Algorithm Class

2017/9/27

Tasks

- **Cloud Web**
  - **Use the given online system to get weak labels**
  - **Split the documents into Related or Not Related**

Tasks

- Task1: Regression
  - Motivation: Calculate the score of the query and document
  - Metric: RMSE MAE
- Task2: Classification
  - Motivation: Calculate the relation of the query and document, related or not related.
  - Metric: Accuracy

Tasks

- Process
  - Pre-process the data, like removing stop words, stemming(Tools: NLTK or StanfordNLP)
  - Splitting the data into train and test
  - Select the feature of document and query, like tf-idf, one-hot, pos, word2vec.
  - Select the algorithm to regress or classificate
  - Train and test with the metrics.

Tasks

- **Reference material**
  - **http://scikit-learn.org/stable/**
  - **https://zhuanlan.zhihu.com/p/20757320**
  - **https://radimrehurek.com/gensim/models/word2vec.html**
  - **https://github.com/rgtjf/Semantic-Texual-Similarity-Toolkits**

Dataset

- **Labeled data file: Hiemstra_LM0.15_Bo1bfree_d_3_t_10_16.res**
- **Document set:**

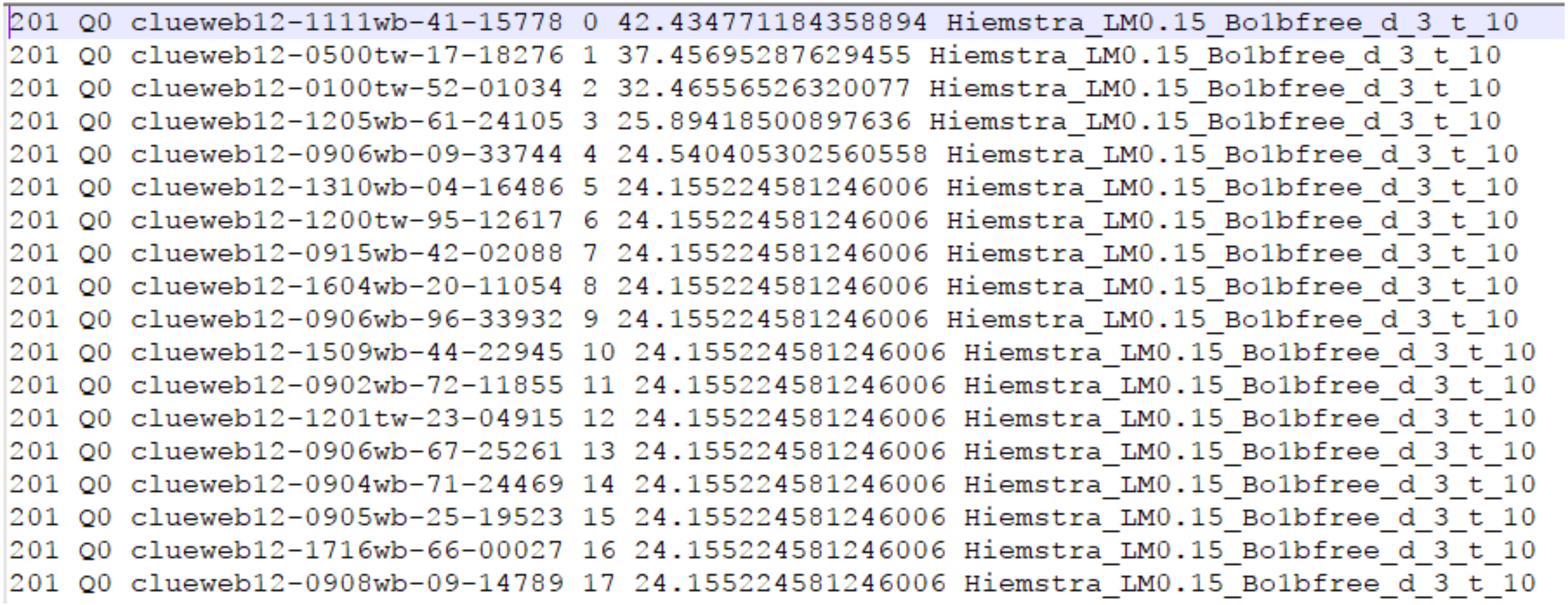
  documents.txt
- **Query set:**

  querys.xml

Dataset

Labeled data file

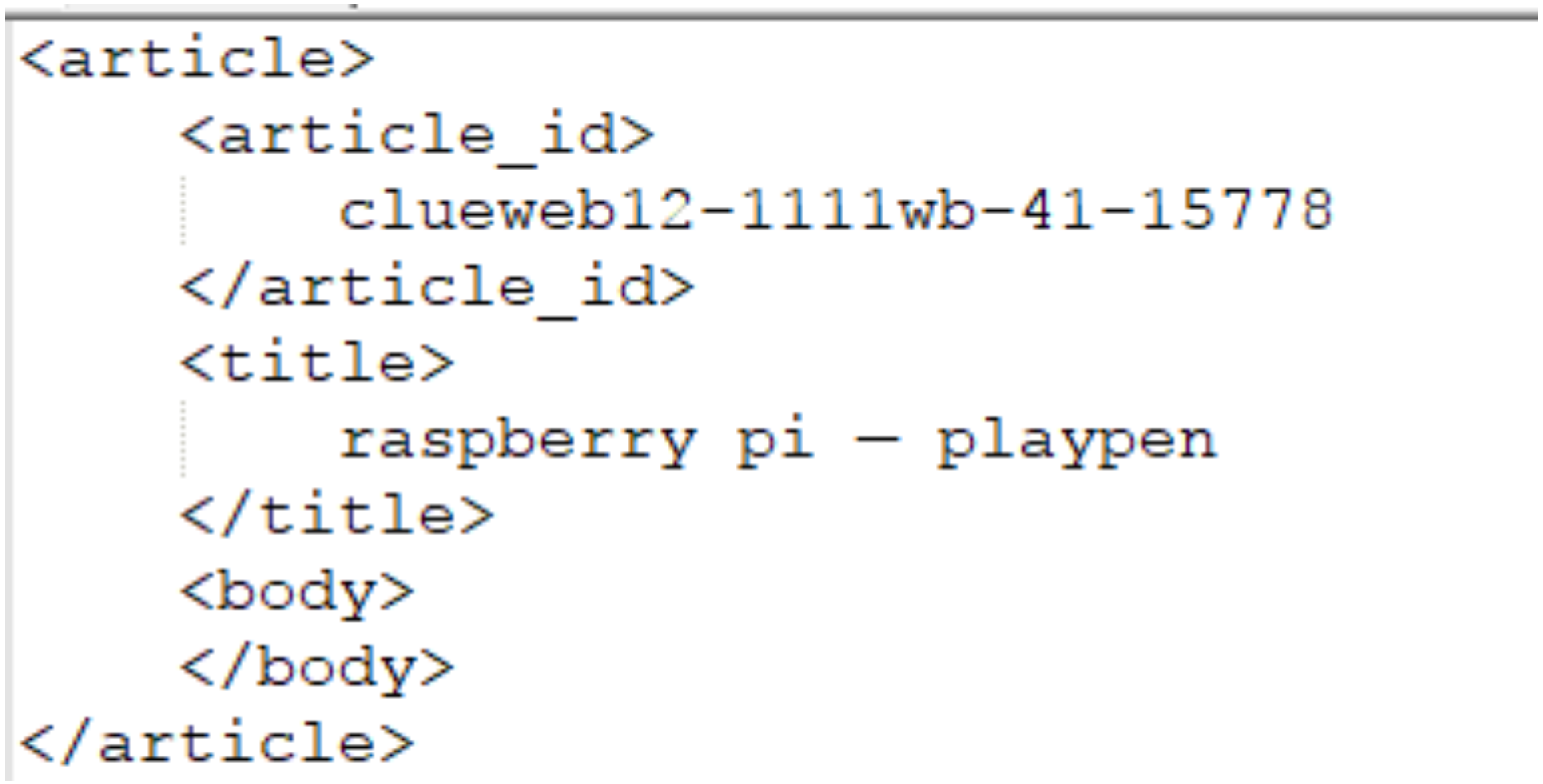
- **First: Query_id**
- **Third: Document_id**
- **Fourth: Score**



```
201 Q0 clueweb12-1111wb-41-15778 0 42.434771184358894 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0500tw-17-18276 1 37.45695287629455 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0100tw-52-01034 2 32.46556526320077 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-1205wb-61-24105 3 25.89418500897636 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0906wb-09-33744 4 24.540405302560558 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-1310wb-04-16486 5 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-1200tw-95-12617 6 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0915wb-42-02088 7 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-1604wb-20-11054 8 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0906wb-96-33932 9 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-1509wb-44-22945 10 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0902wb-72-11855 11 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-1201tw-23-04915 12 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0906wb-67-25261 13 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0904wb-71-24469 14 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0905wb-25-19523 15 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-1716wb-66-00027 16 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
201 Q0 clueweb12-0908wb-09-14789 17 24.155224581246006 Hiemstra_LM0.15_Bo1bfree_d_3_t_10
```

