Text Classification

Text Classification

Using Decision Tree and Maximum Entropy

Student: **Bui Dinh Duong**

Email: dinhduong.bui@gmail.com

September 20th, 2013

Organization

- Introduction
- Objectives
- Text Classification Overview
- Decision Tree
- Maximum Entropy
- Experiment
- Conclusion

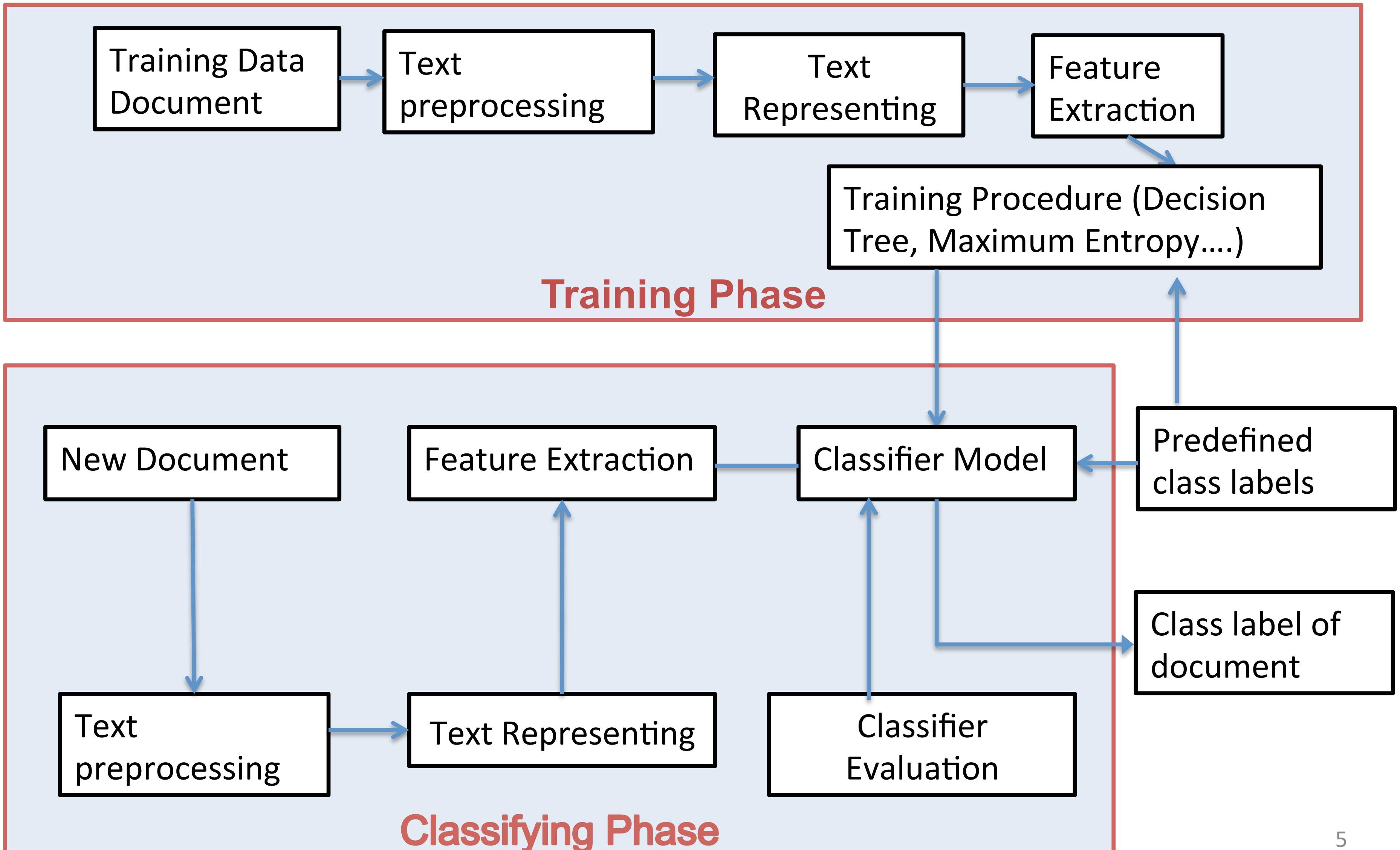
Introduction

- Text classification
  - Assign a document to one or more predefined classes
- Applications
  - E-mail spam filtering
  - Categorize newspaper articles into topics
  - Organize Web pages into hierarchical categories
  - Language identification
- Methods
  - Naive Bayes
  - Maximum Entropy
  - Decision Tree
  - Support Vector Machine (SVM)

