Group23's Meeting Record

1.Project Proposal Meeting[09/03/2019(evening) & 10/03/2019(afternoon)]

● Region choice:

◎ Discuss what we want to do, including recommendation system, Graph Classification and Sentiment Analysis(NLP).

◎ After discussion, choose Sentiment Analysis. Then we find some corresponding blogs to learn here, and the most important blog is:

https://www.cnblogs.com/pinard/category/894695.html. Moreover, find some github

● Database collection and choice:

◎ Find several data sets including some Kaggle dataset of textbook & Yelp & Penn Treebank & Reuters News dataset.

◎ Decide to use Yelp and discuss how to solve the form by python.

● Algorithms

◎ Firstly we want to use RNN or CNN to train our model, then we want to use tf-idf Or Word2vec + some machine learning algorithms.

◎ Find some corresponding libraries

● Job Division

Everyone: Data Collection and Proposal Written; Find some more details of Sentiment Analysis； Try to find more suitable thoughts or algorithms of our project.

◎Zheng Binfan: Whole document consolidation; Web of github;

◎Xie Diangu: Github Website Collection; Understand the sample code of Spark MLlib.

◎Gao Xiaoxin: Github Website Collection; Collect more knowledge about text processing.

● [Additional Part] Proposal Modification

◎The proposal submitted last night is too rough, modify it based on Professor.Wang's suggestion.

2.Final Machine Construction[18/03/2019]

◎ Construct Hadoop and Spark in the fourth machine 71.

◎ Finish the Group connection part of our project.

PS: We just have three numbers after one of our classmates failed to be chosen in this course, so we should construct the fourth machine firstly.

3. Project Kickoff Meeting[25/03/2019]

- **Download Yelp dataset & Use WinSCP to submit to the machine**
- **Separating dataset randomly to different size to do "different size test" in the future.**
- **Data Extraction**
- **Try out simple code based on small dataset just use single machine**
- **Job Division**

Everyone: Do all these part together and debug together.

4. Check-in Meeting[12/04/2019(Evening) 13/04/2019(Afternoon) 14/04/2019(Afternoon)]

- **Deal with problem that only 7 active nodes work while running demo.**
- **Put data into HDFS and convert dataset to correct format.**
- **Modify source codes and successfully run a simple model in Spark**
- **Job Division**

◎Zheng Binfan: communicate with TA to solve problems & Modify the codes

◎Xie Diangu: Modify source codes and test demo

◎Gao Peixing: Run the codes and google the solution of problems and code bugs

5.Check-in Meeting[17/04/2019 & 18/04/2019]

- **Discuss what needs to be shown in the presentation**
- **Prepare the PowerPoint slides for the presentation**
- **Using different algorithms to modify the code(based on reviews.csv) and compare the results & Use the best to train the tips.csv.**
- **Different data size test**
- **Job Division**

◎Zheng Binfan: make slides, modify code, different data size test ◎Xie Diangu: make slides, modify code, configuration design ◎Gao Peixin: make slides, try to apply different algorithms, different data size test

6.Project presentation meeting[19/04/2019& 20/04/2019]

● Demo Record

◎ Make the project demo to make sure everything works well, including tf-idf processing, algorithm training, predicting and computing accuracy.

◎ Check the output.

● Slider Check & Website

◎ Edit the order and the content of the presentation slides. Assign each member’s task of presentation.

◎ Rehearse presentation

◎ Finish the website.

● Job Division

Everyone: Slider Check & Code Check

◎Zheng Binfan & Xiedian: Demo record and checking.

◎Gao Peixin: Website Design & Code Arrangement