

Deep Learningを用いた効率的な特許調査

2018.11.26

－発明の要素毎の根拠箇所抽出と適合判定への応用－

アジア特許情報研究会 1)
花王 知的財産部
安藤俊幸

要旨

最近ではAIの中心技術である各種機械学習のオープンソースライブラリが容易に入手可能である。特許調査担当者の実務的な観点から機械学習を用いた効率的な特許調査の可能性について述べる。先行技術調査ではdoc2vecによる公報文書単位のスコアで査読の優先順位を付け、文単位で発明の要素毎の類似文抽出検討を行い、13種類の教師あり分類アルゴリズムで適合判定を検討した。文単位の類似文抽出で記載の根拠箇所特定の可能性が示せた。動向調査では教師あり機械学習の1次元CNNによる文書分類と教師なしの次元圧縮による文書の可視化検討を行った。文書分類はSDI調査の効率化を目指している。調査目的に応じたアルゴリズムと特徴量の選択が重要である。教師あり機械学習には良質な教師データの準備が重要である。

目次

- 概要 調査目的×アルゴリズム×ドメインデータ
- 特許調査への機械学習適応時の留意点
- 先行技術調査の流れ(進め方)
- doc2vecによる公報(文書)単位の類似度計算
- doc2vecによる発明の要素(文)単位の類似度計算
- Deep Learningの基礎検討
- Deep Learningによる文書分類
- 単語ベクトルの合成による文書のベクトル化検討(3種類)

1

発表の概要: 調査目的 × アルゴリズム × ドメインデータ

調査目的

○ 先行技術調査

←
発明の構成要素毎に
パッセージ検索(記載箇所)

○ 無効資料調査(上記同様)

○ SDI調査

←
査読/ノイズ(2値分類)で
調査効率向上

○ 技術動向調査

- ←
- ・文書分類(自社分類)で
調査効率向上
 - ・次元圧縮で俯瞰・可視化

○ クリアランス調査

→
網羅性重視が必須のため
リスクと調査効率バランス

調査目的
×
アルゴリズム
×
ドメインデータ

調査目的に
合わせた
アルゴリズムと
ドメインデータの
選択と最適化を
行い
学習済モデルを
作成・利用する

自然言語文書の分析方法アルゴリズム

◎ 前処理 クレンジング(不要語除去)
分かち書き(形態素解析)

◎ 単語ベクトル化 word2vec
fastText

◎ 文書・文ベクトル化

- ・One hotベクトル(古典的)
- ・doc2vec
- ・SCDV
- ・因子解析
- ・RNN

◎ 適合判定 査読/ノイズ(2値分類)
・13種類のアルゴリズム
R→Python

◎ 文書分類
・1次元CNN

◎ データセット
・教師データ有 支援用辞書
・教師データ無 専門用語
・化学物質名辞書 等

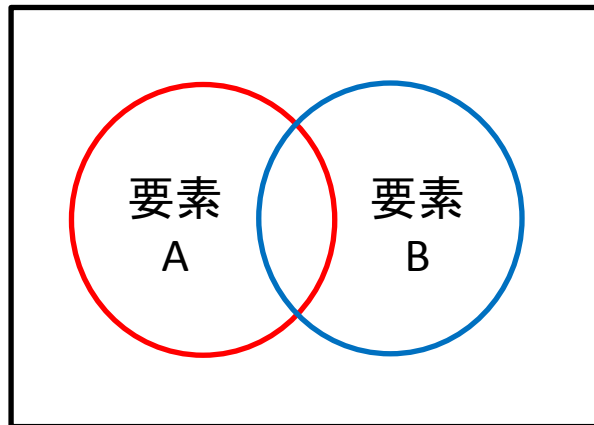
精度
調査効率

再現率
網羅性(漏れ防止)

特許調査への機械学習適応時の留意点

(1) フレーム問題

フレーム問題とは、人工知能における重要な難問の一つで、有限の情報処理能力しかないロボットには、現実には起こりうる問題全てに対処することができないことを示すものである。



そこで枠(フレーム)を作って、その枠の中だけで思考する。

外枠・フレーム

検索において調査範囲を決める
外枠と考えると理解しやすい

(2) ノーフリーランチ定理(NFL 定理)

- ・最適化問題であらゆる問題に適用できる性能の良い**万能のアルゴリズムは無い**
 - ・ある特定の問題に焦点を合わせた専用アルゴリズムの方が性能が良い
- 特許調査に当てはめると調査目的に合った適切な機械学習のアルゴリズムを選択することが重要である。**