Objectives

- Investigate the stages of text classification
- Study Decision Tree method
- Study Maximum Entropy method
- Do experiment in text classification using Weka

Text Classification



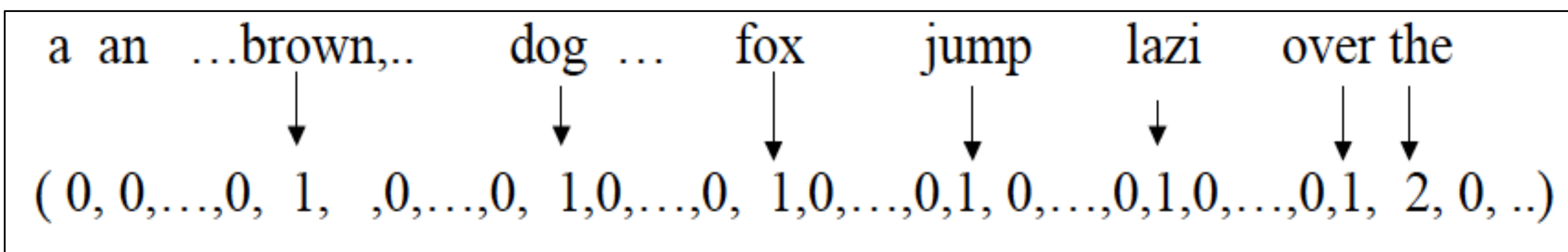
Text Preprocessing

- What is the objective?
  - Reduce the size of data
  - Get only things we need
- How to do?
  - Convert document to lower case
  - Remove words that rely occur in the document
  - Remove special character
  - Remove stop-words (words are not used to classify)
  - Remove suffix, prefix of word to get the root word (“clusters”, “clustering”, “clustered” => cluster)

Text Representing

- What is the objective?
  - Represent text data in a suitable model to process
- How to do?
  - Vector Space Model (most popular method)
    - Each document is represented as a vector of ***word weighting***

For example: “**The brown fox jumps over the lazy dog**”



Feature Extraction

- Word Weighting
  - *Word frequency weighting* and *TF*IDF weighting*: the number of time that a word appears in a document
  - Three values are used to calculate the weighting:
    - Term frequency

      the number of time a word appears in a document
    - Collection frequency

      the number of time a word appears in document collection (whole dataset)
    - Document frequency

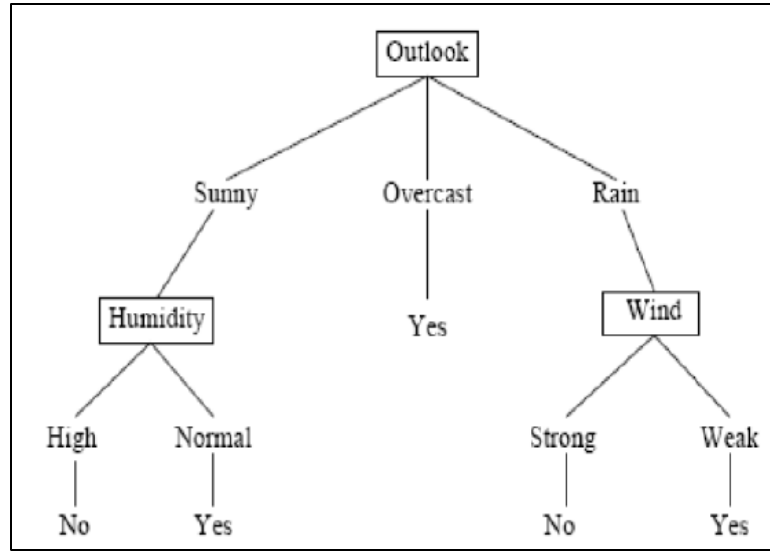
      the number of document contains a word

=> Features: words have highest Word weighting

Classifier Evaluation

- What is the objective?
  - Evaluate quality of the model and its accuracy to know if we can use this model or not
- How to do?
  - **Accuracy**: the proportion of correctly classified objects
  - **Error**: the proportion of incorrectly classified objects
  - **Precision:** the proportion of selected items that the system got right
  - **Recall:** the proportion of the target items that the system selected
  - **Fallout**: the proportion of no targeted items that were mistakenly selected
  - **F-measure**: Precision and Recall are combined

