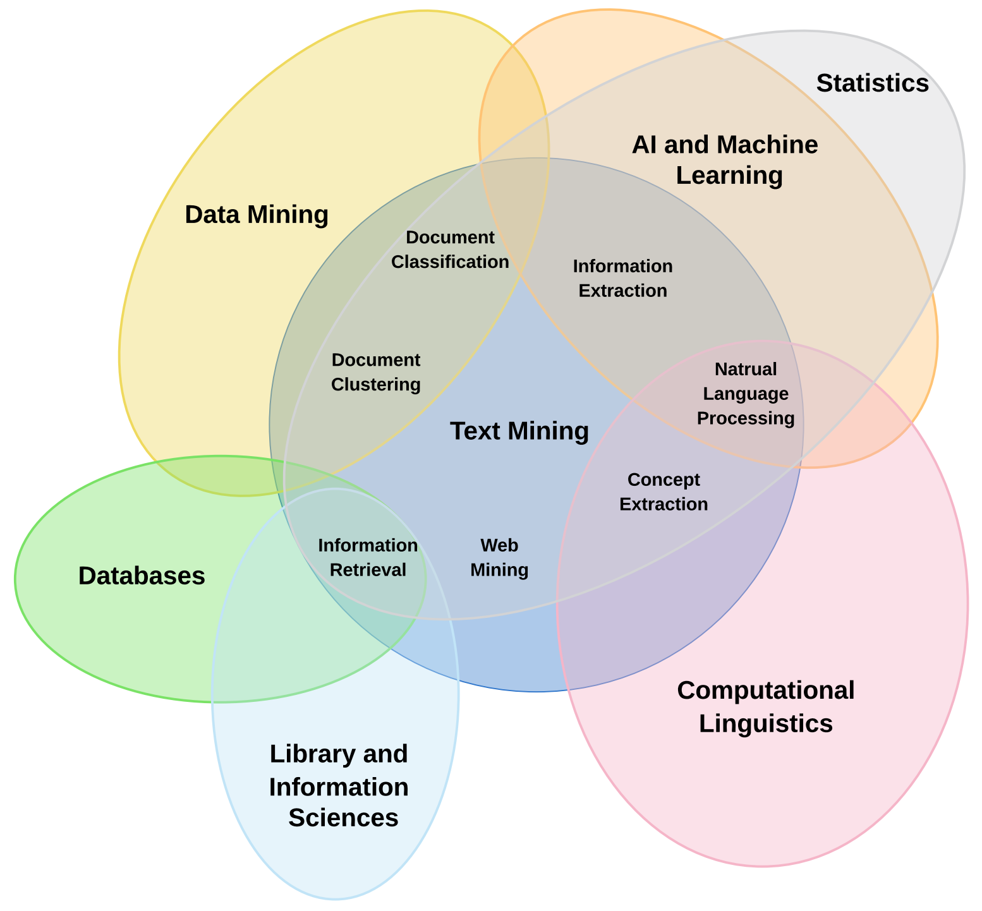
Text Mining

Instructor: Prof. Lu Xiao

Group 5: Xinxia Song, Yunxia Zhao, Tong Cui, Yuhan Liu

Interaction with Other Fields



Milestone in Text Mining

1986 Invention of **ID3 algorithm** (Quinlan)

1996 Introduced the concept of **Named Entity** and **Template Element** (Grishman, Sundheim)

1998 Compared **learning algorithms** for text categorization (Susan Dumais, Mehran Sahami, etc.)

Text categorization with **Support Vector Machines** (Joachims)

