

Business Intelligence and Data Mining

L11 Feature Selection and Text Mining

Prof. Rong Zheng
November. 6, 2012

Agenda

- Feature selection
- Text classification
- Financial news trading

Feature selection

- What are “good” and “bad” features?
 - Good: in differentiating target variable.
 - Bad: redundant, irrelevant
- This is the most frequently asked question in real-world data analysis!!
- Decision tree naturally select good variables.
 - How about other techniques?

Feature selection methods in DM

- Filter approach
 - Apply selection criteria before classification algorithm is applied.
 - Ranking variables based on selection criteria: correlation, Information Gain/Gain ratio, Principle component analysis(PCA)...
- Wrapper approach
 - Classification algorithm is used as a black box to find the best subset of attributes

Wrapper approach

- Selection of an induction algorithm to evaluate different feature combinations
- Selection of a search algorithm to explore the feature combination space
 - How to do sub-space exploration intelligently

Digression on genetic algorithm

Benefit of feature selection

- Improve prediction performance
- Improve efficiencies of prediction
- Better understanding the underlying pattern

Demo on feature selection: case of churn data

Text classification

- Putting documents into different categories based on:
 - Topics
 - Authors
 - Sentiments

Example text: online review

The screenshot shows the Amazon.com website interface. At the top, there's a navigation bar with the Amazon logo, a sign-in prompt, and links for personalized recommendations. Below this is a search bar with 'Arts & Photography' entered. The main content area is titled 'Customer Review' and displays a review by a user named 'Jed'. The review text describes a disappointment with a book, mentioning a 'praying' method and a publisher. To the right, under 'Review Details', there's a star rating breakdown and a price comparison from \$10.99 to \$8.79. At the bottom of the review, there's a 'Permalink' and a 'Was this review helpful to you?' section with 'Yes' and 'No' buttons.

amazon.com Hello. Sign in to get [personalized recommendations](#). New customer? [Start here](#).

Your Amazon.com Today's Deals Gifts & Wish Lists Gift Cards

Shop All Departments Search Arts & Photography GO

Books Advanced Search Browse Subjects Hot New Releases Bestsellers The New York Times® Best Sellers

Customer Review

1,489 of 1,555 people found the following review helpful:

☆☆☆☆☆ **Doesn't work**, March 21, 2008

By [Jed](#)


This book doesn't work. I've tried the "praying" method to get a new Porsche 996 delivered but to no avail. There's nothing in the instructions about not wanting German sports cars but I tried praying for less ambitious things. I gave up when it didn't even get me a Big Mac. In the early part there's a bit about people crossing the desert and being sustained by manna from heaven, so you'd think that it would be able to manage at least a hamburger.

I'm disappointed and will contact the publisher. In the meantime I can't recommend this book as it is clearly faulty.



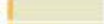


[Permalink](#) | Was this review helpful to you? ([Report this](#))

Review Details

Item

 [Boldtext Pew Bible: Kin](#)

☆☆☆☆☆ (249 customer reviews)