Dataset

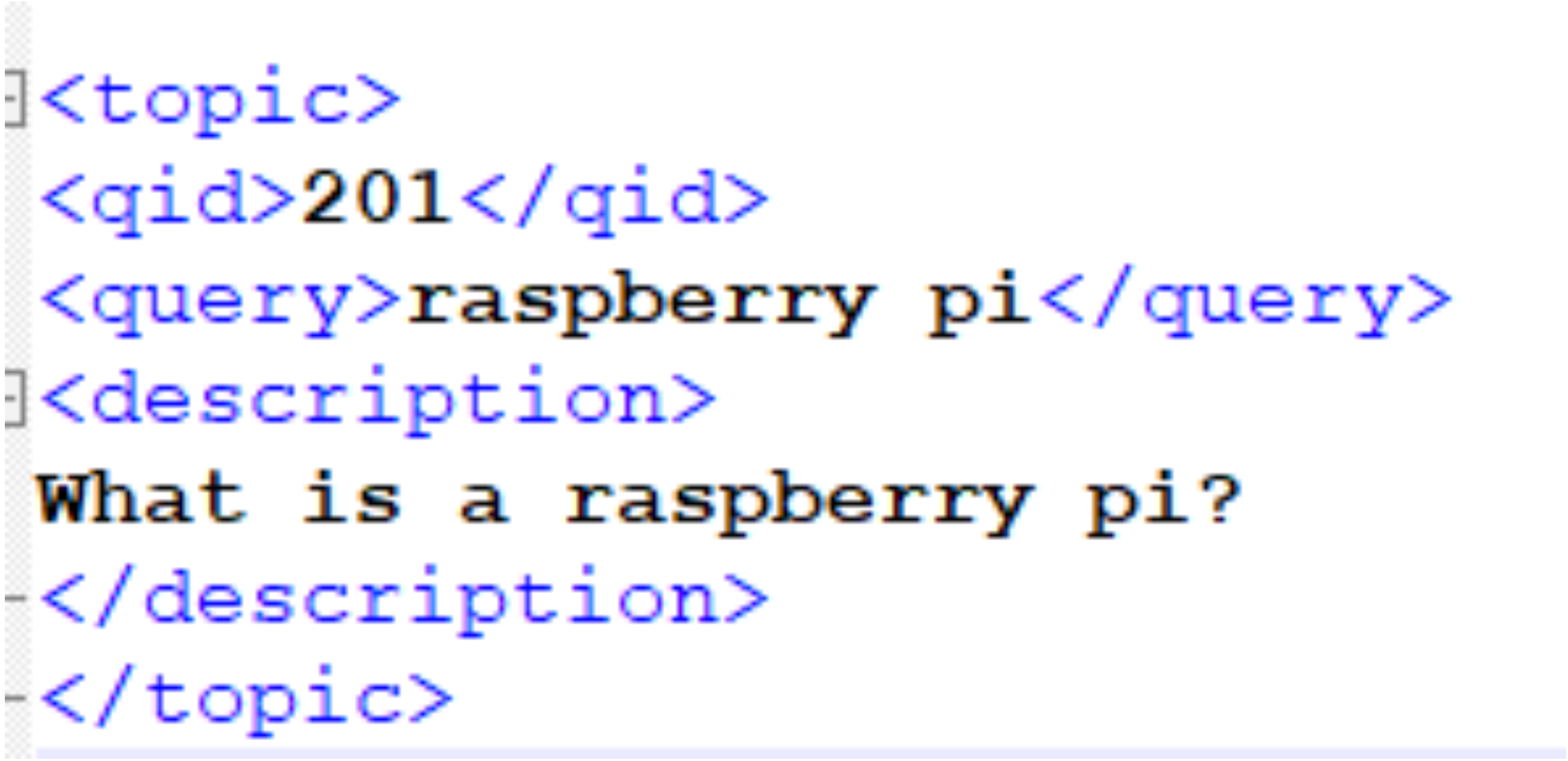
Document set:



```
<article>
    <article_id>
        clueweb12-1111wb-41-15778
    </article_id>
    <title>
        raspberry pi – playpen
    </title>
    <body>
    </body>
</article>
```

