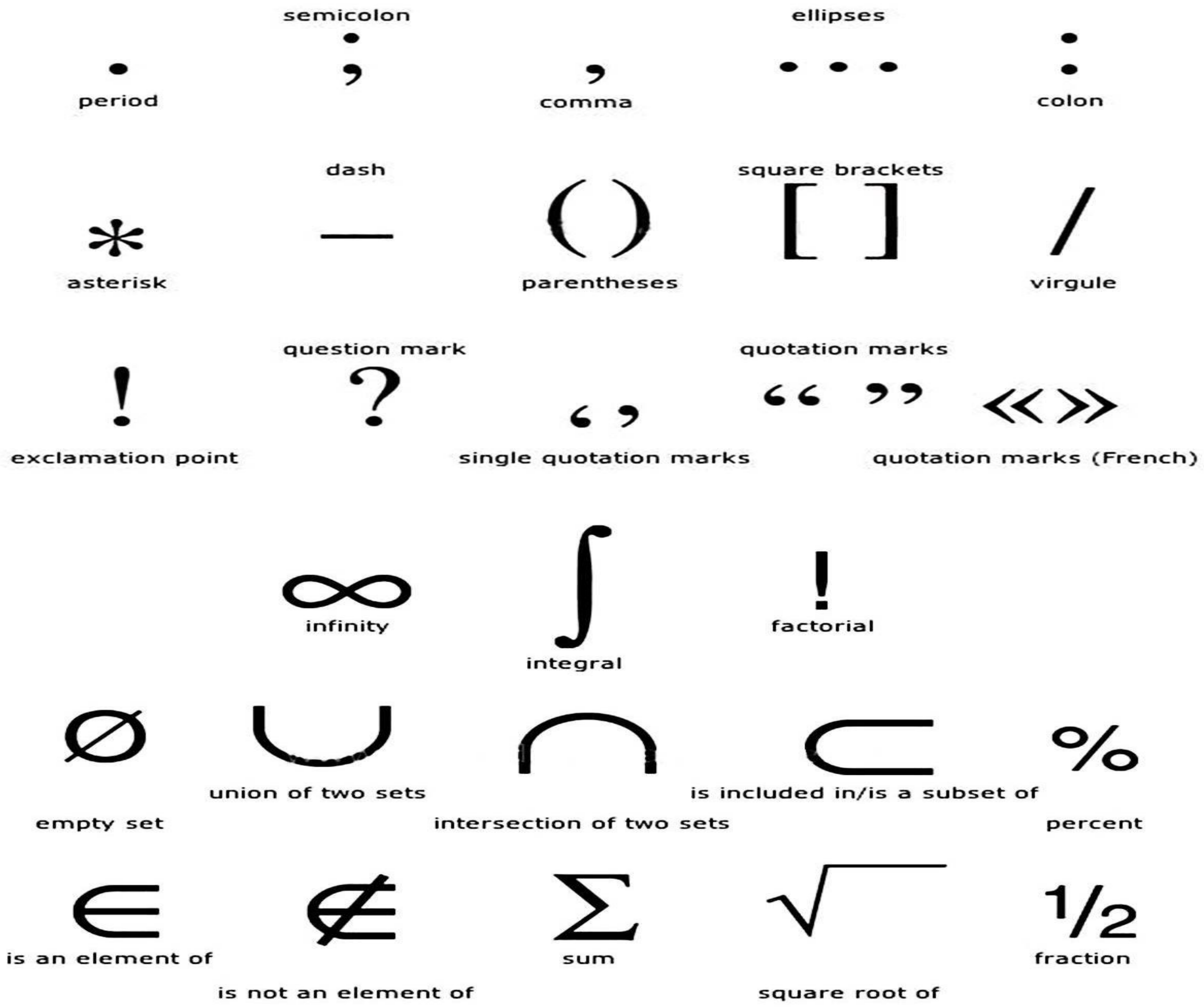
Mr Tsutsumi's fall from grace follows the arrests of several other top executives in Japan as the authorities try to curb the murky business practices which were once widespread in Japanese companies. His determination to stay at the top at all costs may have had its roots in his childhood. The illegitimate third son of a rich father, who made his money buying up property as Japan rebuilt after World War II, he has described the demands his father made. "I felt enormous pressure when I dined with him and it was nothing but pain," Tsutsumi told a weekly magazine in 1987. "He scolded me for pouring too much soy sauce or told me fruit was not for children. He didn't let me use the silk futon, saying it's a luxury." There have been corporate governance issues at some other Japanese companies too. Last year, twelve managers from Mitsubishi Motors were charged with covering up safety defects in their vehicles and three executives from Japan's troubled UFJ bank were charged with concealing the extent of the bank's bad loans.

After Removing Stopwords

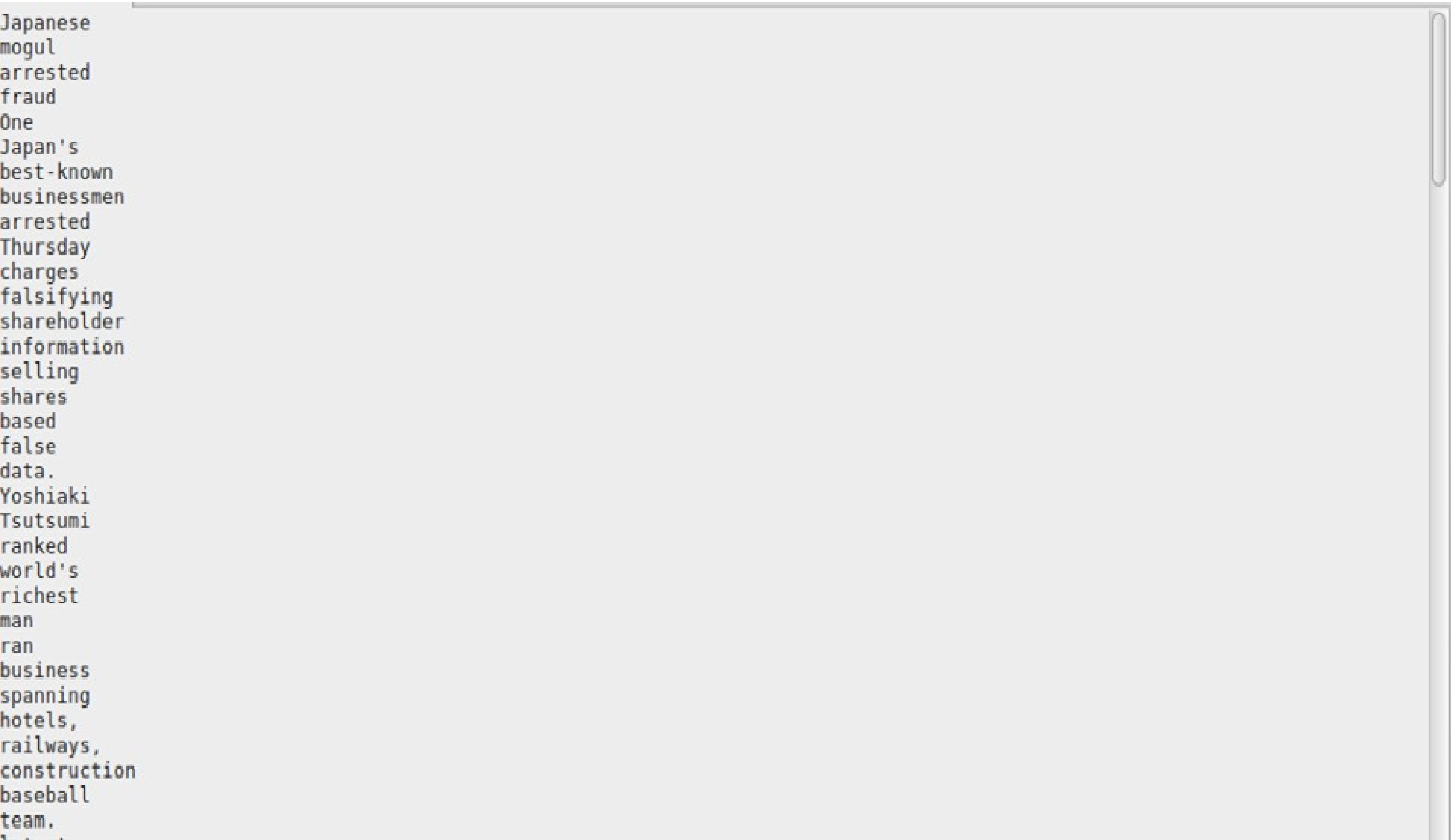


```
Japanese
mogul
arrested
fraud

One
Japan's
best-known
businessmen
arrested
Thursday
charges
falsifying
shareholder
information
selling
shares
based
false
data.

Yoshiaki
Tsutsumi
ranked
world's
richest
man
ran
business
spanning
hotels,
railways,
construction
baseball
team.
latest
series
```