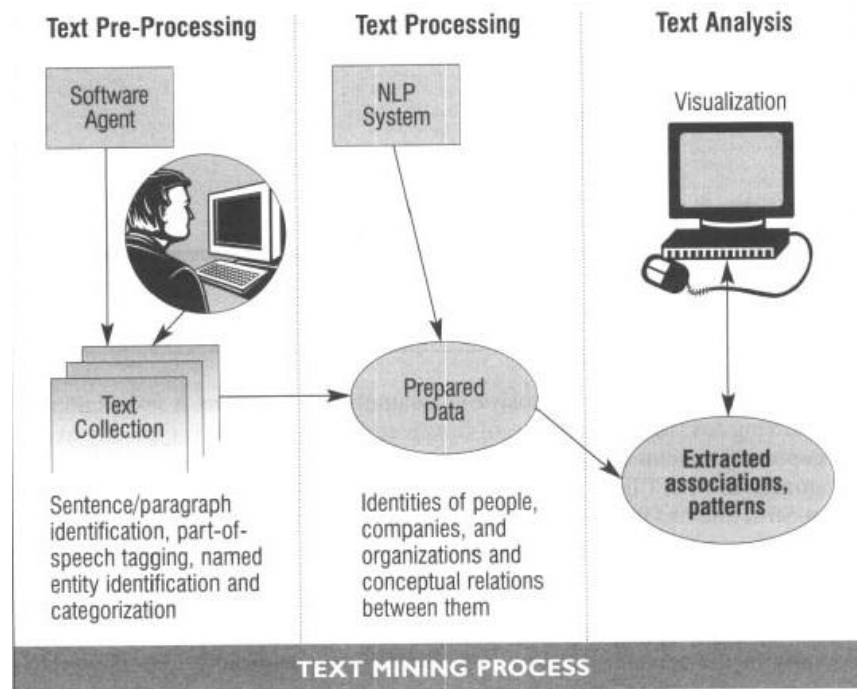
5 star:		(123)
4 star:		(18)
3 star:		(13)
2 star:		(14)
1 star:		(81)

~~\$10.99~~ **\$8.79**

[23 used & new](#) available from \$

Many other online document: news, blogs...

Text Mining: data mining for text



How can we represent text documents in order to apply data mining procedures?

- A document representation aims to capture what the document is about.


Document representation (1)

Each entry in the table represents a document.

*Attribute describes **whether or not** a term appears in the document.*

Data mining is the process of extracting patterns from **data**. As more **data** are gathered, with the amount of **data** doubling every three years, **data mining** is becoming an increasingly important **tool** to transform these **data** into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery...

Statistics is a branch of mathematics concerned with collecting and interpreting **data**. According to other definitions, it is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of **data**.



	Term				
	Data	Mining	Tool	XXX	...
Document 1	1	1	1	0	
Document 2	1	0	0	0	
...

Document representation (2)



Each entry in the table represents a document.

Attributes represent the **frequency** of a term in the document

Usually needs normalized by length.

Statistics is a branch of mathematics concerned with collecting and interpreting **data**. According to other definitions, it is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of **data**.

Data mining is the process of extracting patterns from **data**. As more **data** are gathered, with the amount of **data** doubling every three years, **data mining** is becoming an increasingly important **tool** to transform these **data** into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery...



	Term				
	Data	Mining	Tool	XXX	...
Document 1	6	2	1	0	
Document 2	2	0	0	0	
...

TF*IDF (Term Frequency*Inverse Document Frequency)

- Document frequency: Number of docs that have a term appeared.
- An approach for weighting terms in a document based on each term's frequency in the **document** *and in the corpus* (collection of documents).
- A term would have a higher weight if it is found to be a good descriptor for a particular category.
 - i.e., if it appears frequently in the document but is infrequent in the entire corpus.
- The TF*IDF implements this idea:
 - tf: the higher, the more representative a topic
 - df: the higher, the less differentiating for the topic


Document representation (3)

Each entry in the table represents a document.

Attributes represent the ***tf*idf*** of a term in the corpus

Statistics is a branch of mathematics concerned with collecting and interpreting **data**. According to other definitions, it is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of **data**.

Data mining is the process of extracting patterns from **data**. As more **data** are gathered, with the amount of **data** doubling every three years, **data mining** is becoming an increasingly important **tool** to transform these **data** into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery...



	Term				
	Data	Mining	Tool	XXX	...
Document 1	0.001	0.04	0.005	0	
Document 2	0.002	0	0	0	
...


Text classification

- Popular methods for text:
 - Naïve Bayes
 - Support-vector machines (see previous notes)
 - Rocchio's version of nearest neighbor

Trading stocks on released news:
no way? a way?

