6 Clustering

1. Download the clustering package CLUTO:
   `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview`

2. Use the document collection provided in folder `data`, which is in the appropriate format for CLUTO.

3. Execute the program `vcluster`, requesting 13 clusters, and using single-link:
   `vcluster -clmethod=agglo -crfun=slink docs 13`
   Analyze the output and statistics provided by CLUTO.

4. Execute again the program textttvcluster, using complete-link.
   `vcluster -clmethod=agglo -crfun=clink docs 13`
   What can we say about the results?

5. Complete the program `pip.py` to compute purity and inverse purity of the produced clustering with respect to the original classification of the documents.

6. Use program `pip.py` to compute purity and inverse purity of the clusterings produced by CLUTO using different parameter settings.
   `paste classes doc.clustering.13 | python pip.py`

   Test different clustering methods (option `-clmethod` with values `rb`, `agglo`, `bagglo`, `graph`), link strategies (option `-crfun` with values `slink`, `clink`, `upgma`), and similarity measures (option `-sim` with values `cos`, `dist`, `jacc`)

Which parameter setting produces results more similar to the original classification? What can we say about the used representation of documents? Is there a relation between purity and inverse purity and the internal/external similarity measures provided by CLUTO ?