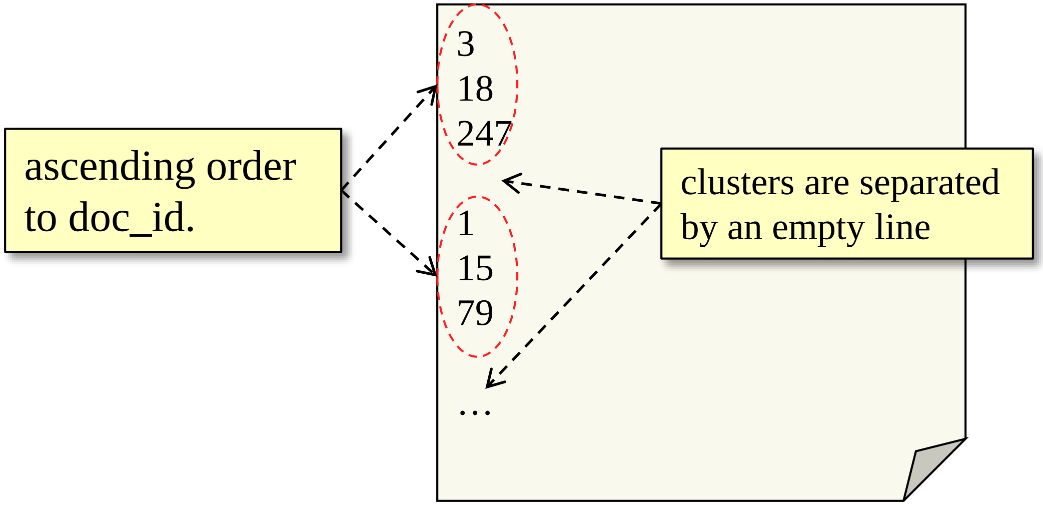
Programming Assignment 4 (1/2)

- **HAC clustering**:
  - Text collection:
    - The 1095 news documents.
  - *K* = 8, 13, and 20.
    - Save each clustering result in a file – *K*.txt (that is, 8.txt, 13.txt, and 20.txt).



```
3
18
247

1
15
79

…
```

ascending order to doc_id.

clusters are separated by an empty line

K.txt

Programming Assignment 4 (2/2)

- TA will evaluate your clustering performances in terms of *precision*, *recall*, and F_1 *metrics*.
- Note:
  - Documents are represented as **normalized** tf-idf vectors.
    - Remind your programming assignment 2.
  - **Cosine similarity** for pair-wise document similarity.
  - Similarity measure between clusters can be:
    - single-link, complete-link, group-average, and centroid similarity.
  - To speed up your clustering … you **MAY** … (15 bonus points)
    - Use HEAP to obtain the cluster pair with maximal similarity.
- Please zip and submit 1.your clustering results (*K*.txt), 2.source code, and 3.a report to TA.
  - 3 weeks to complete, that is, **2019/1/1**.