Bare Demo of IEEEtran.cls for Conferences

Luojie Xiang
Department of Computer Science
Purdue University
West Lafayette, Indiana, USA
Email: xiang7@purdue.edu

Junchao Yan
Department of Computer and Information Technology
Purdue University
West Lafayette, Indiana, USA
Email: yan114@purdue.edu

***Abstract*—The abstract goes here.**

I. 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].

II. Experiment

A. 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**:
> 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.
> Is there a way to check if the uploaded file is an image apart from checking the file extension using PHP?
> **tags**: php image-processing file-upload upload mime-types

B. Hadoop Setup

The hadoop cluster contains six nodes including one masternode and five slavenodes. To set up the hadoop cluster, we first configured the hosts file as shown below.

```
192.168.65.70 masternode
192.168.65.71 slavenode1
192.168.65.72 slavenode2
192.168.65.75 slavenode3
192.168.65.76 slavenode4
192.168.65.77 slavenode5
```

To configure the hadoop accordingly, we updated the `conf/masters` and `conf/slaves` files on the master node as shown below.

`conf/masters` on the master node:

```
masternode
```

`conf/slaves` one the master node:

```
slavenode1
slavenode2
slavenode3
slavenode4
slavenode5
```

In addition, configuration files `conf/core-site.xml`, `conf/mapred-site.xml`, and `conf/hdfs-site.xml` were modified on all the nodes as shown below.

`conf/core-site.xml` on all the nodes:

```
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://masternode:9000</value>
<description>Enter your NameNode hostname
</description>
</property>
<property>
<name>fs.checkpoint.dir</name>
<value>/home/student/DAT500/fs/hdfs/snn
</value>
<description>A comma separated list of paths.
 Use the list of directories</description>
</property>
<property>
<name>hadoop.tmp.dir</name>
<value>/home/student/DAT500/fs/tmp</value>
<description>Comma separated list of paths
</description>
</property>
</configuration>
```

`conf/mapred-site.xml` on all the nodes:

```
<configuration>
<property>
<name>mapred.job.tracker</name>
<value>masternode:9001</value>
<description>Enter your JobTracker hostname
</description>
</property>
<property>
<name>mapred.local.dir</name>
<value>/home/student/DAT500/fs/tmp/mapred/
local</value>
<description>Comma separated list of paths
</description>
</property>
</configuration>
```

`conf/hdfs-site.xml` on all the nodes:

```
<configuration>
<property>
<name>dfs.name.dir</name>
<value>/home/student/DAT500/fs/hdfs/nn
</value>
<description>Comma separated list of paths
</description>
</property>
<property>
<name>dfs.data.dir</name>
<value>/home/student/DAT500/fs/hdfs/dn
</value>
<description>Comma separated list of paths
</description>
</property>
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
</configuration>
```

C. *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

D. *Latent Dirichlet Allocation*

E. *Results*

III. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

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.