公立はこだて未来大学 2016 年度 システム情報科学実習
グループ報告書

Future University Hakodate 2016 System Information Science Practice

Group Report

プロジェクト名

AI するディープラーニング

Project Name

AI Love Deep Learning

グループ名

AI スコープ

Group Name

AI Scope

プロジェクト番号/Project No.

14-A

プロジェクトリーダ/Project Leader

1014041　福田大知　Daichi Fukuda

グループリーダ/Group Leader

1014017　野尻雅音　Masane Nojiri

グループメンバ/Group Member

1014017　野尻雅音　Masane Nojiri
1014041　福田大知　Daichi Fukuda
1014049　齋藤直紀　Naoki Saito
1014116　鈴木才都　Saito Suzuki
1014172　板垣隼基　Junki Itagaki

指導教員

竹之内 高志　永野 清仁　寺沢 憲吾　片桐 恭弘

Advisor

Takashi Takenouchi　Kiyohito Nagano　Kengo Terasawa　Yasuhiro Katagiri

提出日

2017 年 01 月 18 日

Date of Submission

January 18, 2017

概要

近年, 様々な例で人間を模倣できる人工知能が台頭してきている. 人工知能は機械学習などを用いて, 人間が自然に行っている学習能力と同様の知能をコンピュータで実現しようとする技術・手法のことである. 特に, ディープラーニング (深層学習) は画像処理分野で優秀な成果を出している手法である.

本プロジェクトの目標はこれらの機械学習手法を用いて, 人間の思考を模倣・超越することである. そこでメンバー間で目標についてディスカッションをし, プロジェクトをグループ A, グループ B にわけることとした.

ディスカッションの結果, グループ A では配球予想システムの開発, グループ B では人間が操作するよりも, 速く走行できるカーエージェントの開発を目標とした.

野球の配球予想をテーマに選んだ理由として, 現在日本国内での野球人気の凋落が地上波放送の視聴率や放送自体の少なさ等から読み取れること, 2010 年以降の観客動員数も若干ではあるが落ちつつあるということ [2], 野球以外のスポーツ人気の向上 [3] から, 野球に関連するコンテンツで野球に興味をもってくれる人を増やそうと考えたことが挙げられる.

そこで本グループでは, 元プロ野球選手の野村克也氏が, 地上波放送で野球解説を行う際に不定期で表示されていた「野村スコープ」というコンテンツに着目し, それを模倣したコンテンツを機械学習を用いて実装することを考案した. 具体的には各球団の選手データを収集し, それをもとに機械学習を用いて, 「野村スコープ」のような配球予想コンテンツの制作に取り組んだ. また, 制作過程でグループ分けを行い,

- データ収集班 Python を用いてウェブサイトから選手データを抽出するプログラムの作成
- 実装班 Python を用いた機械学習プログラムの作成

の 2 グループに分かれ作業を行った.

前期は各グループでの作業となった. データ収集班はデータ収集プログラムの完成, 実装班はサポートベクターマシーン (SVM) を用いた簡単な予想プログラムの作成を行った. また中間発表後, 後期へ向けて実装方法の見直しや使用データ, 手法の吟味などを行った.

後期では, 各グループでの作業に加え, 実際に集めたデータを使用しながらの作業となった. データ収集班では収集するデータの種類の増加, また指定したデータを取り出すことができるような汎用性の向上を行った. 実装班ではディープラーニングを用いた配球予想プログラムの作成にとりくみ, 実装を進めた. また, 最終発表へ向けた発表資料の作成も並行して行った.

キーワード　人工知能, 機械学習, 深層学習, ディープラーニング

（※文責: 齋藤直紀）

Abstract

Recently, artificial intelligence that is able to imitate human in various example appears. Artificial intelligence is a technique which try to realize intelligence just like learning ability human do naturally using machine learning. Deep learning is a technique which produces an excellent result in a field of image processing.
The goal of this project is to imitate and transcend human's thinking using these techniques of machine learning.
We discuss goal and divide this project into group A and group B.
Group A targets development of system of anticipation of pitch. Group B targets development agent of car that is able to drive faster than human's operation.
The reason why we chose anticipation of pitch includes populality of baseball in Japan declines from decreasing of program rating of TV itself. In addition, it is said that attendance is decreasing slightly from 2010[2] and increase of popularity of sports other then baseball. So, we want to increase the number of people who have an interest in baseball by producing contents involved baseball.
Now, this group focuses "Nomura's Scope" that is indicated sometimes when Katsuya Nomura who was former professional baseball player expounds baseball in terrestrial broadcasting and we devise implementation of contents like it. To be specific, we are approaching making contents of anticipating of pitch just like "Nomura's Scope" using machine learning based on player's data of each baseball teams. Additionaly, we divide us into two groups such as,

- Group of collecting data Making program that extracts data of player from website using Python
- Group of implemention Making program that implements machine learning using Python

We were working in each group in the first semester. Data collection team made a data collection program, and implementation team made the expected program using SVM. In addition, we reviewed implement method and using data for the second semester after the interim presentation.
We were working while using collected data in addition to working at each group. Data collection team improved versatility to increase type of collecting data, and enable to take designated data. Implement team promoted making and implemented program of anticipation of pitch using Deep Learning. In addition, we made presentation materials for final presentation at the same time.

Keyword Baseball, Machine learning, Anticipating of pitch, Python

（※文責: 福田大知）

目次

第 1 章　初めに

1.1　背景

近年 , 機械学習という手法が, 画像認識や自然言語処理, リスク予測など様々な分野で成功を収めている [1]. 機械学習とは人工知能の一分野であり, コンピュータなどの機械が, 学習から行動するための方法を自動的に獲得する方法である [1]. 本グループではこの機械学習を用いて何らかのコンテンツを作ることができないかと考えた.

グループディスカッションの中で機械学習を生かせるコンテンツを話し合った際, 野球の配球予想コンテンツである「野村スコープ」というものを模倣する案が上がった. 他の案とも検討し, 後述の理由から本グループは, 配球予想コンテンツを作成し公開を目標とすることとした.

現在の日本での野球人気は 2010 年以降落ちていると言われている. 理由として野球が 1900 年代のように一般の人々が知っているものから, 野球ファンだけが楽しむものへ変化したためだと考えられる. 根拠として, 地上波放送の減少や視聴率の悪化, 観客動員数の減少 [2], 1990 年代にプロリーグが設立されたサッカーをはじめとした他のスポーツの人気の向上により相対的に野球に触れる人が減っていることが挙げられる. そこで本プロジェクトグループでは, 少しでも野球に興味を持ってくれる人を増やすために, 気軽に使用できる野球人気へ貢献できるコンテンツの開発をしたいと考えた.

