入処言自

門理語然

はじめに...

本課題を取り組むにあたってわかって/できていてほしいもの

- Python 3系が使えること
- TernimalでLinuxコマンド最低限の操作ができること
- 機械学習がなんたるか、雰囲気でわかっていること

説明変数/目的変数/教師データ/テストデータ/分類器の
意味がわかっていればヨシ！

サンプルプログラムetcは以下のGitHubにあるから
ダウンロードしておいてね

https://github.com/mk24601/NLPbasic.git

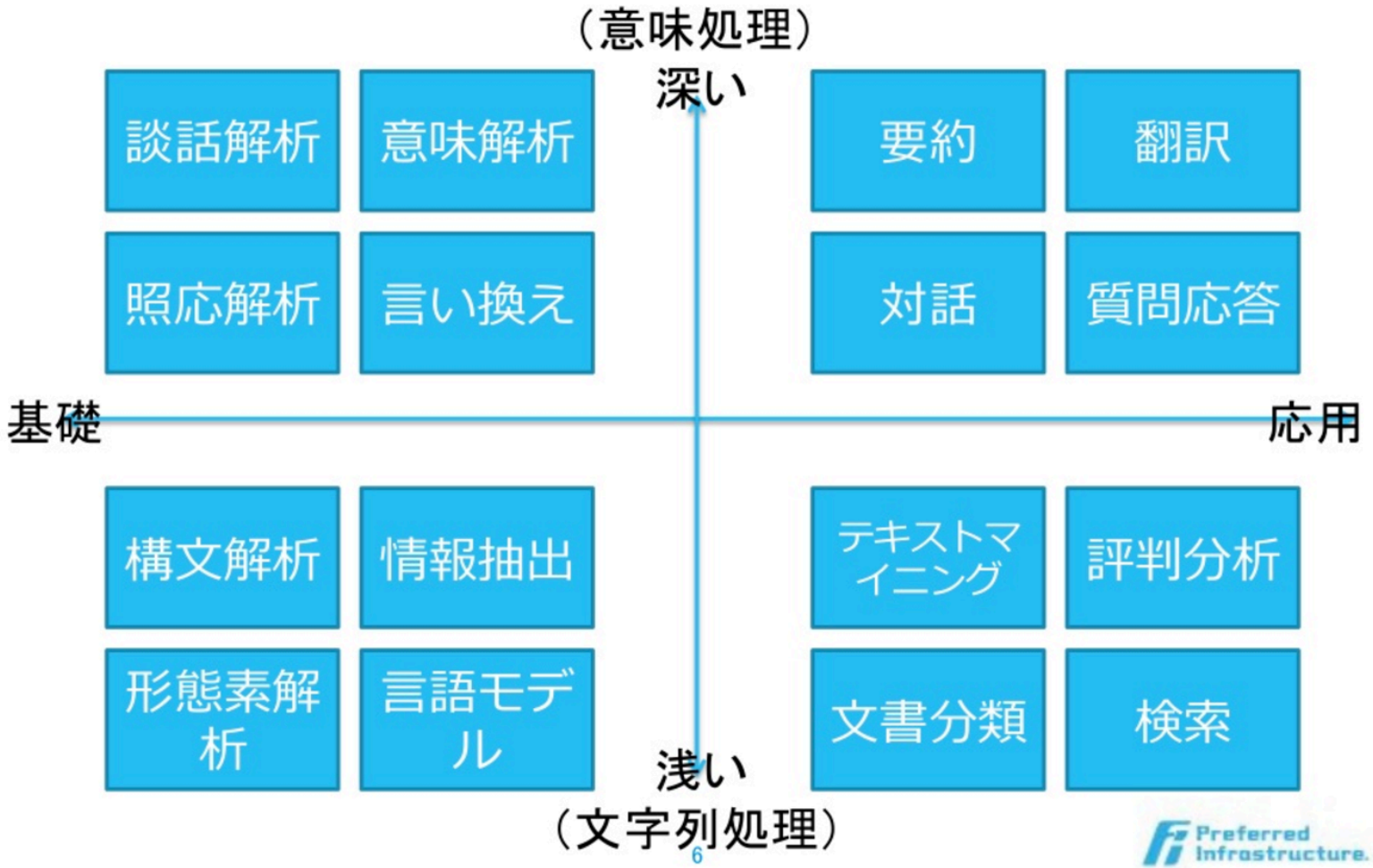
自然言語処理 #とは

自然言語処理（しぜんげんごしょり、英語: natural language processing、略称：NLP）は、人間が日常的に使っている**自然言語**をコンピュータに**処理**させる一連の技術であり、人工知能と言語学の一分野である。

Wikipedia より

- 機械翻訳
- かな漢字変換(IME)
- 感情推定
- 自動要約
- データマイニング (マーケティング)
- 検索エンジン
- スパムメールフィルタ
- 質問応答システム
- 音声認識，音声合成　etc.