ノーフリーランチ定理の名前の由来: 酒場で「**ドリンク注文で昼食無料**」の広告、
実はドリンク代金に昼食も含まれている → 「**教師データ作成、検証**」等の**コストも重要**

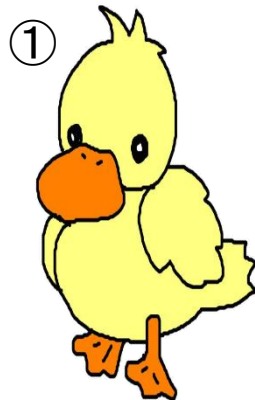
(3) 醜いアヒルの子の定理

「醜いアヒルの子を含むn匹のアヒルがいるとする。このとき醜いアヒルの子と普通のアヒルの子の類似性は任意の二匹の普通のアヒルの子の間の類似性と同じになる」

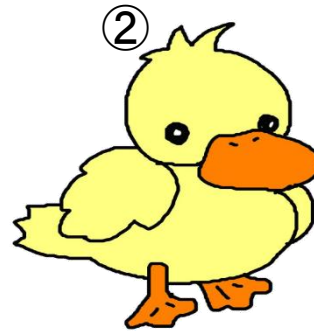
- ・純粋に客観的な立場からはどんなものを比較しても同程度に似ているとしか言えない
- ・各特徴量を全て同等に扱っていることにより成立する定理



醜いアヒルの子



普通のアヒルの子



将来有望な特許を見つけない！？

	体色	背の高さ	横幅	体重	目の開閉	嘴の向き	尾の向き
醜いアヒルの子	灰色	高	細	軽	開	左	右
普通のアヒルの子①	黄色	低	細	軽	開	左	右
普通のアヒルの子②	黄色	低	太	重	開	右	左

特徴量として**体色**に着目すると**類似性**は異なる

(4) シンボルグラウンディング問題

シンボルグラウンディング問題とは、記号システム内の**シンボル**がどのようにして**実世界の意味**と結びつけられるかという問題。記号接地問題とも言う。

現在の「AI」は人間と同じように自然言語を理解しているわけではないことに注意

出願したい明細書から**構成要素**を分析する **着目点**
※醜いアヒルの子の定理

明細書を熟読して**発明内容を理解**し、検索式作成のための
構成要素を決定する

※シンボルグラウンディング問題

予備検索の実行

調査範囲の把握: ※フレーム問題

特許分類(FI、Fターム、IPC)、キーワードの検討
海外の場合(IPC,CPC)

検索戦略立案、検索式作成

調査範囲の決定: ※フレーム問題
着目点: ※醜いアヒルの子の定理

検索式に使用する特許分類、キーワードの抽出
多観点の検索式の検討

スクリーニング過程を詳細に検討し、
機械学習を応用した支援方法(ツール)検討

検索実行、**スクリーニング**

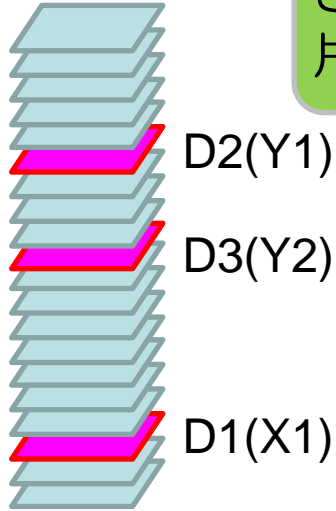
優先順位を決め、効率的にスクリーニングを行う
スクリーニング結果に応じて、検索戦略を再検討

- ・特許性(新規性、進歩性)判断を謳うAI **※シンボルグラウンディング問題、NFL定理**
- ・**人と機械(AI)の役割分担**が重要 ← 本研究のスタンス

スクリーニング課程の現状と理想

現状

どの文献が当たりか判らないので、片端から読み込む



D2(Y1)

D3(Y2)

D1(X1)

AI技術(機械学習、深層学習、自然言語処理)の活用による効率的な特許調査

理想

可能性の高い文献から順番に読みたい

先行文献にスコアが付いていれば、スコアの高い文献から読むことができる。

発明の構成要素毎に根拠個所を見たい

D1(X1)新規性
D2(Y1)進歩性(主)
D3(Y2)進歩性(副)



