# インターンProject agenda

2017/10

Housmart インターン生

萩原　辰哉

# 今回発表するProjectについて

**期間**　　６～１０月

**Project. 1**「マンション画像分類」

**Project. 2**「SEOのためのマンションジャーナル類似記事表示」

# Project.1 マンション画像分類

**目的**

・マンションサイトのファーストビューにふさわしい画像を分類

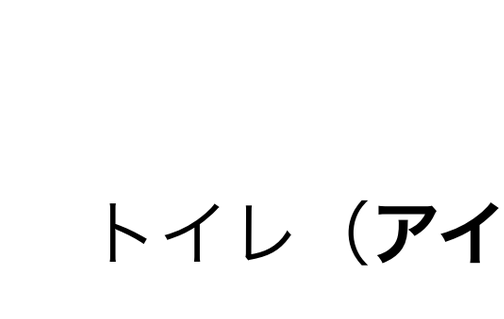
・4種類の画像を判別するアルゴリズムを作成

**今回の対象となる画像（４種類）**

・室内画像、トイレ、キッチン、眺望

トイレ（**アイキャッチにふさわしくない画像**）

→**ファーストビューに配置しない**



室内画像、キッチン、眺望（**キレイな画像**）

→**ファーストビューに配置**

# **CNN** (Convolutional Neural Networks)について

**用途**

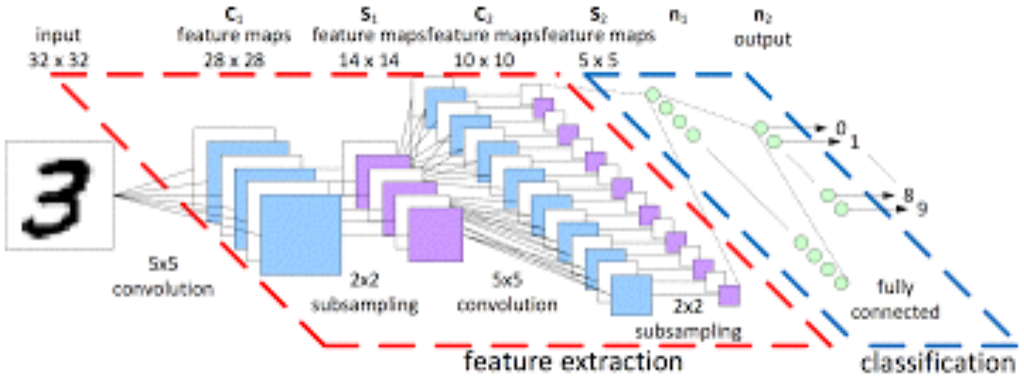
・Googleが発表したディープラーニング用のアルゴリズム

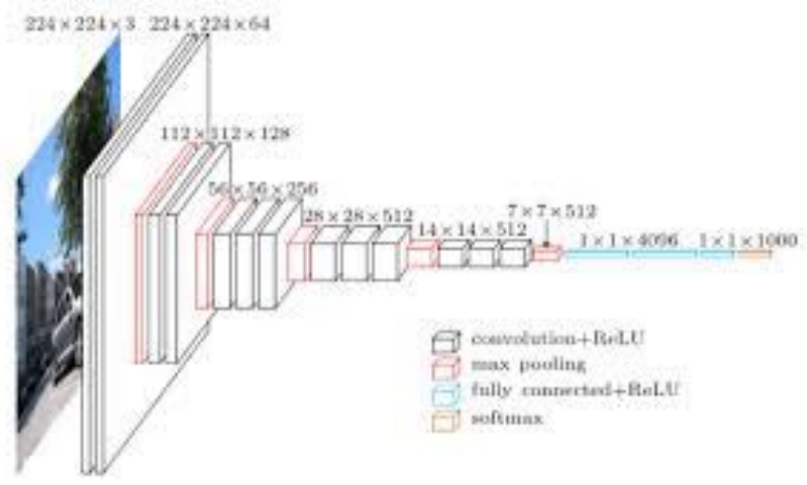
・画像分類に頻繁に使用

**複数のモデルが存在**

「cifar-10」→一般的なモデル

「VGGNet」モデル.etc→高性能なモデル




・通常のモデル「cifar-10」

**特徴**　層が薄い

・高性能モデル「VGGNet」

**特徴**　層が厚い

**今回はVGGNetのモデルを使用**

# CNNの学習過程

**訓練画像とテスト画像を用意**

・訓練画像はGoogle検索で収集→増幅

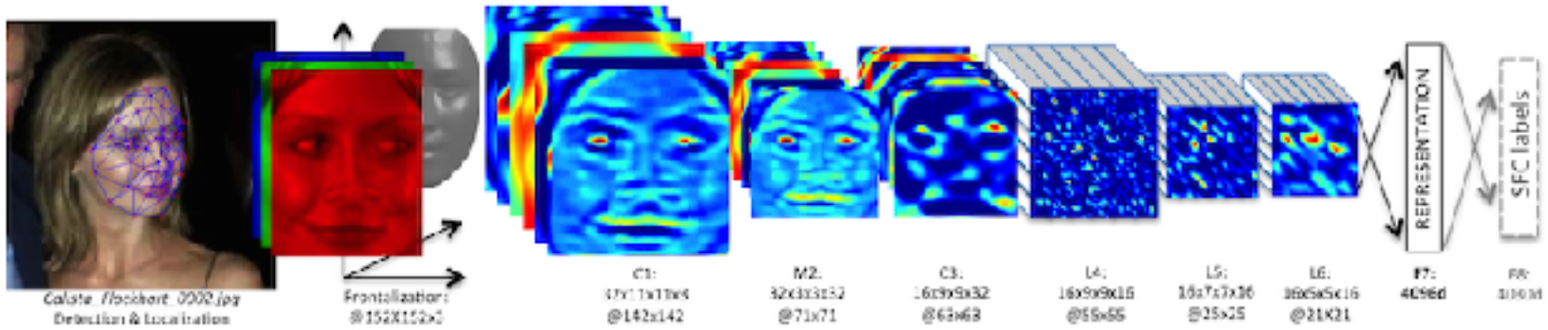
・テスト画像はGoogle検索で収集したもののみ使用

・画像の種類毎にラベルをつける（0,1,2,3）→４種類

**CNNの学習から評価まで**

・訓練画像を学習→テスト画像を判別させる→正解率で評価

# CNNの学習とは？




**画像の特徴（特徴量）を学習**

**例**　コンロと水面台が写っているもの

**→キッチン画像と判別**

# 訓練画像の内訳

**訓練画像　　25300枚**

**Label 0　　6300枚**

玄関 50,　廊下 50,　リビング 150,　寝室 100,　クローゼット 100