（※文責: 齋藤直紀）

1.2　目的

本プロジェクトグループの目標として, プロのキャッチャーを模倣する機械学習を用いた人工知能を開発し, 野村克也氏の「野村スコープ」のようなコンテンツの制作および公開を通して, 野球に興味を持ってくれる人を一人でも増やすことで, 野球人気に貢献したいと考えた. また, プロジェクト活動を通したプログラミング技術や特に機械学習への理解を深めることを目標とした. 既存の配球予想コンテンツとして, 前述の「野村スコープ」がある. 「野村スコープ」とは日本プロ野球で長年キャッチャーとして活躍した野村克也氏が, 地上波での野球中継での解説時にたびたび登場したコンテンツで, 野村克也氏が次はどの球をどのコースに投げるかを予測するものである.

「野村スコープ」は 2009 年以降放送に登場していないので, ここ数年間はコンテンツとして配信されておらず, 現状視聴者が楽しむことはできなくなっている. 本プロジェクトではそのコンテンツをいつでも気軽に楽しめるようにしたいと考えた.

（※文責: 齋藤直紀）

1.3　従来の問題点と課題

前期での問題点として, そもそも「野村スコープ」が 2009 年移行行われていないため, 配球予想というコンテンツ自体が手軽にみられるものとして存在していないことが挙げられる. 前期での課

図 1.1　実際の野村スコープ

題は機械学習を用いるためのデータ収集と, 機械学習への理解を深めるための学習と実装が挙げられた. 後期では, 集めた中からどのデータを使用するかの吟味と, ディープラーニング内のパラメータの調整が精度向上のための課題として挙げられた. この後の展望としては, コンテンツの公開手段の考案や, さらなる精度向上のためのプログラムの改善が挙げられる.

（※文責: 齋藤直紀）

第 2 章　プロジェクト学習の概要

2.1　課題の設定

本プロジェクトグループにおいて, 以下の課題を設定した. まず, グループ全体の課題として,

- 機械学習を用いて, 野村スコープを模倣した配球予想プログラムを作成
- どの試合でも配球予想が見られるようにすること
- インターネット上に公開することで, 誰でも見られるようにすること
- Skype や Evernote などで情報を共有し合い, 進行状況を記録すること
- 配球予想プログラムとデータ収集プログラムの作成には Python を用いること
- ホームページを作成する際には HTML を用いること

次に, プロジェクトを円滑に進めるために, 大まかにデータ収集班と実装班の 2 つの班に分け, 班ごとに課題を設定した. まず, データ収集班において, 以下の課題を設定した.

- データで楽しむプロ野球 [4] のバッターのコース別打率のページの HTML の解析を行い, 必要としている通算のコース別打率がどこに記述してあるかを読み取る
- 読み取った結果を用いて, Python でウェブスクレイピングを行い, 通算のコース別打率を CSV ファイルに出力する. また, NPB（Nippon Professional Baseball）に所属している全選手のデータを収集するために, データで楽しむプロ野球の URL の解析を行い, 解析した結果を用いて全選手のデータを収集する
- 一球速報のデータを抽出し, 球種・コース・結果等を CSV ファイルに保存する

次に, 実装班において, 以下の課題を設定した.

- 簡単な機械学習を用いたプログラムを組むことで, 完成形のイメージをつかみ, 機械学習に対しての知識や理解を深める
- データ収集班が集めたバッター別の通算コース別打率を用いて, 図 2.1 のような, 投げるべき場所には 1 を, 投げてはいけない場所には 0 を出力する簡単な配球予想プログラムを作成すること. その後, 一球速報から抽出した過去の配球のデータをディープラーニングに学習させ, 打者や試合状況と言ったデータを与え次に投げる球を予想する図 2.2 のような配球予想プログラムを作成すること

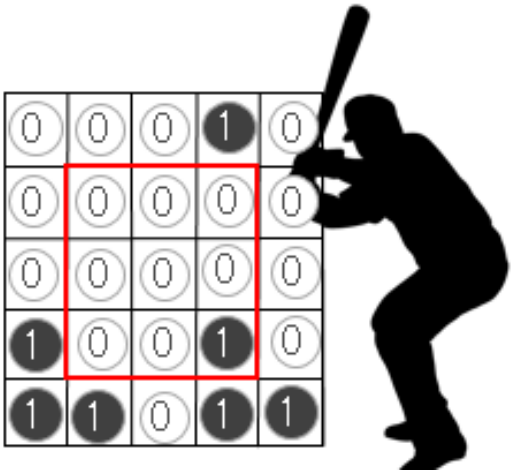

図 2.1　簡単な配球予想プログラム

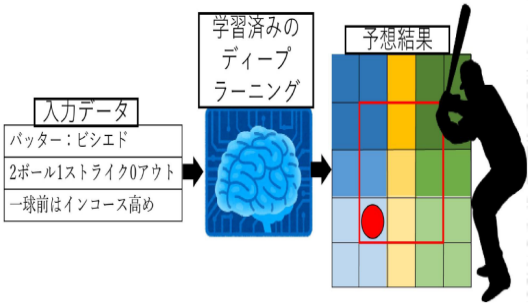

図 2.2　ディープラーニングを用いた配球予想プログラム

（※文責: 野尻雅音）

2.2　前期における到達目標

配球予想コンテンツを作るにあたって, 前期においては, 以下の目標を設定した.

- データ収集班　インターネット上に掲載されているプロ野球に関するデータから, 必要な情報を抽出し, 配球予想プログラムに使用出来るように整形する
- 実装班　コース別打率などの数値から出せるような簡単な配球予想プログラムを作る

（※文責: 野尻雅音）

2.3 後期における到達目標

後期のプロジェクト活動では, 実装班では配球予想プログラムの本格的な実装と, 精度の向上を目的とした. 具体的にはディープラーニングやその他の機械学習の手法から何を用いるのか吟味し実装する作業と, データ収集班から受け取ったデータからどのデータを用いるかの決定を目標として活動をおこなった. データ収集班では前期に作成したプログラムを改良し, データの種類の向上, 指定したデータのみを取得できるといったような汎用性の向上を目標として活動を行った.