自然言語処理 #とは



引用文献)http://sssslide.com/www.slideshare.net/unnonouno/20150316-ipsj-nlp

単語分割

- 日本語の文章は**単語同士が繋がっている**言語(対：スペースで区切られる英語)
- どんな自然言語処理を行うにあたっても、
  情報抽出をするために**単語分割**は必須の作業！

例)すももももももものうち

すもも / も / もも / も / もも / の / うち

単語分割

形態素解析とは

与えられた文を**形態素**(意味を持つ最小の言語単位)単位に区切り、
各形態素に**活用形・品詞などの情報**を付与する処理
(用は、単語分割 + 単語の属性を明らかにすること)

例)すももももももものうち

すもも	も	もも	も	もも	の	うち
名詞 (一般)	助詞 (係助詞)	名詞 (一般)	助詞 (係助詞)	名詞 (一般)	助詞 (連帯体)	名詞 (非自立)

オープンソース形態素解析エンジン MeCabを使ってみよう

- 開発者：奈良先端技術大学 工藤拓 (2002)
- ホームページ：https://taku910.github.io/mecab/

- MaCabのインストール

```
% tar zxfv mecab-X.X.tar.gz
% cd mecab-X.X
% ./configure
 % make
% make check
% su
# make install
```

MeCabの使い方

Terminalから起動してね

```
% mecab ←MeCabを起動
すももももももものうち ←形態素解析したい文章を入力
すもも 名詞,一般,*,*,*,*,すもも,スモモ,スモ
も 助詞,係助詞,*,*,*,*,も,モ,モ
もも 名詞,一般,*,*,*,*,もも,モモ,モモ
も 助詞,係助詞,*,*,*,*,も,モ,モ
もも 名詞,一般,*,*,*,*,もも,モモ,モモ
の 助詞,連体化,*,*,*,*,の,ノ,ノ
うち 名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ EOS
```

MeCabをPythonで使えるようにしよう

- MeCabのPythonバインディングをインストール

  以下のリンクからmecab-python-[バージョン].tar.gz をインストール
  https://drive.google.com/drive/folders/0B4y35FiV1wh7fjQ5SkJETEJEYzlqcUY4WUlpZmR4dDlJMWI5ZUlXN2xZN2s2b0pqT3hMbTQ

- インストール後に以下のコマンドを実行

```
% python setup.py build
% sudo python setup.py install
```

  その際“setup.py” に関するエラーが出た場合の
  修正方法は以下のサイトを参考に
  https://qiita.com/grachro/items/4fbc9bf8174c5abb7bdd