Punctuation marks & Symbols

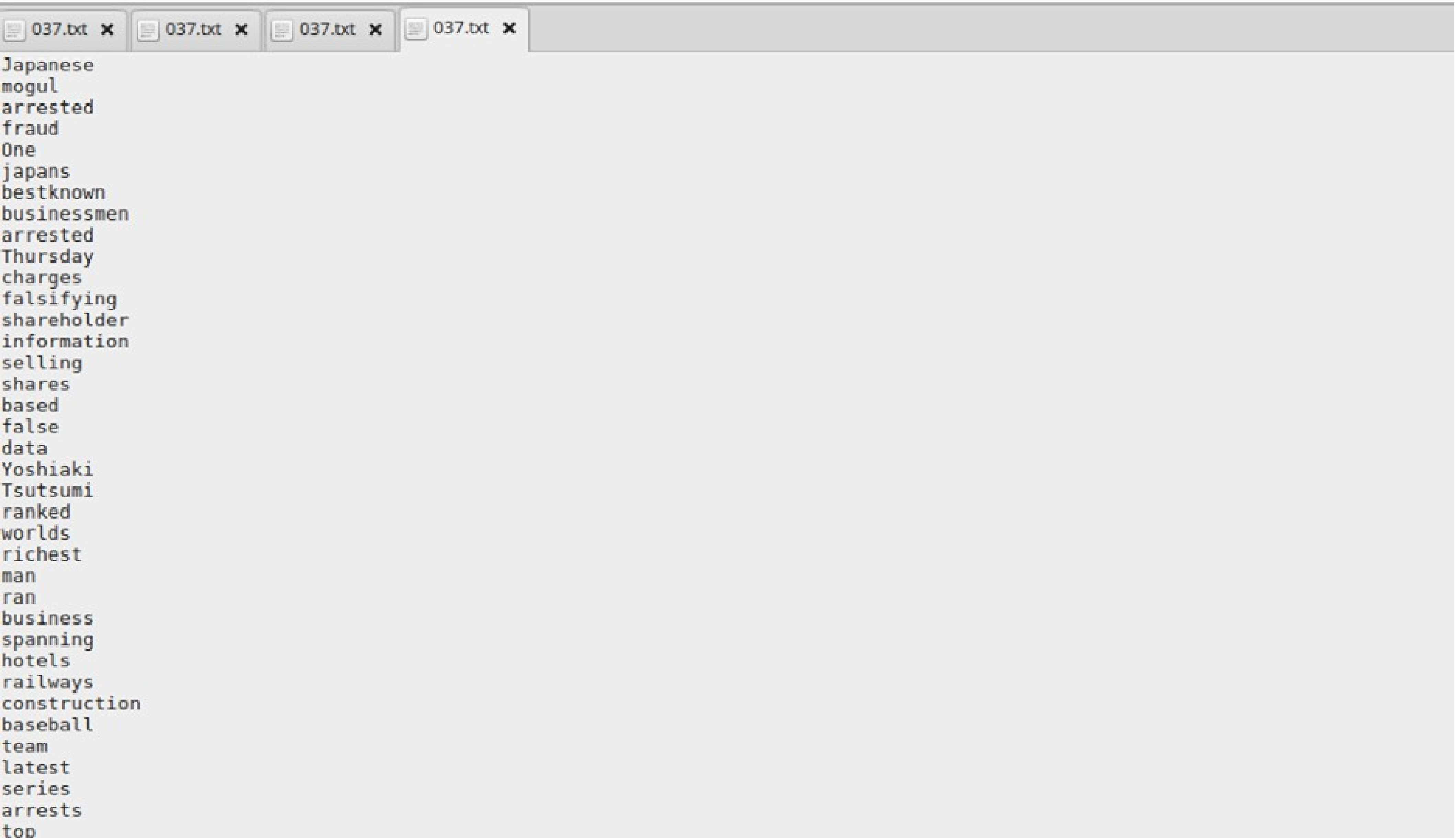


Before Removing punctuation marks and symbols



```
Japanese
mogul
arrested
fraud
One
Japan's
best-known
businessmen
arrested
Thursday
charges
falsifying
shareholder
information
selling
shares
based
false
data.
Yoshiaki
Tsutsumi
ranked
world's
richest
man
ran
business
spanning
hotels,
railways,
construction
baseball
team.
```

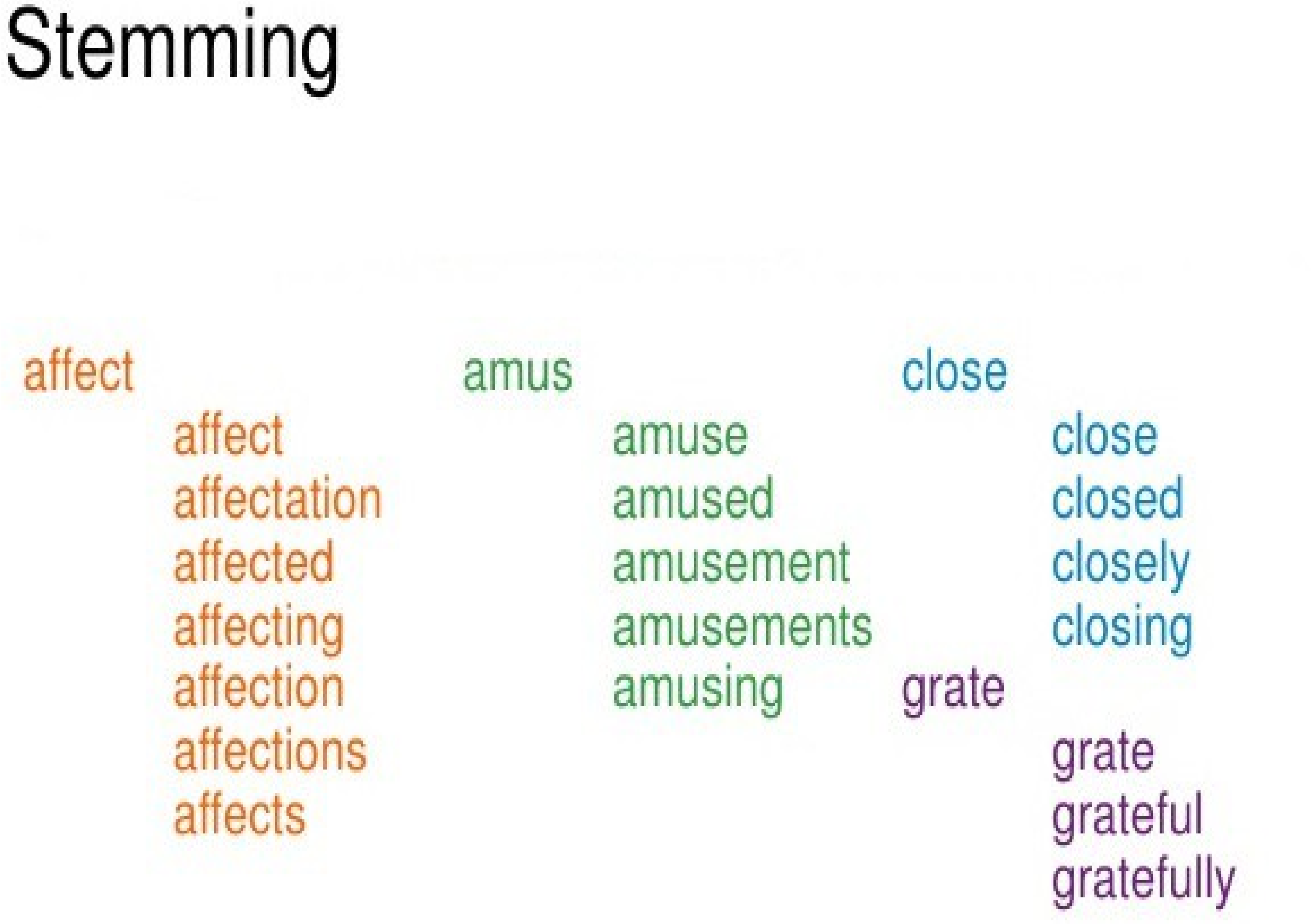
After Removing punctuation marks and symbols



```
Japanese
mogul
arrested
fraud
One
japans
bestknown
businessmen
arrested
Thursday
charges
falsifying
shareholder
information
selling
shares
based
false
data
Yoshiaki
Tsutsumi
ranked
worlds
richest
man
ran
business
spanning
hotels
railways
construction
baseball
team
latest
series
arrests
top
```

