Durham University

Software Project Development

2019/10/22 Group Meeting Report

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Durham University

Group Meeting Summary

- Machine learning
  - Study Tensorflow and machine learning concept
- HPC logs
  - Survey and category HPC logs with limited understanding
- Others
  - Regular group meeting on Thursday 10:00
  - Regular group discussion on Monday 16:00
  - Create a COSMA account
  - Create a online space for document sharing

Durham University

Group Meeting Action Items

- Machine learning
  - Study Tensorflow for text classification
    - https://www.tensorflow.org/hub/tutorials/text_classification_with_tf_hub
    - https://www.xenonstack.com/blog/log-analytics-deep-machine-learning/
- HPC logs
  - List the questions about logs before group meeting
- Others

Durham University

Group Meeting Questions

- Machine learning
  - Which platform will be used to run our machine learning algorithm?
- HPC logs
  - What is our input data and format, from terminal log or email?
  - Do we have separate input files, instead of just one file?
- Others
  - Our schedule (when should we porting the machine learning algorithm)?

Durham University

Software Project Development

2019/10/24 Group Meeting Minutes

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Durham University

Group Meeting Minutes

- Machine learning
  - The targeted platform which will run our machine learning algorithm is COSMA
- HPC logs
  - We can choose and specify the data format as algorithm input and output
  - The “cron” log is generated by routine job results
  - In SSHD log, our algorithm should figure out which is a new illegal ID, or which illegal ID try to access COSMA many times during a period
  - Our algorithm should detect if there is a disk nearly full
- Others
  - We should come out a schedule, including overall status and individual progressing, for this project
  - The next group meeting change from 10 AM 2019/10/31 to 10 AM 2019/10/25

Group Meeting Action Items

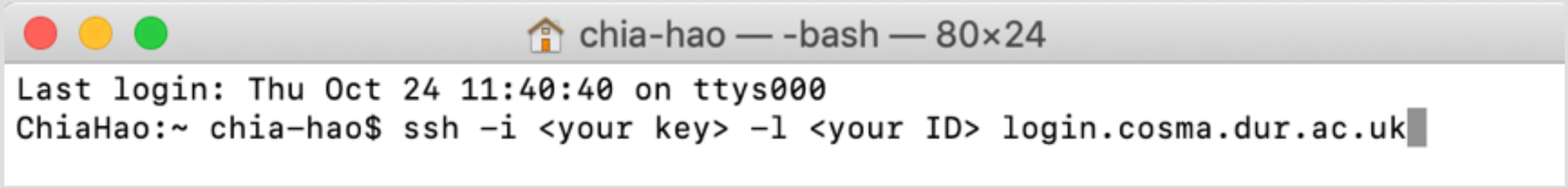
- Machine learning
- HPC logs
- Others
  - Make sure everyone can login to COSMA, please refer to appendix
  - Discuss about our project schedule

Appendix

- Connect to COSMA

Connect to COSMA

- connect to COSMA
  - If you meet connection timeout, please change another network or WIFI
  - Make sure the permission of your key is 600



- PS
  - Your account home is in /comsa/home/durham/<your ID>
  - your account data is in /cosma5/data/durham/<your ID>

Connect to COSMA

- check python version and you will see python version is 2.7.15

chia-hao — dc-li6@login5b:~ — ssh -i ~/Desktop/Software_Development_Proj...

```
[dc-li6@login5b ~]$ python -v
```

Durham University

Connect to COSMA

- check if new version of python is available, and you will see pythonconda3/4.5.4 which we are going to ues

chia-hao — dc-li6@login5b:~ — ssh -i ~/Desktop/Software_Development_Proj...

```
[dc-li6@login5b ~]$ module available
```

Connect to COSMA

- Unload old version python
- Load new version python

```
[dc-li6@login5b ~]$ module unload python/2.7.15
[dc-li6@login5b ~]$ module load pythonconda3/4.5.4
[dc-li6@login5b ~]$
```

Connect to COSMA

- Check python version again and the current python version is 3.6.2

chia-hao — dc-li6@login5b:~ — ssh -i ~/Desktop/Software_Development_Proj...

```
dc-li6@login5b ~]$ python3 -v
```

Connect to COSMA

- Import tensorflow

chia-hao — dc-li6@login5b:~ — ssh -i ~/Desktop/Software_Development_Proj...

```
>>> import tensorflow as tf
```

Durham University

Software Project Development

2019/10/25 Group Meeting Minutes

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Minutes

- Machine learning
- HPC logs
  - We will have a log directory in our COSMA account
- Others
  - We need to submit a formative report in week 6, at most 2500 words
  - We need to submit final report, poster, and code in the beginning of second term
  - Submitted code includes README, logbook, submitted in git (make sure comment is clear when commit code)

	Poster	Code
20%	design clarity	feature
20%	writing style(amount of text, amount of information)	code complexity
20%	scientific method	correctness, robustness
20%	approach, development method	ease of installation
20%	result	performance

Durham University

Group Meeting Minutes

- Others
  - We basically use COSMA5
  - If we need some software which is not installed in COMSA5, email to cosma_support@durham.ac.uk
  - The next group meeting is 11 AM 2019/11/05

Durham University

Group Meeting Action Items

- Machine learning
  - Survey machine learning algorithm for text classification
- HPC logs
  - Check log directory in our COSMA account
- Others
  - Create git account(github.com)
  - Discuss about our project schedule

Durham University

Software Project Development

2019/10/31 Group Meeting Report

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Durham University

Group Meeting Summary

- Machine Learning
  - The process of analyzing text information belongs to the area of Natural Language Processing
  - The task in this project is to detect error message, which is a binary classification task
  - We tend to give a weight score for each message, which we use to sort messages based on how likely they were error message
  - Example: improper message detection in website
- HPC logs
  - We are going to classify according to content(text), not file
- Others
  - Discuss the report for week 6
  - Set the schedule

Durham University

Group Meeting Summary

- Reference
  - https://www.youtube.com/watch?v=PJ_kx9-OPgc&list=PLLDNv7dYom9mHgqZaLyfvD0p5IDeab1YG
  - https://automationlogic.com/log-classification-a-comparison-of-machine-learning-approaches-part-two/
  - https://medium.com/tensorflow/text-classification-using-tensorflow-js-an-example-of-detecting-offensive-language-in-browser-e2b94e3565ce
  - https://towardsdatascience.com/my-first-machine-learning-project-designing-a-hate-speech-detecting-algorithm-56ab32f10833
  - https://medium.com/isiway-tech/deep-nlp-for-hate-speech-detection-25eed707997
  - https://www.futurice.com/blog/hate-speech-detection?fbclid=IwAR0BtSNmzYnpzEiZWYJzTng0FAsgNOGDuQyzYktqf65KDQolyfTKBFm_5bE
  - https://towardsdatascience.com/how-to-do-text-classification-using-tensorflow-word-embeddings-and-cnn-edae13b3e575
  - https://logz.io/blog/machine-learning-log-analytics/

Durham University

Schedule

Week	W03 17-21	W04 28-3	W05 4-10	W06 11-17	W07 18-24	W08 25-1	W09 2-8	W10 9-15	W11 16-22	W12 23-29	W13 30-5	W14 6-12
Log Survey												
Algorithm Survey												
Midterm Report												
Algorithm Coding												
Algorithm Test												
Prepare Report												

Durham University

Group Meeting Action Items

- Machine learning
  - Describe the concept of algorithm in the midterm report
- HPC logs
  - We need to label, weight each error, fail, …. manually
  - connect the error into reason
    - msg 0-222(Jingwen Cai)
    - msg 223-445(Zexu Jiang)
    - msg 446-669(Marah Jaber)
    - msg 670-894(Chia-Hao)
- Others

Durham University

Group Meeting Questions

- Machine Learning
  - Any comment about our concept of algorithm
- HPC logs
  - COSMA has disk quota limit and file quota limit?
  - If COSMA raises a load is warning but recover after a while, do we need to take care this event?
- Others

Durham University

rkhunter Daily Run on login5a.pri.cosma7.alces.network

```
--------------------- Start Rootkit Hunter Update ---------------------
[ Rootkit Hunter version 1.4.6 ]

Checking rkhunter data files...
  Checking file mirrors.dat                                    [ No update ]
  Checking file programs_bad.dat                               [ No update ]
  Checking file backdoorports.dat                              [ No update ]
  Checking file suspscan.dat                                   [ No update ]
  Checking file i18n/cn                                        [ No update ]
  Checking file i18n/de                                        [ No update ]
  Checking file i18n/en                                        [ No update ]
  Checking file i18n/tr                                        [ No update ]
  Checking file i18n/tr.utf8                                   [ No update ]
  Checking file i18n/zh                                        [ No update ]
  Checking file i18n/zh.utf8                                   [ No update ]
  Checking file i18n/ja                                        [ No update ]

---------------------- Start Rootkit Hunter Scan ----------------------
Warning: The SSH and rkhunter configuration options should be the same:
         SSH configuration option 'PermitRootLogin': yes
         Rkhunter configuration option 'ALLOW_SSH_ROOT_USER': unset

----------------------- End Rootkit Hunter Scan -----------------------
```

Durham University

cosma-system: Schedule event from TSM client

```
---- Checking TSMTAPE log file for scheduled backups on 22/10/19 ----

22/10/19   02:10:39 Incremental backup of volume '/cosma/home'
22/10/19   02:10:39 Incremental backup of volume '/cosma/local'
22/10/19   02:22:26 ANS1802E Incremental backup of '/cosma/local' finished with 1 failure(s)
22/10/19   03:20:51 ANS1802E Incremental backup of '/cosma/home' finished with 2 failure(s)
22/10/19   03:20:51 --- SCHEDULEREC STATUS BEGIN
22/10/19   03:20:51 Total number of objects inspected:   16,964,601
22/10/19   03:20:51 Total number of objects backed up:       21,467
22/10/19   03:20:51 Total number of objects updated:             15
22/10/19   03:20:51 Total number of objects rebound:              0
22/10/19   03:20:51 Total number of objects deleted:              0
22/10/19   03:20:51 Total number of objects expired:          1,493
22/10/19   03:20:51 Total number of objects failed:               3
22/10/19   03:20:51 Total number of objects encrypted:            0
22/10/19   03:20:51 Total number of objects grew:                 0
22/10/19   03:20:51 Total number of retries:                    153
22/10/19   03:20:51 Total number of bytes inspected:           1.93 TB
22/10/19   03:20:51 Total number of bytes transferred:         9.10 GB
22/10/19   03:20:51 Data transfer time:                      325.62 sec
22/10/19   03:20:51 Network data transfer rate:           29,289.56 KB/sec
22/10/19   03:20:51 Aggregate data transfer rate:          2,264.41 KB/sec
22/10/19   03:20:51 Objects compressed by:                       0%
22/10/19   03:20:51 Total data reduction ratio:               99.54%
22/10/19   03:20:51 Elapsed processing time:                01:10:11
22/10/19   03:20:51 --- SCHEDULEREC STATUS END
22/10/19   03:20:51 --- SCHEDULEREC OBJECT END DAILY_GPFS_BACKUP 22/10/19   02:00:00
22/10/19   03:20:51
```

Durham University

Final Report Grading

	Poster	Code
20%	design clarity	feature
20%	writing style(amount of text, amount of information)	code complexity
20%	scientific method	correctness, robustness
20%	approach, development method	ease of installation
20%	result	performance

Durham University

Software Project Development

2019/11/05 Group Meeting Minutes

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Durham University

Group Meeting Summary

- Machine Learning
  - First step is going to detect error, warning…directly
  - Further improvement
    - check if the failures will be recovered by COSMA itself
      - Need to add TIME domain information into algorithm
    - Re-learn if the fail information is not important and should not report in the future
- HPC logs
  - Discuss several logs
  - We can select parts of logs for our algorithm and prove the concept
  - The units is a sentence, therefore, we are going to detect line by line
- Others
  - Next meeting is 2019/11/14, 10:00 AM

Durham University

Group Meeting Action Items

- Machine learning
  - Describe the concept of algorithm in the midterm report
- HPC logs
  - We need to define the scope of logs in our report
- Others

Durham University

Schedule

Week	W03 17-21	W04 28-3	W05 4-10	W06 11-17	W07 18-24	W08 25-1	W09 2-8	W10 9-15	W11 16-22	W12 23-29	W13 30-5	W14 6-12
Log Survey												
Algorithm Survey												
Midterm Report												
Algorithm Coding												
Algorithm Test												
Prepare Report												

Durham University

Final Report Grading

	Poster	Code
20%	design clarity	feature
20%	writing style(amount of text, amount of information)	code complexity
20%	scientific method	correctness, robustness
20%	approach, development method	ease of installation
20%	result	performance

Durham University

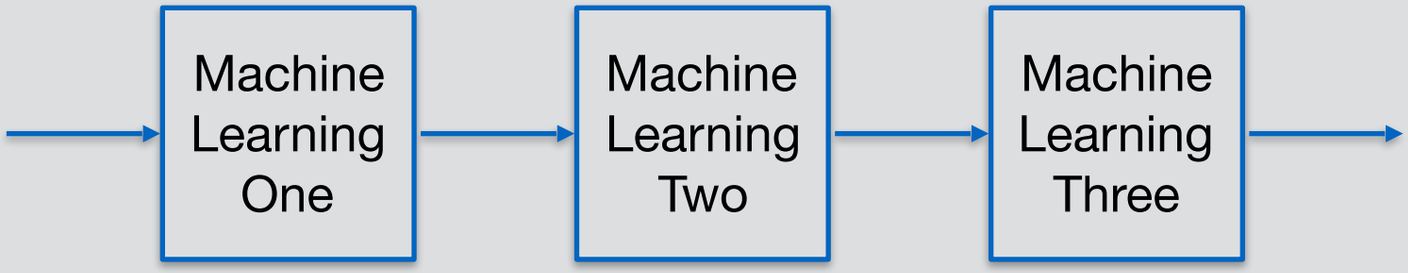
Software Project Development

2019/11/14 Group Meeting Report

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Report

- Machine learning
  - Three stage implementation



\	One	Two	Three
Input	COSMA logs	Output of Machine Learning One	Output of Machine Learning Two
Output	Failure or not	How important the failure is.	Recovery by COSMA itself or not
Algorithm	supervised learning generalized linear classifier SVM,Naive Bayes Classifier,Decision Tree	Term weighting approach Tf-idf	Retraining

  - Run an example on COSMA(Use neural network and adam optimization)
    - https://towardsdatascience.com/how-to-do-text-classification-using-tensorflow-word-embeddings-and-cnn-edae13b3e575

Durham University

- Finish report
- Keep Labeling Logs
- Log Qs:

  -For “Unmatched Entries”:if the entries too much should it become an emergency?