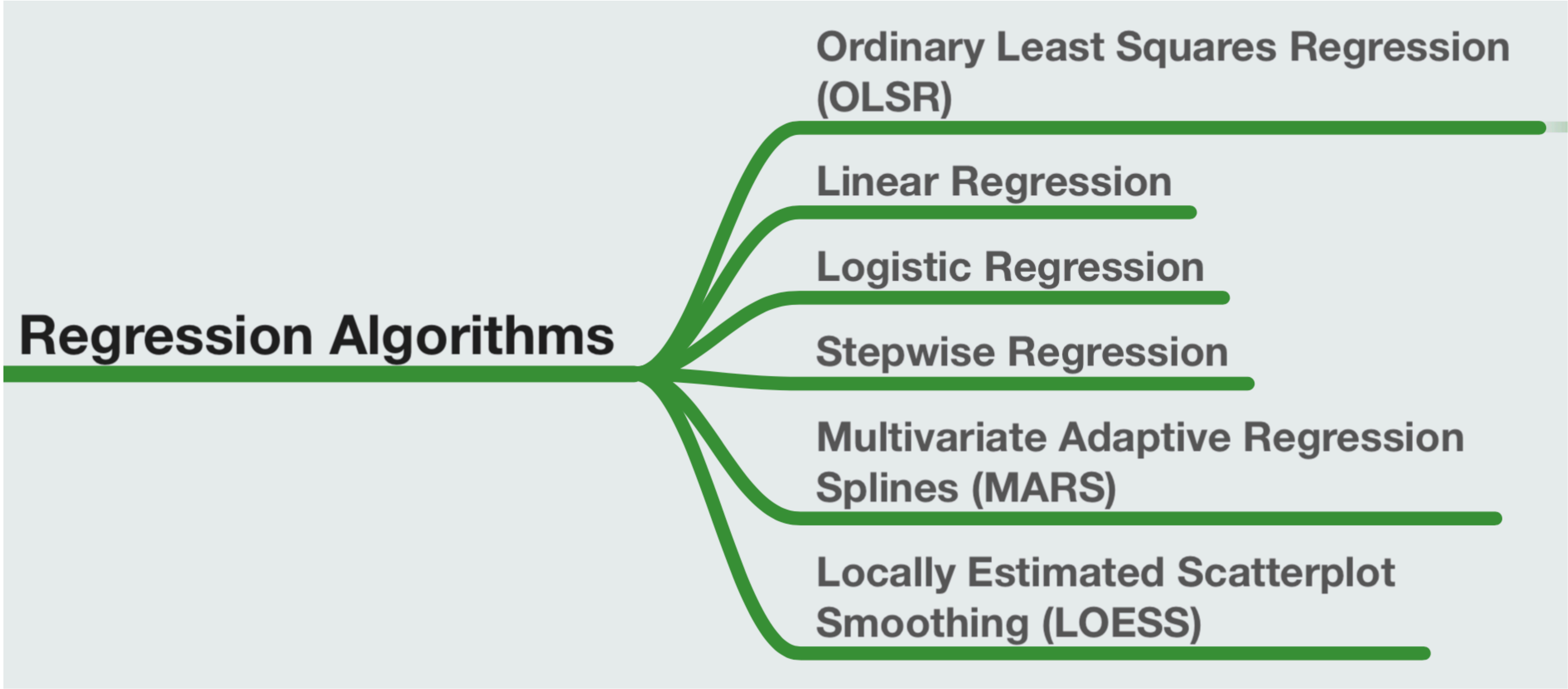
Dataset

Query set:



```
<topic>
<qid>201</qid>
<query>raspberry pi</query>
<description>
What is a raspberry pi?
</description>
</topic>
```

Regression Algorithms



Regression Algorithms

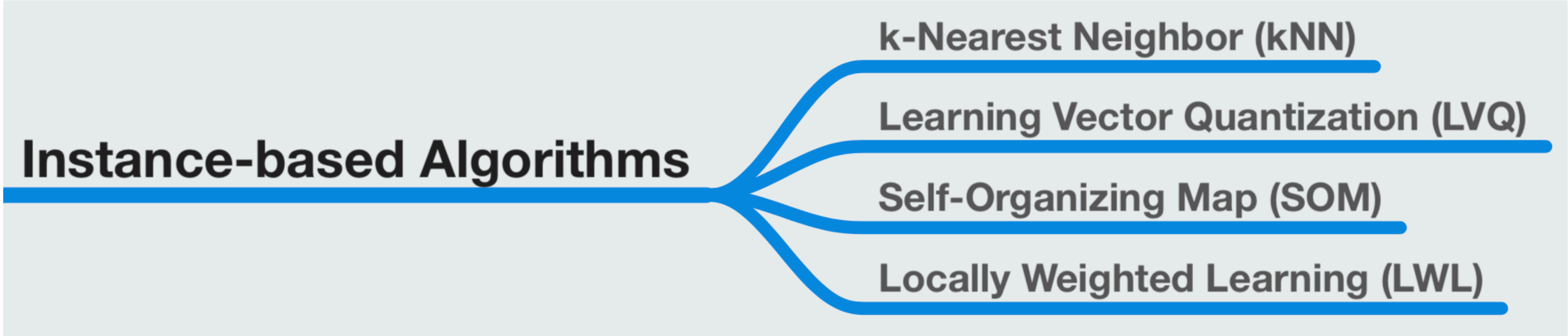
- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

Regression Algs

