naoya

“みんかぶ”のコメントデータ取得プログラム仕様書

目次

[概要 3](#_Toc479954408)

[Pythonファイル 4](#_Toc479954409)

[出力フォルダとファイル 5](#_Toc479954410)

[使い方 6](#_Toc479954411)

[処理概要 7](#_Toc479954412)

[記録用CSVの項目内容 9](#_Toc479954413)

[出力CSV（コメントデータ）の項目内容 11](#_Toc479954414)

# 概要

”みんかぶ”サイトにて、日経225の株価に対する予想コメントをクローリング&スクレイピングしている。

各企業で対象としているウェブページのサンプルは

コメント一覧のページ

<http://minkabu.jp/stock/1332/pick>

コメント詳細のページ

<http://minkabu.jp/stock/1332/pick/10000000001982843>

の二種類



イメージ図

# Pythonファイル

・C\_update.py

　クローリング用のPythonファイル

・S\_update.py

　スクレイピング用のPythonファイル

・nikkei225-stock-prices.txt

　クローリングで対象とする企業がのっているテキストファイル

　このファイルを参照してクローリングを行っているので必須。

# 出力フォルダとファイル

・Output

　クローリングした結果のhtmlファイルを記録するフォルダ

　このディレクトリ下に各企業の名前でフォルダが作成され、

　そのディレクトリ下にC\_update.py実行した日付でフォルダが作成される

・Output/“企業名”/”ファイル実行日時”/”~.html”

　各企業の一覧ページのクローリング結果

・Output/“企業名”/”ファイル実行日時”/comment/”~.html”

　各企業の詳細ページのクローリング結果

・result

　このディレクトリ下に各企業の名前でフォルダが作成される

　そのディレクトリ下にS\_update.py実行した日付でフォルダが作成される

・result/クローリング更新情報.csv

クローリングした結果について、コメントの日付を記録するファイル

・result/スクレイピング更新情報.csv

　スクレイピングした結果について、コメントの日付を記録するファイル

・result/クローリング更新情報\_前回.csv

・result/スクレイピング更新情報\_前回.csv

　一回前に実行したときの結果を記録するファイル

# 使い方

C\_update.pyを実行

更新がある企業についてのみクローリングを行う

S\_update.pyを実行

前回のスクレイピング以降、一度でもクローリングが行われた場合、スクレイピングを行う

結果はresult/”各企業の名前”/“各企業の名前”.csv

に、各企業のすべてのコメントが記録されている。

もし、何らかの理由で記録を取り直したい場合は

Outputフォルダとresultフォルダを全て削除する

# 処理概要

(初回実行時)