  -Different “Disk Filling Up”, are they the same weight?

  -”The directories listed above were most likely created by a logwatch run that failed to complete successfully. If so, you may delete these directories.”

  -”X-Authentication-Warning”

  ...

- Other Qs:

  -The influence of choosing different granularity(5/10 lines per unit):

  Parallel Computing/ Post-Processing(go back to find more information)

Durham University

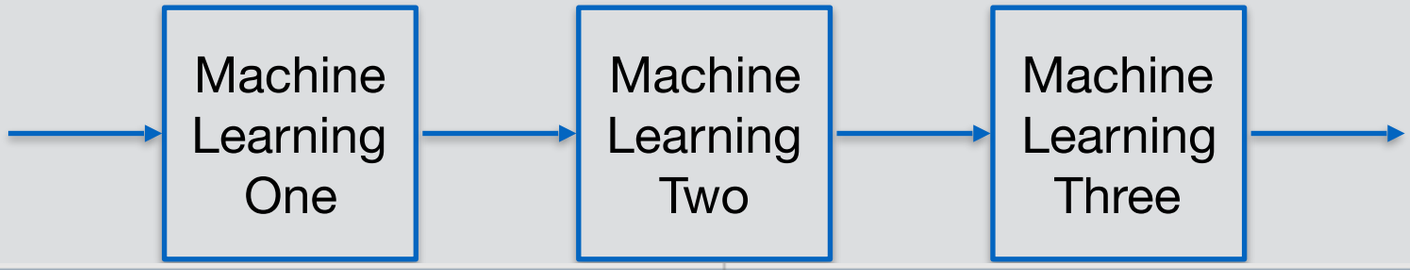
Software Project Development

2019/11/14 Group Meeting Minutes

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Minutes

- Machine learning
  - Three stage implementation



\	One	Two	Three
Input	COSMA logs	Output of Machine Learning One	Output of Machine Learning Two
Output	Failure or not	How important the failure is.	Recovery by COSMA itself or not
Algorithm	supervised learning generalized linear classifier SVM,Naive Bayes Classifier,Decision Tree	Term weighting approach Tf-idf	Retraining

  - Use a pre-stage to classify logs into different types
  - Apply different strategy or granularity for each type of logs
  - Generate artificial logs to improve the prediction quality
- Others
  - Discuss the meaning of logs
  - Final report is individual report

Durham University

Group Meeting Action Item

- Machine learning
  - Make sure the implementation result of line-by-line binary classification is correct
    - Program: /cosma/home/durham/dc-li6/share/line_classify
  - Modify the file classification example for the pre-stage
    - Program: /cosma/home/durham/dc-li6/share/file_classify
  - Check if we can leverage parts of the article title classification program
    - Program: /cosma/home/durham/dc-li6/share/article_title_classify
- Others
  - Next meeting: 10:00 AM, 11/21, OC103

Durham University

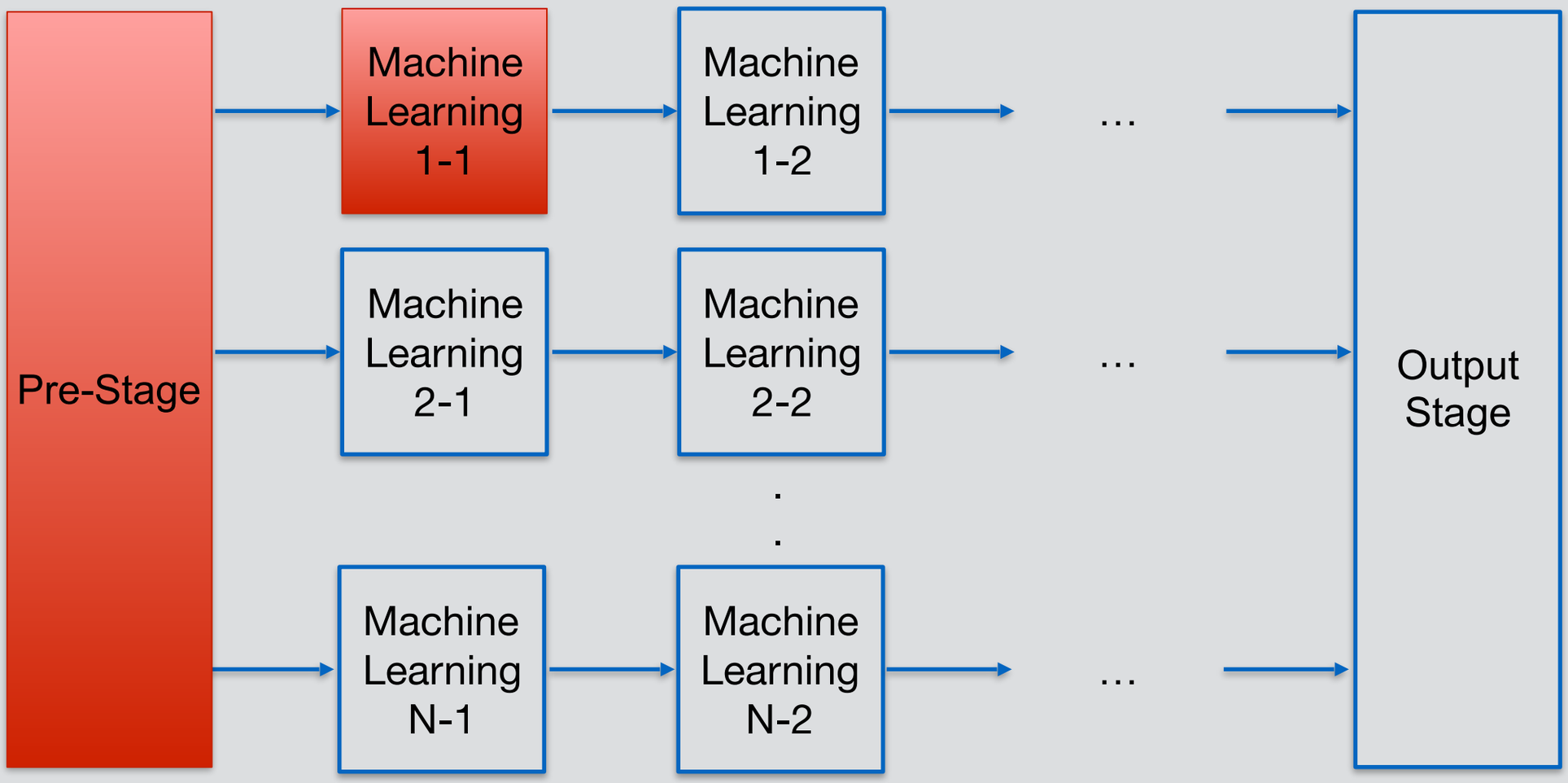
Software Project Development

2019/11/21 Group Meeting Report

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Report

- Algorithm block diagram



- We divide our group into two sub-groups, and each sub-group focuses on one stage
  - Sub-group one for pre-stage: Zexu, Marah
  - Sub-group two for stage 1-1: Jingwen, Chia-Hao

Stage 1-1

- Purpose: recognize keywords in logs line by line
- Input: one category in the pre-stage output
- Output: files contain different types of keyword
- Result:
  - Select 800 logs
  - Train data : test data = 8 : 2
  - Without text augmentation

```
 Performance Report

[[   61     0     0     0     0     0     0]
 [    0   419     0     0     2     0     0]
 [    0     0    56     0     0     0     0]
 [    0     0     0    14     0     0     0]
 [    0     0     1     0 33936     0     0]
 [    0     0     0     0     0    24     0]
 [    0     0     0     0     0     0   294]]
              precision    recall  f1-score   support

 disconnect       1.00      1.00      1.00        61
      error       1.00      1.00      1.00       421
       fail       0.98      1.00      0.99        56
    illegal       1.00      1.00      1.00        14
       pass       1.00      1.00      1.00     33937
  unmatched       1.00      1.00      1.00        24
    warning       1.00      1.00      1.00       294

avg / total       1.00      1.00      1.00     34807

Accuracy: 1.0
```

Durham University

Stage 1-1

——Error Types

Unmatched:

Unmatched Entries/ connect failed

Unmatched Entries/ gethostbyaddr failed

Unmatched Entries/ hostname lookup failed

Unmatched Entries/ Failed to create session: Connection timed out

Unmatched Entries/ Address already in use

Unmatched Entries/ network unreachable

Unmatched Entries/ Name or service not known

Durham University

Stage 1-1

——Error Types

Error:

1 lines must begin with a keyword or a filename

Error ID/ Error Code/ Fibre Channel ports

ErrorRetry

Requests with error response codes

Checking Cosma 5/ Error: timeout

LED error（DIMM 9/ Fault/ SysBrd Vol Fault / CPU 2 PECI/ CMOS Battery/ DIMM 14）

SR ErrCorode

Kernel Errors Present / HEST /Enabling Firmware First mode for corrected errors

Network Read Write Errors

Error processing/ file not found

Kernel Errors Present/ ACPI Error

Error Sequence Number

JS_AMPD_MEDIUM_ERROR

Recipient Errors

DUE TO EXCESSIVE ERRORS

An error was detected by a disk drive

Durham University

Stage 1-1

——Error Types

Warning:

Authentication-Warning

Disk Filling up

Segmentation Faults

General Protection Faults

Client-SSL-Warning: Peer certificate not verified

Running nightly diskusage warnings

Current Load is WARNING

Warning Event Notification

The SSH and rkhunter configuration options should be the same

On warning

Total Processes is WARNING

Warning from cosma6 - Warning LOG_ST_POOL_CHANGE

Warning from cosma6- Error LOG_ST_MI_PD_FAILED

Swap Usage is WARNING

cosma-system/ Warning: DISK_DETECTED_ERROR

EVENT SEVERITY:Warning

Durham University

Stage 1-1

——Error Types

Failed:

Failed logins

Setting tty modes failed

Update failed

scheduled backup/ Total number of objects failed

Scratch volume mount request denied – mount failed

failed to complete successfully

Others:

Received disconnect

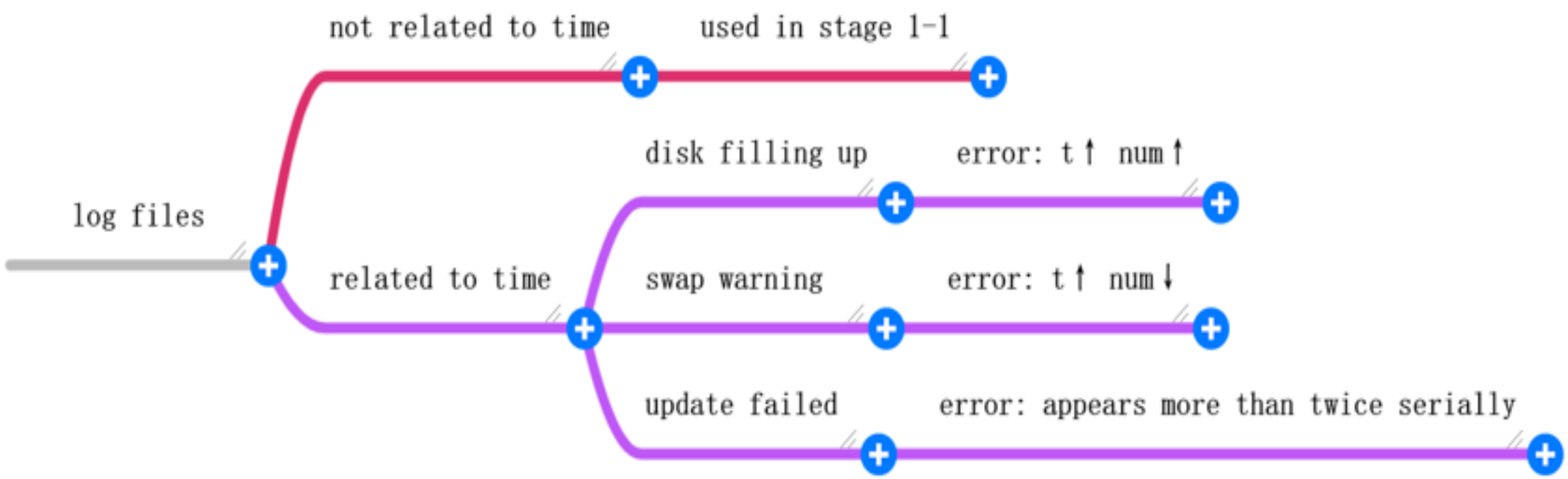
Illegal users

Durham University

Stage 1-1

- Action item:
  - Consider text augmentation in stage 1-1
    - Prepare text augmentation program and rerun the machine learning algorithm
    - Evaluate the effect of text augmentation
  - Consider the further improvement in stage 1-2
    - Give each abnormal event different weight

Pre-Stage


not related to time
used in stage 1-1
disk filling up
error: t↑ num↑
log files
related to time
swap warning
error: t↑ num↓
update failed
error: appears more than twice serially

Durham University

Pre-Stage Analysis

- Data input and pre-processing
  - ❏ CSV file, contains : category, filename, log subject, log content
  - ❏ It's important to have the data balanced, at least 100 example for each category
- Maybe it's enough to use the log subject!
- If not,
  - ❏ we need Data Cleaning such as removing the log data in files before the subject.
  - ❏ need to check if our logs are distinct enough
  - ❏ give different weighting to words based on their importance to the log. for example, using TF-IDF weighting.