Decision Tree



- The first node is root node
- Internal nodes are attribute tests
- Leaf nodes are class label
- Many algorithms ID3, C4.5, CART, CHAID, MARS in decision tree
- ID3 uses Entropy and Information Gain
- Pruning
  - The pruning step is to avoid **over fitting**
- Cross-validation
  - To maximize the accurate classification of classifier tree model

Decision Tree (ID3)

- Entropy
  - Entropy is a measure the impurity/inhomogeneity of a data set
  - Entropy is the indicator of how much information inside a data set.

$$Entropy(S) = \sum_{i=1}^{C} -p_i \log_2 p_i$$

Where:

- S is the set of training data
- C number of class labels
- p_i: the rate of elements belong class ci

Decision Tree (ID3)

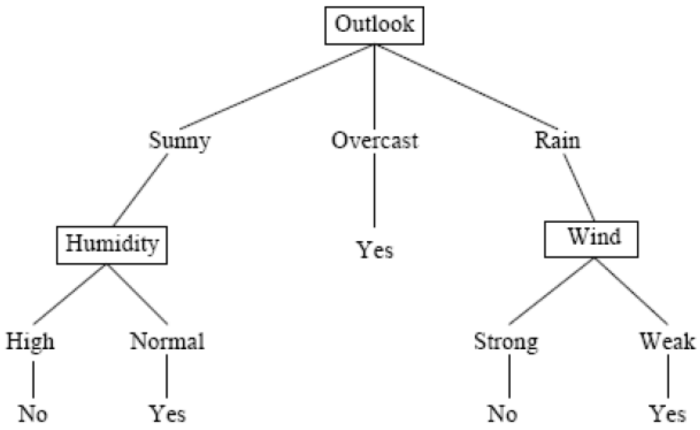
- Information Gain
  - Information gain is the measures of reducing entropy in S by an attribute in S

$$Gain(S,A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{S} Entropy(S_v)$$

Where:

- Value (A): set of A values
- Sv: subset of S

Decision Tree (ID3)



- General steps of ID3 Algorithm
  1. From the dataset S,

  calculate the entropy of every attribute

  1. Split the set S into subsets using the attribute for which entropy is minimum (or information gain is maximum)
  2. Expanding decision tree by adding a node containing that attribute
  3. Recursive on subsets using remaining attributes

Maximum Entropy

- Main idea
  - Satisfy constraints
  - Probability distribution of model is most uniform
- What is the constraint?
  - From training data
  - Constraint : If a document contains the word “professor”, it has a 40% chance of probability distribution in faculty class

$$f_i(\vec{x}_j, c) = \begin{cases} 1, & if\ w_{ij} > 0\ and\ c = 1 \\ 0, & otherwise \end{cases}$$

Wij is the word weighting of word i in document j

Maximum Entropy

- Loglinear Model
  - Use to classify document in Maximum Entropy

$$p(\vec{x}, c) = \frac{1}{Z}\prod_{i=1}^{K} \alpha_i^{f_i(\vec{x},c)}$$

Where :

- K is the number of constraints
- Z: is a constant
- α_i is the weight of f_i

  - Compute $p(\vec{x}_{new}, 1)$ and $p(\vec{x}_{new}, 0)$.

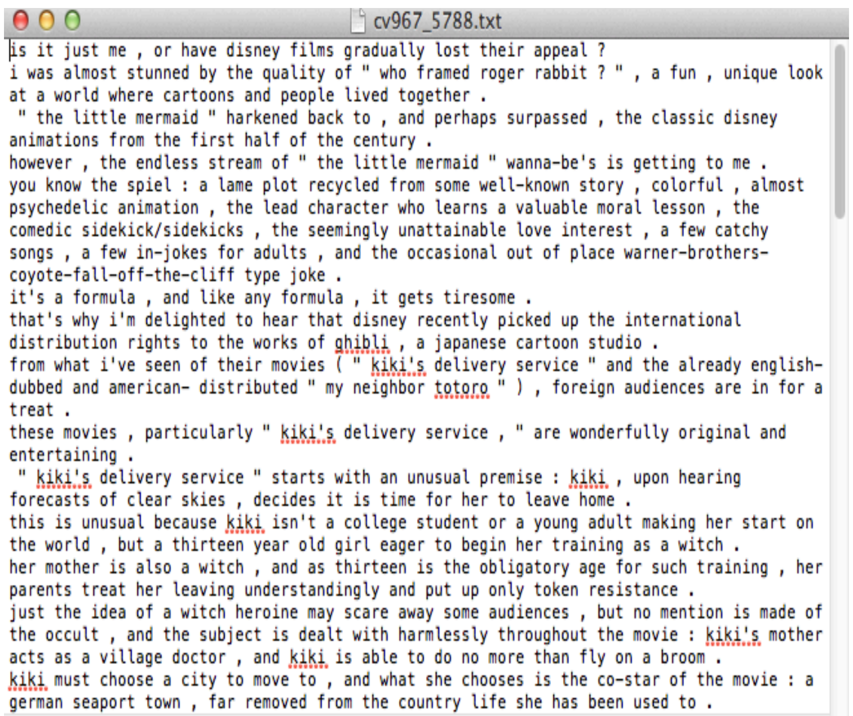
  New document belong to class which has higher probability

