##################################################################Create by: ***Davidyu***

***Last update : 2019/06/17***

***Version: V0***

##################################################################

# Version setup

## For download

In this version all the data download and save to the **database are the whole data, which means that we download all the dataset and truncate the table and store all the data,** need to update to the incremental update.

# Script rules

-- BASH files

## Configure file in

davidyu\_cfg.py

-save it to the python library

-create project in data\_path, analysis\_path, download\_path

## Run python file

For loop, make a python module

And run it in shell

*/home/davidyu/stock/scripts/davidyu\_stock/scripts/download/day\_history/v1*

**for i in $(seq 0 $stk\_index\_count)**

**do**

**echo $i**

**sed -r 's/TheInput/'$(echo $i)'/g' download\_data\_insert.py > to\_download.py**

**echo "python todownload"^[[>0;136;0c**

**`python to\_download.py >> $Thelog & echo $! > pidfile.txt`**

**sleep 5s**

**done**

**echo "finished" >> $Thelog**

**./run.sh > run.log 2>&1 &**

## Run log name –> to Create Running Log

Script\_name + Date + extention (py/r/scala…) + “.log

1. Run program and save the log to the current dir
2. Move the log file to the log dir

$script\_raw\_name"\_"$DATE"\_scala.log"

## Source code in shell

source ~/.bashrc

cd `dirname $0`

curr\_dir=`pwd`

programName=${0##\*/} ### filename without type e.g. test.sh > test

source $shell\_function\_dir"create\_log.sh"

## Run with log and PID

nohup ./run.sh > run.log & echo $! > pidfile.log

## Hive & Spark setup

Set mapreduce.reduce.memory.mb

Set mapred.reduce.tasks

Set io.sort.spill.percent

Set mapred.job.shuffle.input.buffer.percent

Set mapred.job.shuffle.merge.percent

Set mapred.compress.map.output

Set hive.input.format=org.apache.hadoop.hive.ql.io.CombineHiveInputFormat

Set hive.exec.dynamic.partition=true

Set hive.exec.dynamic.partition.mode=nonstrict

Set mapreduce.input.fileinputformat.split.maxsize=256000000

Set mapreduce.input.fileinputformat.split.minsize.per.node

Set mapreduce.input.fileinputformat.split.minsize.per.rack

# Script framework

## Download

### Source functions

***DIR : scripts/functions***

***-ForDownload.py***

#### generate\_stock\_index.py

'''

input : i stock index raw input '601398'/ '917'

output : 601398.ss

000917.sz

'''

#### stk\_index\_list\_gen.py

'''

functions : get the stock index list without "3XXXXX" chuangyeban

output: list

list of all stock index

'''

#### CxExtractor.py

Extract contents from a html

#### combine\_csv\_in\_folder.py

#### to\_hive.scala

save the dataframe to hive

### day\_history

1. read stk index list
2. loop to download to data dir i.e. 601398.csv
3. get all csv to “all.csv”
4. put into hive

## Analysis

### Historical day data

#### Linear regression

/scripts/functions

**Class LinearReg**

**-def single\_linear\_reg.py one column get the slope**

#### get\_technical\_indicators

analysis/day\_history/test\_AI

--MACD

--Bollinger Bands

### All news

#### Content summarize 🡪 TextRank

# Important Scripts -- for system

## Start jupyter

# 如果你的系统是装在服务器上，没有图形界面，需要远程连接的则可以加上如下参数

jupyter notebook --ip 0.0.0.0 --no-browser

## Check if chinese is in the string

zhmodel = re.compile(u'[\u4e00-\u9fa5]')

## Python

### @Property 可读写属性 / 只读属性

### 未完成--多重继承，装饰器

### CNTK

Keras 🡪 base source code ，need to check

<https://github.com/microsoft/CNTK>

cntk notebooks.azure.com

### Some point

Np.where(condition,x,y) if satisfied the condition, return x else y

### Bp network

<https://blog.csdn.net/weixin_40432828/article/details/82192709>

#### matrix transpose

-need no check numpy.transpose

# Outside data

## -新华字典全量数据

<https://github.com/pwxcoo/chinese-xinhua.git>

DIR '/home/davidyu/gits/chinese-xinhua/data'

## -中国大陆 31 个省份 1978 年至 2019 年一千多万工商企业注册信息，包含企业名称、注册地址、统一社会信用代码、地区、注册日期、经营范围、法人代表、注册资金、企业类型。

-<https://github.com/imhuster/Enterprise-Registration-Data-of-Chinese-Mainland/blob/master/Enterprise-Registration-Data/csv/README_CSV.md>

- /home/davidyu/stock/outside\_data/Enterprise-Registration-Data-of-Chinese-Mainland/csv

# Important Data

## 中国外汇&黄金储备

http://www.safe.gov.cn/safe/whcb/index.html

# Gits

## 情感分析

https://github.com/SeanLee97/xmnlp.git

## the algorithms

<https://github.com/TheAlgorithms>

## 面试笔记

https://github.com/imhuay/Algorithm\_Interview\_Notes-Chinese.git

# News

安邦

## 氢能

2019/06/15 日本燃料电池车和家用燃料电池发展至今，一再燃料电池和氢能技术开发灵与走在了始接前列，已卡法了日本市场应用最广泛，世界范围内技术最领先的产品。日本是一个能源极度匮乏的国家不得不重视氢能源的开发，以实现真正的能源安全和能源独立。日本政府在 能源基本计划 中将氢能源定位位与电力和热能并列的核心二次能源，病提出建设 氢能社会 的远景，希望通过氢燃料电池实现氢能在家庭 工业 交通 等灵与的应用，构建氢能产业的应用体系和消费体系。

## --2019年5月 工业产出

2019/06/18 今年5月，中国的工业产出增长降至逾17年的最低水平，一位这始接第二大经济体正在奋力应对经济下行压力。国家统计局6月15日发布的数据表明，今年5月份，中国规模以上工业增加值同比增长5%，这是自2002年2月以来的最低水平，低于前一个月的5.4%。5月的数据还显著低于路透社钓场的经济学家所预测的5.5%。

## 房地产

2019/06/19 央企大量撤出房地产行业显示资金逃离的迹象。近期，房地产行业产权转让项目明显增多，而且转让方多为氧气。根据北京产权交易所公开信息梳理统计，5月以来，共有20个房地产项目进行股权转让，而今年前4个月只有2个，去年4月至12月，9个月的时间里总共仅有10个

# PA

Query-> 去特殊字符 – 纠错- 分词- 代词消解- 词性标注- 实体识别-主谓宾提取 – 同义词替换 –专名归一化 – 预处理结果

# Papers

## LSTM

### long short term memory hochreiter 1997

## AutoML

## **NLP**

### ELMo

<https://blog.csdn.net/triplemeng/article/details/82380202>

# Book

## The links

/home/davidyu/gits/book

# People

## 黄学东 微软-语音识别

## 刘知远 清华-NLP