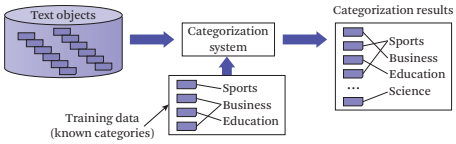
Text Categorization
- Problem and Motivation

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

Text Classification

- A standard supervised learning problem
  - Goal: classify a new document into one of the pre-given categories
  - Training data: labeled document examples



Problem Definition

- Input:
  - a document d
  - a fixed set of categories $C = \{c_1, c_2, \ldots, c_J\}$
- Output: a predicted category $c \in C$ for the document d

Email Spam Detection

From: "Laraine" <Claribel@garancs.net>
Subject: Original pilules at factory prices
Date: March 26, 2013 12:05:00 PM EDT
To: <ud-aec@UDel.Edu>

USPS - Fast Delivery Shipping 1-4 day
Best quality drugs
Professional packaging
100% guarantee on delivery
Best prices in the market
Discounts for returning customers
FDA approved productas
35000+ satisfied -customers

==========
If you can't click on link, please click "no spam" or copy and paste it to address bar
==========

http://wawuotu.inhealth-pro.ru

Examples of Text Classification Problem

- Authorship Identification
  - Classify a document as the one written by different authors
- Sentiment analysis
  - Classify a review to positive or negative
- Automatic essay grading
  - Classify a student essay to A,B,C or D.
- Help desk management system
  - Classify an email according to technical staff's specialty such as Mac, Window, etc.
- …

Classification methods (1): manually created rules

- Rules based on combinations of words or other features
  - In email spam detection,
    - If an email's sender is in a black-list, the email will be classified as spam.
    - If an email's message contains "dollars" and "have been selected", this email will be classified as spam.
- If rules are carefully identified and refined by experts, the accuracy can be high.
- But it requires lots of human efforts to build and maintain the rules.

UNIVERSITY of DELAWARE UNIVERSITY of DE

Classification Methods (2): Supervised Machine Learning

- Input:
  - a document d
  - a fixed set of categories $C = \{c_1, c_2, \ldots, c_J\}$

A training set of m hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$

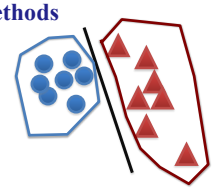
a learned classifier: f(d) → c

- Output: a predicted category $c \in C$ for the document d

8

UNIVERSITY of DELAWARE UNIVERSITY of DE

Automatic Classification Methods



- Generative classifiers
  - Naïve Bayes classifier
- Discriminative classifiers
  - Rocchio classification
  - KNN (K nearest neighbors)
  - Support Vector Machines (SVM)

9

UNIVERSITY of DELAWARE UNIVERSITY of DE

Feature Selection

UNIVERSITY of DELAWARE UNIVERSITY of DE

Feature selection→ Document Representation

f(I love this movie. It is sweet and full of humors. The dialogue is great. I love the main actress. She is so pretty. ☺ I would recommend it to my friends. I have watched it multiple times, but I would not mind watching it again.)=c

UNIVERSITY of DELAWARE UNIVERSITY of DE

Bag of Words Representation (Unigram)

f( )=c

I	8
movie	3
love	3
great	4
…	…

UNIVERSITY of DELAWARE UNIVERSITY of DE

Use only selected words as features

f(I **love** this movie. It is **sweet** and full of humors. The dialogue is **great**. I **love** the main actress. She is so **pretty**. ☺ I would **recommend** it to my friends. I have watched it **multiple** times, but I would not mind watching it **again**.)=c

Use only selected words as features