機械(AI)がなぜ根拠と判定したか理由を知りたい 7

先行技術調査の事例検討

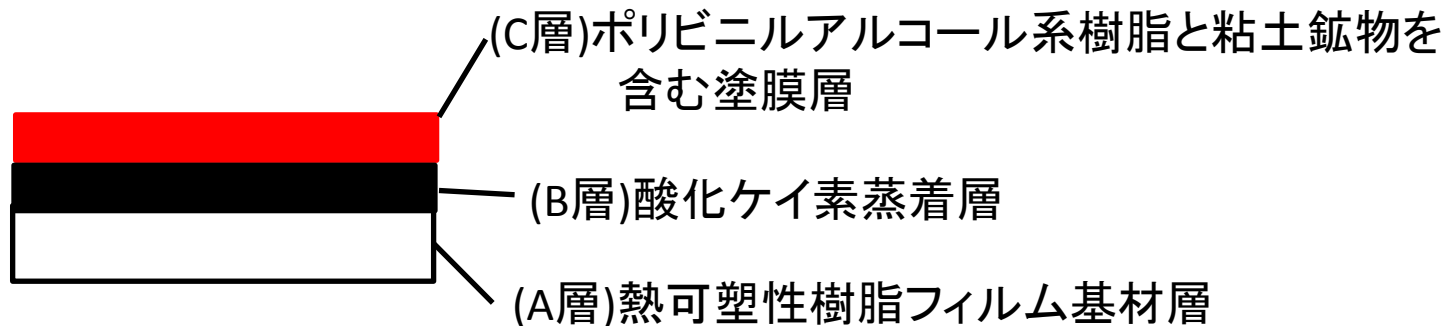
YEARBOOK2017
YEARBOOK2018

特許検索競技大会2016 化学・医薬分野
出題内容:【問2】問題文概要(2/3)

【特許請求の範囲】

【請求項1】

熱可塑性樹脂フィルム基材層(A層)、酸化ケイ素蒸着層(B層)、ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層(C層)が他の層を介して又は介さずにこの順に積層されてなることを特徴とするガスバリア性包装用フィルム。



ガスバリア性包装用フィルム

図4. 特許検索競技大会2016の化学・医薬分野の問2

分散表現(単語埋め込み)とは

← 固定長、数百次元、密ベクトル

分散表現(あるいは単語埋め込み)とは、単語を高次元の実数ベクトルで表現する技術
近い意味の単語を近いベクトルに対応させるのが分散表現の基本
ベクトルの足し算が意味の足し算に対応する「加法構成性」などを中心に、
理論や応用の研究が進んでいる。 例: **王様 - 男 + 女 = 女王** **King - Man + Woman = Queen**
(岩波 データサイエンス vol.2 [特集] 統計的自然言語処理 — ことばを扱う機械)

・局所表現(local representation)

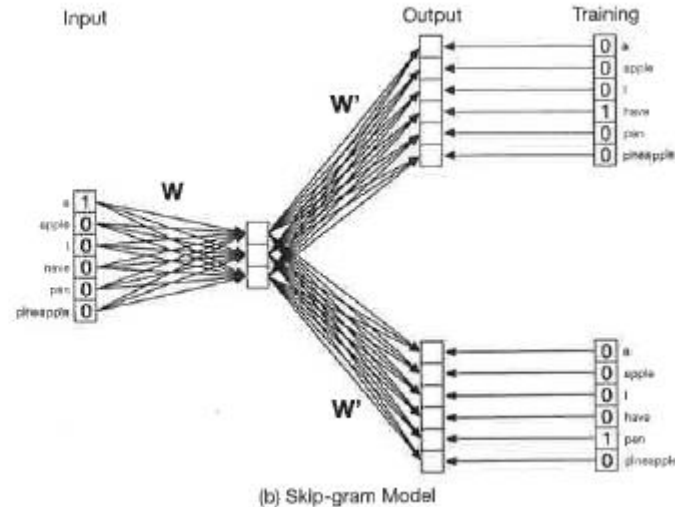
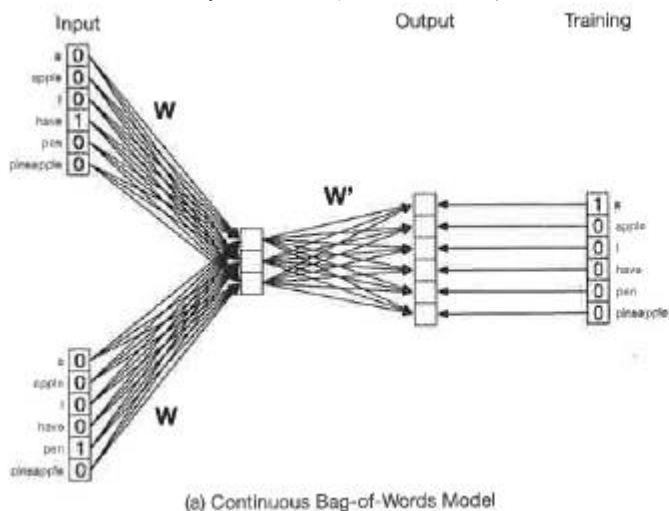
各単語(固有ID)に1つの次元 → 単語数(種類数)の高次元ベクトル(one hotベクトル)