A news about HSBC

[HSBC Global Site](#)[HSBC home page](#)[Contact HSBC](#)[HSBC site map](#)

HSBC  **The world's local bank**

Search [GO](#)

[About HSBC](#)[Newsroom](#)[Investor Relations](#)[Sustainability](#)[Careers](#)[Personal Banking](#)[Business and Corporate](#)[Global Banking and Markets](#)[Private Banking](#)[Internet banking](#)

You are here: [Home](#) > [Newsroom](#) > [News](#) > [News Archive 2009](#)
> [HSBC strengthens emerging markets focus as the Group CEO moves to Hong Kong](#)

Newsroom

News

[News Archive 2009](#)

[News Archive 2008](#)

[News Archive 2007](#)

[News Archive 2006](#)

[News Archive 2005](#)

[News Archive 2004](#)

[News Archive 2003](#)

[News Archive 2002](#)

Speeches

Awards and rankings

Media contacts

Media Kit

HSBC strengthens emerging markets focus as the Group CEO moves to Hong Kong

**** Group CEO also to become Chairman of HSBC's Asia business ****
**** HSBC Holdings to remain domiciled in the UK ****

25 September 2009

HSBC, the world's leading international and emerging markets bank, is to relocate the principal office of the Group Chief Executive to Hong Kong, in line with its stated strategy to focus on emerging markets and the unique international connectivity that HSBC's global network offers its 100 million-plus customers around the world. The move further positions the Group for the shift in the world's centre of economic gravity from West to East, while HSBC's continued strong presence in major developed markets reflects the increasingly interconnected nature of the global economy and the profile of the Group's customers.

The Group was founded in Hong Kong and Shanghai in 1865 and remains the largest international bank in the region. Operating from Hong Kong, the hub for HSBC's Asia-Pacific business, the Group Chief Executive, Michael Geoghegan, will be located in the Group's strategically most important region, with a focus on ensuring its growth potential is fully realised. The Group Chief Executive will also assume responsibility for developing Group strategy in agreement with the Group Chairman and for recommendation to the Board. As chair of HSBC's executive management team, the Group Management Board, he will continue to drive company performance within Board agreed strategic goals and commercial objectives.

Related links

[Media enquiries to
pressoffice@hsbc.com](mailto:pressoffice@hsbc.com)

Useful tools

[Print this page](#)[Increase font](#)[Decrease font](#)[Site map](#)

Latest News

13 Oct 2009
[The Saudi British Bank Third Quarter 2009 Results - Highlights](#)

07 Oct 2009

Good news !?



Bad news !?