Maximum Entropy

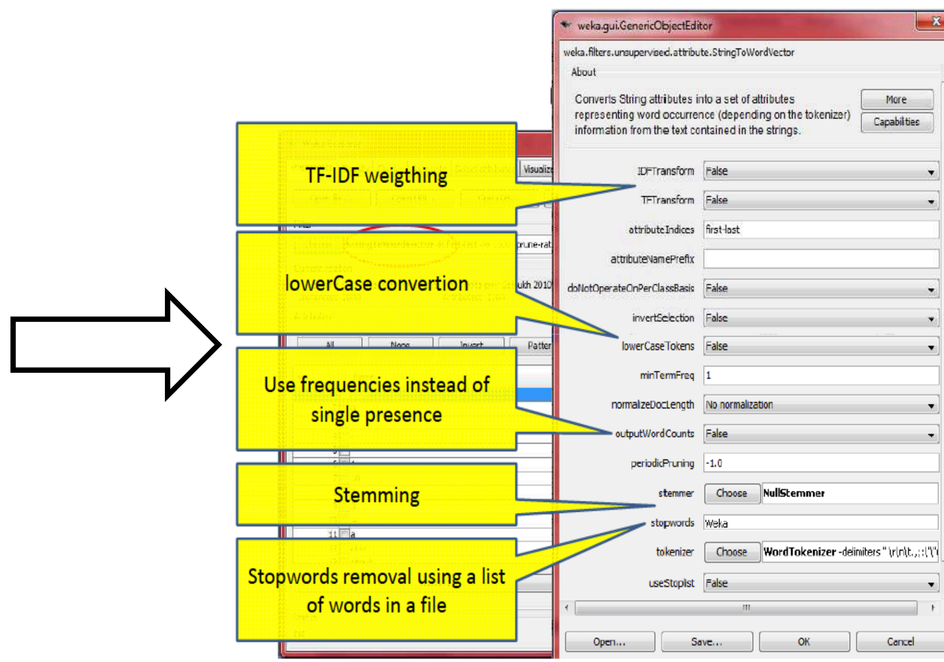
- Generalized iterative scaling (GIS)
  - Use to find α_i in the Loglinear Model
  - GIS find probability distribution which has maximum entropy of Loglinear Model

Experiment

- Dataset : 1000 negative movie reviews and 1000 positive movie reviews
- Text Preprocessing