Stemming process

Getting the root form of the word



Stemming

- affect
  - affect
  - affectation
  - affected
  - affecting
  - affection
  - affections
  - affects
- amus
  - amuse
  - amused
  - amusement
  - amusements
  - amusing
- close
  - close
  - closed
  - closely
  - closing
- grate
  - grate
  - grateful
  - gratefully

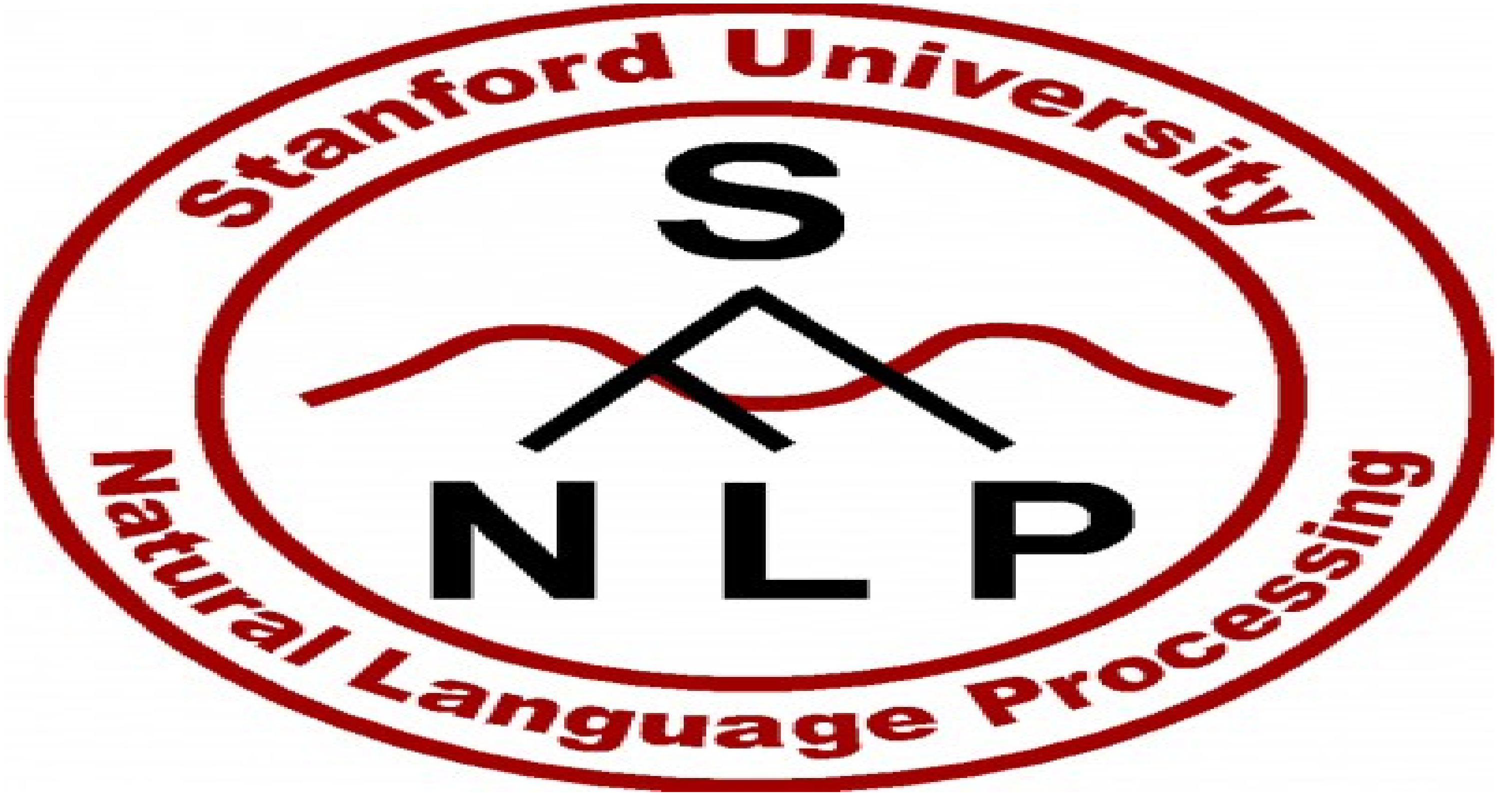
Stemmer

Porter Stemmer VS Stanford CoreNLP

Stanford CoreNLP wins !!