（※文責: 齋藤直紀）

第 3 章　課題解決のプロセス

3.1　プロジェクト内における課題の位置づけ

グループ A では捕手が投手に要求するコースと球種を予想して試合中に 60% 以上一致することを目標としている．具体的にはプロ野球の 1 試合に投げられる球数が 100〜150 球であるため，その 60% である 60〜90 球以上の一致を目標としている．また予想した配球を web ページや twitter で公開することも目標の 1 つである．

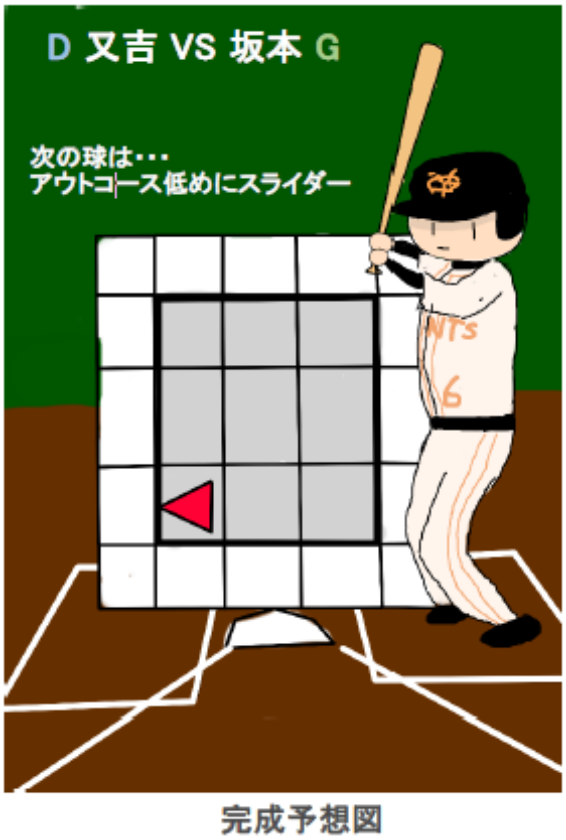

図 3.1　配球予想プログラム完成予想図

（※文責: 福田大知）

3.2　課題解決の方法

前述の目標のために，まず事前学習としてタイタニック号の乗客データから生存予測をする機械学習のプログラムの項目ごとの解読をグループで行った．その後グループ内でデータ収集班と実装班の 2 つの班に分かれた．データ収集班では「データで楽しむプロ野球」のページの HTML を Pandas や Pyquery などのライブラリを用いて必要な情報が載っている場所を探し，CSV ファイルとして書き出すプログラムの作成を目標として活動した．この必要な情報とは，打者のコース別

打率や球種別打率のデータのことである．また実装班では機械学習, ディープラーニングを学習する過程として教師あり学習のモデルでサポートベクターマシーン (以下 SVM) による分類器の作成を目標として活動した．SVM とは, 教師あり学習を用いるパターン認識手法の 1 つであり, 二値分類や回帰に用いるものである．

（※文責: 福田大知）

3.3 中間発表までの開発

3.3.1 データ収集班

データ収集班では野村スコープの再現をするために, 機械学習に学習させるための野球についてのデータを収集するプログラムを Python で作成した．前期ではコース別打率を収集するプログラムと球種別打率を収集するプログラムの 2 つのプログラムを作成した．

コース別打率収集プログラムは, 主に urllib2 と pandas と PyQuery という Python ライブラリを用いて作成した．このプログラムは初めにデータで楽しむプロ野球のコース別打率の表がある Web ページの URL を指定し, その URL が無効でないときに HTML からデータ収集を行った．まず, ページタイトルから選手名とチーム名を抽出した．次に, その Web ページ上の複数の表から通算コース別打率の表だけを抜き取り, その表の中身をセル毎に分けてリスト変数に入れた．その後, 1 セルずつ中身が存在するかを確認し, あるのならそのセルを 1 行 3 列として書き込み, ないのなら None として 1 行で書き込んだ．以上を繰り返して, URL を 1 つずつ実行することでデータで楽しむプロ野球のサイト上にある全ての選手のコース別打率を URL 中の ID が若い順に追加し, チーム別で CSV ファイルへ保存した．

球種別打率収集プログラムは, 基本はコース別打率収集プログラムと同様な手順である．しかし, 球種の数が欠けたページが存在するという問題があったため一部異なる部分がある．この問題を解決するために “ストレート”, “スライダー”, “フォーク”, “シュート”, “シンカー”, “カーブ”, “カットボール”, “チェンジアップ”, “特殊球” の順で保存する順番を決め, もしもページにある球種が存在しなかったらその球種の行を None で書きこむ, となるようにプログラムを作成した．後の保存についてはコース別打率と同様である．図 3.2 が実際に使用したデータて楽しむプロ野球の打率ページである．

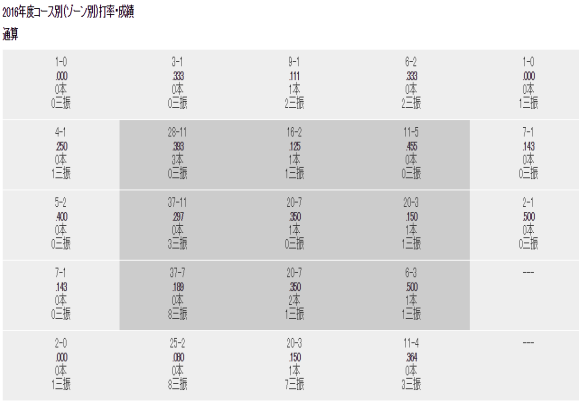

図 3.2　実際に使用したデータで楽しむプロ野球の打率ページ

（※文責: 鈴木才都）

3.3.2 実装班

実装班では配球を予想するプログラムを Python で作成した. 前期ではデータ収集班に集めてもらったコース別打率のデータに, 自らラベルを付けることで打者が苦手とするであろうコースを予想するプログラムを作成した.

この配球予想プログラムは, ディープラーニングを学ぶ過程として, scikit-learn と numpy という Python のライブラリを使用し, 2 値分類の問題として考え, 分類機として SVM を用いた. この分類機に訓練させるデータとして当時首位打者の巨人の坂本選手のコース別打率を使用した. これらのデータに対して実装班独自の判断で, 投げても良いなら “1”, 悪ければ “0” というラベルを付けた. この状態で予想させたい選手のデータを読み込ませることで結果が出力される. 出力される結果は, 投げても良いコースが “1”. 投げてはいけないコースが “0” と出力される. しかしこの状態では, もともと打率が低いボールゾーンを予想してしまう結果が増えてしまった. これを解決するために各コース別打率に加えて, 各コースの打数とヒット数を追加した. これにより打率の低いストライクゾーンと打数に対してヒット数の少ないボールゾーンを予想するようになった.