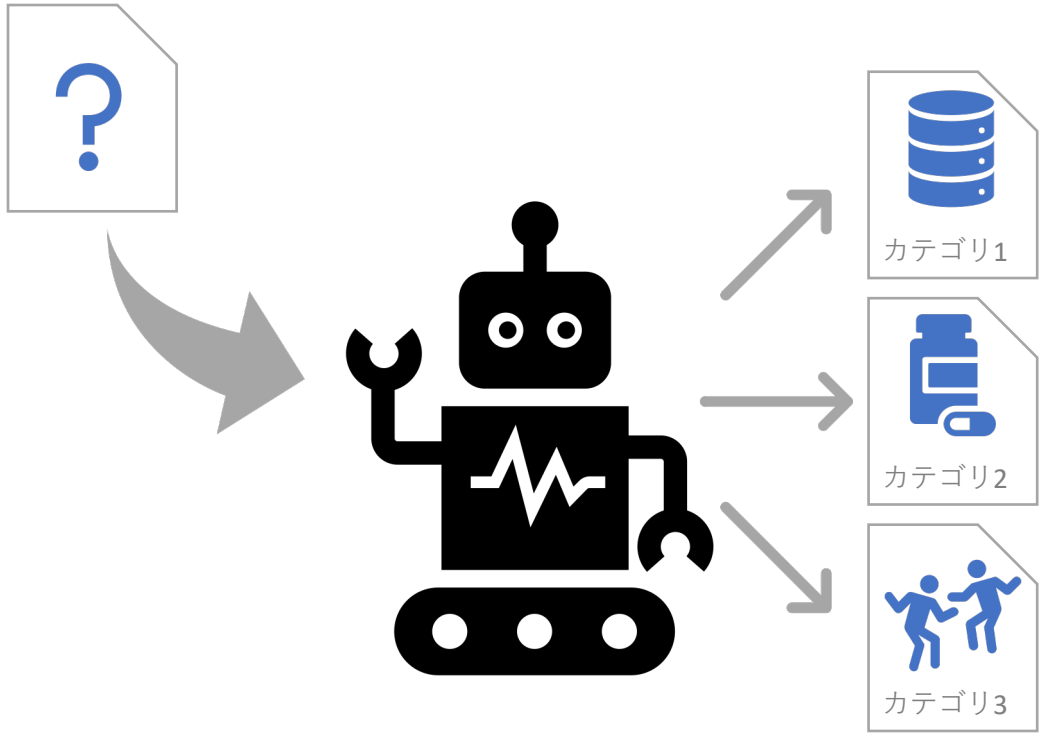
MeCabをPythonで使えるようにしよう

- **注意：**pip install MeCabによるインストールはお勧めしません
  - なぜかnode.surface()が正しく動かないので...

- インストールが完了したら、
  サンプルプログラム Macabexp.py を走らせてみよう

課題

機械学習による
テキスト自動分類に
挑戦！



データセットとタスク

- **Baseconnect.in**上にある企業の説明文データセット
- 3カテゴリの企業データ計1,500件

x 500 **x 500** **x 500**

タスク：説明文からカテゴリを自動で推定する分類器の実装！

(説明変数) (目的変数)

データセット(companies.csv)の概要

	index	company	category	training /test	description
文書0	0	日立製作所	IT	training	情報・通信システム、社会・産業システム、電子装置・システム、建設機械、...
文書1	1	日本電気株式会社	IT	training	政府や官公庁など、公共機関向けのネットワーク技術および、センサ技術やデ...
文書2	2	株式会社リコー	IT	training	主として、複合機やファクシミリ、プロジェクターといったオフィス用品や産業...
⋮					
文書1498	1498	株式会社ＳＣＲＥＥＮ...	event	test	マニュアル作成の他に販促ツール制作、展示会、Web、広報・IR支援や多言語...
文書1499	1499	株式会社群馬トヨタ総...	events	test	住宅展示場イベントなどのイベントの企画および設営のほか、テントや机...

カラムの説明

index：ただの数字(あるだけで今回は使わない)
company：会社の名前(あるだけで今回は使わない)
category：各会社のカテゴリ“IT” “medical” “event” のいずれか。それぞれ500件ずつ
training/test：“training” “test” のいずれか。詳しくは課題3で説明
description：各会社の説明文