**Label 1　　6000枚**

キッチン 300

**Label 2　　7000枚**

トイレ 400,　風呂 150,　洗面化粧室 150

**Label 3　　6000枚**

眺望 300枚

# テスト画像の内訳

**テスト画像　　540枚**

**Label 0　　150枚**

玄関 30,　廊下 30,　リビング 30,　寝室 30,　クローゼット 30

**Label 1　　115枚**

キッチン 115

**Label 2　　160枚**

トイレ 80,　風呂 40,　洗面化粧室 40

**Label 3　　115枚**

眺望 115

# 成果

**正解率　87.8%**

・540枚中、474枚を正しく判別

**各ラベル毎の正解率**

・Label 0 （室内類）　96.1%

・Label 1 （キッチン）　77.3%

・Label 2（トイレ類）　92.5%

・Label 3（眺望）　93.9%

**キッチン画像の**

**正解率が低い**

# Why ?

**原因**

・室内画像と似ているため、Label 1の「室内」画像と間違えて判別されたため

**キッチン画像を選ぶときの基準**

- 「コンロと水面台」が写っていること
- それ以外の特徴は室内と同じでも良い＝室内画像と似ている

# 見分けがつきにくい

キッチン画像

室内画像

# Project.1まとめ

・マンションサイトにふさわしい画像を判別するアルゴリズム

・ディープラーニング用アルゴリズムCNNを使用

→モデルは高性能のVGGNetを使用

・正解率は**87.8%**　→　かなり良い結果

・改善点はキッチン画像と室内画像が間違えやすい点

# Project.2
# SEOのためのマンションジャーナル類似記事表示

**目的**

・マンションジャーナルの記事を上位表示するため

・記事同士の類似度を表示させるアルゴリズムを作成

**使用するアルゴリズム**

・Doc2Vec

**なぜ類似記事表示がSEOに有効か？**

・ホームページ内の回遊率を上げる点でSEO（検索上位表示）に有効

# ホームページ内の回遊率を上げるとは？

## 回遊率とは

・どのくらい一人のユーザーがサイト内のページを閲覧したか

記事内に類似記事があれば、ユーザーがそれを閲覧

ページ閲覧回数が増える＝回遊率UP