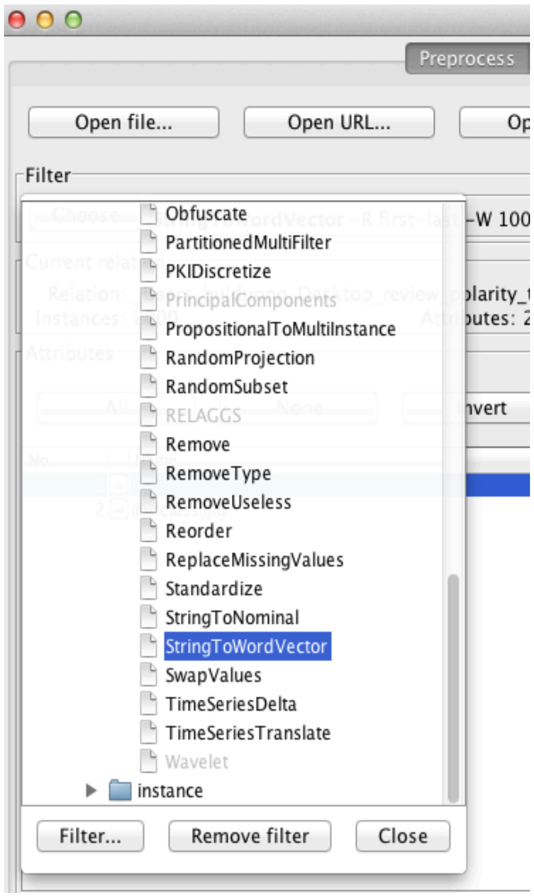

cv967_5788.txt

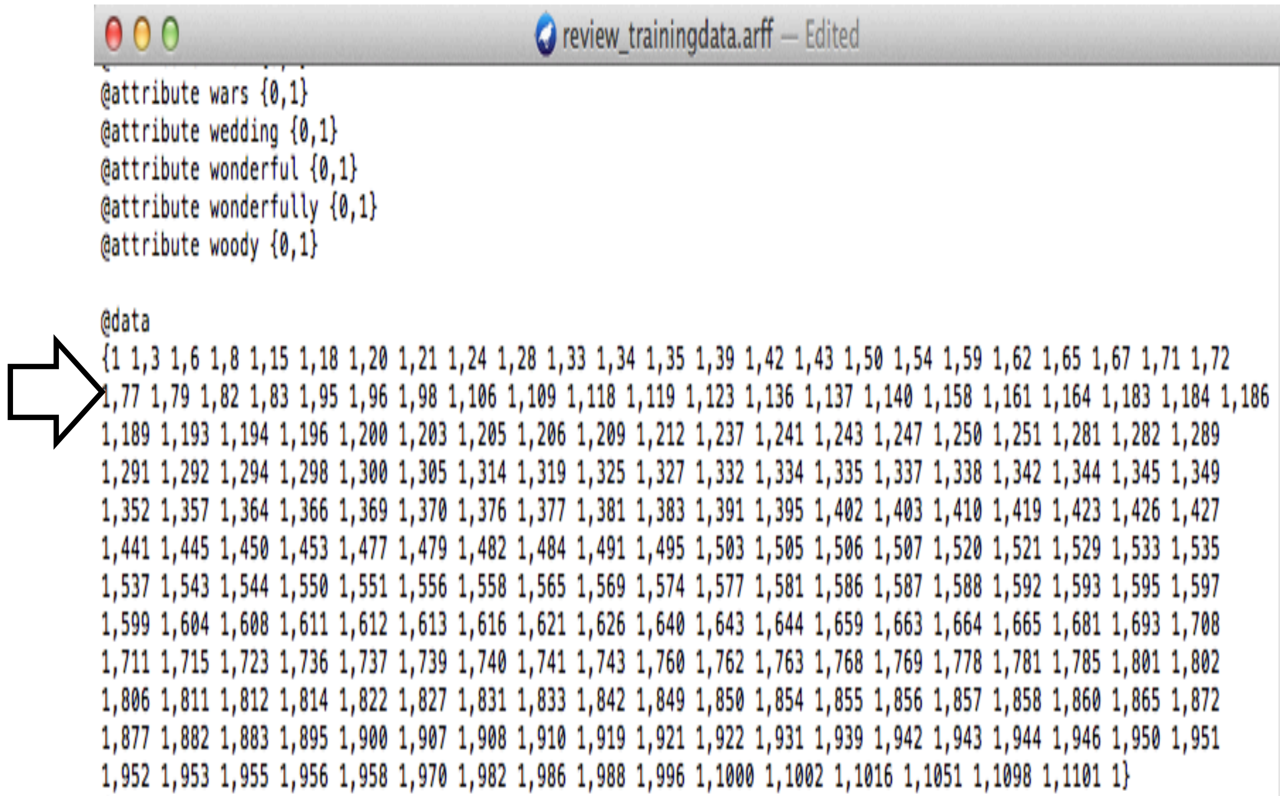
is it just me , or have disney films gradually lost their appeal ?
i was almost stunned by the quality of " who framed roger rabbit ? " , a fun , unique look at a world where cartoons and people lived together .
" the little mermaid " harkened back to , and perhaps surpassed , the classic disney animations from the first half of the century .
however , the endless stream of " the little mermaid " wanna-be's is getting to me .
you know the spiel : a lame plot recycled from some well-known story , colorful , almost psychedelic animation , the lead character who learns a valuable moral lesson , the comedic sidekick/sidekicks , the seemingly unattainable love interest , a few catchy songs , a few in-jokes for adults , and the occasional out of place warner-brothers-coyote-fall-off-the-cliff type joke .
it's a formula , and like any formula , it gets tiresome .
that's why i'm delighted to hear that disney recently picked up the international distribution rights to the works of ghibli , a japanese cartoon studio .
from what i've seen of their movies (" kiki's delivery service " and the already english-dubbed and american- distributed " my neighbor totoro ") , foreign audiences are in for a treat .
these movies , particularly " kiki's delivery service , " are wonderfully original and entertaining .
" kiki's delivery service " starts with an unusual premise : kiki , upon hearing forecasts of clear skies , decides it is time for her to leave home .
this is unusual because kiki isn't a college student or a young adult making her start on the world , but a thirteen year old girl eager to begin her training as a witch .
her mother is also a witch , and as thirteen is the obligatory age for such training , her parents treat her leaving understandingly and put up only token resistance .
just the idea of a witch heroine may scare away some audiences , but no mention is made of the occult , and the subject is dealt with harmlessly throughout the movie : kiki's mother acts as a village doctor , and kiki is able to do no more than fly on a broom .
kiki must choose a city to move to , and what she chooses is the co-star of the movie : a german seaport town , far removed from the country life she has been used to .