Durham University

Pre-Stage Analysis

- Extracting Features
  - we will use the Bag-of-words approach model: In this model, a text (such as a sentence or a document) is represented as multiset of its words, disregarding grammar and even word order but keeping multiplicity.
  - Apply Term Frequency, Inverse Document Frequency, tf-idf.

Durham University

Pre-Stage Analysis

- Model training
  - try different models and choose from them. such as: Logistic Regression, Random Forest classification

- Model evaluation
  - part of the logs will be used to evaluate the model.

Durham University

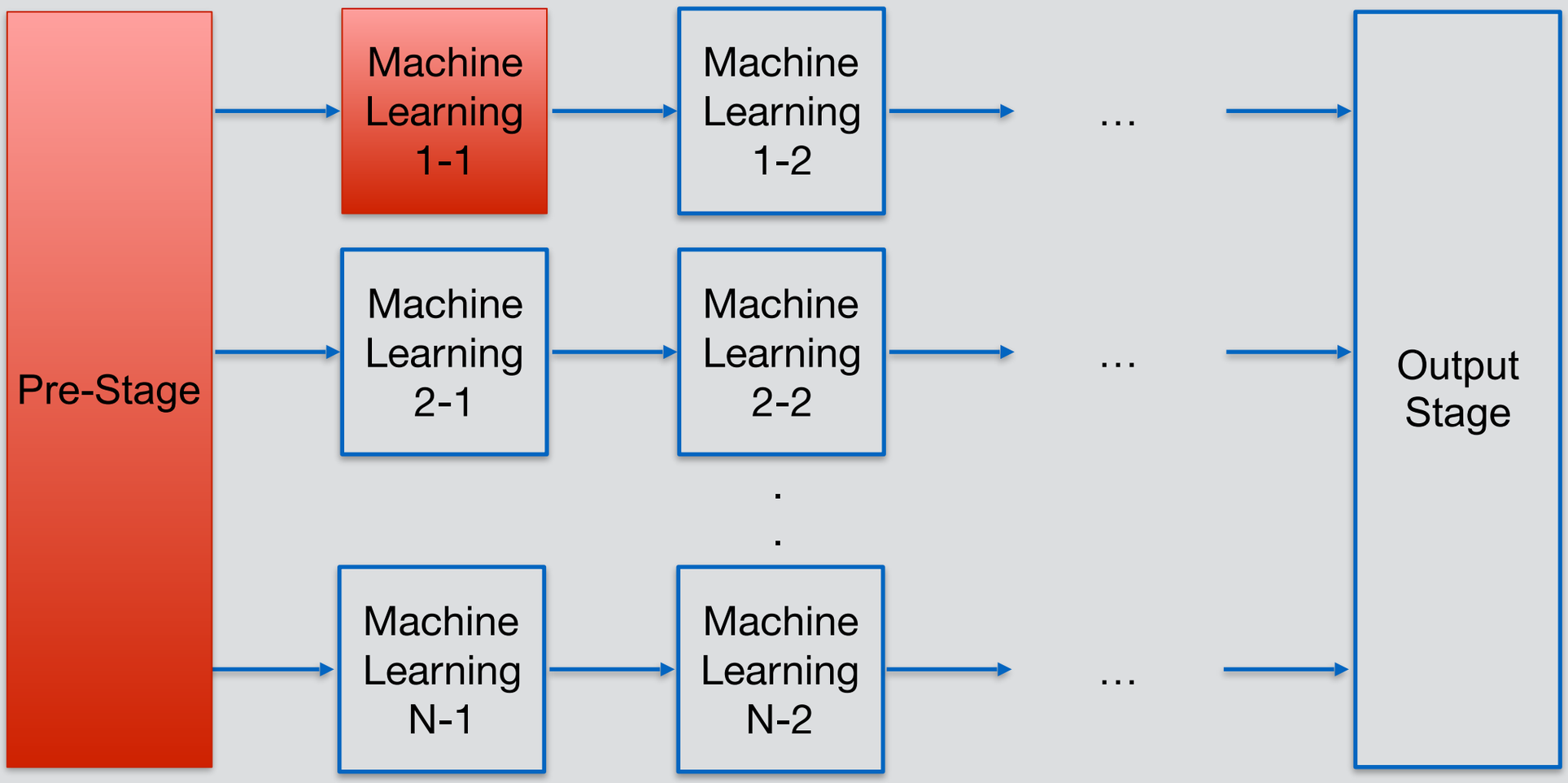
Software Project Development

2019/11/21 Group Meeting Minutes

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Minutes

- Algorithm block diagram



- We divide our group into two sub-groups, and each sub-group focuses on one stage
  - Sub-group one for pre-stage: Zexu, Marah
  - Sub-group two for stage 1-1: Jingwen, Chia-Hao

Durham University

Group Meeting Minutes

- Check the weak point of stage 1-1
  - no error, no failures …
- Consider low-case, upper-case, and mix case of keywords
  - ERROR, error, Error
- Consider the usage of database for new and large logs
- Compress log
  - Store the change part of logs only
- Handle the log has both time dependency and non time dependency abnormal message

Durham University

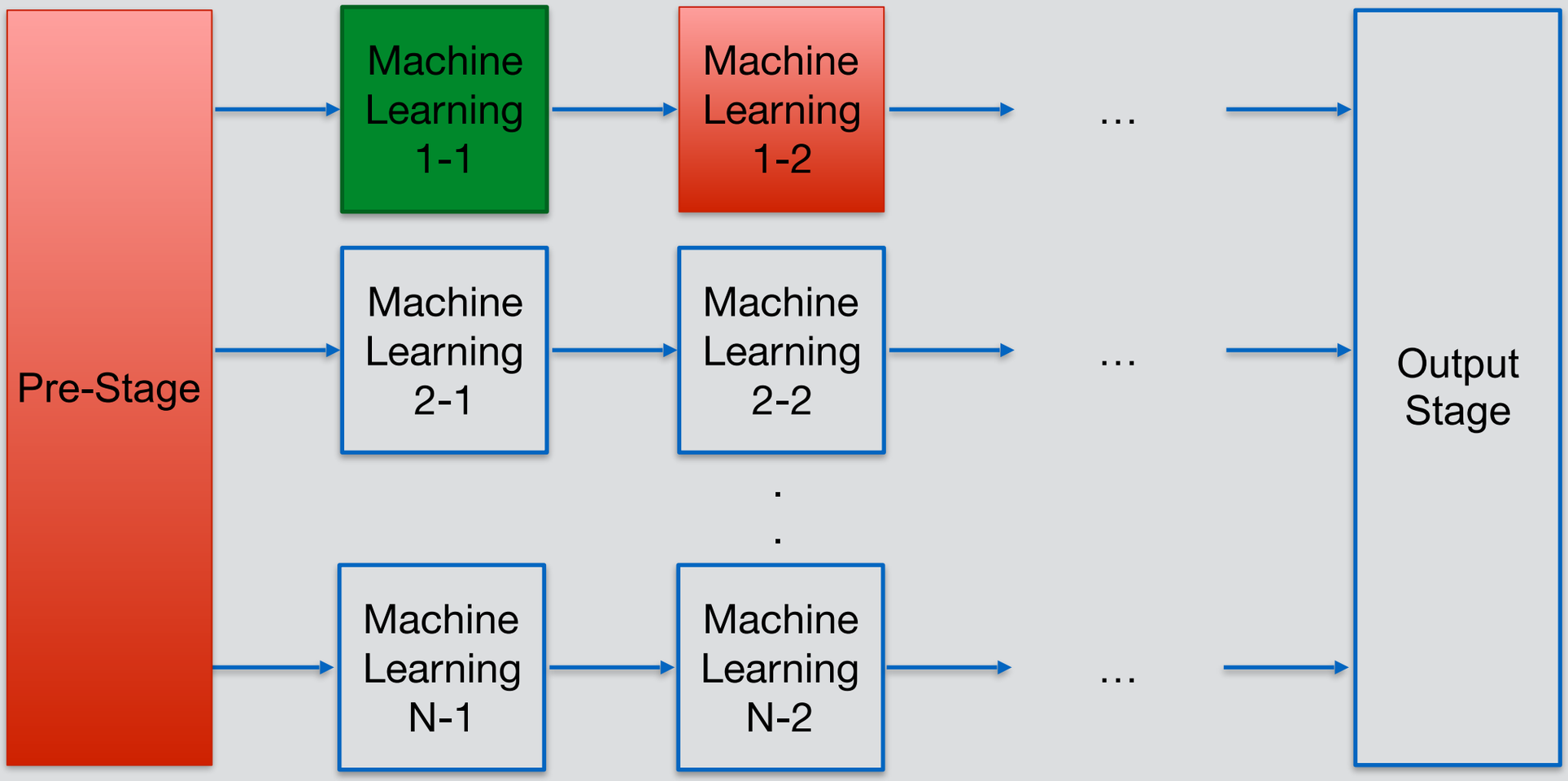
Software Project Development

2019/11/28 Group Meeting Report

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Report

- Algorithm block diagram



- We divide our group into two sub-groups, and each sub-group focuses on one stage
  - Sub-group one for pre-stage: Zexu, Marah
  - Sub-group two for stage 1-1 and stage 1-2: Jingwen, Chia-Hao

Durham University

Stage 1-1

- Purpose: recognize abnormal events in logs line by line
- Input: one category in the pre-stage output
- Output: each type of abnormal events has one text file
  - text file format: predicted result, correct answer, original text, text location

```
error   error   10/25/19   05:12:08      ANE4005E (Session: 851, Node: TSMTAPE)  Error processing       ../logs/msg.885,line_number=6198
error   error    error: connect_to xyanxkqbyfntqyb: unknown host (Name or service not known) : 1 time(s)        ../logs/msg.230,line_number=365
error   error       error (network unreachable) resolving 'mirror.sax.uk.as61049.net/A/IN': 2001:dc3::35#53: 2 Time(s)  ../logs/msg.043,line_number=828
error   error   10/25/19   06:04:03      ANE4005E (Session: 851, Node: TSMTAPE)  Error processing       ../logs/msg.885,line_number=7504
error   error       error (network unreachable) resolving 'mirror.veriteknik.net.tr/A/IN': 2001:500:2f::f#53: 2 Time(s) ../logs/msg.043,line_number=882
```

- Result:
  - Input all logs
  - Train data : test data = 8 : 2
  - Compare the effect of text augmentation (artificial log)

Durham University

Stage 1-1

- We create 104919 lines of artificial logs via text augmentation program
- The prediction result is more accurate with text augmentation(left) than without text augmentation(right)
- Text augmentation drawback: increase 50% computing time

Performance Report

```
[[  318     0     0     0     0     0     0     0]
 [    0  4415     0     0     0     1     0     0]
 [    0     0  8669     0     0     0     0     0]
 [    0     0     0  8330     0     2     0     0]
 [    0     0     0     0    15     0     0     0]
 [    0     0     0     0     0 42507     0     0]
 [    0     0     0     0     0     0    37     0]
 [    0     0     0     0     0     0     0   537]]
              precision    recall  f1-score   support

 disconnect       1.00      1.00      1.00       318
      error       1.00      1.00      1.00      4416
       fail       1.00      1.00      1.00      8669
    failure       1.00      1.00      1.00      8332
    illegal       1.00      1.00      1.00        15
       pass       1.00      1.00      1.00     42507
  unmatched       1.00      1.00      1.00        37
    warning       1.00      1.00      1.00       537

avg / total       1.00      1.00      1.00     64831
```

Performance Report

```
[[   66     0     0     0     0    12     0     0]
 [    0   702     0     0     0     2     0     0]
 [    0     0    37     0     0     2     0     0]
 [    0     0     0    30     0     4     0     0]
 [    0     0     0     0    16     0     0     0]
 [    0     0     0     0     0 42633     0     0]
 [    0     0     0     0     0     0    31     0]
 [    0     0     0     0     0     0     0   312]]
              precision    recall  f1-score   support

 disconnect       1.00      0.85      0.92        78
      error       1.00      1.00      1.00       704
       fail       1.00      0.95      0.97        39
    failure       1.00      0.88      0.94        34
    illegal       1.00      1.00      1.00        16
       pass       1.00      1.00      1.00     42633
  unmatched       1.00      1.00      1.00        31
    warning       1.00      1.00      1.00       312

avg / total       1.00      1.00      1.00     43847
```

Durham University

Stage 1-2

- Purpose: recognize which abnormal events is important
- Input:
  - The classification result from stage 1-1
  - User define high priority keywords
- Output: rank all abnormal events according to importance
- Method:
  - KMeans (https://towardsdatascience.com/applying-machine-learning-to-classify-an-unsupervised-text-document-e7bb6265f52)
  - TFIDF (https://www.youtube.com/watch?v=ZOYYrTDN6N0)
  - Logistic Regression (https://medium.com/@ertuodaba/string-matching-using-machine-learning-with-python-matching-products-of-getir-and-carrefoursa-f8ce29d2959f)

Durham University

Pre-Stage Analysis

- Data input and pre-processing
  - ❏ CSV file, contains : category, filename, log subject, log content
  - ❏ It's important to have the data balanced, at least 100 example for each category
- Maybe it's enough to use the log subject!
- If not,
  - ❏ we need Data Cleaning such as removing the log data in files before the subject.
  - ❏ need to check if our logs are distinct enough
  - ❏ give different weighting to words based on their importance to the log. for example, using TF-IDF weighting.