1999 Improved **AdaBoost algorithm** and created enhanced decision trees (Robert E. Schapire, Yoram Singer)

2000 Introduced **BoosTexter** (Robert E. Schapire, Yoram Singer)

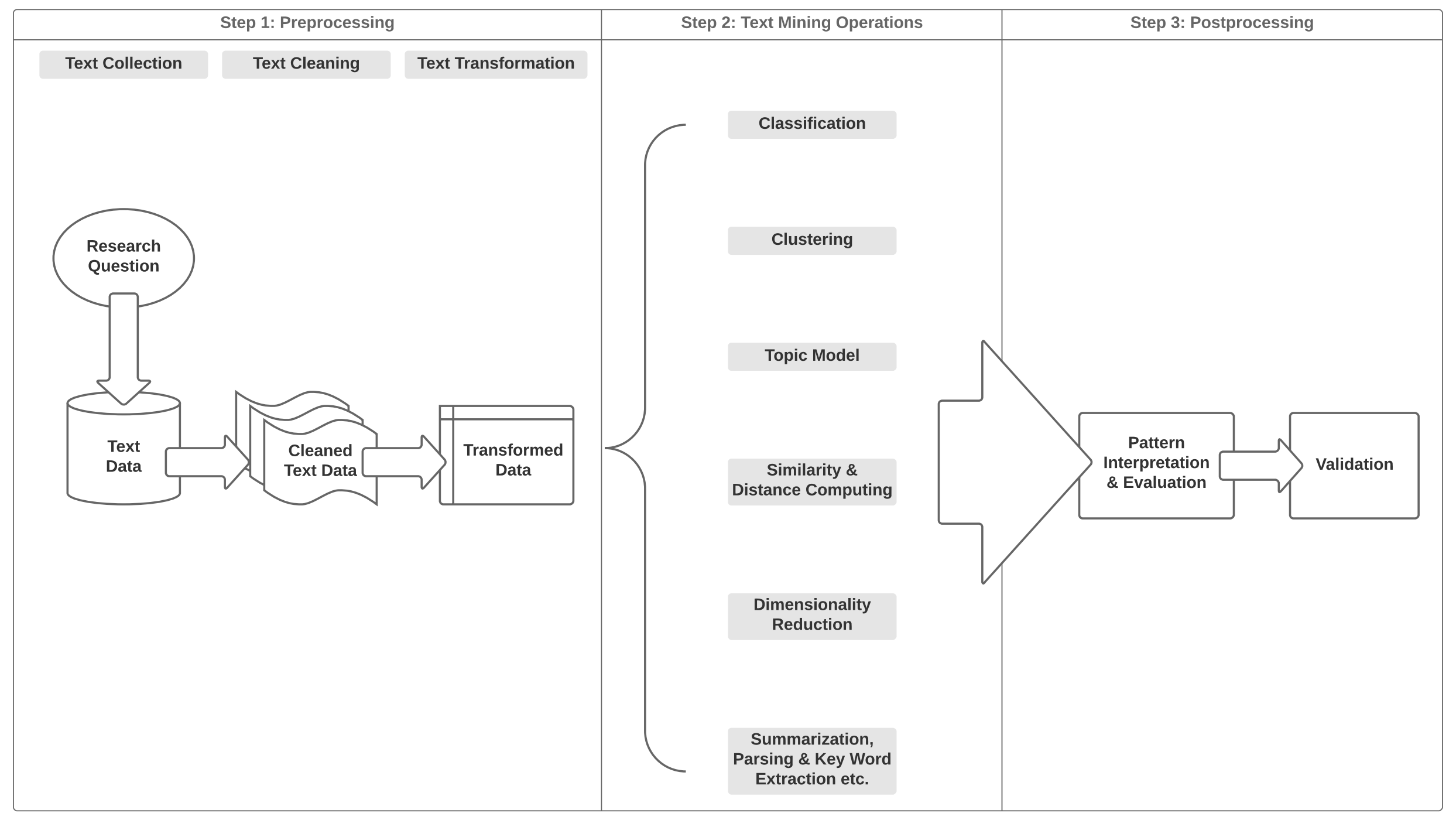
Testified that the accuracy of learning text classifiers can be improved by blending a large number of unlabeled documents with a small number of labeled training documents (Tom Mitchell, etc.)

2002 Applied **machine learning** in automated text categorization and evaluated the performance (Fabrizio Sebastiani)

2010 Applied **n-gram** on sentiment analysis and opinion mining in Twitter (Alexander Pak, Patrick Paroubek)

2014 Applied **Convolutional Neural Networks** to Sentence Classification (Yoon Kim)

Process of Text Mining



Related Algorithms

1. **Naïve Bayes**
The core formula is Bayes formula. Initially we may have a corpus or a training dataset that can tell the prior probabilities. Then compute and compare the posterior probabilities of different classes. At last pick up the maximum and its corresponding class is what we want.

2. **Decision Tree**
Turn the sentences into feature sets. Each node of the tree is a test of some features of the training instance.

3. **Support Vector Machine**
In Perception we use a hyperplane to divide all the points in a space exactly into two classes. In SVM we maximize the distances between those points which are closest to the hyperplane and the hyperplane.

4. **CNN**
Firstly, we vectorize the sentences. And several sentences compose a matrix. Then we use a convolutional matrix, a pooling layer and a activate function to turn the input matrix to a value. At last, we classify the text depending on its value.

5. **Clustering**
Turn the text into a vector, i.e. the presence and absence of a word in a sentence is represented by 1 and 0. Then use a similarity function to find groups of similar texts in a set of texts.

Tasks

1. Question Answering	- Answering questions in open domain - Answering questions about news and articles - Answering questions about certain subjects
2. Text Summarization	- Creating a title for a text - Creating a summary of a text
3. Machine Translation	- Translating a text document - Converting audio to text - Converting text to audio
4.Caption Generation	- Describing an image - Creating a caption for a video
5. Speech Recognition	- Transcribing speeches - Issuing commands to the radio while driving
6. Language Modeling	- Spelling correction - Handwriting recognition - Machine translation - Text generation
7. Text Classification	- Sentiment analysis - Spam filtering - Language identification - Genre classification

Real World Applications

1. **Digital Libraries** (Green-stone international digital library)
- Extract what or who a document mentions
- Understand what topics it deals with
- Group papers in the same field into one category

2. **Healthcare**
- Gain symptoms on the similar diseases, because sometimes they are hard to classify
- How symptoms vary in different gender and ages
- Evaluating the effectiveness of medical treatments

3. **Social Media** (twitter, facebook, instagram)
- Do people love my advertisement? And why?
- Why people hate someone in this event?
- Do people support the policy change? And Why?

4. **Business Intelligence** (Amazon, Yelp, Booking)
- Customers' attitude toward to new products, like it or not?
- Which part I can improve customers' satisfaction? Product or service?
- Why people have no interest about my products? Size problem, color or material?

netowl

AYLIEN

rapidminer

References

[1] Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text Mining in Organizational Research. Organizational Research Methods, 21(3), 733–765.
[2] Talib R et al (2016) Text mining-techniques applications and issues. Int J Adv Comput Sci Appl 7(11):414–418
[3] Zampieri, Marcos & Malmasi, Shervin & Nakov, Preslav & Rosenthal, Sara & Farra, Noura & Kumar, Ritesh. (2019). Predicting the Type and Target of Offensive Posts in Social Media. 1415-1420. 10.18653/v1/N19-1144.
[4] https://machinelearningmastery.com/applications-of-deep-learning-for-natural-language-processing/