CO544 Machine Learning and Data Mining

Sampath Deegalla

5.11.2018

Text Classification

- To assign one or more categories to text documents
- Automatic vs. Manual
  - Cost, Speed, Accuracy
- Applications
  - Email categorization
  - News
  - Web pages to subject categories
  - Research papers to subject categories

Approaches

- Manual
  - Read each document and classify it
- Automatic
  - The machine learning approach
- ML approach
  - Build classifiers from characteristics of documents
  - Manually assign labels to the training set
  - Classifiers used to predict category of a new document

Document Representation

- Documents
  - Training and test instances are documents
- Document Representation
  - Documents must be represented in a way that is understood by the classifier

What is a document?

Dems take House as GOP clings to Senate, CNN projects

POSTED: 2:38 a.m. EST, November 8, 2006

Supporters of the Democratic Party cheer incoming results at a party in Washington, D.C.

Previous Next

ADVERTISER LINKS
- Distance Learning
- Fresh Flowers
- Women's Apparel

STORY HIGHLIGHTS
- NEW: Republican Jim Talent concedes tight Missouri Senate race
- Democrats must win both undecided races to control Senate, CNN projects
- Democrats will win control of House of Representatives, CNN projects
- Rep. Nancy Pelosi, set to be House speaker, challenges President Bush over Iraq

More on CNN TV: Track the races and get the results with Larry King and his guests. Live now.

Adjust font size:

(CNN) -- Democrats will take control of the House of Representatives for the first time since the 1994 Republican revolution, while control of the Senate hangs in the balance, CNN projects.

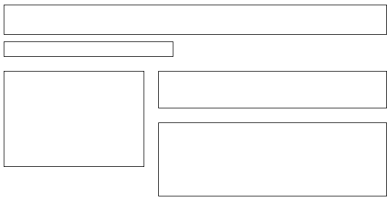
Democratic challengers have picked up four seats in the Senate, CNN projects. Republicans would need to take just one of the two remaining competitive races to keep control of the chamber. Results are still too close to call in Montana and Virginia.

Republican Sen. Jim Talent of Missouri conceded defeat to Democrat Claire McCaskill early Wednesday after a hard-fought race, telling supporters that "the headwind was just very, very strong this year."

Taken from cnn.com

What is a document?

- Text content
- Document structure
- Facts
- Document layout



Bag-of-Words representation

- Compare words, disregard order of words in the text
- Example:
  - D1 the dog sat on the floor
  - D2 the man sat on the floor
  - D3 the man walked his dog

Bag-of-Words

	DOG	FLOOR	HIS	MAN	ON	SAT	THE	WALKED
D1								
D2								
D3								

Bag-of-Words

	DOG	FLOOR	HIS	MAN	ON	SAT	THE	WALKED
D1	1	1	0	0	1	1	2	0
D2	0	1	0	1	1	1	2	0
D3	1	0	1	1	0	0	1	1

Term frequencies

Example

2832	THE
1527	OF
1433	TO
1013	SAID
398	TONNES
233	GOLD
202	AS,ARE,AN
…	
1	ACCEPTS,ACCEPTING,ABSENT,ABIDE

Stop words

- Words that occur often in a document may be a reasonable indicator of the document's topic, but not if the words in many documents
- These words are called "stop words"
- Examples
  - a, about, above, according, across, after, afterwards, again, against, albeit, all, …
- These words must be treated specially (often removed)

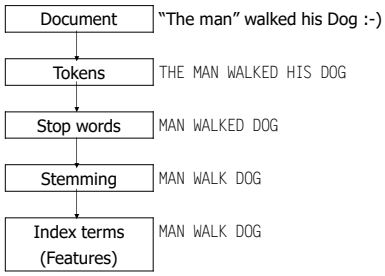
Stemmers

- Form of words
  - House, Houses
- Derivational (Derive one word from another)
  - health, healthy
- Stemmers
  - English: Porter Stemmer (remove suffixes)