（※文責: 板垣隼基）

3.4 中間発表のポスターとスライド作成

3.4.1 ポスター

中間発表では Adobe Illustrator を用い, A 班 B 班合同でのポスターを一枚製作した. 完成は中間発表直前のプロジェクト中であり, 準備不足だったと言わざるを得ない.

（※文責: 齋藤直紀）

3.4.2 スライド

スライド作成は冒頭の共通部分を除き, A グループ, B グループでそれぞれ個別に作成し, 最終的にそれらを統合する方針で決定した.

A グループにおいては, まず大まかな発表の流れを話し合い, ホワイトボード上に絵コンテを作成し, それを PowerPoint で再現する形でスライドを作成した. スライド作成とポスター作成を平行して行ったため, ポスターとスライドの表現に細かなズレが生じ, 教員に指摘されその都度修正するなど, あまり効率的ではなかった.

プログラムで行っている処理を出来るだけ理解してもらうため, 入力データや出力データを表や図を用いて表現するように工夫をした. 今回は, A グループは Windows, B グループは Mac でスライドを作成したため, スライドの統合やフォントの統一, 細かい場所の修正などに時間や手間が掛かってしまった.

スライドが完成次第, 実際のプレゼンテーションと同様の形式を取って練習を行った. 教員を交えて練習を行い, 不適切な表現や不足している内容がないかなどの確認を行っていただいた. スライドの修正の回数が多く, 最終的に実際の発表で使うスライドが完成したのは, 中間発表会の直前

であり, 発表練習は若干不足していた.

（※文責: 野尻雅音）

3.5 中間発表

2016 年 7 月 8 日, 公立はこだて未来大学内にて中間発表が行われた. 中間発表に向けてポスターの制作や発表用のプレゼンテーション用のスライドの制作を行った. 発表時には, 制作したポスターの展示, スライドを用いた説明のためプロジェクターとスクリーンを用意した.

（※文責: 福田大知）

3.5.1 中間発表評価シートのまとめ

中間発表で 69 人から受けた評価は, 表 3.1 に示す結果となった.

評価	発表技術	発表内容
1	0	0
2	0	0
3	2	1
4	2	2
5	5	3
6	12	7
7	16	16
8	15	20
9	7	9
10	3	5
無回答	7	6
平均点	7.032	7.476

表 3.1 中間発表での評価

中間発表では多数の意見や質問が寄せられた. まず, 発表内容に関しての意見, 質問を以下に示す.

- 野球の配球を予測するというのは, 面白いと思いました.
- グループ A の成果物がどんなものなのかいまいち伝わってこなかった.
- ディープラーニングの意味が理解できた.
- 野球のやつがとても面白そうでした. 難しいことを一般の人でもわかりやすいようにしててとてもよかったです. 実際に使えるようになってほしいと思いました.
- 野球選手の選考基準を伝えたほうがいいと感じた.
- 具体的にどのようなデータが配球決定に使われているのかというのがはっきり示されていた.
- 果たして得られた結果が有効であったのか否か.

- 最新かつ話題の技術をそれぞれ興味のあることに適用しようとしていることがよく分かった.
- グループ A はいろいろ甘い部分がある気がしたが, モノになれば面白いものだと思った.
- 打倒野村氏と言っていたが活動結果から何をしたいのか, 人間を超えてどうするのか, 目的が見えない.
- 成果物がどのような社会貢献につながるのか, 収益性が見えない.
- 野球に関するデータとは具体的にどういうデータだったのだろうか.
- なぜ最初からディープラーニングを使わず, SVM を選択したかがわからなかった.
- 背景的な情報がわかりやすい.
- 野球は Deep でなくてもよいのでは.
- 将来的にテレビ局に売り込みに行くのか.
- 残り半年でできるのかが気になった.

次に発表技術に関しての意見, 質問を以下に示す.

- プレゼン時に無駄な動きが多い.
- スクリーンか手元の端末しか見ていない.
- ディープラーニングの説明がわかりやすく, 目的も伝わった.
- 全体的に聞こえやすく, スライドもわかりやすい.
- もう少し大きな声で発表したほうがいい.
- はきはきしゃべっていた.

（※文責: 福田大知）

3.5.2 反省

中間発表後には, 中間発表の質疑応答の際に寄せられた質問や評価者フィードバックの意見を中心に反省会を行った. 中間発表時の段階では開発の構想がまだまだ不完全であり, 質問や意見もその点を突かれたものが多数であったため, 夏季休業期間中にはさらに構想を練ってくることが課題となった.

（※文責: 福田大知）

3.6 最終発表までの開発

3.6.1 データ収集班

後期プロジェクトでは, 配球予想プログラムのデータの向上を目標とした. その中でデータ収集班は, 前期に集めたデータだけでなく, データの種類の増加や実際の試合に近いデータの収集が求められた. そのため, データ収集班は goo 一球速報 [5] からデータを収集するプログラムを Python で作成した. 数ある野球情報サイトの中から goo 一球速報を選択した理由として, 一球毎のコースが見られること, ボールカウントやランナー等の試合中の状況を知ることができること等が挙げられる.

後期での開発にあたってプログラムの開発環境を変更した. Python2.7 の実行環境を Anaconda に変更すると, 様々なパッケージやモジュールが初めから導入されていて使いやすく, また, 新

たなパッケージの導入なども手軽である，という情報を得られたため，Python2.7 の実行環境を Anaconda を用いたものに変更した．

プログラム開発について，まず，データ収集班は前期に作成したプログラムを goo 一球速報用に改造して製作した．図 3.3 が使用した goo 一球速報の画像である．

図 3.3　実際に使用した goo 一球速報のページ

しかし，goo 一球速報の Web ページは前期にスクレイピングした Web ページと違い，DOM(Document Object Model) が動的に生成されるため，うまくソースコードを読み込めず収集に失敗した．次に，動的に生成された DOM を読み込み，スクレイピングするために 2 つのパッケージを導入した．1 つ目に，Selenium という Web ブラウザ操作の自動化などに用いられるパッケージを導入した．2 つ目に，スクレピングを高速化するために PhantomJS というヘッドレスブラウザのパッケージを導入した．この 2 つのパッケージを用いることで，goo 一球速報の Web ページのソースコードを読み込み，操作することで，スクレイピングすることに成功した．また，雑多な情報から整然としたデータを得られるようにするため，Beautiful Soup というソースコードを解析し，扱いやすい形に変換する機能を持つパーサーのパッケージを導入した．