Durham University

Pre-Stage Analysis

➔ need to check if our logs are distinct enough

The logs are not distinct, most of the logs contain different error types.

➔First method: using a label for each file failed. it is difficult for the algorithm to distinguish the files. and it's even more difficult to give a label for the training data set. since they contain more different errors.

➔Second method: Multi label Text Classification.

https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/

https://www.kaggle.com/roccoli/multi-label-classification-with-sklearn

https://towardsdatascience.com/multi-label-text-classification-with-scikit-learn-30714b7819c5

Durham University

Pre-Stage Analysis

Multi-label text classification:

- using multiple labels for each log file

input: labels (list), filename, log subject, log content

- Data cleaning: removing numbers, empty lines, whitespaces...

Data cleaning: remove popular words as they will give noise ? wasn't able to download the library (ntlk)

Durham University

Pre-Stage Analysis

Multi-label text classification:

- Converting text to features:

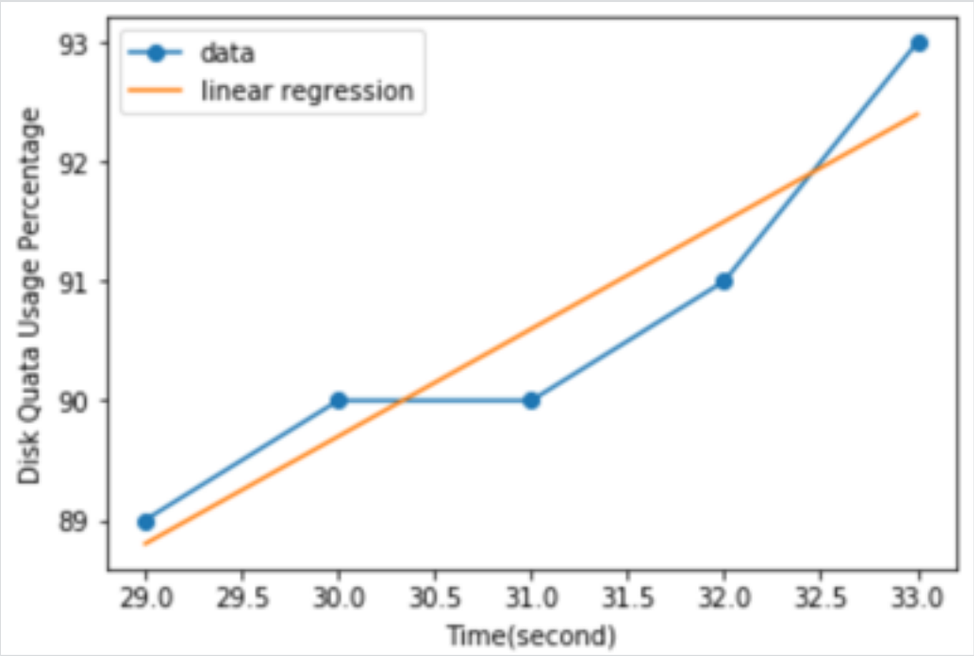
  -TF-IDF, Bag of words, word2vec, GloVe, ELMo

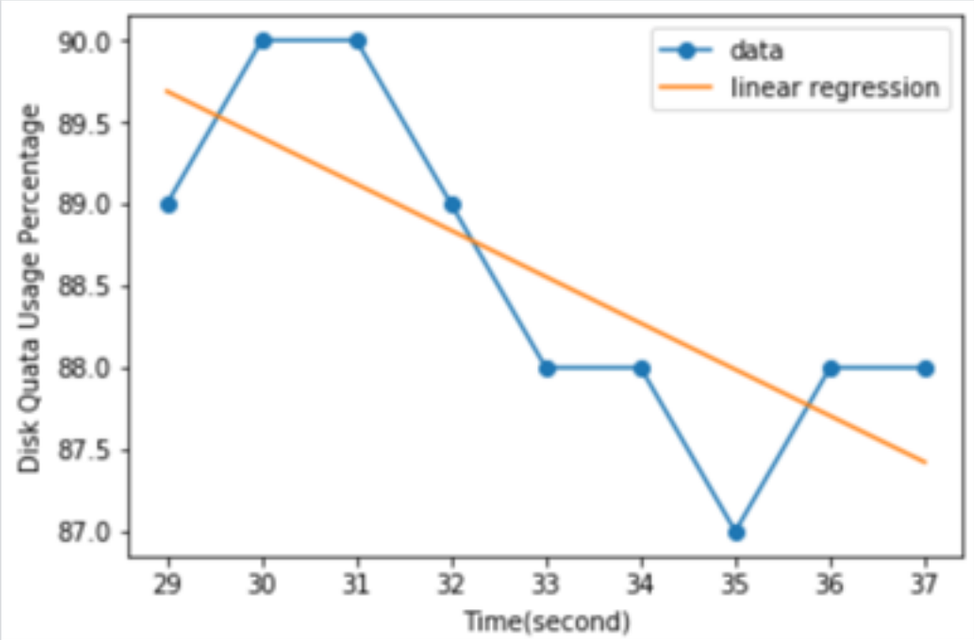
  ➔ we used TF-IDF
- Build a prediction model:

  logistic regression model, OneVSRestClassifier
- Output files with predicted tags

Stage 2-1

- Use linear regression to check the tendency of input figures
- The algorithm will alert operator if the following two condition are satisfaction
  - The latest figure is larger than threshold value (above 90%)
  - The figure will hit maximum value in the future (hit 100%)
- Example:





Durham University

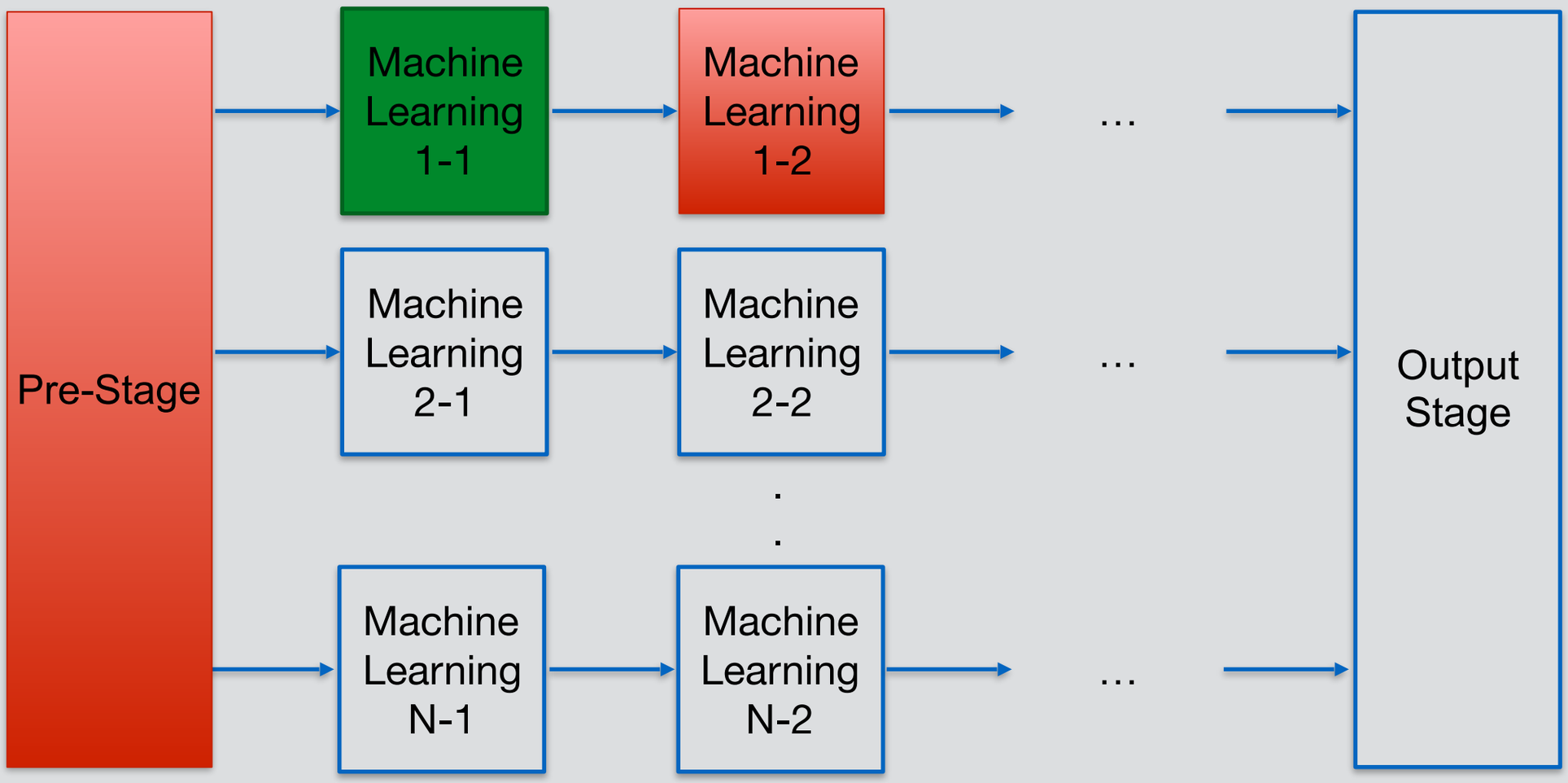
Software Project Development

2019/11/28 Group Meeting Minutes

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Minutes

- Algorithm block diagram



- We divide our group into two sub-groups, and each sub-group focuses on one stage
  - Sub-group one for pre-stage: Zexu, Marah
  - Sub-group two for stage 1-1 and stage 1-2: Jingwen, Chia-Hao

TBD

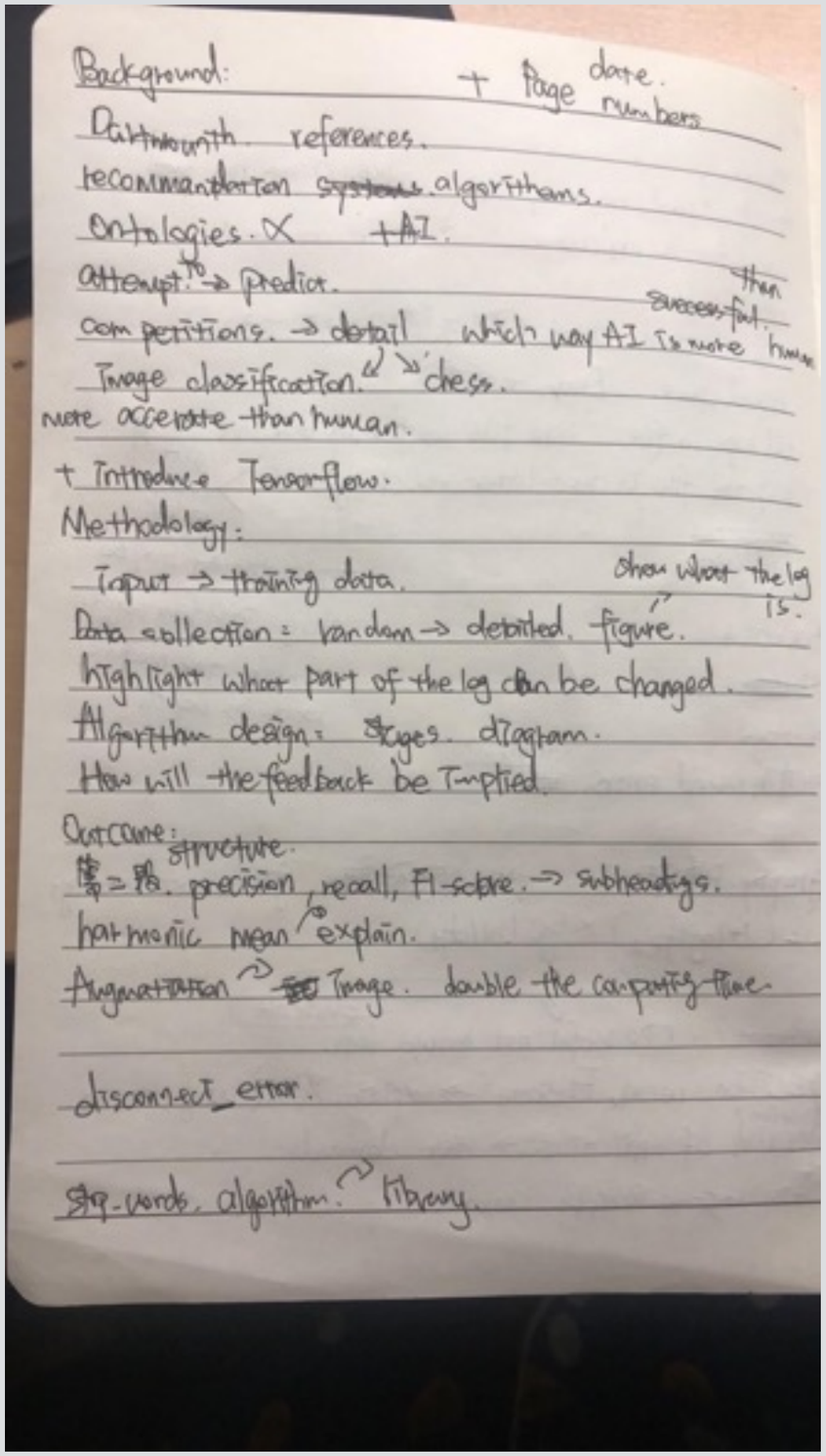
Ongoing

Finished

Durham University

Group Meeting Minutes

- Review the week 6 report
- Sub-group one is going to implement multi-label classification for pre-stage
- Sub-group two will move from stage 1-1 to stage 1-2
- Consider how to implement stage 2-1 for time dependency abnormal events