- **Group: total 6 persons; 1 person / algorithm**

- **Content**
- **Homework**
  - **Split the documents into Related or Not Related**
  - **Classification**

Instance-based Algorithms



Instance-based Algorithms

k-Nearest Neighbor (kNN)

Learning Vector Quantization (LVQ)

Self-Organizing Map (SOM)

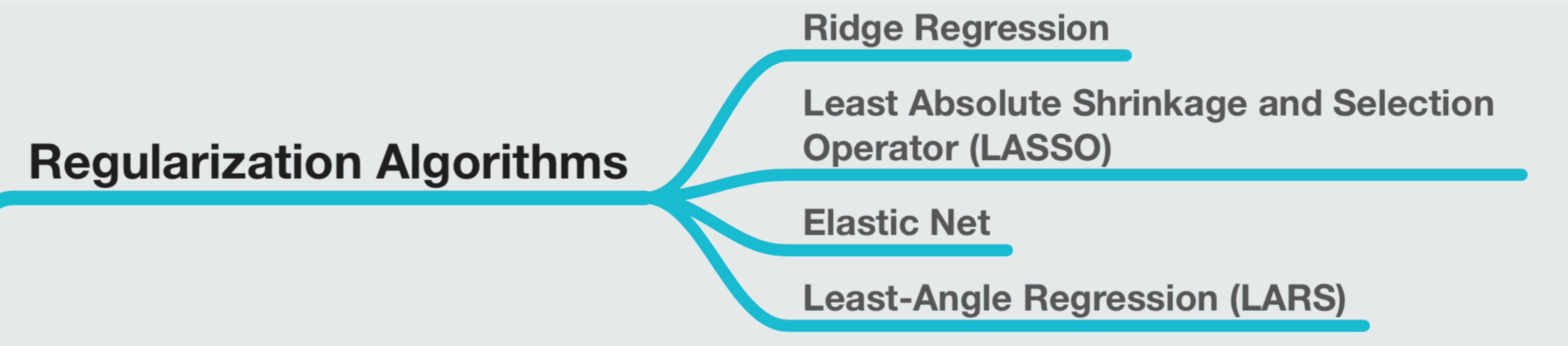
Locally Weighted Learning (LWL)