Web スクレイピングをするために，Web ページと URL の構造を解析した．goo 一球速報の URL に含まれる sj_PageID を調べると，Ga(試合 ID)_(イニング数)_(表か裏)ball という様な規則性があることがわかった．例として，sj_PageID=GA1622123_03_Tball の場合，1622123 が試合 ID , 03 がイニング数，T が表となる．表か裏は T と B で表現されているが，これは Top と Bottom のことであると推測できる．また，URL で isNext=true と指定すると，そのページの一番最初から表示できることがわかったため，これを利用して開発を行った．goo 一球速報の Web ページでは，盗塁や選手交代など，通常の投球場所が表示されない特殊なケースが存在した．そのため，ソースコードのタグの ID から投球場所が表示されている画面とそうでない画面を判別した．

プログラムで収集したデータについて，goo 一球速報の動的な DOM から機械学習に用いるためのデータとして，バッターの名前，キャッチャーの名前，ボールカウント，ストライクカウント，アウトカウント，球種，球速，ピッチャーの投球場所，ランナーの有無を収集した．ランナーの有無は，ランナーがいたら 1 , いなければ 0 の 2 進法で表し，1 塁を 1 桁目，2 塁を 2 桁目，3 塁を 3 桁目で表現した．また，Web ページを操作するための条件付けのために，打席内での球数，試合終了時表示する部分のソースコード，特殊な画像の有無，チーム名，裏表などを収集した．これらは，試合終了時のページや盗塁の時にだけ表示される画像などを実際に目で見て確認することで，ページ操作の条件付けに用いることを決定した．

今回作成した Web スクレイピングプログラムの流れはおおまかに 3 つのステップに分かれる．

まず, DOM から機械学習に使うためのデータとページ操作などのスクレイピングに用いるためのデータを収集する. 次に, スクレイピング用に収集したデータで条件付けを行い, ページを操作する. その後目的のページに到達したら, そのページから機械学習に用いるためのデータを収集し保存する. 以上の 3 ステップを繰り返すことで goo 一球速報から自動で機械学習に用いるデータを収集するプログラムを実装した. 投球場所の判別は表示されるボールの画像の座標を用いて行った. 投球場所の座標の範囲が, 横は 0px から 160px, 縦は 0px から 228px であることがわかった. ここで, それぞれの最大ピクセル数を 5 で割った値, 横は 32, 縦は 45.6 で投球場所の座標を割り int 型に変換して少数点以下を切り捨てることで, 横縦それぞれ 0 から 5 までの数字として示した.

また, goo 一球速報のプログラムが大方完成した後, 前期に作成したコース別打率を収集するためのプログラムと合体させた. 1 つのプログラム内で同時に 2 つのページの情報を収集することで, 2 つのプログラムを動かす場合よりも時間と手間が縮小された.

（※文責: 鈴木才都）

3.6.2 実装班

後期からディープラーニングの実装に取り掛かった.

実装に伴い, 開発環境を一部変更した. 後期から Jupyter notebook というツールを使用し始めた. これによりディープラーニングのプログラムで出力される訓練データでの正答率, と誤差関数, 予想するときの正答率, と誤差関数を簡単にグラフとして表すことができ, 視覚的に精度の向上等が見やすいからである.

まず最初に web 上に掲載されていたプログラム [7] をもとに 3 層のニューラルネットワークを作成した. ニューラルネットワークを作成するにあたり, Python には多くのライブラリが存在する. その中でも今回は Chainer を使用した. Chainer は GPU を使って高速に学習させれることができ, また, インターネット上に多くの Chainer を使用したディープラーニングのプログラムが掲載させており, 参考にしやすいという利点がある.

前期では 2 値分類であったのに対し, どこのコースに投げるのかという 1 点の予測にさせるために出力を 25 値分類にした. 入力に必要なデータはコース別打率, ボールカウント, ストライクカウント, アウトカウントとした. 学習させるデータは, 予想させたい選手が所属している球団の 3 試合分, 600〜700 球を用意した.

しかし, 予想の精度があまり上がらなかった. そのため予想の方法を 25 値分類から, 縦と横をそれぞれ 5 値分類で予想させ, 組み合わせることで, ある 1 点を予想させる方法に切り替えた. これに加え. 入力データにランナーの有無と 1 球前にどこに投げられたかという情報を追加した. またデータを交流戦を除いた全試合分追加した.

精度をさらに上げるために, ネットワークの調整を図った. ネットワークは入力層, 隠れ層, 出力層の 3 種類から成り立っている. この隠れ層は 1 層以上により構成されており, この層の数によっても精度は上下する. 今回はネットワークを調整する過程で隠れ層を 2 層にした. 次に誤差関数を最小化させるために勾配降下法というアルゴリズムの最適化手法を選ぶことになった. この最適化手法には多くの種類が存在しており, 最初は Adam という手法を使用していた. 他の手法での正答率を見るために, RMSpropGraves, MomentumSGD, AdaDelta という 3 つの手法と Adam を比べてみたところ, それぞれにおいて正答率に大きな差はなかったが, 収束する速度が最も早かった RMSpropGraves に最適化手法を変更した. この RMSpropGraves には学習率というパラメータがあり, このパラメータも精度に大きくかかわってくる. ネットワークを作る際にこのような精

度にかかわってくるハイパーパラメータは多く存在し, これらの調整が必要となってくる. このハイパーパラメータを調整する手法としてグリッドサーチ, ランダムサンプリング, ベイズ最適化とある. 今回はランダムサンプリングを使用した. グリッドサーチは 1 回の施行時間が長く, ディープラーニングではハイパーパラメータの量が多いため不向きであるので今回実装は見送った. ベイズ最適化は実装が間に合わなかったため使用することができなかった. ランダムサンプリングはパラメータを無作為に抽出する方法で探索の分布や幅を簡単に変えることができる. この手法でハイパーパラメータの調整を行った.

この結果できたネットワークと, 新たに Python のライブラリの 1 つであるランダムフォレストを使って配球を学習させた. こちらも縦, 横でそれぞれ 5 値分類で予想させ, ある 1 点を予想するようにしている. さらに比較対象として野球経験者の人間にも同じように予想させた. 今回対象としたのは 9 年間の野球経験がありピッチャーの経験もある人物を対象とした. 表 3.1 がそれぞれの予想結果である. 前期の時点ではこの予想に加えて投げる球種についての予想も計画していたが, コースに対する予想の精度を向上させることに集中するために断念した.