Durham University

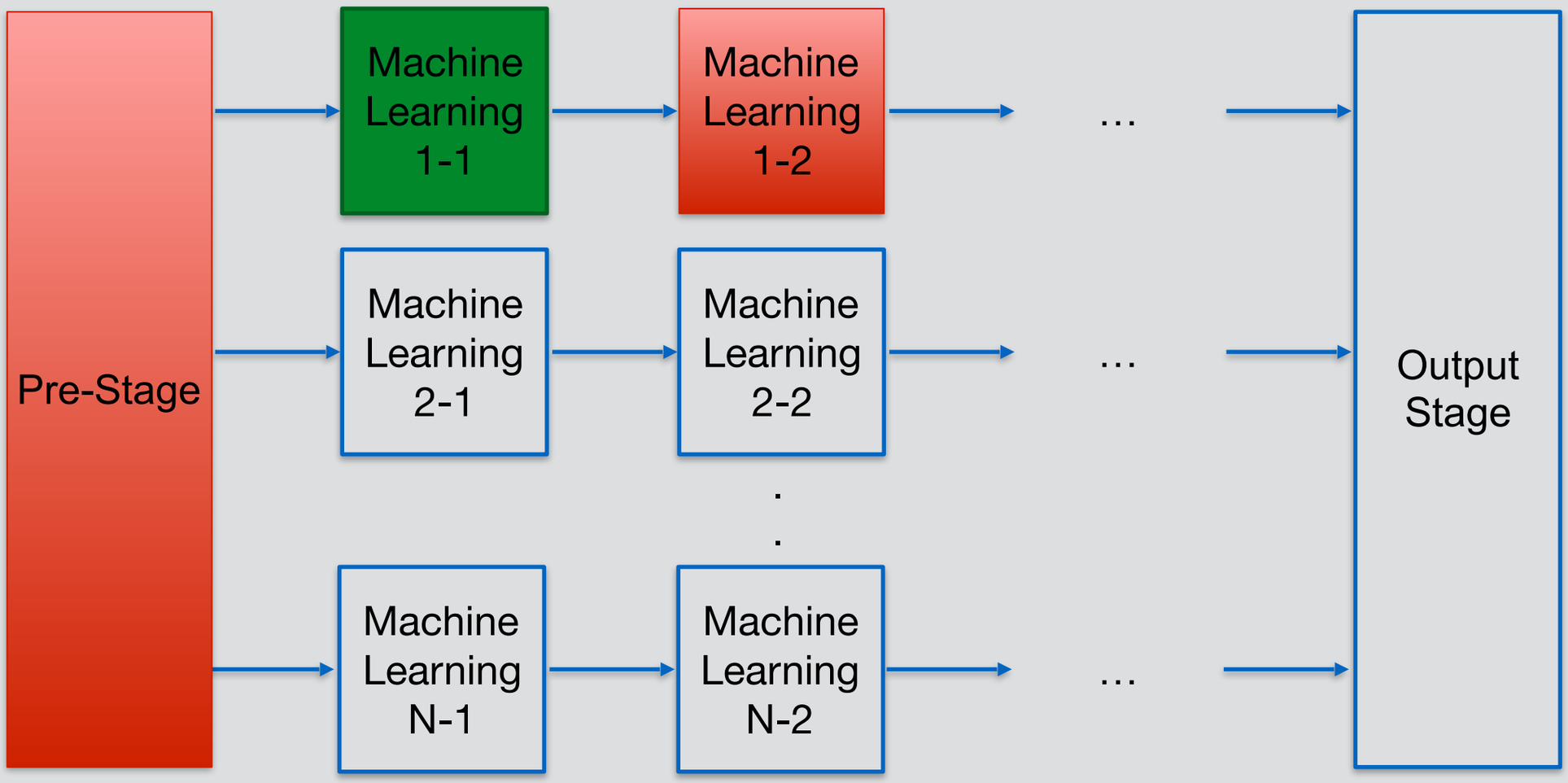
Software Project Development

2019/12/05 Group Meeting Report

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Report

- Algorithm block diagram



- We divide our group into two sub-groups, and each sub-group focuses on one stage
  - Sub-group one for pre-stage: Zexu, Marah
  - Sub-group two for stage 1-1 and stage 1-2: Jingwen, Chia-Hao

TBD

Ongoing

Finished

Schedule

Week	W03 17-21	W04 28-3	W05 4-10	W06 11-17	W07 18-24	W08 25-1	W09 2-8	W10 9-15	W11 16-22	W12 23-29	W13 30-5	W14 6-12
Log Survey												
Algorithm Survey												
Midterm Report												
Algorithm Coding												
Algorithm Test												
Prepare Report												

- 12/05 - 12/13 for summative assignment
- 12/13 - 12/22 for algorithm implementation
- 12/23 - 01/12 for algorithm test and individual report(including two posters)
- Next group meeting is on 12/16 or 12/17 ?

Durham University

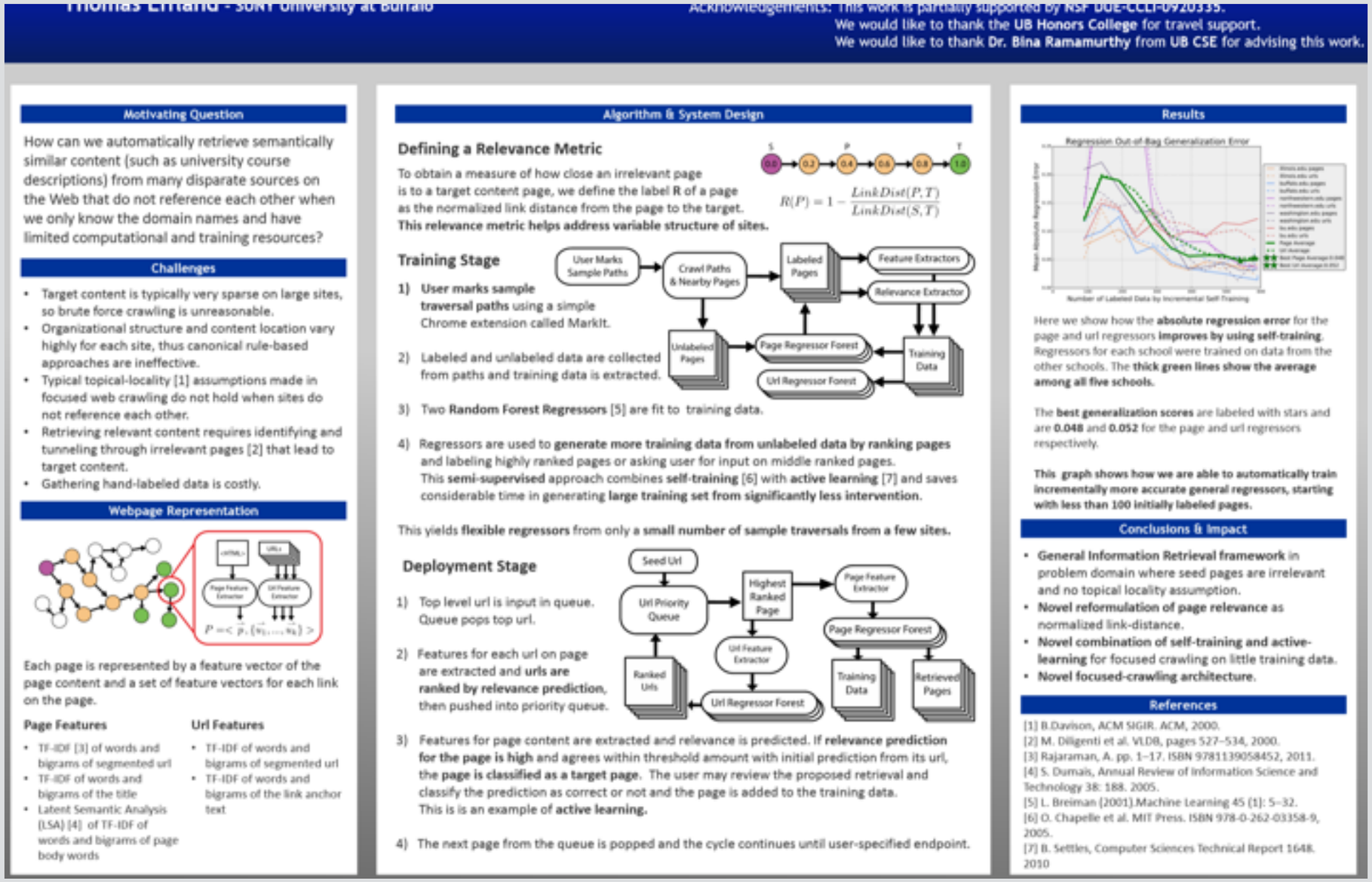
Final Report Grading

	Poster	Code
20%	design clarity	feature
20%	writing style(amount of text, amount of information)	code complexity
20%	scientific method	correctness, robustness
20%	approach, development method	ease of installation
20%	result	performance

Durham University

Poster

an example:



Thomas Lirtard - SUNY University at Buffalo

Acknowledgements: This work is partially supported by NSF DUE-CCLI-0920335.
We would like to thank the UB Honors College for travel support.
We would like to thank Dr. Bina Ramamurthy from UB CSE for advising this work.

Motivating Question

How can we automatically retrieve semantically similar content (such as university course descriptions) from many disparate sources on the Web that do not reference each other when we only know the domain names and have limited computational and training resources?

Challenges

- Target content is typically very sparse on large sites, so brute force crawling is unreasonable.
- Organizational structure and content location vary highly for each site, thus canonical rule-based approaches are ineffective.
- Typical topical-locality [1] assumptions made in focused web crawling do not hold when sites do not reference each other.
- Retrieving relevant content requires identifying and tunneling through irrelevant pages [2] that lead to target content.
- Gathering hand-labeled data is costly.

Webpage Representation

Each page is represented by a feature vector of the page content and a set of feature vectors for each link on the page.

Page Features

- TF-IDF [3] of words and bigrams of segmented url
- TF-IDF of words and bigrams of the title
- Latent Semantic Analysis (LSA) [4] of TF-IDF of words and bigrams of page body words

Url Features

- TF-IDF of words and bigrams of segmented url
- TF-IDF of words and bigrams of the link anchor text

Algorithm & System Design

Defining a Relevance Metric

To obtain a measure of how close an irrelevant page is to a target content page, we define the label R of a page as the normalized link distance from the page to the target. **This relevance metric helps address variable structure of sites.**

$$R(P) = 1 - \frac{LinkDist(P,T)}{LinkDist(S,T)}$$

Training Stage

1) **User marks sample traversal paths** using a simple Chrome extension called MarkIt.
2) Labeled and unlabeled data are collected from paths and training data is extracted.
3) Two **Random Forest Regressors** [5] are fit to training data.
4) Regressors are used to **generate more training data from unlabeled data by ranking pages** and labeling highly ranked pages or asking user for input on middle ranked pages. This **semi-supervised** approach combines **self-training** [6] with **active learning** [7] and saves considerable time in generating **large training set from significantly less intervention.**

This yields **flexible regressors** from only **a small number of sample traversals from a few sites.**

Deployment Stage

1) Top level url is input in queue. Queue pops top url.
2) Features for each url on page are extracted and **urls are ranked by relevance prediction,** then pushed into priority queue.
3) Features for page content are extracted and relevance is predicted. If **relevance prediction for the page is high** and agrees within threshold amount with initial prediction from its url, the **page is classified as a target page.** The user may review the proposed retrieval and classify the prediction as correct or not and the page is added to the training data. This is is an example of **active learning.**
4) The next page from the queue is popped and the cycle continues until user-specified endpoint.

Results

Here we show how the **absolute regression error** for the page and url regressors **improves by using self-training.** Regressors for each school were trained on data from the other schools. The **thick green lines show the average among all five schools.**

The **best generalization scores** are labeled with stars and are **0.048** and **0.052** for the page and url regressors respectively.