Experiment

- Text representing





Feature and word weighting

Experiment (result)

```
|   |   |   worst = 1
|   |   |   |   bring = 0
|   |   |   |   |   tom = 0
|   |   |   |   |   |   details = 0: neg
|   |   |   |   |   |   details = 1
|   |   |   |   |   |   |   -- = 0: pos
|   |   |   |   |   |   |   -- = 1: neg
|   |   |   |   |   tom = 1
|   |   |   |   |   |   come = 0: neg
|   |   |   |   |   |   come = 1: pos
|   |   |   |   bring = 1
|   |   |   |   |   see = 0: neg
|   |   |   |   |   see = 1
|   |   |   |   |   |   usually = 0
|   |   |   |   |   |   |   america = 0: pos
|   |   |   |   |   |   |   america = 1: neg
|   |   |   |   |   |   usually = 1: neg
|   |   wonderfully = 1
|   |   |   red = 0: pos
|   |   |   red = 1: neg
|   stupid = 1
|   |   bob = 0
|   |   |   into = 0
|   |   |   |   perfect = 0
|   |   |   |   |   certainly = 0: neg
|   |   |   |   |   certainly = 1
|   |   |   |   |   |   - = 0: pos
|   |   |   |   |   |   - = 1: neg
|   |   |   |   perfect = 1: pos
|   |   |   into = 1: neg
|   |   bob = 1
|   |   |   10 = 0: pos
|   |   |   10 = 1: neg
```

Classifier Model (ID3)

```
Correctly Classified Instances         247               61.75   %
Incorrectly Classified Instances       153               38.25   %
Kappa statistic                          0.235
Mean absolute error                      0.3825
Root mean squared error                  0.6185
Relative absolute error                 76.5    %
Root relative squared error            123.6932 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.61      0.375      0.619     0.61      0.615      0.618    neg
                 0.625     0.39       0.616     0.625     0.62       0.618    pos
Weighted Avg.    0.618     0.383      0.618     0.618     0.617      0.618
```

Result of classifying phase (training data 66%, test 34%)

Conclusion

- Achievements
  - Understand the stages of text classification
  - Understand Decision Tree method
  - Understand Maximum Entropy method
- Future works
  - Continue researching other methods
  - Programming a tool with decision tree to classify document

Reference

- Christopher D.Manning, Hinrich Schutze, **“Foundations of Statistical Natural Language Processing”**
- Kamal Nigam, John Lafferty, Andrew McCallum, **“Using Maximum Entropy for Text Classification**”, In IJCAI-99 Workshop on Machine Learning for Information Filtering
- Kostas Fragos, Yannis Maistros, Christos Skourlas, “**A Weighted Maximum Entropy Language Model for Text Classification**”, NLUCS 2005: p.55-67
- Tom M. Mitchell, "**Machine learning**", Published by McGraw-Hill, Maidenhead, U.K., International Student Edition, 1997. ISBN: 0-07-115467-1

Thank You!