	縦横 5 分割	ランダムフォレスト	25 分割	学生
縦方向	27%	24%	×	23%
横方向	30%	25%	×	28%
25 分割	10%	8%	7%	5%

表 3.2　出力結果の比較

（※文責: 板垣隼基）

3.7　最終発表のポスターとスライド作成

3.7.1　ポスター

最終発表では中間発表と同じく Adobe Illustrator を用い, A グループ B グループ合同でのポスターを一枚, および A グループのみの説明ポスターを一枚作製した. 内容としては, 概略, 用いた手法, プログラム実装, データ収集の概要, まとめと改善案である. 完成はこちらも最終発表直前のプロジェクト後であり, もう少し余裕を持てなかったのかという反省がある. 各項目においては実現できた項目を強調し, わかりやすい説明を心がけた. また中間発表での反省を生かして, なるべく図解をしていくことを目標とし作成を行った. 中間発表時よりは文字数を抑え, 図を多くすることはできたが, まだまだ改善の余地はあったと考えられる.

（※文責: 齋藤直紀）

3.7.2　スライド

最終成果発表でのスライドは, 大まかな流れや構成などは中間発表時のスライドを参考にし, 作成を行った. また, 作成方法も中間発表時と同じように A グループ, B グループでそれぞれ個別に作成し, それらを統合する方式を取った.

共通知識であるディープラーニングの解説は中間発表時のスライドを変更, 修正し用いた. A グ

ループにおいては, 野球の基礎知識などといった中間発表時と同じ事を発表する項目は, 中間発表時のスライドを元に作成することで, 可能な限り作業量を減らすことを心掛けた. また, 完成したプログラムで出力する結果をコンテンツとして認識出来る程度に可視化できなかったことや, プログラムの処理がブラックボックスのようなものであることを考慮した結果, 聴講者に出来るだけ理解しやすく伝えるために, 図を中間発表時よりも多く用いる事を心掛けた. 図 3.4 は実際に使用した図の一例である. その為, 詳細は口頭で説明を行うことにした.

その後, このスライドを用いて発表練習を行った. まず学生のみで練習を行い, 内容や表現に間違いが無いかを確認し, タイムを測り実際の発表の際に時間をオーバーしないように原稿の量の調整を行った. その後教員を交えて練習を行い, 改めて内容や表現に間違いが無いかを確認してもらい, 指摘された修正点をグループ内で話し合いを行って修正作業を行った. しかし, 今回も最終成果発表会直前で修正を行ったため, 原稿の変更が間に合わなかったりと本番でアドリブを交えて発表を行ったため, 万全の体制とは言えない状態であった.

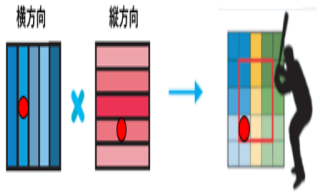

図 3.4　スライドに使用した図の例

（※文責: 野尻雅音）

3.8　最終発表

2016 年 12 月 9 日, 公立はこだて未来大学内にて最終成果発表が行われた. 最終成果発表に向けてポスターの制作や発表用のプレゼンテーションの資料の制作を行った. 中間発表と同様に, 発表時には制作したポスターの展示, スライドを用いた発表のためプロジェクターとスクリーンを用意した.

（※文責: 福田大知）

3.8.1 最終発表評価シートのまとめ

最終発表で 76 人から受けた評価は, 表 3.3 に示す結果となった.

評価	発表技術	発表内容
1	0	0
2	0	0
3	0	0
4	1	0
5	1	5
6	12	7
7	19	11
8	26	25
9	12	19
10	5	9
無回答	0	0
平均点	7.631	7.960
中間発表との差	+0.599	+0.484

表 3.3 最終発表での評価

成果発表では多数の意見や質問が寄せられた. まず, 発表内容に関しての意見, 質問を以下に示す.

- 野球について, 自分は個人的に好きだったのでわかりやすかった.
- 機械学習感がない. でも題材としては面白い.
- まず配球に本当にデータ間の相関があるのか気になります. ネットワークモデルを RNN に変えるとどうなるか興味があります.
- これから先, さらに精度を上げることを考えていてよいと思います.
- 自分も野球に興味があるので面白かった. 身近なことを研究していてよいと思った. もっと予測ができるようになれば野球の実況が面白くなりそう.
- ディープラーニングはすごいことが分かった. 後半, 難しい言葉が多くてよくわからなかった.
- よく勉強していて, 成果が表れている. 結果に関する考察, なぜうまくいくのか, 行かないのかがあるともっと良いでしょう.
- 野球は入力項目が少なすぎる気がします.
- 目標設定と今後の展望がわかりやすかった.
- プロを超えたらすごいと思いました.
- ディープラーニングをよく知らない人でもよくわかる内容になっていたと思う.
- deep learning にしては少し課題が易しすぎたかもしれませんが次のプロジェクトに期待したいです.
- また応用的な発表をお待ちしています.
- プロのキャッチャーによっても人によって配球パターンが変わる上, プロのキャッチャーで

も最善の配球ができている確証が取れないのではないか. 内容は非常に面白かった.

- 配球予想は野球ゲームに応用できるようになるといいですね. プロの配球に挑戦という形で. 精度に関してはキャッチャーそれぞれによって配球が変わると思うので選手を絞るべきだと思います.
- 面白い題材を取り上げてよく勉強していると思います.
- 実際に数値を出して成果があることを見せていることが良かったです.
- どのように学習させたか, これがどのくらいすごいディープラーニングなのかがちょっとわかりにくかったです.
- It's not clear what you actuarry did. I wanted some video or pictures of experimental set up.
- もっと目的を達成するための条件を考えていたほうが良かったと思う.
- 野球の配球の知識がなかったので説明してくれて助かった.
- キャッチャーによって配球は変わるため, ノイズとなるデータが多そう.
- ディープラーニングの細かい説明から発表の内容に含まれていたのでとてもわかりやすい発表だった. 学習データも 135000 球というデータ量を使っていたことからとても努力していたことが伝わった
- 今後の将来を担う分野であり, 期待しています.
- 無知な人は理解できないのではないかと思った.
- 野村スコープと比較して欲しかった.
- システムがとても良くできていた. 評価基準が野尻君なのは主張が弱い.
- 野村スコープは何 % くらいなのか気になる.
- 野尻君の判定が当たる確率が高いのか低いのかわからないので, 野尻君のデータが欲しい.
- 結果として野尻君の予想を超えていて良かったと思う.
- 25 値化や進んだ距離を基にしている等の発表だったが, その根拠を発表していただきたかった.
- 専門用語が多くてわかりにくい.
- 野球の 10% がどのくらいすごいのかわかりやすい. 対人戦でも同じようにできるのか気になる.