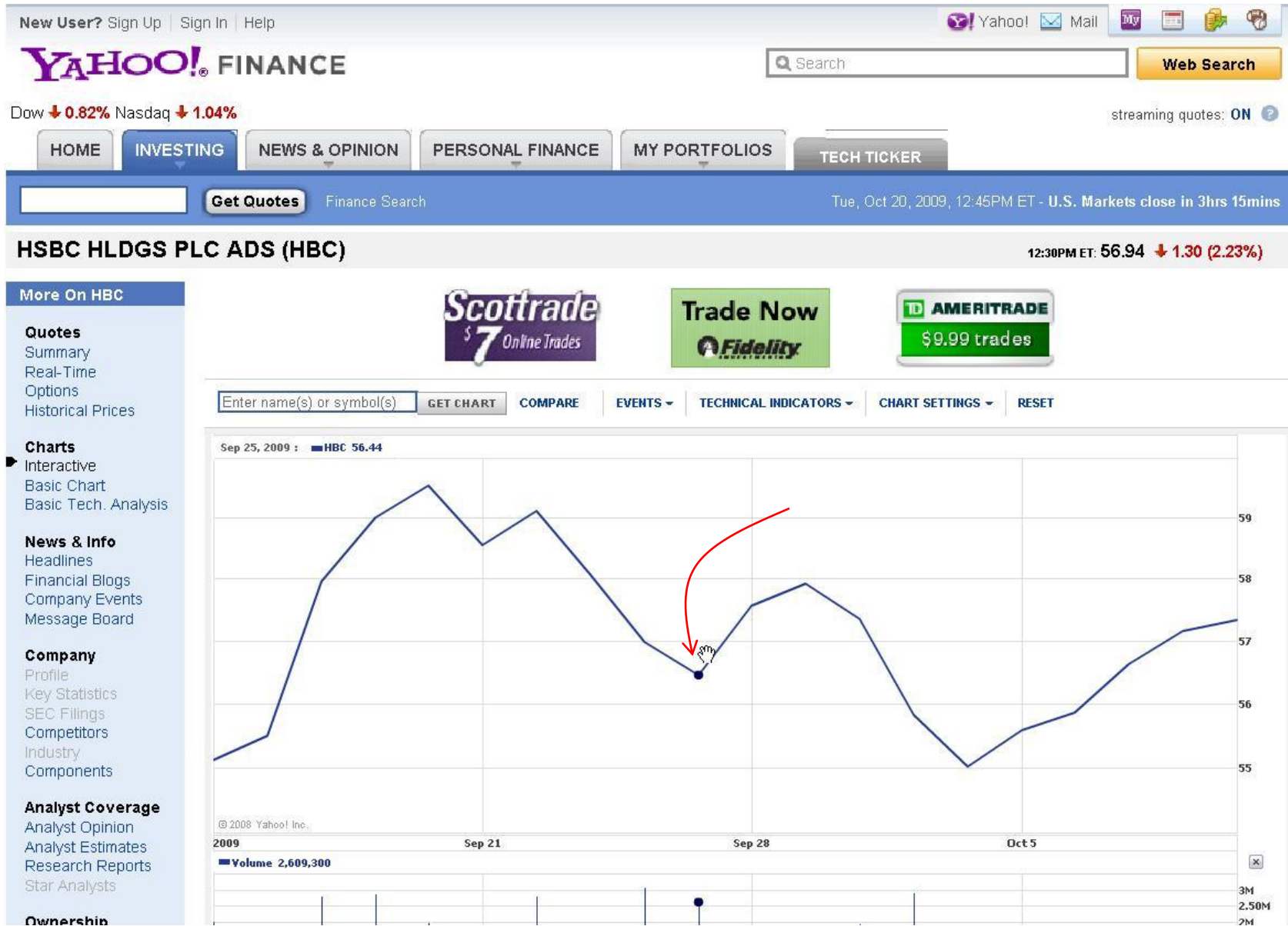
A business problem

- How the stock market will react to different news?
 - If the news has been priced in ...
 - If competitors benefit from the same good news...
 - If the news is biased....
 - ...

Too many factors...