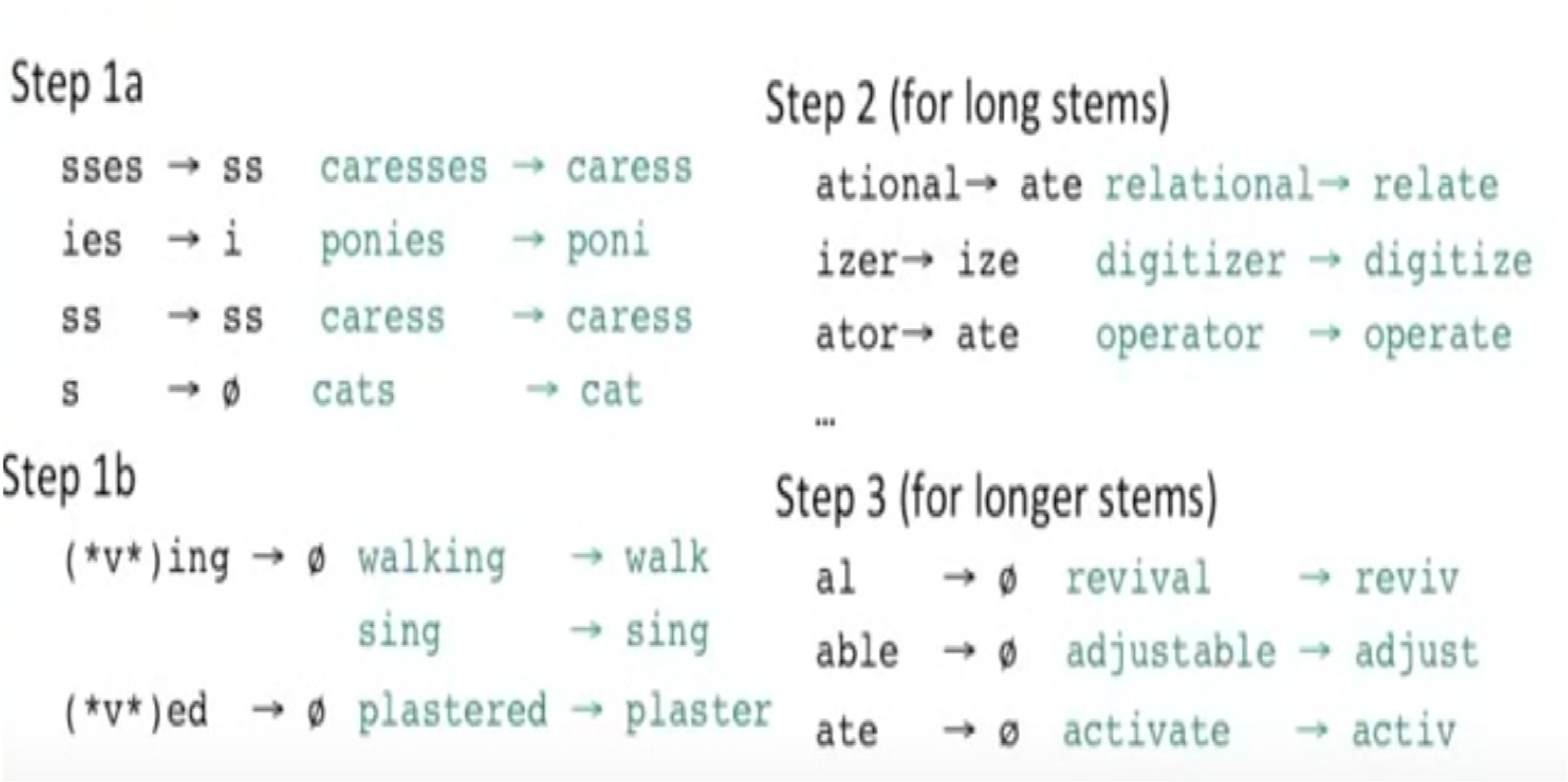
Stanford CoreNLP

- One of the most powerful integrated systems for NLP
- Uses complex techniques for stemming



Porter Stemmer

- The most common English stemmer
- It's simple



Step 1a

sses	→	ss	caresses	→	caress
ies	→	i	ponies	→	poni
ss	→	ss	caress	→	caress
s	→	ø	cats	→	cat

Step 1b

(*v*)ing	→	ø	walking	→	walk
			sing	→	sing
(*v*)ed	→	ø	plastered	→	plaster

Step 2 (for long stems)

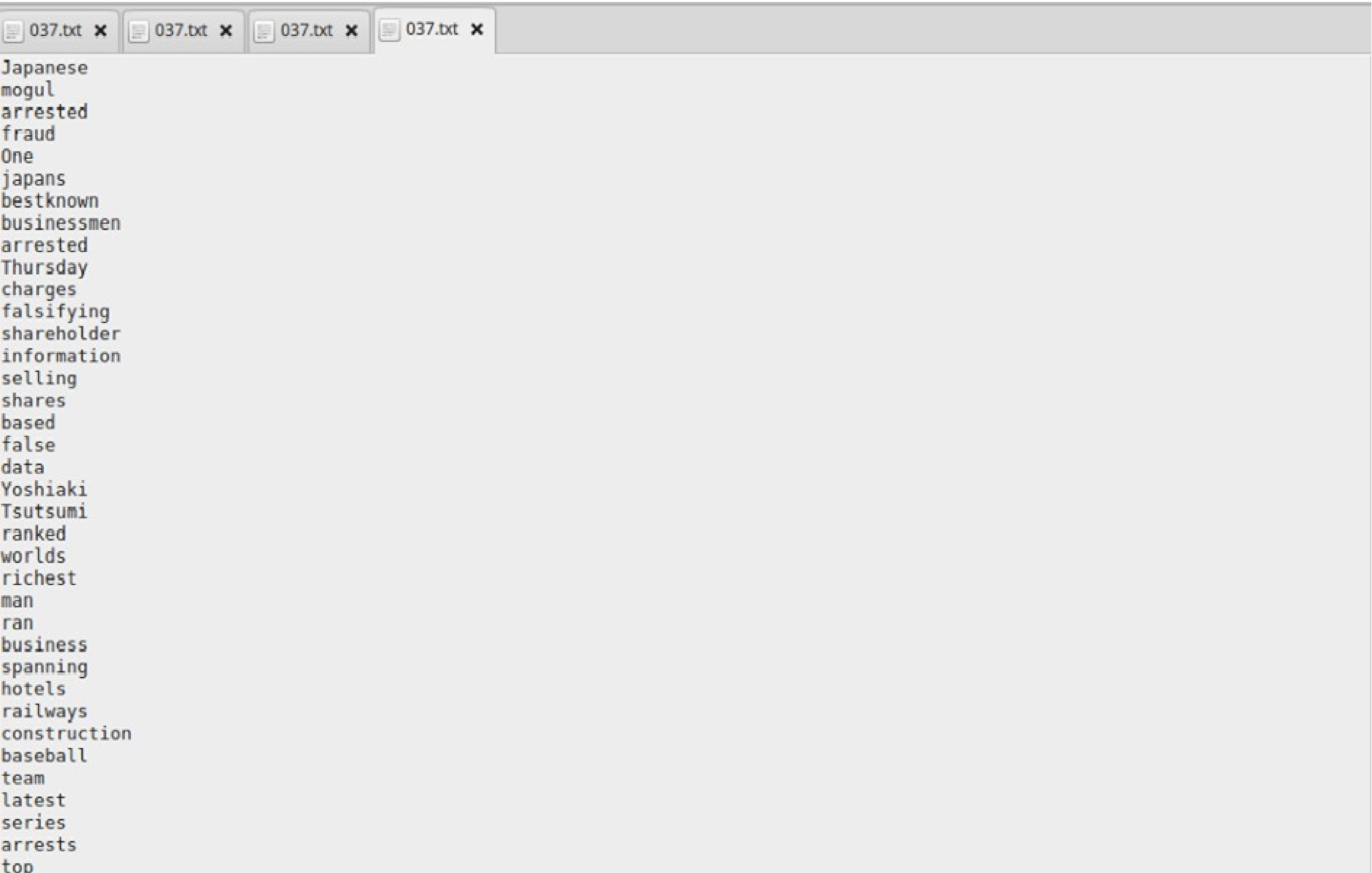
ational	→	ate	relational	→	relate
izer	→	ize	digitizer	→	digitize
ator	→	ate	operator	→	operate

...