テキスト自動分類のために必要な処理は？

課題(1)
形態素解析

医薬品を中心に、化粧品や健康食品・一般医療機器などの

↓

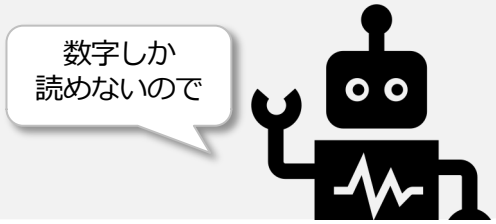
医薬品/を/中心/に/、/化粧品/や/健康食品/・/一般/医療機器/など/の

課題(2)
文書のベクトル化

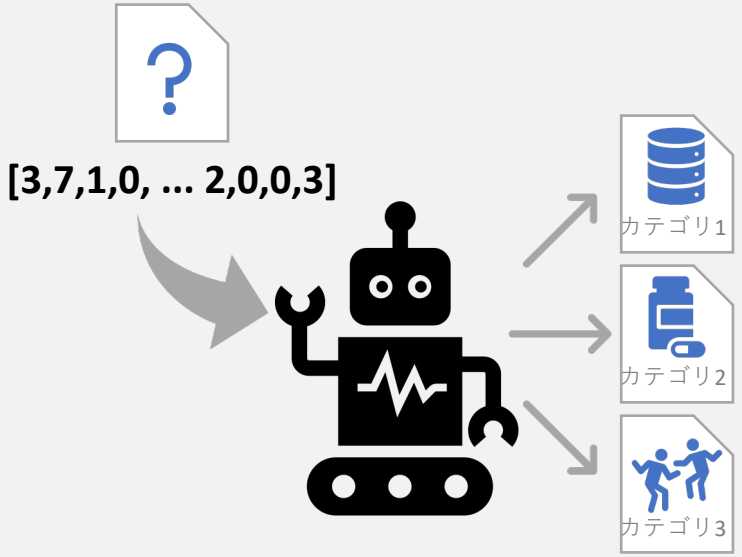
医薬品/を/中心/に/、/化粧品/や/健康食品/・/一般/医療機器/など/の

↓

[3,7,1,0, … 2,0,0,3]



課題(3)
機械学習による テキストの自動分類



課題(1)形態素解析

MeCabを使って、データセットの各文章を
二種類の方法で単語分割(形態素解析)してみよう！

1. 普通の単語分割(companies_tokenized1.csv)
2. 名詞・動詞・形容詞以外を削除し、
   動詞・形容詞は原型に統一した単語分割(companies_tokenized2.csv)

医薬品を中心に、化粧品や健康食品・一般医療機器などの製造を行っている。

1. 医薬品 を 中心 に 、 化粧品 や 健康食品 ・ 一般 医療機器 など の 製造 を 行っ て いる
2. 医薬品 中心 化粧品 健康食品 一般 医療機器 製造 行う いる

課題(1)完成のイメージ

完成のイメージ(companies_tokenized1.csv)

	index	company	category	training /test	description
文書0	0	日立製作所	IT	training	情報 ・ 通信システム 、 社会 ・ 産業 システム 、 電子 装置 ・ システム 、 建設機械
文書1	1	日本電気株式会社	IT	training	政府 や 官公庁 など 、 公共機関 向け の ネットワーク 技術 および 、センサ 技術 や
文書2	2	株式会社リコー	IT	training	主として 、 複合機 や ファクシミリ 、 プロジェクター といった オフィス用品 や 産
⋮					
文書1498	1498	株式会社ＳＣＲＥＥＮ...	event	test	マニュアル 作成 の 他 に 販促 ツール 制作 、 展示会 、 Web 、 広報 ・ IR 支援 や
文書1499	1499	株式会社群馬トヨタ総...	events	test	住宅展示場 イベント など の イベント の 企画 および 設営 の ほか 、 テント や 机

“住宅展示場␣イベント␣など␣の␣イベント...

形態素を半角スペースで区切って、一つの文字列として保存してね

課題(2)

機械学習による分類を行うためには、
各文書を何かしらの方法で**ベクトル表現**する必要がある

[3,7,1,0,9,1 … 0,0,2,0,0,3]

今回は課題(1)で分かち書きしたコーパスを用いて、
各文書に**どの単語が何回出現するかをカウントする行列**
(= Bag of Words) を作成することによって、
文書のベクトル化を行います

課題(2)完成のイメージ

- 最初に**コーパス全体に何種類の単語が存在するか(＝語彙数n)**を知る必要がある
- 完成の概要(companies_bow1.csv / companies_bow2.csv)

n個

	index	company	category	training /test	クラウド	ネットワーク	薬	使う	病院	…	ビジネス	ベトナム	バイオ
文書0	0	日立製作所	1	training	2	3	0	1	0		1	0	0
文書1	1	日本電気株式会社	1	training	1	1	0	3	0		1	1	0
文書2	2	株式会社リコー	1	training	1	0	0	2	0		0	1	0
⋮													
文書1498	1498	株式会社ＳＣＲ…	0	test	0	0	3	4	2		2	0	0
文書1499	1499	株式会社群馬…	0	test	1	0	2	3	1		1	0	2

各単語のカラム表示はしてもしなくてもよい

(重要なのは分布なので)

課題(3)

サンプルプログラム “classification.py” を動かして、
機械学習による自動分類が行われる様子を見てみよう！