Instance-based Algs

- Group: total 4 persons; 1 person / algorithm

- Content
- Homework
  - Split the documents into Related or Not Related
  - Classification

Regularization Algorithms



Regularization Algorithms

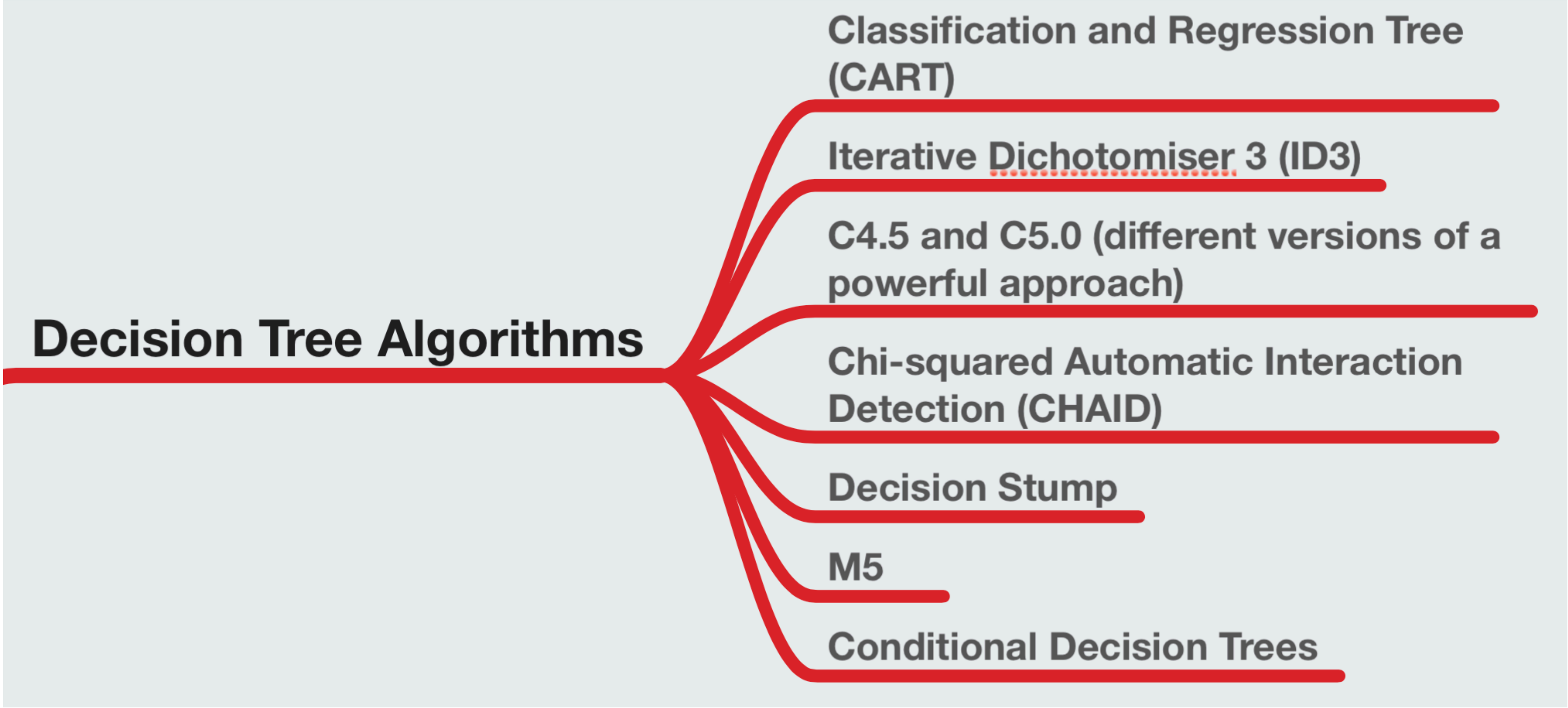
- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)