Step 3 (for longer stems)

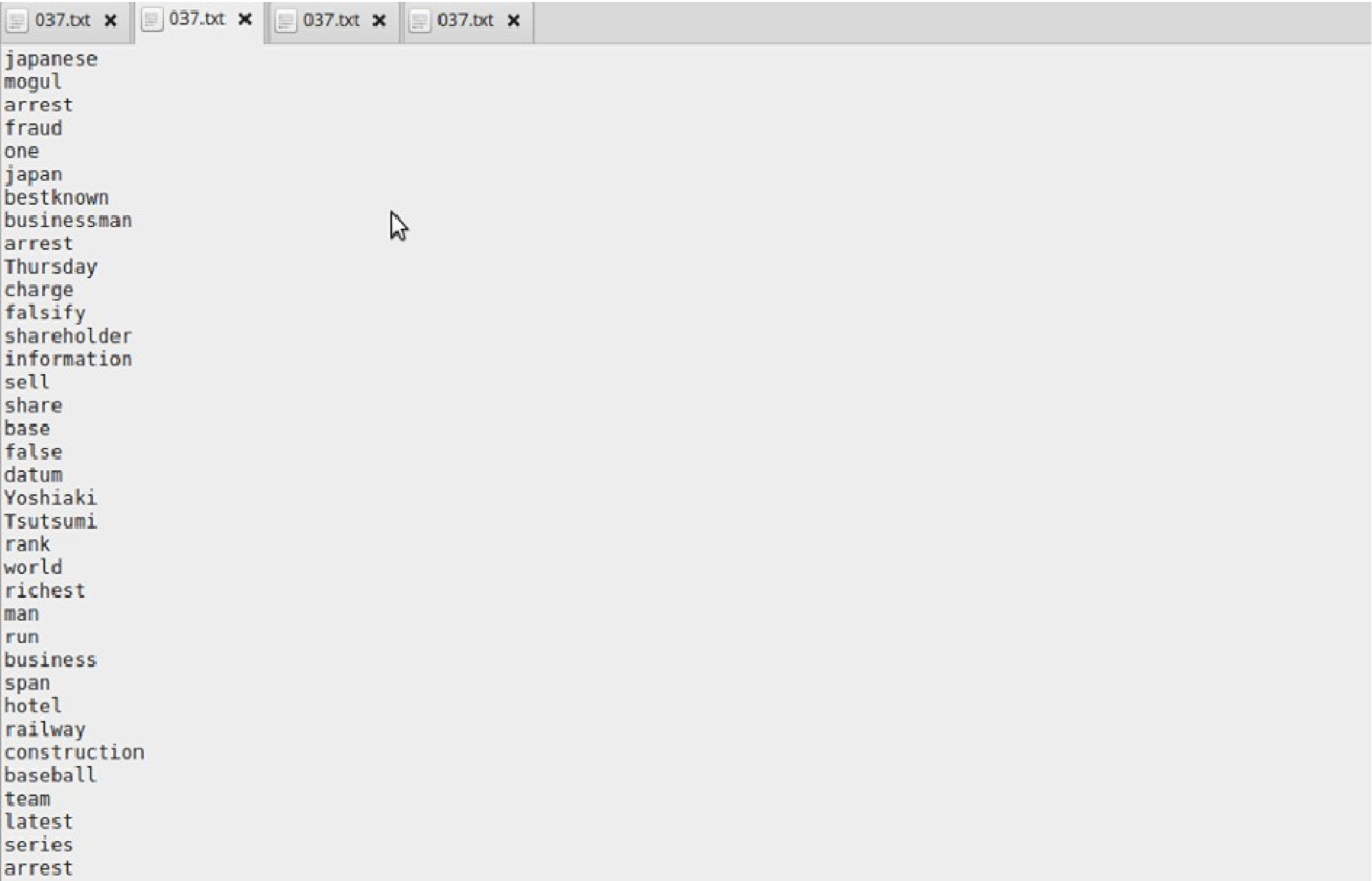
al	→	ø	revival	→	reviv
able	→	ø	adjustable	→	adjust
ate	→	ø	activate	→	activ

Before Stemming



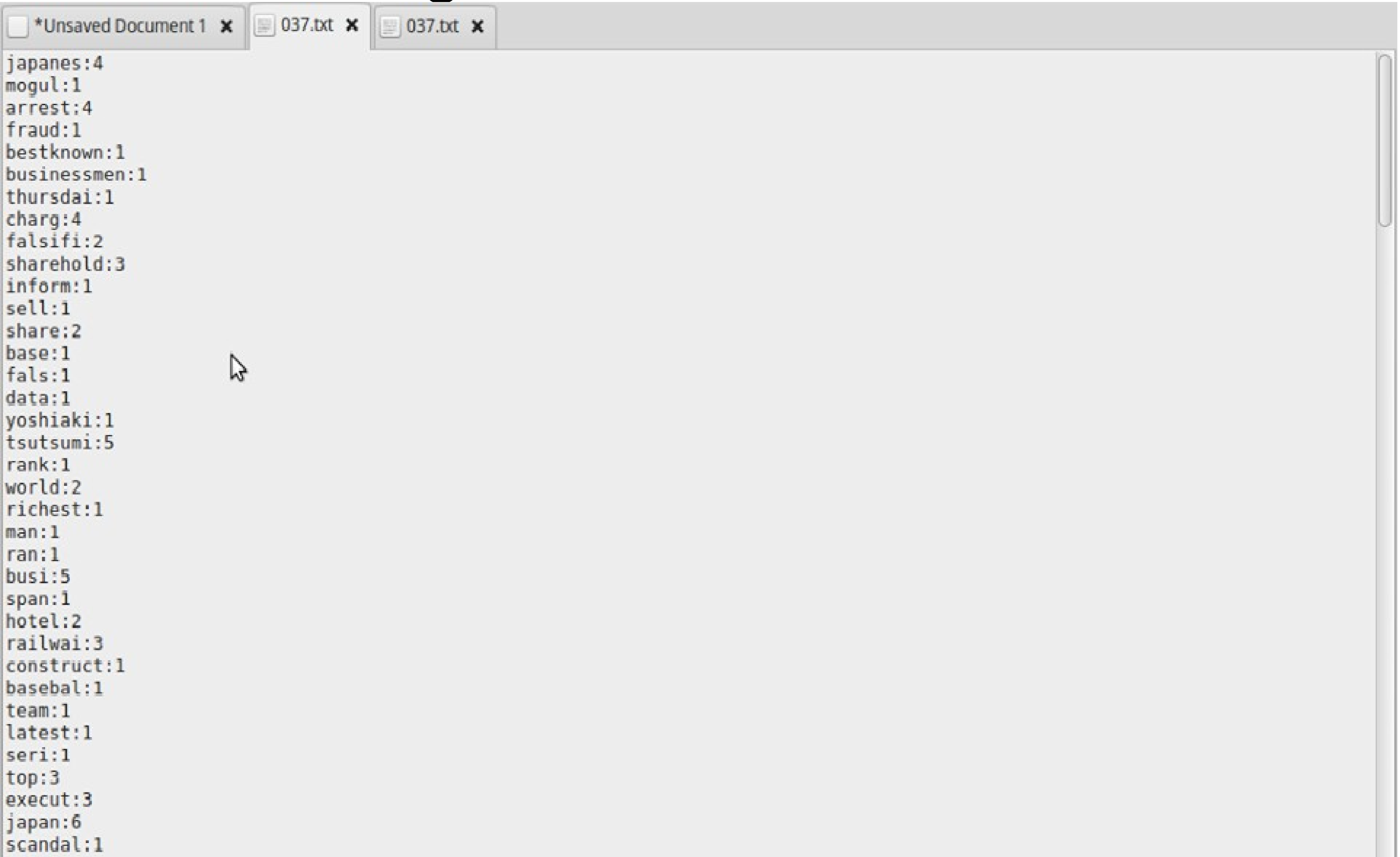
```
Japanese
mogul
arrested
fraud
One
japans
bestknown
businessmen
arrested
Thursday
charges
falsifying
shareholder
information
selling
shares
based
false
data
Yoshiaki
Tsutsumi
ranked
worlds
richest
man
ran
business
spanning
hotels
railways
construction
baseball
team
latest
series
arrests
top
```