This graph shows how we are able to automatically train incrementally more accurate general regressors, starting with less than 100 initially labeled pages.

Conclusions & Impact

- **General Information Retrieval framework** in problem domain where seed pages are irrelevant and no topical locality assumption.
- **Novel reformulation of page relevance** as normalized link-distance.
- **Novel combination of self-training and active-learning** for focused crawling on little training data.
- **Novel focused-crawling architecture.**

References

[1] B.Davison, ACM SIGIR. ACM, 2000.
[2] M. Diligenti et al. VLDB, pages 527–534, 2000.
[3] Rajaraman, A. pp. 1–17. ISBN 9781139058452, 2011.
[4] S. Dumais, Annual Review of Information Science and Technology 38: 188. 2005.
[5] L. Breiman (2001).Machine Learning 45 (1): 5–32.
[6] O. Chapelle et al. MIT Press. ISBN 978-0-262-03358-9, 2005.
[7] B. Settles, Computer Sciences Technical Report 1648. 2010

Q:

- content: focus on algorithm / whole idea / outcome / teamwork?
- when it comes to some terms ,do we need to explain them in the poster(whether the poster should be as simple as possible or not-depend on the reader?)

Stage 1-2

- Purpose: recognize which abnormal events is important
- Input:
  - The classification result from stage 1-1
  - User define high priority keywords
- Output: list all keywords about abnormal events
- Result:
  - We can know that which keywords are highly related to abnormal events via TFIDF analysis
    - Disconnect type abnormal event is related to “authentication”
    - Fail type abnormal event is related to “update” and “session”
  - However, some meaningless words, such as bye, ac, durham, dat, com…, affect the result
  - we need to consider using stop-words to enhance the result

```
[keywords in disconnect : ['authentication', 'bus', 'bye', 'disconnect', 'disconnected', 'polkitd', 'preauth', 'received', 'time', 'user']
[keywords in error : ['ac', 'cosma', 'dur', 'error', 'network', 'resolving', 'session', 'time', 'uk', 'unreachable']
keywords in fail : ['anri', 'checking', 'dat', 'failed', 'file', 'number', 'session', 'time', 'total', 'update']
keywords in illegal : ['authentication', 'com', 'disconnecting', 'illegal', 'ip', 'time', 'times', 'undef', 'unknown', 'users']
keywords in unmatched : ['begin', 'end', 'entries', 'error', 'network', 'sshd', 'time', 'unmatched', 'unreachable', 'user']
keywords in failure : ['authentication', 'begin', 'bst', 'failures', 'oct', 'sshd', 'time', 'unknown', 'user', 'wed']
keywords in warning : ['ac', 'cosma', 'dur', 'durham', 'franz', 'id', 'oct', 'received', 'support', 'uk']
```

Durham University

Stage 1-2

- Action item:
  - Consider using stop-words
  - Add user defined high priority keywords
  - Output the result as a file

Durham University

Pre-Stage

➔ classify files into different error types using MultiLabel Text classification

Status: Implementation is ready, the output Accuracy is 83%

How to improve: More data is required, currently the data used in training the model is not enough and unbalanced.

Durham University

Pre-Stage

- Action item:
  - Using more data (balanced)
  - Output error types into different folders

Durham University

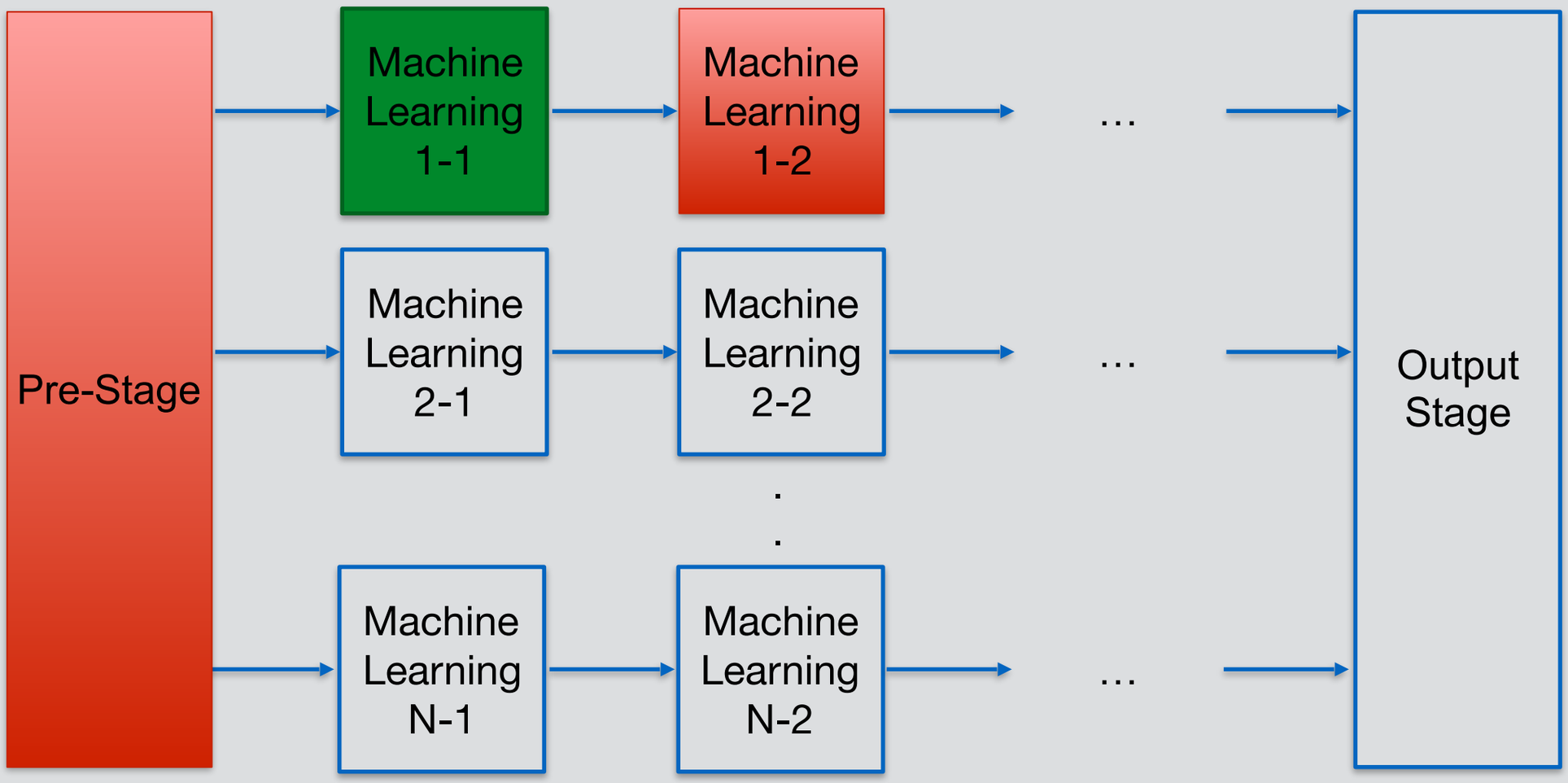
Software Project Development

2019/12/05 Group Meeting Minutes

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Minutes

- Algorithm block diagram



- We divide our group into two sub-groups, and each sub-group focuses on one stage
  - Sub-group one for pre-stage: Zexu, Marah
  - Sub-group two for stage 1-1 and stage 1-2: Jingwen, Chia-Hao

TBD

Ongoing

Finished

Durham University

Group Meeting Minutes

	Poster	Code
20%	design clarity	feature
20%	writing style(amount of text, amount of information)	code complexity
20%	scientific method	correctness, robustness
20%	approach, development method	ease of installation
20%	result	performance

- Code
  - Error handling, loop usage, good instruction and comments
  - Execute few times without exception
- Posters
  - Introduction to whole idea
  - Section: algorithm, outcome, teamwork, glossary, references
  - Do not have too much text
  - Title includes team member name, contact, date, and project name

Durham University

Group Meeting Minutes

- Stage 1-2 action item:
  - Consider using stop-words
    - Be careful if NLTK is only suitable for general English articles, not for computer logs
    - Consider modify stop-words in NLTK
  - Add user defined high priority keywords
  - Output the result as a file
- Pre-stage action item:
  - Using more data (balanced)
    - generate artificial logs
  - Output error types into different folders
- Other:
  - How to visualize the output of our algorithm
  - Next meeting: 12/16, 10:00AM, OC103

