Project Report

Luojie Xiang, Junchao Yan

1 Introduction

In this project, we will extract the topics from a corpus collected from Stack Overflow using Latent Dirichlet Allocation (LDA) on a hadoop cluster. LDA is a statistical model for discovering underlying topics from a collection of documents [1].

2 Experiment

2.1 Dataset

The dataset of this project is obtained from Kaggle (www.kaggle.com), which is a platform for data analysis and prediction competitions. The data that we use are posted by Facebook for a keyword extraction competition. The dataset consists the files both for training and testing, of which the training file contains four columns id, title, body, and tags. In this project, only the title and body are used to extract the topics.

Example:

```
id: 1

title: How to check if an uploaded file is an image without mime type?

content:

<p>I'd like to check if an uploaded file is an image file (e.g png, jpg,
jpeg, gif, bmp) or another file. The problem is that I'm using Uploadify
to upload the files, which changes the mime type and gives a 'text/octal'
or something as the mime type, no matter which file type you upload.</p>

<p>Is there a way to check if the uploaded file is an image apart from
 checking the file extension using PHP?</p>

tags: php image-processing file-upload upload mime-types
```

2.2 Data Cleaning

Describes the steps we take to do data cleaning.

remove content within `<code>`
remove tags `"<[^>]*>"`
Remove punctuation `"[!@#$%^&*()-\_:;"'?/.,<>]"`
Remove new lines `"\n" to ""`
Remove multiple white spaces `"\s+" to " "`
Remove words shorter than length 3
Turn all letters to lowercase

2.3 Hadoop Setup

Describes how we set up our 6 node hadoop cluster.

2.4 Latent Dirichlet Allocation

2.5 Results

3 Conclusion

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.