After Stemming



```
japanese
mogul
arrest
fraud
one
japan
bestknown
businessman
arrest
Thursday
charge
falsify
shareholder
information
sell
share
base
false
datum
Yoshiaki
Tsutsumi
rank
world
richest
man
run
business
span
hotel
railway
construction
baseball
team
latest
series
arrest
```

Calculating the frequency of each remaining word within each class file



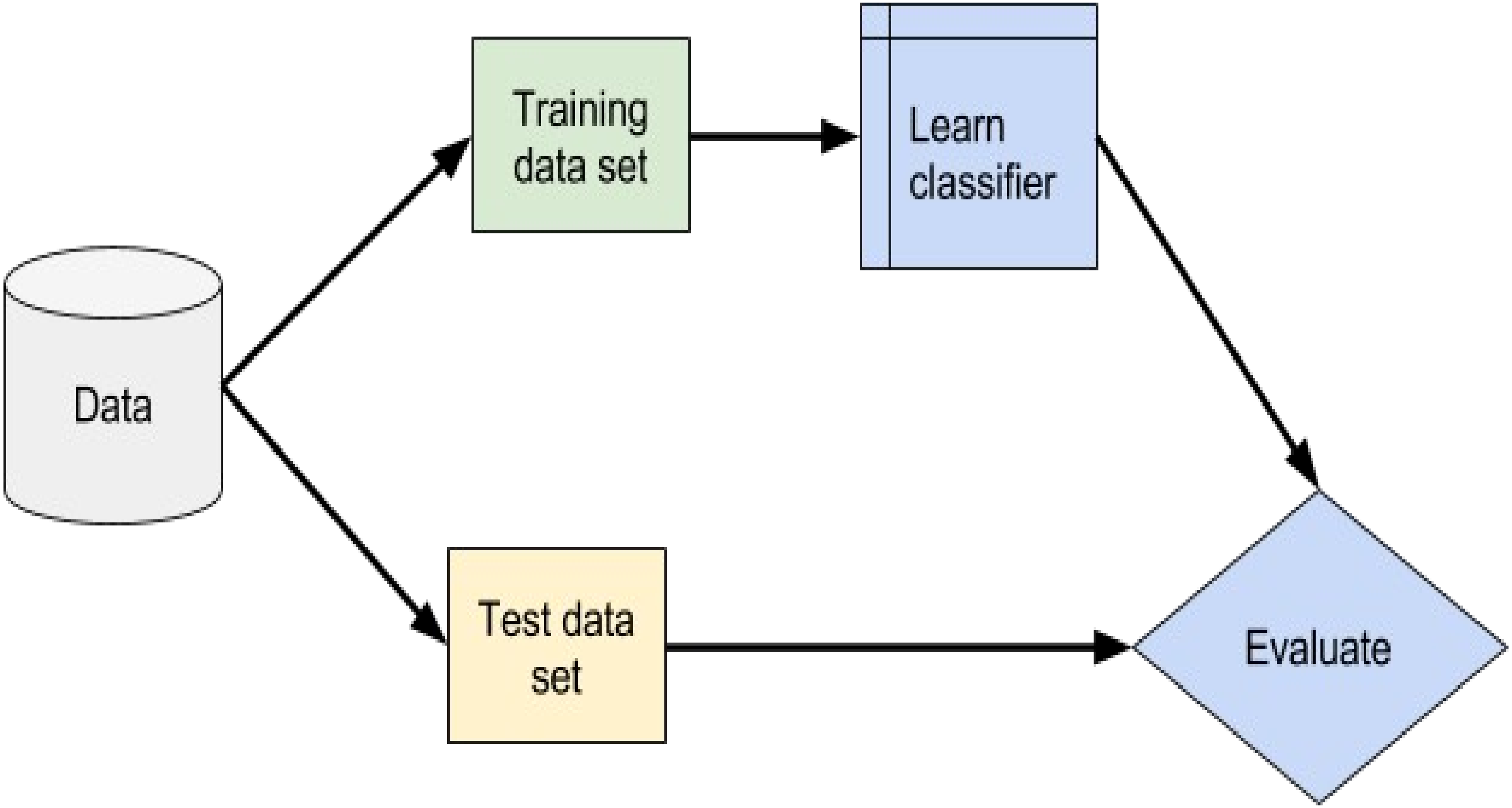
```
japanes:4
mogul:1
arrest:4
fraud:1
bestknown:1
businessmen:1
thursdai:1
charg:4
falsifi:2
sharehold:3
inform:1
sell:1
share:2
base:1
fals:1
data:1
yoshiaki:1
tsutsumi:5
rank:1
world:2
richest:1
man:1
ran:1
busi:5
span:1
hotel:2
railwai:3
construct:1
basebal:1
team:1
latest:1
seri:1
top:3
execut:3
japan:6
scandal:1
```

Getting only the words with highest frequency to represent the file

- This step plays a vital role in classification accuracy

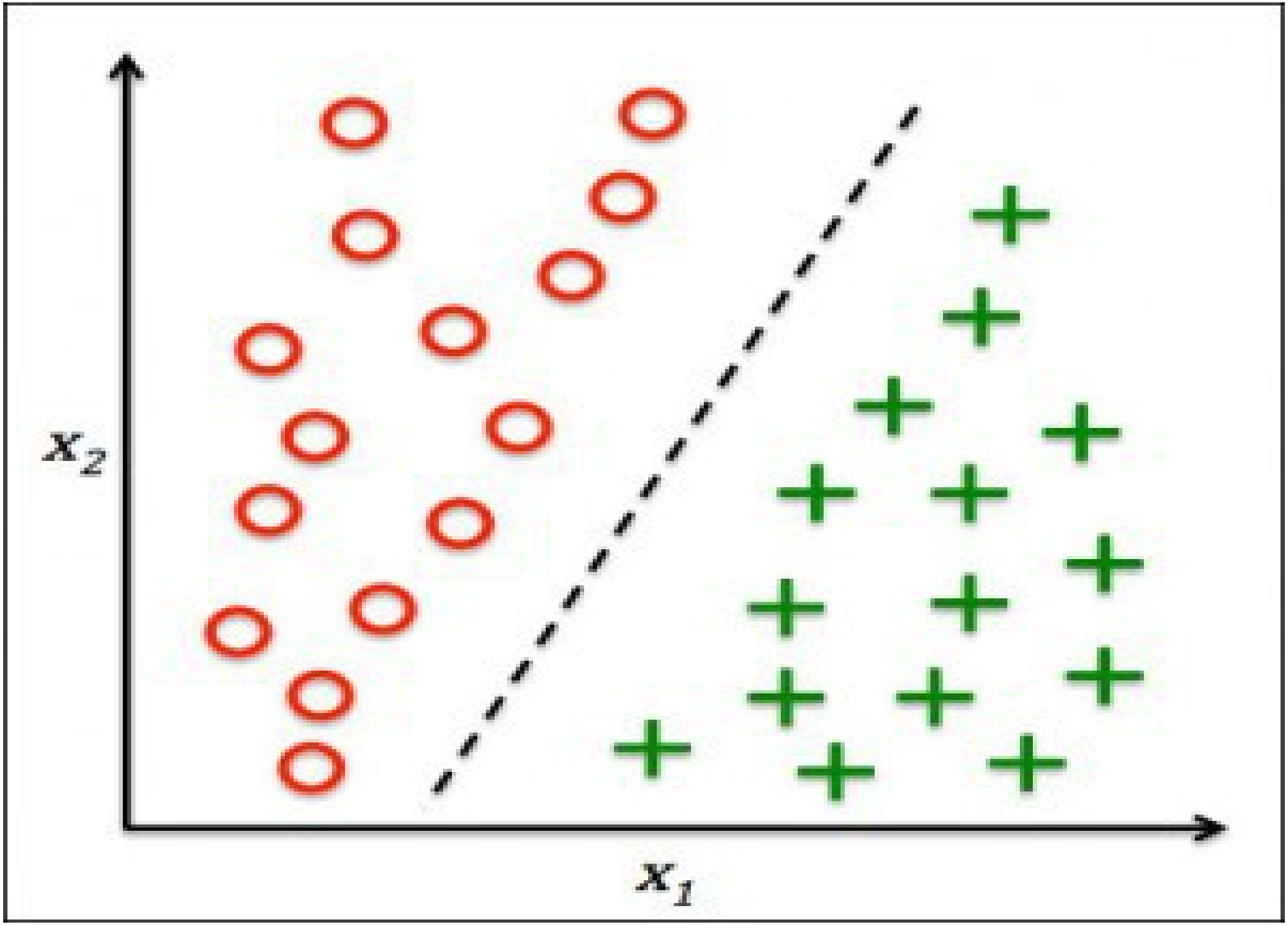
- Determination of the minimum frequency is a real challenge !!