クローリング

C\_update.pyを実行

Outputディレクトリとresultディレクトリが作成される

Outputディレクトリに、クローリングした結果が

一覧ページ

/Output/"企業名"/"ファイルを実行した日付"/番号.html

詳細ページ

/Output/"企業名"/"ファイルを実行した日付"/comment/番号.html

で記録される。

また、各企業のクローリングの結果について

/result/クローリング更新情報.csv

に記録される。

スクレイピング

S\_update.pyを実行

各企業でクローリングしたファイルの詳細ページを参照し、該当項目をスクレイピングする

resultディレクトリにスクレイピングした結果が

/result/"企業名"/"ファイルを実行した日付".csv

/result/"企業名"/”企業名”.csv

の、二つの名前で記録される。

また、各企業のスクレイピングの結果について

/result/スクレイピング更新情報.csv

に記録される。

（更新時）

クローリング

C\_update.pyを実行

/result/クローリング更新情報.csv

の、”クローリン更新コメント時間\_最新”項目を参照し、各企業のコメントが更新されているか確認する。

それ以降の処理は初回実行時と同様

ただし、”クローリン更新コメント時間\_最新”の日付よりも前の日付を取得したら処理を終了する。

/result/クローリング更新情報.csv

も更新。前回の内容は

/result/クローリング更新情報\_前回.csv

に、保持しておく。

スクレイピング

S\_update.pyを実行

/result/スクレイピング更新情報.csv

の、”前回取得した最新フォルダ”よりも新しいフォルダからのhtmlのみスクレイピングを行う。

後の処理は初回実行時と同様。

ただし、

/result/スクレイピング更新情報.csv

の“スクレイピング更新コメント時間\_最新”項目を参照し、それよりも後のコメントのみ記録する。

/result/スクレイピング更新情報.csv

も更新。前回の内容は

/result/スクレイピング更新情報\_前回.csv

に、保持しておく。

# 記録用CSVの項目内容

クローリング更新情報.csv

|  |  |  |
| --- | --- | --- |
| 項目名 | データ例 | 内容 |
| 企業名 | 日本水産 | 企業名 |
| クローリン更新コメント時間\_最新 | 2016/04/01 0:00 | 今回クローリングしたページの中で、最も新しい投稿の時間。  投稿の時間は”みんかぶ”にコメントされた時間。 |
| クローリン更新コメント時間\_最古 | 2016/04/01 0:00 | 今回クローリングしたページの中で、最も古い投稿の時間。 |
| クローリング参照コメント時間 | 2016/03/25 0:00 | 前回のクローリングでの  クローリン更新コメント時間\_最新 |
| クローリング更新結果 | True | 今回のクローリング結果が問題ないかどうか  初回実行時はFirst  更新成功時はTrue  失敗時はFalse  更新内容がないときは  Not update  が記録される |

スクレイピング更新情報.csv

|  |  |  |
| --- | --- | --- |
| 項目名 | データ例 | 内容 |
| 企業名 | 日本水産 | 企業名 |
| スクレイピング更新コメント時間\_最新 | 2016/04/01 0:00 | 今回スクレイピングしたクローリングのhtmlの中で、最も新しい投稿の時間。  投稿の時間は”みんかぶ”にコメントされた時間。 |
| クローリン更新コメント時間\_最古 | 2016/04/01 0:00 | 今回スクレイピングしたページの中で、最も古い投稿の時間。 |
| スクレイピング参照コメント時間 | 2016/03/25 0:00 | 前回のスクレイピングでの  スクレイピング更新コメント時間\_最新 |
| 前回取得した最新フォルダ | 2016-03-31-13-28-38 | 前回スクレイピングを行った際に参照したクローリングフォルダのうち最新のもの |
| スクレイピング更新結果 | True | 今回のスクレイピング結果が問題ないかどうか  初回実行時はFirst  更新成功時はTrue  失敗時はFalse  更新内容がないときは  Not update  が記録される  Falseが記録されている場合は確認をしたほうがいい |

# 出力CSV（コメントデータ）の項目内容

|  |  |  |
| --- | --- | --- |
| 項目名 | データ例 | 内容 |
| 概要 | "ソフトバンクグループ(9984)の売り予想。  徳川慶喜 さんの売買予想。 目標株価： 7,196円 期間： 中期(数週間~数ヶ月) 理由： チャート 割高" | コメントページの概要  万が一、この内容とデータ内容が違う場合は要確認 |
| 企業名 | ソフトバンク | 企業名 |
| 予想日 | 2017年04月7日 | 予想が投稿された日付 |
| 予想時刻 | 23:23 | 予想が投稿された時刻 |
| 売買 | 売り | 売りの予想か買いの予想か |
| ユーザ名 | 徳川慶喜さん | “みんかぶ”内でのユーザ名 |
| 所持ポイント | -2.73 | ユーザが所有しているポイント  詳細は”みんかぶ”ページを参照 |
| 登録時株価 | "8,062.0" | 予想が登録された時点での株価 |
| 取得ポイント | +1.23 | 今回の予想で取得したポイント |
| 目標株価 | "7,196" | 今回の予想での目標株価 |
| 予想期間 | 中期(数週間～数ヶ月) | 予想期間 |
| 理由 | チャート | 理由　選択 |
| コメント | 米国発の大きなNEWSがある事に期待して、、、います!! | 各予想に対する予想している人間のコメント  ただし、空白の場合がある |
| データ取得時株価 | "8,008.0" | クローリングが行われた時点での株価 |
| データ取得日 | 04月10日 | クローリングを行った日付 |
| データ取得時間 | 15:27 | クローリングを行った時間 |
| 参照html |  | このコメントが記録されているhtmlファイルのパス |