Big Data: Supervised Learning Lab Application: Spam Filtering

Vibhav Gogate

The University of Texas at Dallas

SPAM FILTERING

- Given: A set of emails
- To do: Classify each email as either "spam" or "ham."

Steps in Supervised Learning: Revisited

1. Determine the representation for “x,f(x)” and determine what “x” to use

   Feature Engineering

2. Gather a training set (not all data is kosher)

   Data Cleaning

3. Select a suitable evaluation method
4. Find a suitable learning algorithm among a plethora of available choices

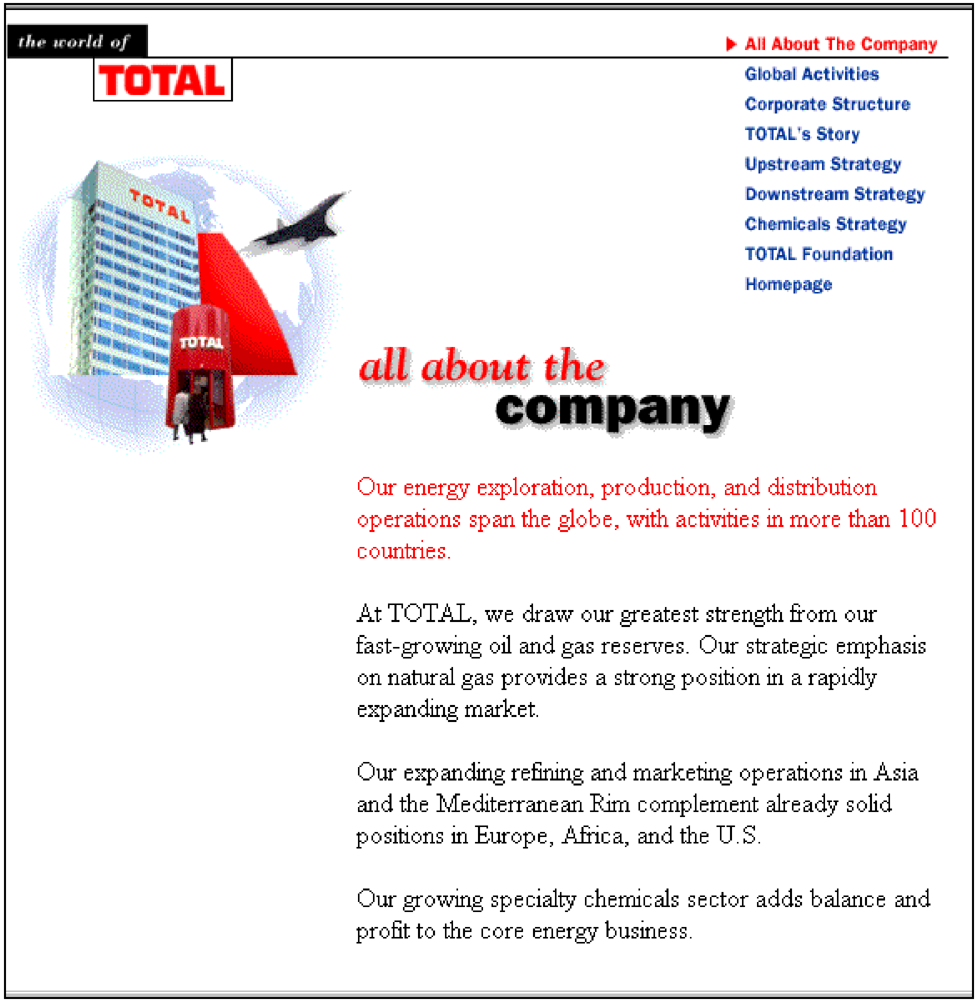
SPAM FILTERING: WHAT FEATURES TO USE?

- First swipe at the problem
  - Features **X** are word sequence in the email.
    - X_i for i^{th} word in the email
- Each Email has at least 1000 words, $\mathbf{X}=\{X_1,\ldots,X_{1000}\}$
  - X_i represents i^{th} word in the email, i.e., the domain of X_i is entire vocabulary, e.g., Webster's Dictionary (+ some more), 10,000 words, etc.
- Size of the space: $10{,}000^{1000} = 10^{4000}$
- Atoms in Universe: 10^{80}
  - We may have a problem…

SPAM FILTERING: WHAT FEATURES TO USE?

- Bag of Words Model
  - **Position of the word in the email does not matter**
  - Ignore the order of words
  - Sounds really silly, but often works very well!
- For each word in the Dictionary
  - Count how many times the word appears in the email
- Each Email = A vector/array of pairs of the form (w,#) where "w" is the word and "#" is the number of times "w" appears in the email

Bag of Words Approach



all about the **company**

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

EVALUATION

- Given a set of emails which are already classified as Spam/ham
- Try different algorithms
  - Perform 10-fold Cross-Validation
  - Choose one or a collection based on their accuracy and F1 score
- Use WEKA
  - A tool for machine learning
- Key Step: Transform your data into ARFF format

WEKA’S ARFF FORMAT

```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

Using WEKA

- Usage:
  - java -Xmx1000M -jar ~/weka/weka.jar
- Load data from directory “data”
- Run Different Classifiers
  - Best F1-score?
  - Classifier with the best F1-score?
  - Options used for Classifier with the best F1-score?