Text Clustering: Evaluation

ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

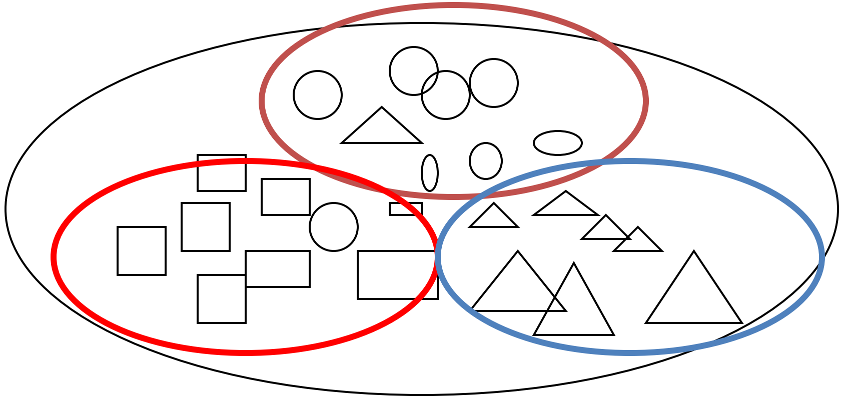
Overview

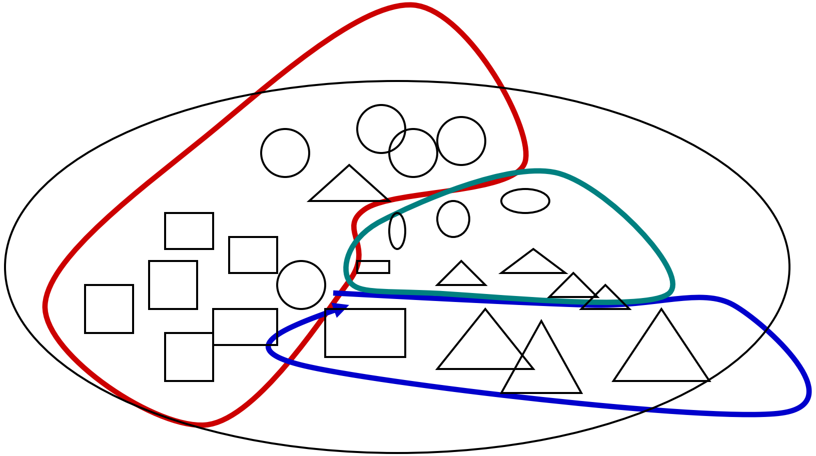
- What is text clustering?
- Why text clustering?
- How to do text clustering?
  - Generative probabilistic models
  - Similarity-based approaches
- **How to evaluate clustering results?**

The “Clustering Bias”

- Any two objects can be similar, depending on how you look at them!
- A user must define the **perspective** (i.e., a “**bias**”) for assessing similarity!

Basis for evaluation





Direct Evaluation of Text Clusters

- Question to answer: How close are the system-generated clusters to the ideal clusters (generated by humans)?
  - "Closeness" can be assessed from multiple perspectives
  - "Closeness" can be quantified
  - "Clustering bias" is imposed by the human assessors
- Evaluation procedure:
  - Given a test set, have humans to create an ideal clustering result (i.e., an ideal partitioning of text objects or "gold standard")
  - Use a system to produce clusters from the same test set
  - Quantify the similarity between the system-generated clusters and the gold standard clusters
  - Similarity can be measured from multiple perspectives (e.g., purity, normalized mutual information, F measure)

Indirect Evaluation of Text Clusters

- Question to answer: how useful are the clustering results for the intended applications?
  - “Usefulness” is inevitably application specific
  - “Clustering bias” is imposed by the intended application
- Evaluation procedure:
  - Create a test set for the intended application to quantify the performance of any system for this application
  - Choose a baseline system to compare with
  - Add a clustering algorithm to the baseline system ➔ “clustering system”
  - Compare the performance of the clustering system and the baseline in terms of any performance measure for the application

Summary of Text Clustering

- Text clustering is an unsupervised general text mining technique to
  - obtain an overall picture of the text content (exploring text data)
  - discover interesting clustering structures in text data
- Many approaches are possible
  - Strong clusters tend to show up no matter what method used
  - Effectiveness of a method highly depends on whether the desired clustering bias is captured appropriately (either through using the right generative model or the right similarity function)
  - Deciding the optimal number of clusters is generally a difficult problem for any method due to the unsupervised nature
- Evaluation of clustering results can be done both directly and indirectly

Suggested Reading

- Manning, Chris D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2007. (Chapter 16)