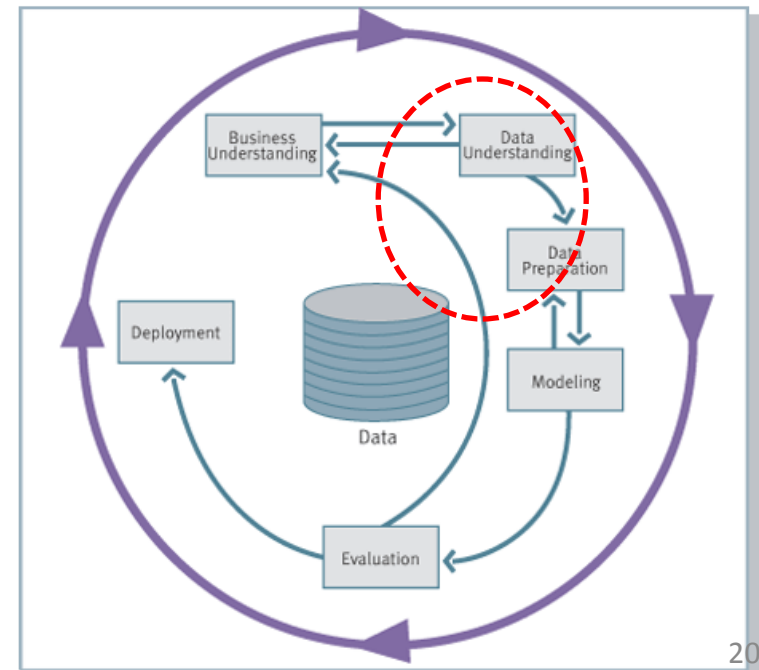


Stock price change



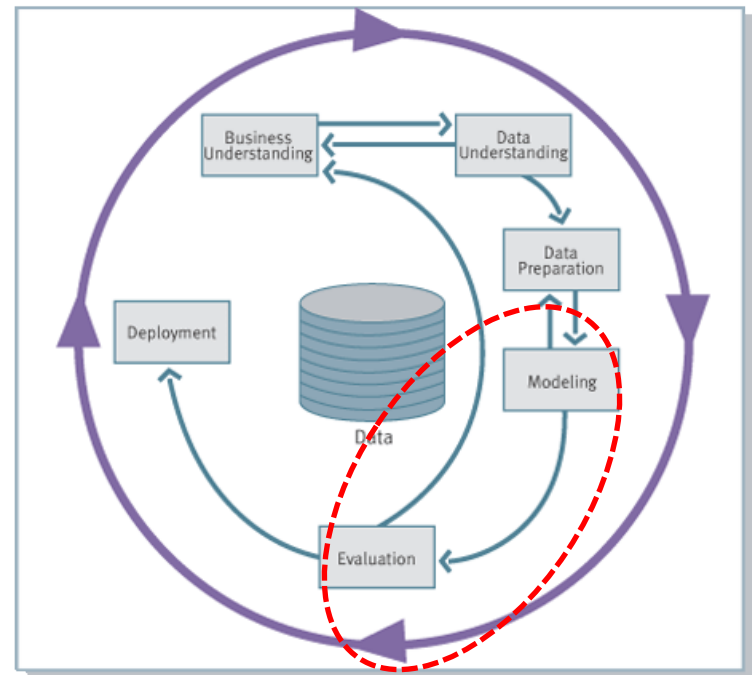
Supervised data mining problem

- What is the target variable?
- What are the attributes?
- Do we have data ready yet?
 - Value of target variable?
 - Attributes?



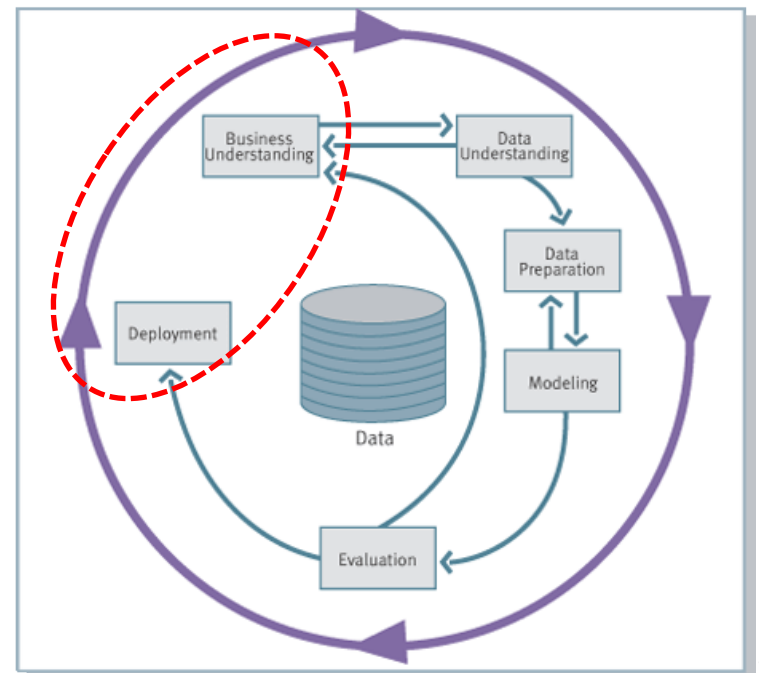
Supervised data mining problem

- What inductive technique should we use?
- How can we know if our model works well?



Supervised data mining problem

- Where in the business process shall we **use** the model?
- Will the model keep working well?



Demo: trading on news