次に発表技術に関しての意見, 質問を以下に示す.

- よく練習していますが, 時間が短かったですね. もう少し技術面を説明するスライドを増やしてもよかったのでは.
- コンセプトがシンプルで分かりやすかった.
- 原稿を見てなくてよかったと思います.
- 全体的に声量は問題ないように感じた. ただ人によっては聞き取りやすさが変わるので, しゃべる人は統一したほうがいいと思う.
- 声が大きく, 発表の内容も分かりやすいです. 映し出された画面を見て話すより, 終始観覧側を見て話すよう心がけると表情等が伝わりよいと思います. またスライドもよかった.
- 発表者は１２人にしたほうが見やすいのではないかと感じました.
- 声が小さいのとスライドに無駄が見えた. ポスターの位置に不思議さを感じた.
- スライドの中身は伝えたいことを重点に置いていてフォントの色を使って区別させていてわかりやすかった. 声もはっきりと聞き取れ, プレゼンに集中できました.

- スライドに大量の文字があり文字が小さく見にくいところもあった. 手を使ってスライドのある部分を示していたが, ポインターなどを使って後ろに影を作らないようにしたほうが良いと思う.
- 聞き取りやすいがスライドを見過ぎだと思った. ジェスチャーが良いと思う.
- 野球の 10% がどのくらいすごいのかわかりやすい. 対人戦でも同じようにできるのか気になる.
- プレゼンが工夫されていてよかった.
- 早口で聞き取りづらかった. 声量は良かった.
- 良い発表だとは思うが, 具体例があるともっと良かった.
- 板垣君, 発表慣れしていますね.
- もう少し大きい声だと良い. スライドや話の内容はわかりやすい.
- 発表スピードなどが良い, わかりやすい発表.
- プロの予想を手に入れられれば良いですね.
- 精度を上げるための方法が明確でわかりやすかった.
- 個人的に興味のある内容で面白かった

（※文責: 福田大知）

3.8.2 反省

最終発表後には, 最終発表の質疑応答の際に寄せられた質問や評価者フィードバックの意見をもとに反省会を行った. 最終発表時の成果物に対する意見や質問を受け止め, 来年度のプロジェクトに期待するとともに報告書の執筆に活かしていきたいと思った.

（※文責: 福田大知）

第 4 章　グループ内のインターワーキング

4.1　各人の課題の概要とプロジェクト内における位置づけ

- 野尻雅音 (グループリーダー, データ収集班)

  グループリーダーとして, グループ全体の進捗状況を把握し, 班員に指示を出した. データ収集班としては, 「データで楽しむプロ野球」 の HTML を解析し, Python を用いて通算コース別打率のデータを抽出し, CSV ファイルに保存するプログラムを作成した. また後期では, 班員が開発した goo 一球速報から必要なデータを抽出するプログラムに前述したコース別打率を抽出するプログラムを組み込み, 全自動で配球予想に必要なデータを収集するプログラムの開発を行った. 中間発表会と最終成果発表会においては, A グループのスライドを作成し, それに応じた原稿を書き上げた. また B グループの作成したポスターと結合させ., 最終的にプレゼンテーションで用いるスライドを作成した.

  （※文責: 野尻雅音）

- 福田大知 (データ収集班)

  まず野球に関する知識が少なく今後のグループの活動に支障や遅れが出ると考えたため, 野球について個人的に学んだ. またデータ実装班としてメンバーが作成した「データで楽しむプロ野球」のページの HTML を解析し, 必要な知識を抽出するプログラムを実際に動かして CSV ファイルに保存することでプログラム作成者と連携して班内で効率良く作業を行った. 中間発表会においては, 作成されたポスターの英訳を主に担当した. 後期では, データ収集班として goo 一球速報のデータをプログラム完成までの間手動で集め, 完成後もプログラムを動かしてデータの収集に勤しんだ. またメインポスターの英訳も務め, 質疑応答の内容を記録し文書に起こした.

  （※文責: 福田大知）

- 鈴木才都 (データ収集班)

  主にデータ収集プログラムの開発を行った. 前期では, Web サイト「データで楽しむプロ野球」の HTML を解析して CSV ファイルとしてデータを保存するプログラムを Python で実装した. 後期では, Web サイト「goo 一球速報」から動的に生成された DOM を解析し, 自動でデータを抽出, 保存するプログラムを開発した. 最終発表にむけて, プロジェクト 14 の A グループ, B グループ共通部分のスライドを作成した. また, 発表では A グループのスライド前半部分を発表した.

  （※文責: 鈴木才都）

- 齋藤直紀 (実装班)

  web サイトで掲載されている情報を参考に, 収集班から貰ったデータを整理, 必要な情報だけを抽出し, 空欄を随時補填するプログラムを作成した. また板垣と協力し SVM のプログラムの作成, 理解にも努めた. 中間発表の際は A グループ側及び全体でのポスター制作を担当した. 後期では, データ収集班からデータを随時受け取り, プログラムと対応させて実際

の動作を行った．また最終発表時には A グループでの発表及びポスター制作を担当した．

（※文責: 齋藤直紀）

- 板垣隼基 (実装班)
  前期では web サイトで掲載されているプログラム [6] をもとに SVM を用いた 2 値分類での配球予想プログラムの作成し, ディープラーニングの作成にも取り掛かった．後期では 3 層, 4 層のニューラルネットワークを作成した．基本的に一度インターネット上に掲載されているプログラムを自分の環境で実行, 動作を確認しプログラムを理解するところか始めた．その後本グループの目的と会うようにプログラムを改変, 調整することでネットワークを作り上げた．また, ディープラーニングに必要なパラメータを調整するためにランダムサーチやグリッドサーチについても学んだ．また比較対象を作るために SVM とランダムフォレストを用いた配球予想プログラムの作成にも取り掛かった．しかし SVM を用いた配球予想プログラムはデータ量が多すぎ, 計算時間が膨大になってしまい実装することができなかった．中間発表, 及び最終発表では A グループのプログラム部分についての発表を担当した．

  （※文責: 板垣隼基）

第 5 章　成果物の現状

本グループの最終的な成果物は, Python で作成したデータ収集プログラムと, 同じく Python で作成したディープラーニングを用いた配球予想プログラムの二つである.