・分散表現(distributed representation)

各概念(単語)は複数のニューロンで表現される
各ニューロンは複数の概念の表現に関与する

← **スパース(疎)ベクトル
数万次元**

word2vecのニューラルネットワーク



分布仮説に基づいた文脈中の単語の重み学習 (word2vec)

分布仮説

- ・類似する文脈でよく使われる表現は似た意味を持つ
- ・単語の意味はその周辺単語の分布により知ることができる

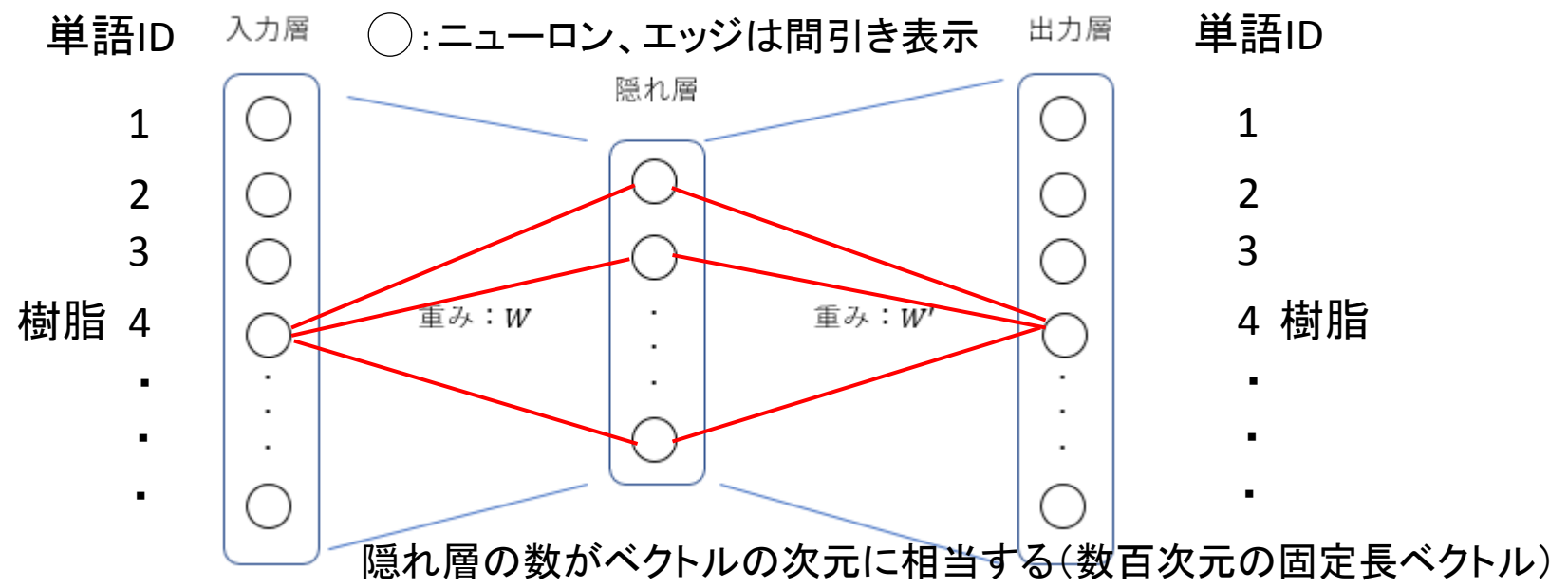
学習例

熱可塑性樹脂フィルム基材層、酸化ケイ素蒸着層、ポリビニルアルコール系樹脂...

