RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

“One cannot guess how a word functions. One has to look at its use and learn from that."
Ludwig Wittgenstein

Text Mining

Presented by Khrystyna Bochkay

Rutgers Business School

Motivation for Text Mining

- **Most of the business data is text**
  - Technical documents
  - Corporate documents
  - News stories
  - Web pages and blogs
  - Emails
  - Books
  - Digital libraries
  - Customer reviews and complaint letters
- **Growing rapidly in size and importance**

Definition and Applications

- ***Text mining*** is the process of combing through countless pages of plain-language digitized text to find useful information that has been hiding in plain sight.

- First developed—as a labor-intensive manual discipline—in the 1980s, text mining has become ever more efficient as computing power has increased.

- Relevant today to a large number of different businesses in practice and research.

  - classification, clustering, e-mail and news filtering, association and prediction

Challenges

- Very large textual databases
- Unstructured form of documents
- Extremely large number of features to analyze:

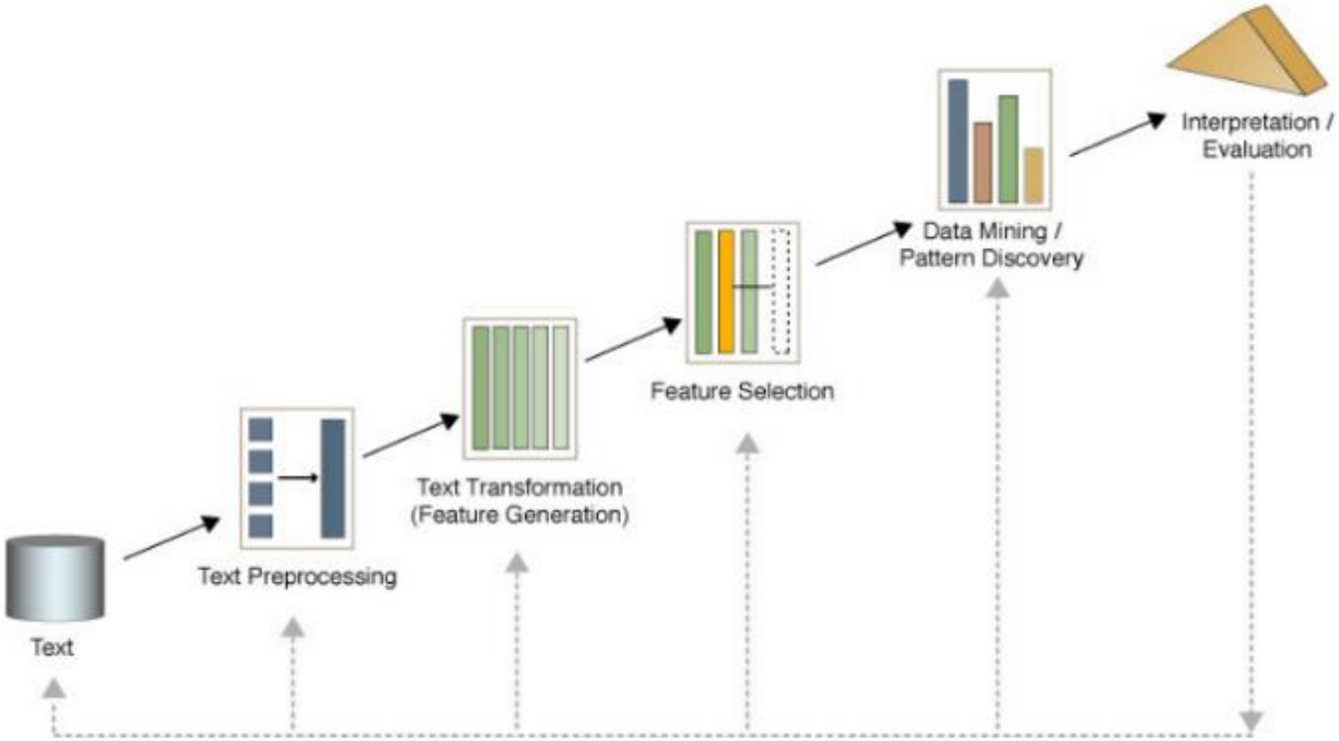
  - millions of words and word combinations in a language;
- Complex and subtle relationships between concepts in text

  - “Sales increased in November” vs. “Jump in sales in November”
- Word ambiguity and context sensitivity

  - war = invasion
  - depression (a mental state) or depression (an economic state)
- Noisy data

  - typos, different writing styles, etc.

Text Mining Process



Text
Text Preprocessing
Text Transformation (Feature Generation)
Feature Selection
Data Mining / Pattern Discovery
Interpretation / Evaluation

- Source: Fan, W., Wallace, L., Rich, S. and Zhang, Z. Tapping into the Power of Text Mining. Communications of ACM, 2005.