（※文責: 齋藤直紀）

5.1　データ収集プログラム

データ収集プログラムでは, バッターの名前, コース別打率, ランナーの状況, ボールカウント, 投げたコース, 球種, キャッチャーの名前, ピッチャーの名前を Web サイト「goo 一球速報」, および「データで楽しむプロ野球」から自動取得できるプログラムを作成した. その結果, 2016 年のプロ野球におけるほぼ全試合分のデータを用意することに成功した. ここでの「ほぼ」という記述については, 「goo 一球速報」において, 打者がアウトになった際のコースが表示されないという仕様のため, 全データ取得が不可能であったことが理由である.

（※文責: 齋藤直紀）

5.2　配球予想プログラム

配球予想プログラムは上記の通り, ディープラーニングを手法として採用しており, 当初は 25 コースから一か所を予想する予定であった. しかし, この方法は精度が良くないと考え, より精度を向上させるためにコースを上下方向, 内外角それぞれ 5 等分したものをかけあわせて予想を導き出す構造とした. ただし今回は, バッターの左右打席を考慮しておらず, 内外角の予想は不十分であった可能性が指摘される.

（※文責: 齋藤直紀）

5.2.1　使用したデータ

収集したデータ約 14 万球のうち, 135,000 球を学習データ, 5,000 球をテストデータとして使用し, データの内訳は, バッターのコース別打率, ボールカウント, 一球前のコースとした. 上記のデータ収集班で収集したデータの中から一部をピックアップして使用している理由としては, 入力として与えても大きく変化が起きないと判断したためである. 例としてランナーの状況などを入力として与えても予想結果に変化が現れなかったことが挙げられる. しかしこれはデータの与え方の問題も考えられるので考慮が必要な項目となった.

（※文責: 齋藤直紀）

5.2.2　各手法の正答率

プログラムの実行結果として, ディープラーニングでは最終的な分割したものを掛け合わせる手法での正答率は内外方向が約 27%, 上下方向が約 30%, 25 分割での予想が約 10% となった. また, 比較対象としてランダムフォレストと SVM, ディープラーニングよる一括での 25 分割予想, またグループリーダーの野尻が実際に試合を視聴しながら予想を行ったものを用意した. ランダムフォレストでは内外方向が約 24%, 上下方向が約 25%, 25 分割が約 8% という結果になり, ディープラーニングの方が若干ではあるが高いという結果が得られた. SVM は, データが膨大なため計算速度に時間がかかるという理由で行うことができなかった. ディープラーニングによる一括での予想は約 7% と実装した手法の方が優れていることを実証する結果となった. グループリーダー野尻による予想は, 内外方向は約 23%, 上下方向は約 28%, 25 分割では 5% となった. このことから, 長い野球歴を持つアマチュアの選手よりは精度が良いという結論を得ることとした. 以上が成果物として完成した 2 つのプログラムの説明である. また, 各手法での正答率を表 5.1 にまとめた.

	縦横 5 分割	ランダムフォレスト	25 分割	学生
縦方向	27%	24%	×	23%
横方向	30%	25%	×	28%
25 分割	10%	8%	7%	5%

表 5.1　出力結果の比較

（※文責: 齋藤直紀）

第 6 章　課題と今後の展望

本グループでは, プロ野球における配球予想プログラムを完成させることができた. 現状として, 正答率は約 10 %を記録し, アマチュアの野球選手の予想よりも高いという成果を出すことができた. 課題として, さらなる正答率の向上が挙げられる. そのためには, 打者や捕手, そして投手を区別することが重要だと考えた. 今回のプログラムは開発途中に捕手に ID を振り分け, 捕手ごとの配球を区別しようと考えた. しかし, 知識や調査が足りなく, ID の振り分け方, ディープラーニングに対する読み込ませ方に問題があったため, 最終的なプログラムでは捕手の区別は行っていない. これらを改善し, 打者や捕手, 投手を区別することで, 捕手がその打者だけに行う配球などの要素を学習出来ると考えた. また投手の利き腕や, 打者の打席の左右等も入力データに含めること, ネット上のデータだけではなく, リアルタイムで試合映像からデータを抽出し, それを利用することなどが課題として挙げられた. 次に, 今後の展望としては, コースだけの予想では無く, 投手が投げる球種も予想することが挙げられる. そのためには, その投手が投げる球種データ, 打者の球種別打率データ, 一球前に投げた球種データ等の入力が必要だと考える. また, 当初目標として掲げていたリアルタイムで配球を予想し, その結果をウェブ上で公開することで, 全世界のプロ野球ファンが楽しめるようなコンテンツになると考える.

（※文責: 野尻雅音）

第 7 章　まとめ

本グループでは, ネット上に掲載されているプロ野球に関するデータをウェブスクレイピングで収集し, 試合における捕手の要求するコースを予想するプログラムを作成した. 目標であるプロのキャッチャーの配球を模倣することは, 正答率約 10 %という結果から完全に達成できたとは言えないが, アマチュアの野球選手の予想の正答率は上回ることが出来た. 一方で, ディープラーニングに入力するデータの形や入力方法に問題点が残っている事がわかっており, これらの問題点を解決していくことでより良いものへと発展させることが可能であると考える. まとめとして, 本プロジェクト内における本グループは, グループ全体の目標に対して一定の成果を上げることができ, プロジェクト学習においてグループメンバー 5 人全員が十分に貢献することが出来た.

（※文責: 野尻雅音）

参考文献

[1] イラストで学ぶディープラーニング, 山下隆義, 講談社
[2] 日本野球機構, http://npb.jp/statistics/
[3] 統計からみたスポーツの今昔, 総務省, http://www.stat.go.jp/data/topics/topi640.htm
[4] データで楽しむプロ野球, http://baseballdata.jp
[5] goo 一球速報, http://sports.goo.ne.jp/baseball/npb/
[6] Qiita , 金貨が本物かどうか見極める, http://qiita.com/ynakayama/items/33231bf23d40d1c1f344
[7] Qiita , 【機械学習】ディープラーニング フレームワーク Chainer を試しながら解説してみる, http://qiita.com/kenmatsu4/items/7b8d24d4c5144a686412