**SEOに有利**

# Doc2Vecとは？

**文章Aに対するその他の文章の類似度を表示**

例：マンションジャーナル記事Aに対して、

　　マンションジャーナル内の記事（残り1658記事）から、

最も類似した上位5記事の「記事名＆記事番号」と

「類似度」を表示

# 類似度表示の例

**最も類似してる上位5記事を表示**

・該当コード(topnで表示数を指定できる)

・model.docvecs.most_similar([対象の文章], topn=5)

(‘969マンションジャーナルのiPhoneアプリがリリースされました！', 0.8666160702705383),

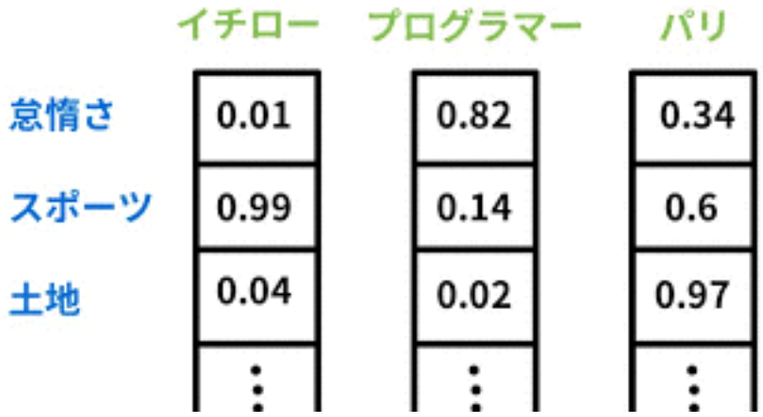
('672【カウル】Morning Pitchに登壇させて頂きました', 0.8381995558738708),

('160「広報の佐藤さん@ハウスマート」はじめました！', 0.801512121542358),

('519【物件リクエスト機能】カウルなら殆どのマンションをご紹介可能です',

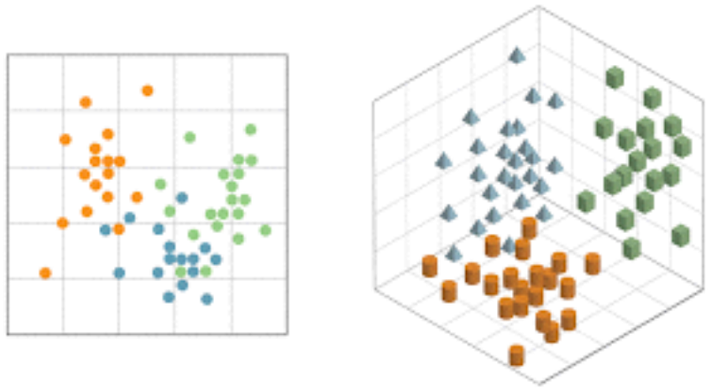
0.8014494180679321), ('1051dp-magjam', 0.8001388907432556),

# Doc2Vecの原理





**単語をベクトル（数値）で表示**



**文章を類似度で表示**

# 結果

**CSVファイルで提出**

・縦1659、横20の表（1659×20）

・縦→マンションジャーナルの全1659記事の

タイトル＆記事番号（1659）

・横→各記事に最も類似した上位10記事の

類似度とタイトル＆記事番号（10×2=20）

# Project.2 まとめ

・マンションジャーナルの類似記事を表示するアルゴリズムの作成

・類似記事表示はユーザーの回遊率を上げる → SEOに有利

・アルゴリズム　→　Doc2Vecを使用

・本プロジェクトの結果　→　CSVファイルで提出

# まとめ

## Project

・「画像分類」と「マンションジャーナルの類似記事表示」

## Project.1 画像分類

・高級マンションのファーストビューにふさわしい画像を配置するため

## Project.2 マンションジャーナル類似記事表示

・SEO（検索上位表示）のため

**→いずれもHousmartのサービスに組み込み予定**