Durham University

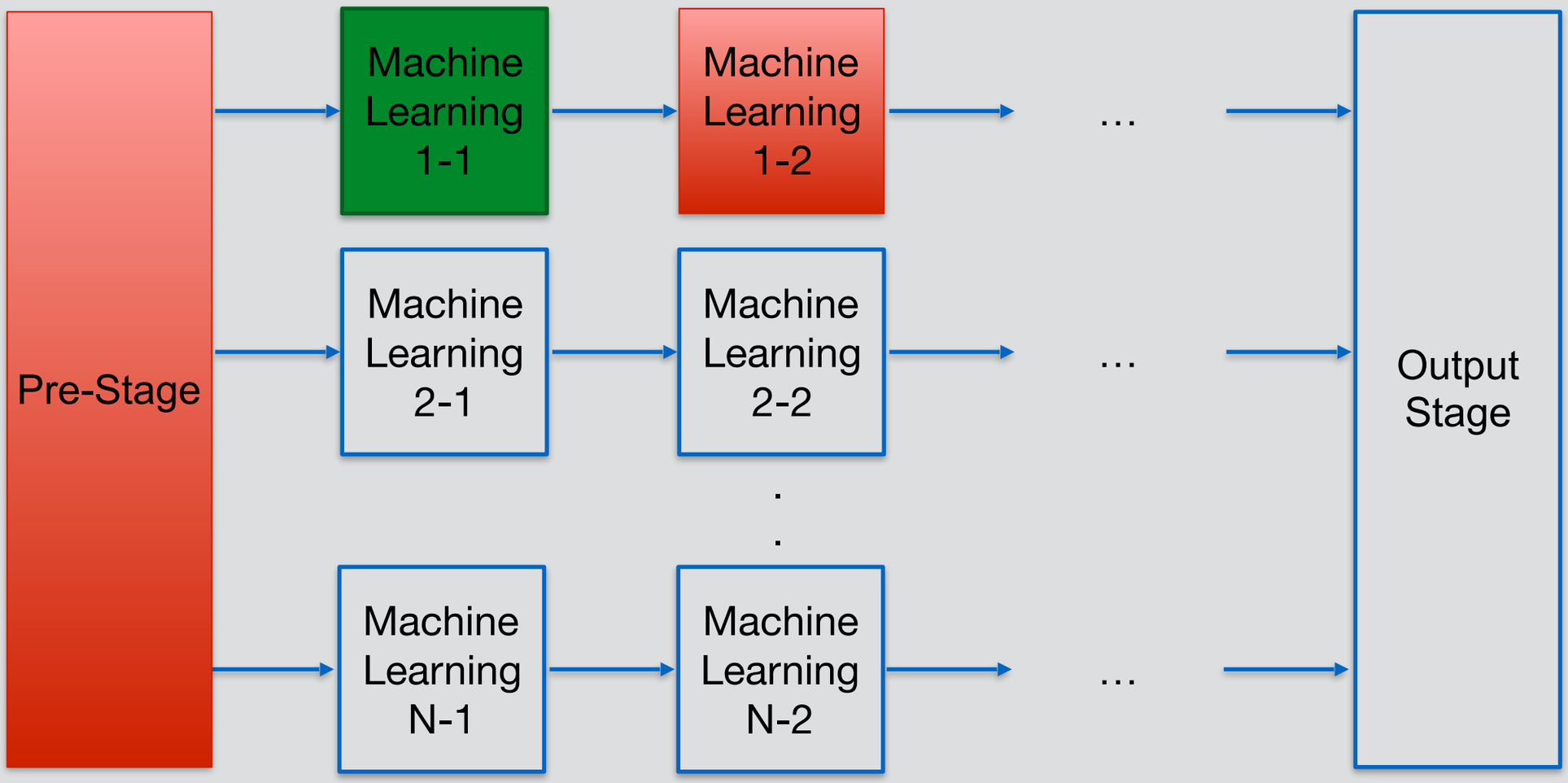
Software Project Development

2019/12/16 Group Meeting Report

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Report

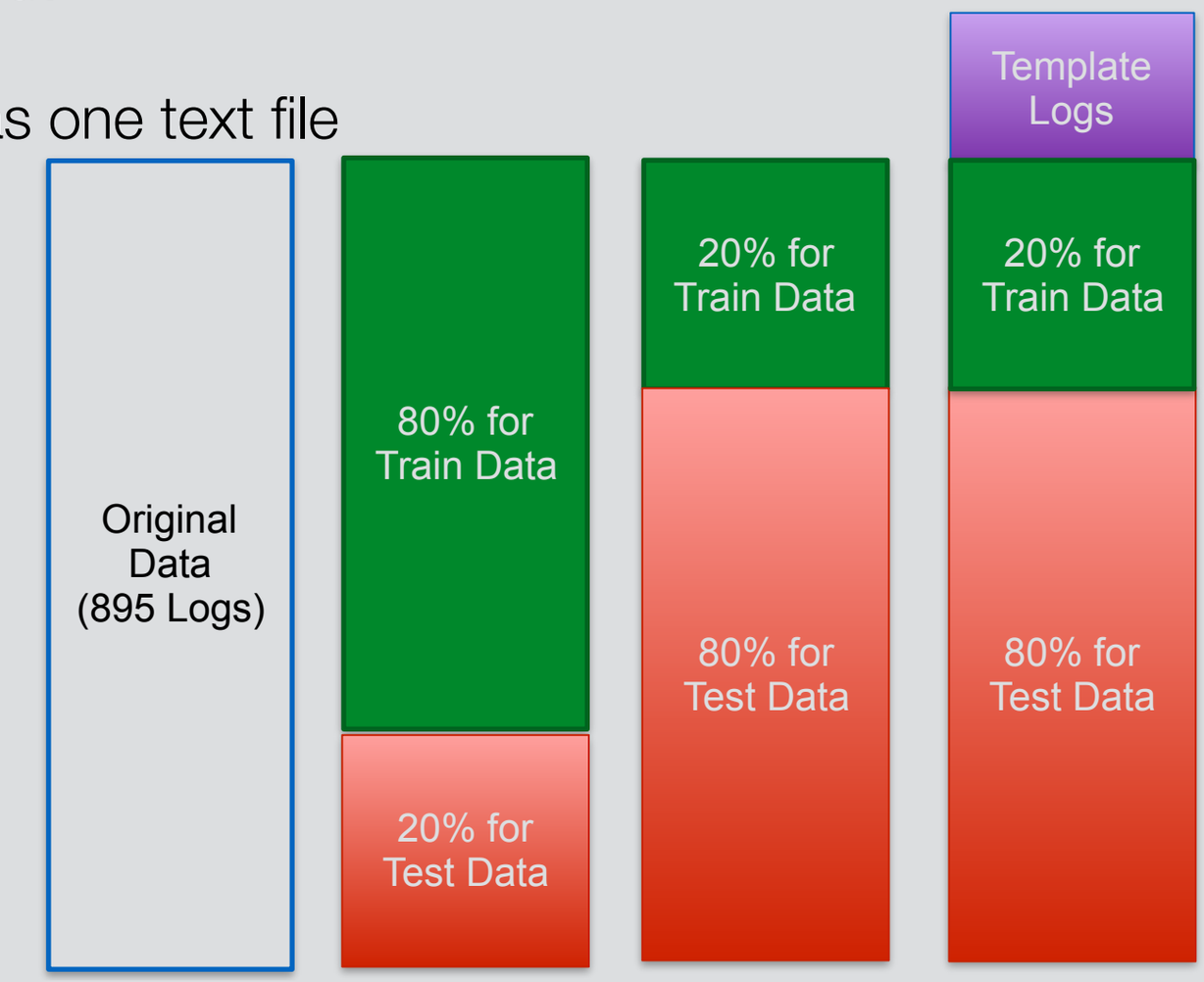
- Algorithm block diagram



- We divide our group into two sub-groups, and each sub-group focuses on one stage
  - Sub-group one for pre-stage: Zexu, Marah
  - Sub-group two for stage 1-1 and stage 1-2: Jingwen, Chia-Hao

Stage 1-1

- Purpose: recognize abnormal events in logs line by line
- Input: one category in the pre-stage output
- Output: each type of abnormal events has one text file
- Problem: training time is too long
- Solution: consider train with general logs
  - Huge logs may not have more information
  - Generate 25 basic general logs as templates
  - Use templates and 20% of original data as train data
  - Run time reduce from 2 hours to 10 minutes
  - PS: predict wrong number is the number of abnormal event which is classified to normal



Training Time	2 hours	5 minutes	10 minutes
Predict wrong Number	20	119	41

Durham University

Stage 1-2

- Purpose: recognize which abnormal events is important
- Input:
  - The classification result from stage 1-1
  - User define high priority keywords (example: log_raid_uncorrected_bad_block)
- Output: list all keywords about abnormal events
  - error is related to “log_raid_uncorrected_bad_block”, “network”, and “session"
  - fail is related to “update”

	Stop-words	User Priority List
Top	NO	NO
Middle	YES	NO
Down	YES	YES

```
keywords in disconnect : ['authentication', 'bus', 'disconnected', 'freedesktop', 'object', 'org', 'policykit', 'polkitd', 'preauth', 'time']
keywords in error : ['ac', 'cosma', 'dur', 'error', 'network', 'resolving', 'session', 'time', 'uk', 'unreachable']
keywords in fail : ['checking', 'cosma', 'dat', 'failed', 'file', 'number', 'session', 'time', 'total', 'update']
keywords in failure : ['authentication', 'bst', 'cosma', 'failure', 'failures', 'oct', 'sshd', 'time', 'user', 'wed']
keywords in warning : ['ac', 'cosma', 'dur', 'durham', 'franz', 'id', 'oct', 'received', 'support', 'uk']
keywords in illegal : ['begin', 'com', 'illegal', 'net', 'sshd', 'time', 'times', 'undef', 'unknown', 'users']
keywords in unmatched : ['begin', 'end', 'entries', 'error', 'failed', 'host', 'sshd', 'time', 'unmatched', 'user']
```

TOP

```
keywords in disconnect : ['authentication', 'bus', 'disconnected', 'en_gb', 'freedesktop', 'org', 'policykit', 'polkitd', 'preauth', 'time']
keywords in error : ['error', 'file', 'mirror', 'net', 'network', 'node', 'resolving', 'session', 'time', 'unrehable']
keywords in fail : ['checking', 'dat', 'failed', 'file', 'number', 'objects', 'session', 'time', 'total', 'update']
keywords in failure : ['authentication', 'begin', 'bkup', 'dc', 'failure', 'failures', 'incremental', 'sshd', 'time', 'user']
keywords in warning : ['esmtp', 'franz', 'hermes', 'id', 'localhost', 'majordom', 'received', 'set', 'support', 'warning']
keywords in illegal : ['begin', 'com', 'illegal', 'net', 'sshd', 'time', 'times', 'undef', 'unknown', 'users']
keywords in unmatched : ['begin', 'end', 'entries', 'error', 'failed', 'host', 'sshd', 'time', 'unmatched', 'user']
```

Middle

```
keywords in disconnect : ['authentication', 'bus', 'disconnected', 'en_gb', 'freedesktop', 'org', 'policykit', 'polkitd', 'preauth', 'time']
keywords in error : ['error', 'file', 'log_raid_uncorrected_bad_block', 'mirror', 'net', 'network', 'resolving', 'session', 'time', 'unrehable']
keywords in fail : ['checking', 'dat', 'failed', 'file', 'number', 'objects', 'session', 'time', 'total', 'update']
keywords in failure : ['authentication', 'begin', 'bkup', 'dc', 'failure', 'failures', 'incremental', 'sshd', 'time', 'user']
keywords in warning : ['esmtp', 'franz', 'hermes', 'id', 'localhost', 'majordom', 'received', 'set', 'support', 'warning']
keywords in illegal : ['begin', 'com', 'illegal', 'net', 'sshd', 'time', 'times', 'undef', 'unknown', 'users']
keywords in unmatched : ['begin', 'end', 'entries', 'error', 'failed', 'host', 'sshd', 'time', 'unmatched', 'user']
```

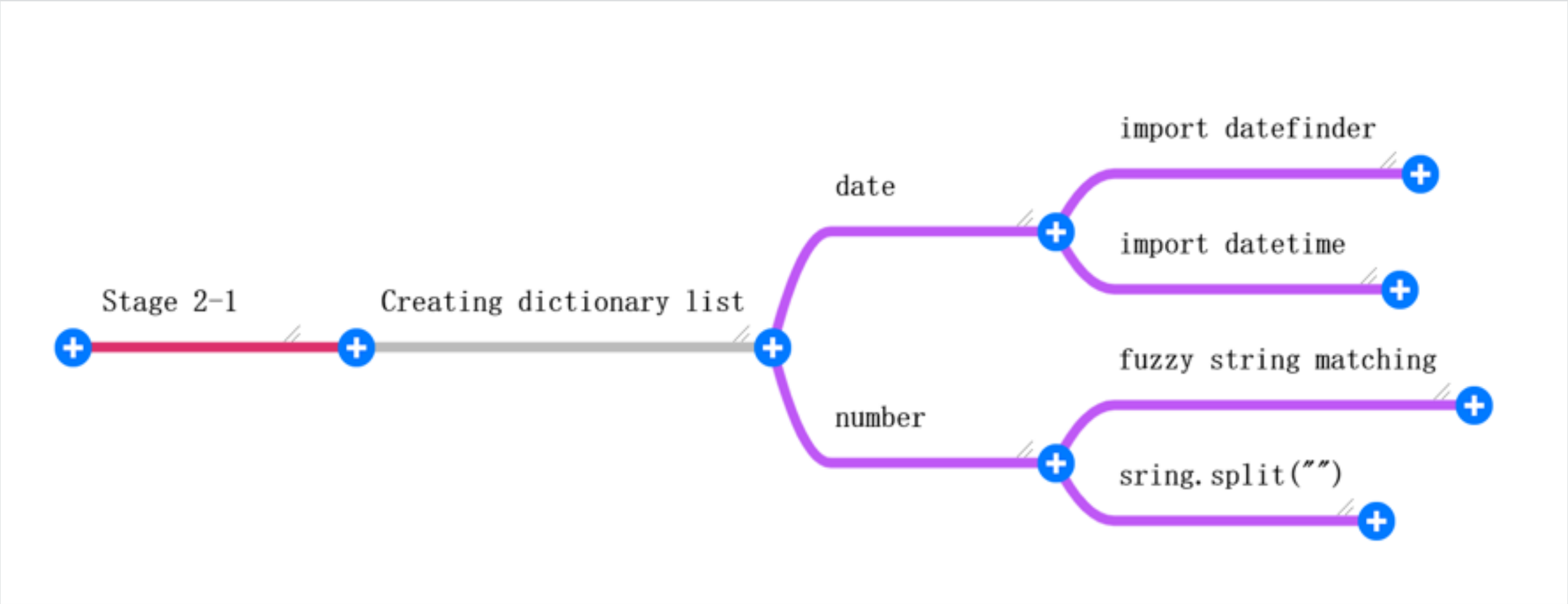
Down

Durham University

Action Item

- Stage 1-1
  - Consider to apply stop-words to check whether the incorrect number will reduce
- Stage 1-2
  - review stop-words
    - check which stop-word in NLTK is suitable for this project
    - Review logs and add stop-word for this project
- Others
  - Integrate two stages
  - Improve readability of the program result

Stage 2-1



dic_list = [{'date': '2019-10-15 03:15:12', 'number': '91%'},

{'date': '2019-10-16 06:30:22', 'number': '92%'},

……]

Durham University

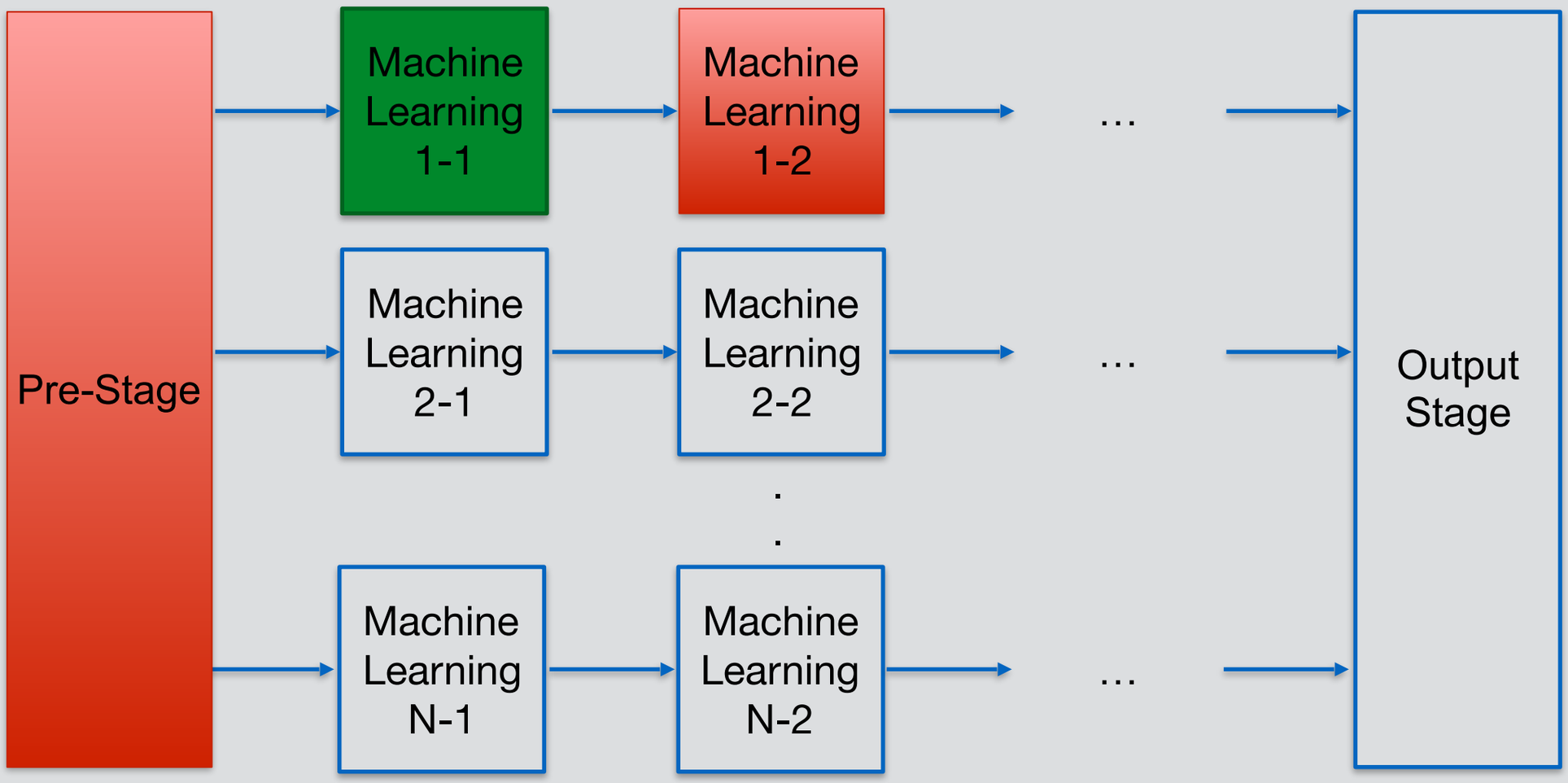
Software Project Development

2019/12/16 Group Meeting Minutes

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

Group Meeting Minutes

- Algorithm block diagram



- We divide our group into two sub-groups, and each sub-group focuses on one stage
  - Sub-group one for pre-stage: Zexu, Marah
  - Sub-group two for stage 1-1 and stage 1-2: Jingwen, Chia-Hao