Classification (1)

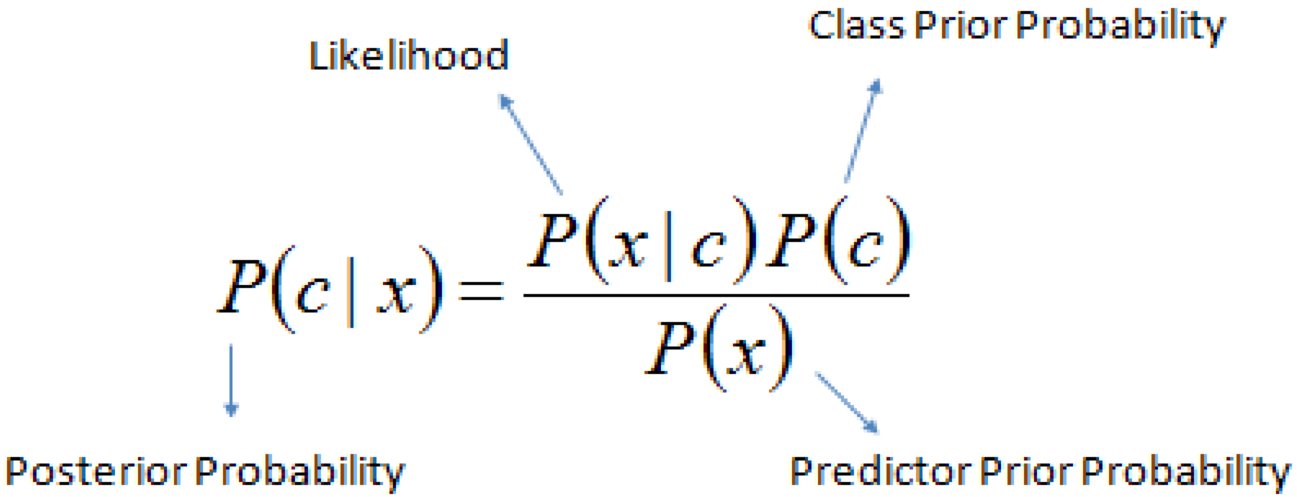


Classification (2)

scatter plot with fitted Decision boundary



Naive Bayes Classifier



Likelihood

Class Prior Probability

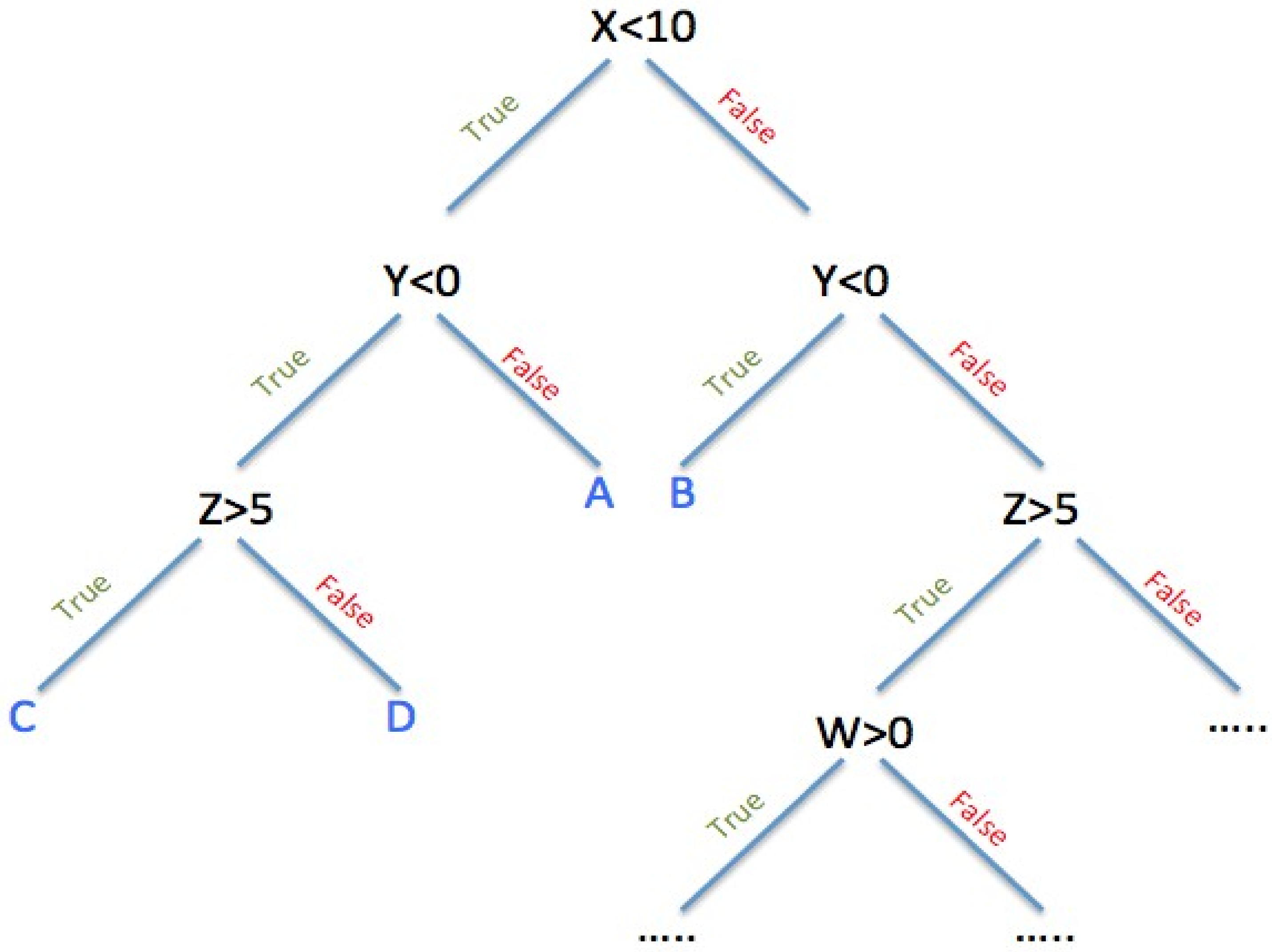
$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Decision tree classifier



Accuracy measurement (1)

