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