Durham University

Action Item

- Stage 1-1
  - Consider to apply stop-words to check whether the incorrect number will reduce
  - Plot percentage of training data vs accuracy rate
  - Plot percentage of training data vs training time
- Stage 1-2
  - review stop-words
    - check which stop-word in NLTK is suitable for this project
    - Review logs and add stop-word for this project
  - Integrate two stages
  - Improve readability of the program result
- Stage 2-1
  - Implement the purposed algorithm
- Others
  - HPC: parallel computing system
  - Exascale computing: at least one exa flops calculation per second
  - Next meeting: TBD

Durham University

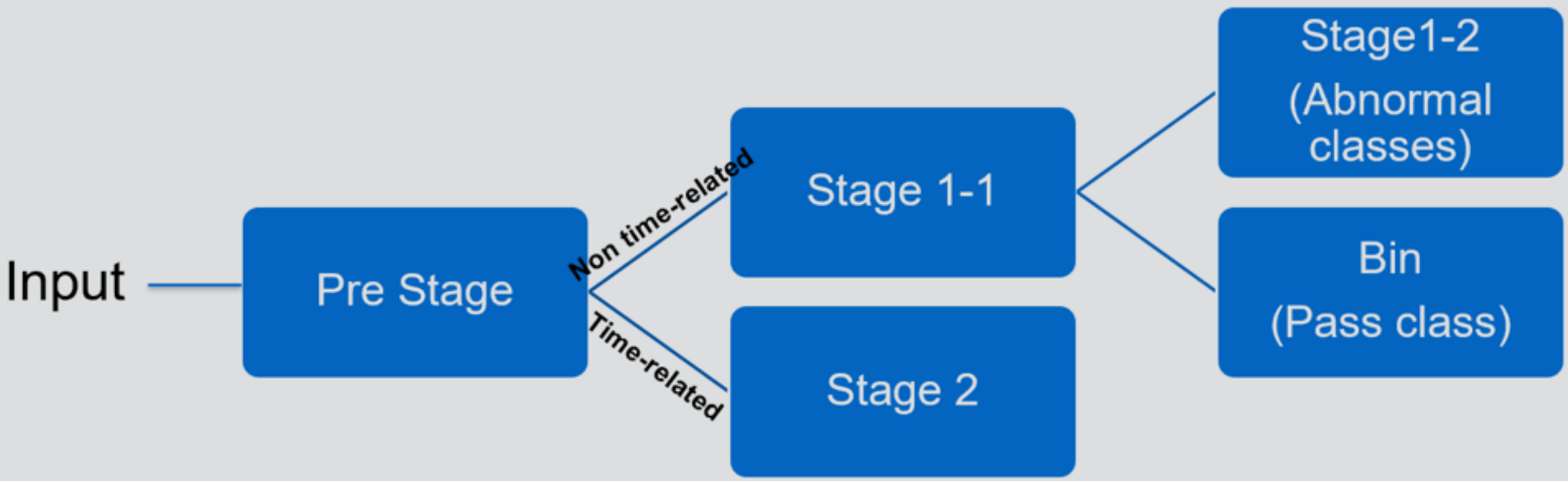
Software Project Development

2020/1/9 Group Meeting Report

Jingwen Cai
Zexu Jiang
Marah Jaber
Chia-Hao Li

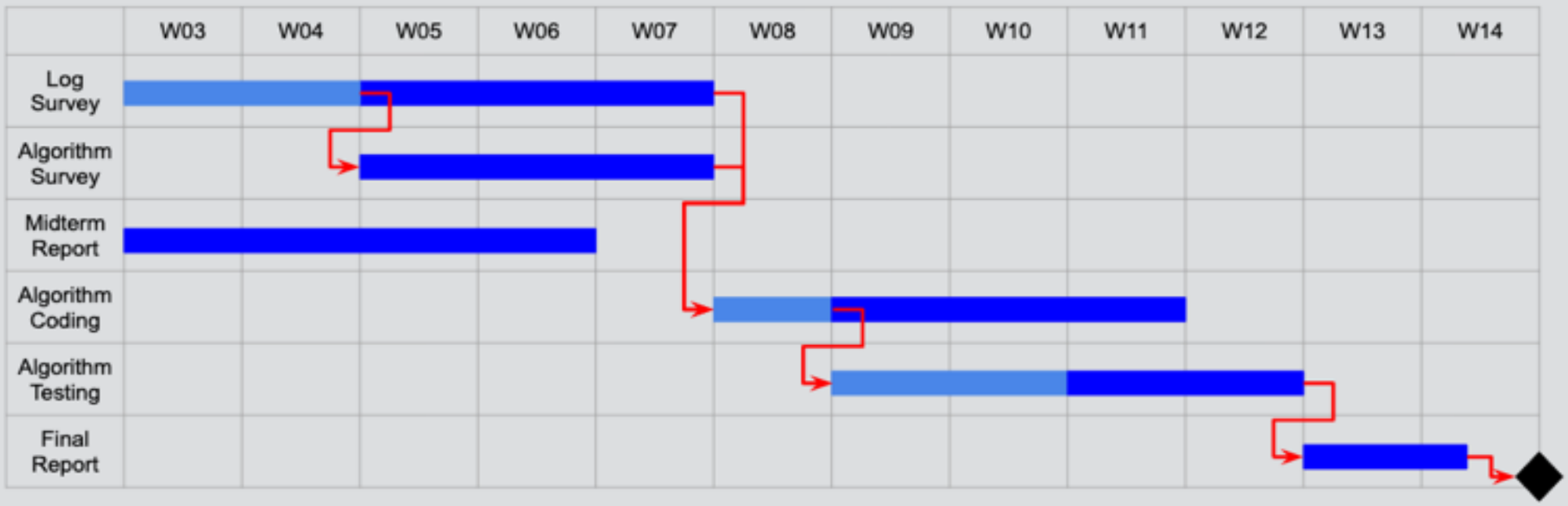
Proposed Method

- Algorithm block diagram



Proposed Method

- Project schedule

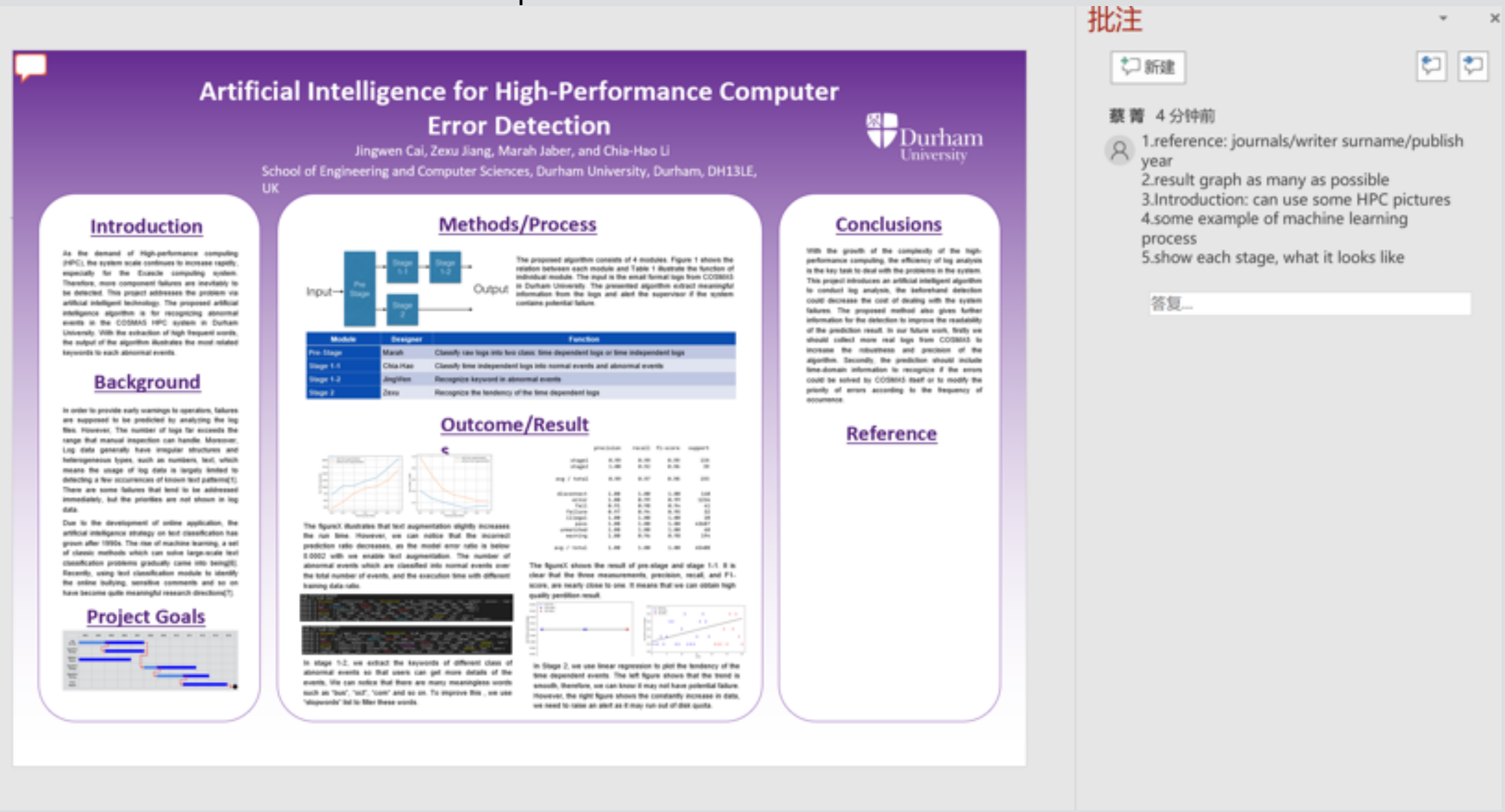


- Function and owner of each stage

	Owner	Function
Pre-Stage	Marah	Classify raw logs into two class: time dependent logs or time independent logs
Stage 1-1	Chia-Hao	Classify time independent logs into normal events and abnormal events
Stage 1-2	JingWen	Recognize keyword in abnormal events
Stage 2	Zexu	Recognize the tendency of the time dependent logs

Poster

- The current status of our poster:



- We are trying to refine the language and make the words less.

Q & A

- The content of our individual report should contain questions below?

Report

The last deliverable is an individual deliverable, i.e. each student has to submit one report. The submission is summative, i.e. you will receive an individual mark, worth 1/3 of the overall module. This deliverable will be due in the last week of Michaelmas term or the first week of Epiphany term. Refer to DUO or the course handbook for more information.

Please submit a report (PDF format, font size 11pt, maximum of 2,500 words) discussing the following items:

- Own contribution towards group success (25/100 marks)
- Reflection on the group dynamics and own contribution towards productive, healthy working environment (25/100 marks)
- Reflection on development process, i.e. has the chosen development plan been followed, have changes become necessary, and have internal deadlines been met (25/100 marks)
- Lessons learned for your personal project (25/100 marks)

Suggestion: do not put much introduction about the whole AI as draft report, pay attention on the parts above. But should illustrate the whole project idea, methods and results(maybe as an introduction).

- For the first question, should everyone describe “the motivation and result of the own stage” or else?

Suggestion: Yes, motivation of the whole project, and the section of own contribution.

Q & A

- How should we understand the second question?

Suggestion: May be around the following questions:

1. How well the group work go on?
2. Whether everyone contributed and did their work?
3. How happy are you in this group?
4. What do you think of working around them?

Be aware do not put actual names of team members in the report, use person 1,2…

- For the third question, should everyone describe “if the development of the own stage follow project schedule” or else?

Yes

- For the last question, should everyone describe “the understanding of AI technology, the comment on group organization, and the idea of project management” or else?

Suggestion: Describe how you came up with the schedule and the structure of the program.

- Should we also describe the whole idea/methodology in the report (structure)?

Same as Q1

- Do we have a poster presentation?

Waiting for advisor’s reply.

- About code:

Should have something like an instruction to tell the marker to follow it so that can test code successfully.