Regularization Algs

- **Group: total 4 persons; 1 person / algorithm**

- **Content**
- **Homework**
  - **Split the documents into Related or Not Related**
  - **Classification**

Decision Tree Algorithms



Decision Tree Algorithms

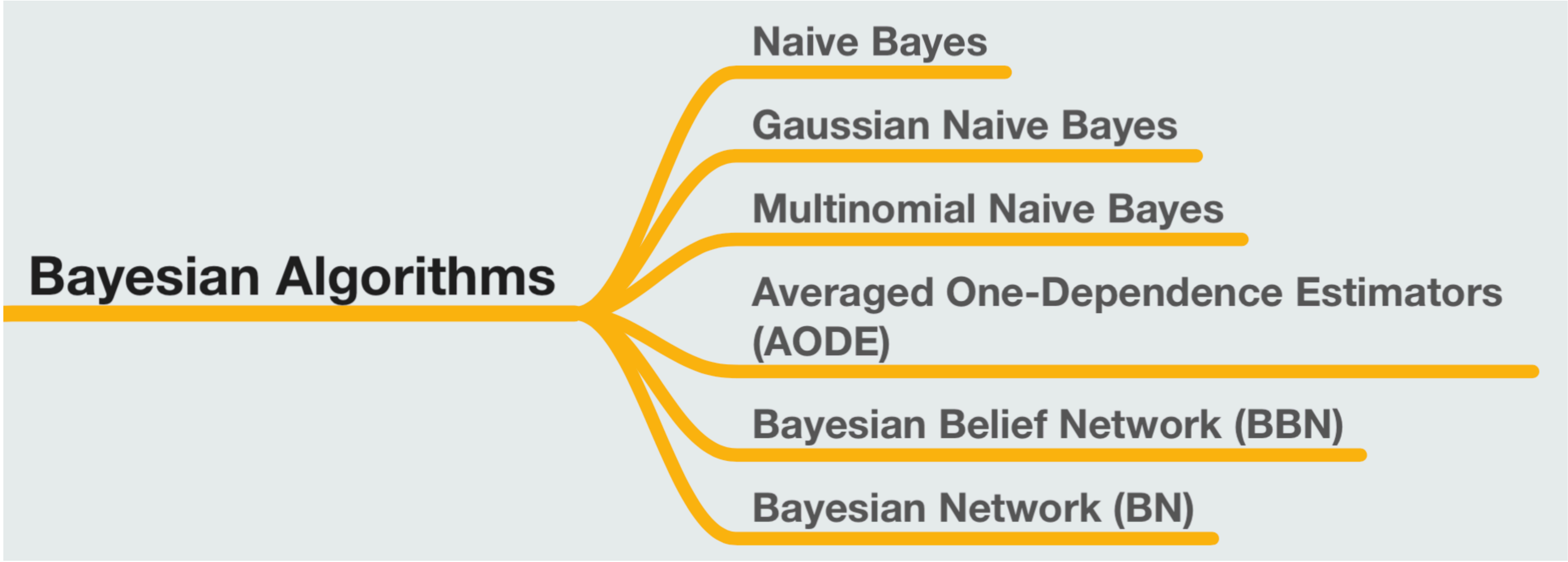
- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees