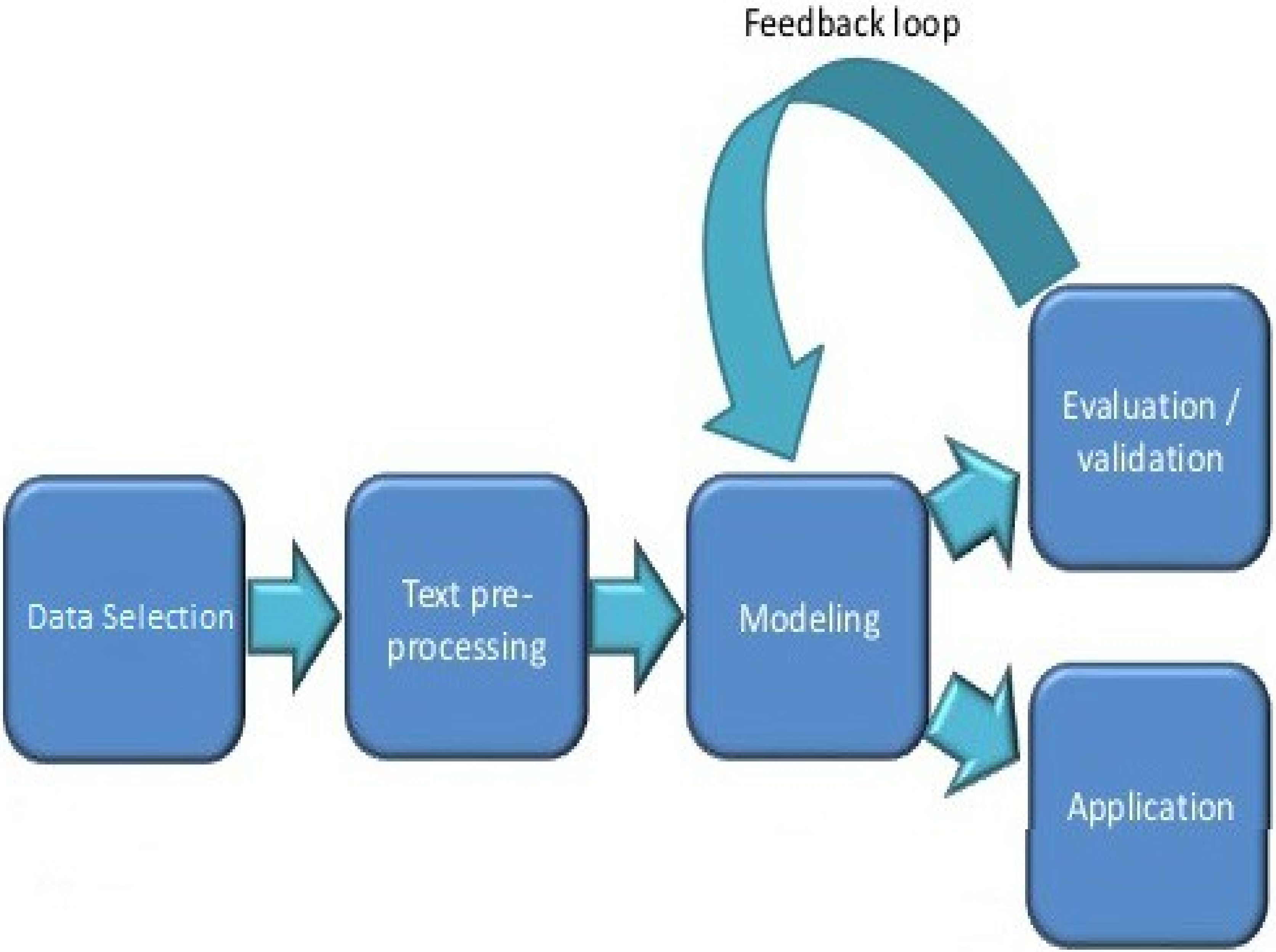
- Calculating the percentage of success rate according to previous steps .

- Some changes could be happened to get the required accuracy

Accuracy measurement (2)

Possible changes

- Minimum frequency for the words that represent each class
- Testing more than one classifier
- Testing more than one Stemmer

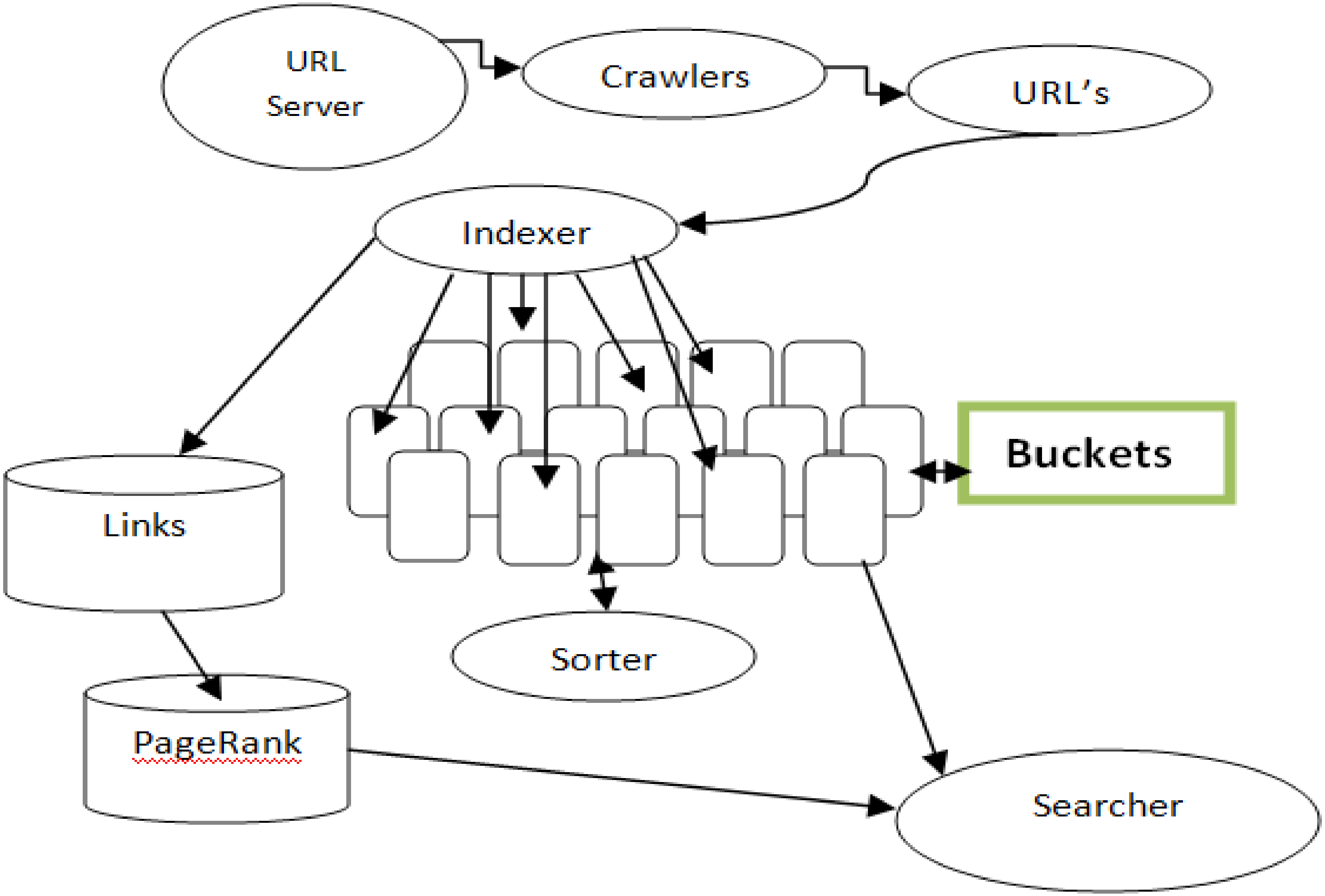

Feedback loop
Data Selection
Text pre-processing
Modeling
Evaluation / validation
Application

Testing And Validation

- 20% of the Dataset will be considered as a testing Dataset

- A software tool will be developed to predict the class of each testing case to evaluate the success rate

Applications : Search engine indexing


URL Server
Crawlers
URL's
Indexer
Buckets
Links
Sorter
PageRank
Searcher