Data Mining Assignment - Clustering

Sam Pinxteren, Prof. Calders

Deadline: May 18, 23:59

This assignment document consists of 3 sections; the project description, the tasks and questions, and the notices. Please read all carefully before beginning your work.

1 Project Description

This assignment involves clustering 2100 short texts. These texts are Wikipedia lead texts, the first paragraphs which describe a certain concept. You'll find the lead texts in the file *articles.tsv*. The format of this file is as such:

```
Title<tab>Text
Title<tab>Text
...
```

However you load this data, make sure you end up with 2100 articles.

Note that this assignment is smaller than the others. Because of this, it also counts towards your grade less. It counts for 10 points (as opposed to 20) towards the grade for the practical section of this course.

2 Tasks and Questions

Complete the following tasks and report your process together with the answers to each question in a pdf-file.

1. How many clusters do you expect the data to have? Clearly describe your process of finding the number of clusters. How confident are you, or can you be, in this answer?
2. Cluster the given articles using one of the method you have seen in class. Give a brief overview of the clustering methods you have seen and why each would be a good or bad choice for this dataset.
3. Write your clusters to a file called *clusters.tsv*, numbering each of your clusters. The format of the file is similar to that of the *articles.csv* file:

   ```
   Title<tab>Cluster Number
   Title<tab>Cluster Number
   ...
   ```

   Make sure this file has 2100 lines, one for each text in the original file.

3 Minimal Requirements

These points indicate the minimal work needed to pass this assignment. In other words, you should pass the assignment if you've completed these tasks correctly. If you finish only these minimal requirements and make mistakes you might not pass. Gaining higher grades requires creativity on your part, this generally consists of putting above minimal effort into solving the given problems.

- Give a clear answer to *every* question in the second section. Make sure it is easy for us to find these answers in your report.
- Try multiple methods for solving the cluster count and clustering problems and compare results, or give very good reasons for why certain methods are superior.
- Use the method you have deemed best to create a clustering and add the correctly formatted *clusters.tsv* file to your submission.
- For any answer or statement (which is not a part of common knowledge or part of the course) describe the process and reasoning to come to this. Your report is essentially a scientific document.

4 Notices

Your final submission should be a zip-archive called submission.zip which contains:

1. Your report, a PDF file called *report.pdf*. Although it is not a requirement, we recommend you use LaTeX.
2. The code you have written to solve the questions. The report should stand on its own as your submission. The code should not be necessary for us to evaluate your work and should never be part of the report.
3. A file called *clusters.tsv*.

Important Notes:

- Any file with the wrong file name will not be considered, including capital letters and file extensions. Please double check that this is correct.
- Upload your submission to the assignments section on blackboard for this course and assignment before the announced deadline, any delay will result in a score of 0 for this assignment.
- Do not include any of the provided data in your submission. Clearly state where the dataset needs to be placed for the code to work.
- If you have any questions or remarks, feel free to send an email to: sam.pinxteren@uantwerpen.be.