Decision Tree Algs

- **Group: total 7 persons; 1 person / algorithm**

- **Content**
- **Homework**
  - **Split the documents into Related or Not Related**
  - **Classification**

Bayesian Algorithms



Bayesian Algorithms

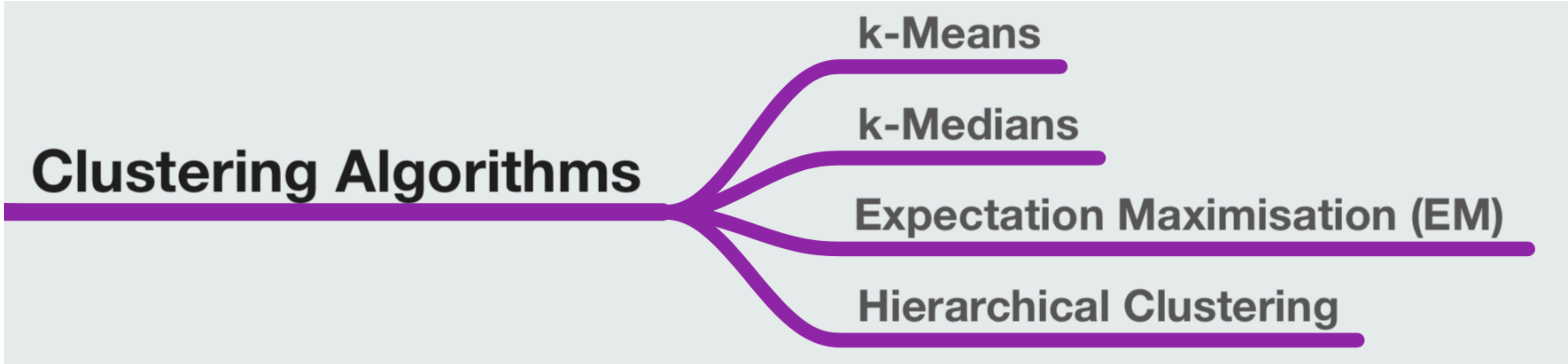
- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)

Bayesian Algs

- Group: total 7 persons; 1 person / algorithm; exception: 2 persons / BN

- Content
- Homework
  - Split the documents into Related or Not Related
  - Classification

Clustering Algorithms



Clustering Algorithms

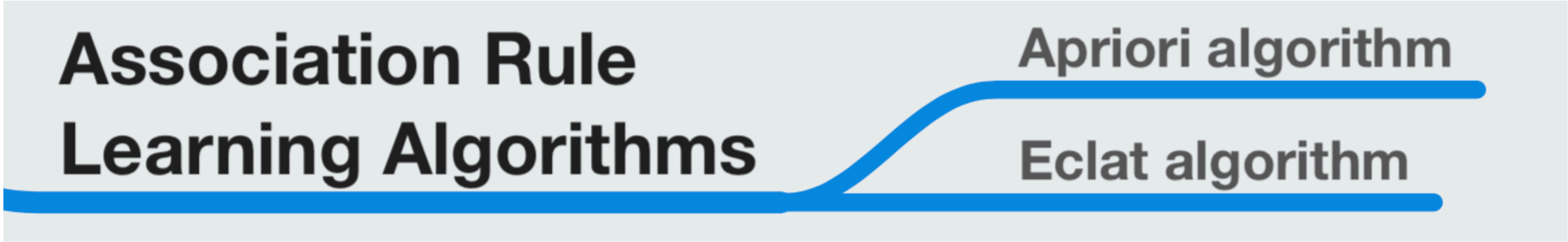
- k-Means
- k-Medians
- Expectation Maximisation (EM)
- Hierarchical Clustering

Clustering Algs

- Group: total 4 persons; 1 person / algorithm;

- Content
- Homework
  - Split the documents into Related or Not Related
  - Classification

Association Rule Learning Algorithms



Association Rule Learning Algorithms

Apriori algorithm

Eclat algorithm

Association Rules Learning Algs

- Group: total 2 persons; 1 person / algorithm

- Content
- Homework
  - Split the documents into Related or Not Related
  - Classification