Index Terms

- Words
  - Words as featues
- Phrases
- N-grames
  - Sequences of N words

Extracting index terms (Feature construction)



Term selection (Feature selection)

- Reduce feature set to a smaller set that maximize the effectiveness
- Two approaches – Wrapper and Filter
- **Wrapper**
  - Identify the smaller set by using the same ML algorithm that will be used for classification (Add or remove features iteratively, cross-validate to evaluate)
- **Filter**
  - Select the feature set by a function that measures the importance of the terms for the categorization task (Ex. Information Gain)
  - Computationally less expensive that Wrapper

Assignment

- Write you own code for document representation
- Construct a classifier for the training set using favorite ML algorithm.
- Predict class labels of the testing set with the selected classifier

Submission

- A small report (about 5 pages) where you describe your representation, your algorithm and how you trained your classifier.
- Predicted labels for the test set.

Assignment

- Two-class categorization problem
- Training set: 200 training instances
- Testing set: 100 test instances
- Each document is one line of text
- Fields are separated by the tab '\t' character
- <CLASS> \t <TITLE> \t <DATE> \t <BODY>
- CLASS is either +1 or -1

Positive Example

- +1 JAPAN FIRM PLANS TO SELL U.S. FARMLAND TO JAPANESE
MORIOKA, Japan, March 12 - A Japanese real estate company said it will launch a campaign to sell land in U.S. Farming areas to rich Japanese, Higashi Nippon House said it would offer around 2,200 acres of land in Illinois, California, Florida and Indiana from early April to gauge response. It set up International Farm Corp of America in Chicago last September to oversee the operation. American farmers would continue as working tenants and part of the profits from harvests of rice, corn, soybean and oranges would go to the Japanese investors as rental, Japanese Agriculture Ministry officials told Reuters sales were limited to farmers to keep land in agricultural use. "Two years ago, I began to seek my own farmland in Japan," said Isao Nakamura, president of Higashi Nippon. "However, sale of Japanese farmland is strictly controlled by the government, so I began to look for the land in the U.S to make my dream to own farm land come true." Nakamura said hundreds of companies exist in the U.S. To sell farmland to investors as more and more farmers face difficulties due to the recession in U.S. Agriculture. REUTER

Negative Example

- -1 NORTH BH SETS ONE-FOR-FIVE OFFER FOR NORGOLD FLOAT
MELBOURNE, March 12 - North Broken Hill Holdings Ltd & lt;NBHA.ME > (NBH) said it will offer one & lt;Norgold Ltd > share for every five NBH shares in the float of its newly created gold offshoot. The 20 cent par-value shares will be offered at 22 cents to shareholders registered April 3, NBH said in a statement. Norgold's issued capital will be 240.5 mln shares, of which 63 pct will be held by NBH after 89 mln are issued to shareholders to raise 19.6 mln dlrs, it said. Norgold will take control of a portfolio of precious metal exploration and pre-development interests held by NBH. The major gold deposit to be acquired by Norgold is 100 pct of the Bottle Creek deposit, west of Leonora in Western Australia, NBH said. Production of gold from the project, at an annual rate of 35,000 ounces, is scheduled to begin early in 1988. Norgold will also have a 10 pct stake in the Coronation Hill gold/platinum project in the Northern Territory and 43 pct of the Poona copper/gold project in South Australia. Other gold exploration interests to be acquired by Norgold are in Western Australia, Queensland, New South Wales and Tasmania, NBH said. REUTER

Links

- See Moodle course page for details
- egrep for Linguists, Nikolaj Lindberg
  - http://www.ida.liu.se/~lensa/nikolaj/egrep_for_linguists.html
- Machine Learning Software
  - WEKA http://www.cs.waikato.ac.nz/ml/weka/
  - scikit-learn http://scikit-learn.org/

- Acknowledgement: Rickard Coster, DSV, Stockholm University, Sweden