1 2 3 4 5 6 7 8 9 10 11 8 12 13 4
熱/可塑/性/樹脂/フィルム/基/材/層/酸化/ケイ素/蒸着/層/ポリビニルアルコール/系/樹脂

ウィンドウ幅: 5

- ①注目単語の前後の周辺単語を学習/予測する
- ②周辺単語から注目単語を学習/予測する



色の表現方法の局所表現 (one hotベクトル) とRGB表示

色の表現方法

局所表現に相当 色番号=単語ID
色名称=単語

インデックスカラー

RGB表示

←分散表現に相当

色番号 色名称

整数表示

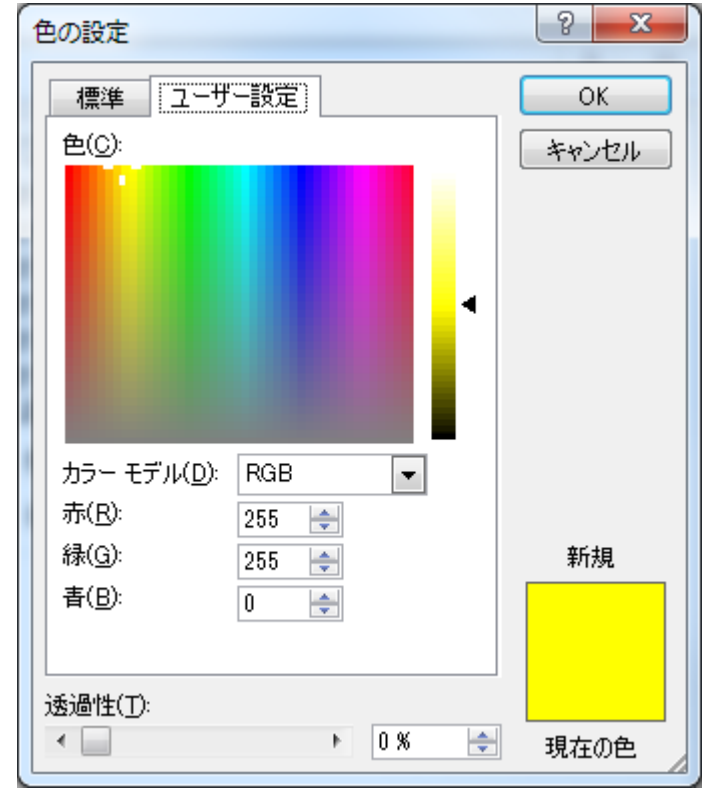
実数表示

- 1 赤 Red
- 2 青 Blue
- 3 緑 Green
- 4 黄色 Yellow
- 5 赤紫 Magenta
- 6 水色 Cyan

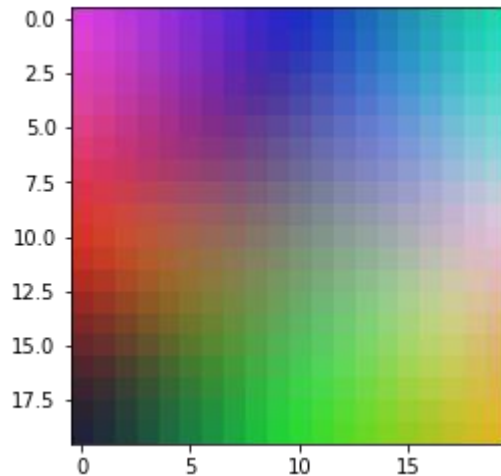
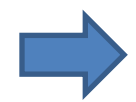
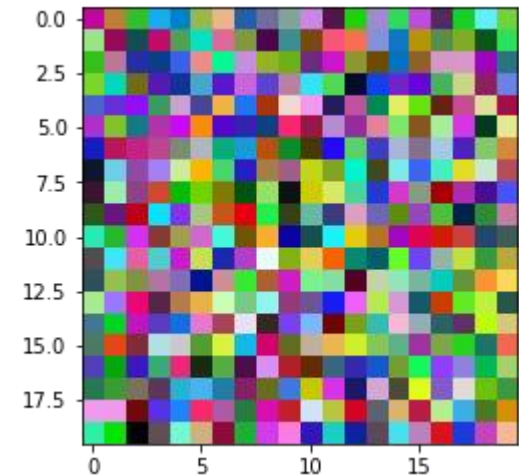
色番号	色名称	整数表示	実数表示
1	赤 Red	255,0,0	1.0, 0.0, 0.0
2	青 Blue	0,255,0	0.0, 1.0, 0.0
3	緑 Green	0,0,255	0.0, 0.0, 1.0
4	黄色 Yellow	255,255,0	1.0, 1.0, 0.0
5	赤紫 Magenta	255,0,255	1.0, 0.0, 1.0
6	水色 Cyan	0,255,255	0.0, 1.0, 1.0

↑
色名称の数の次元数
フルカラー $256^3 \doteq 1600$ 万

3次元の固定長ベクトル



RGB表示データを使った自己組織化マップ(SOM)による並び替え

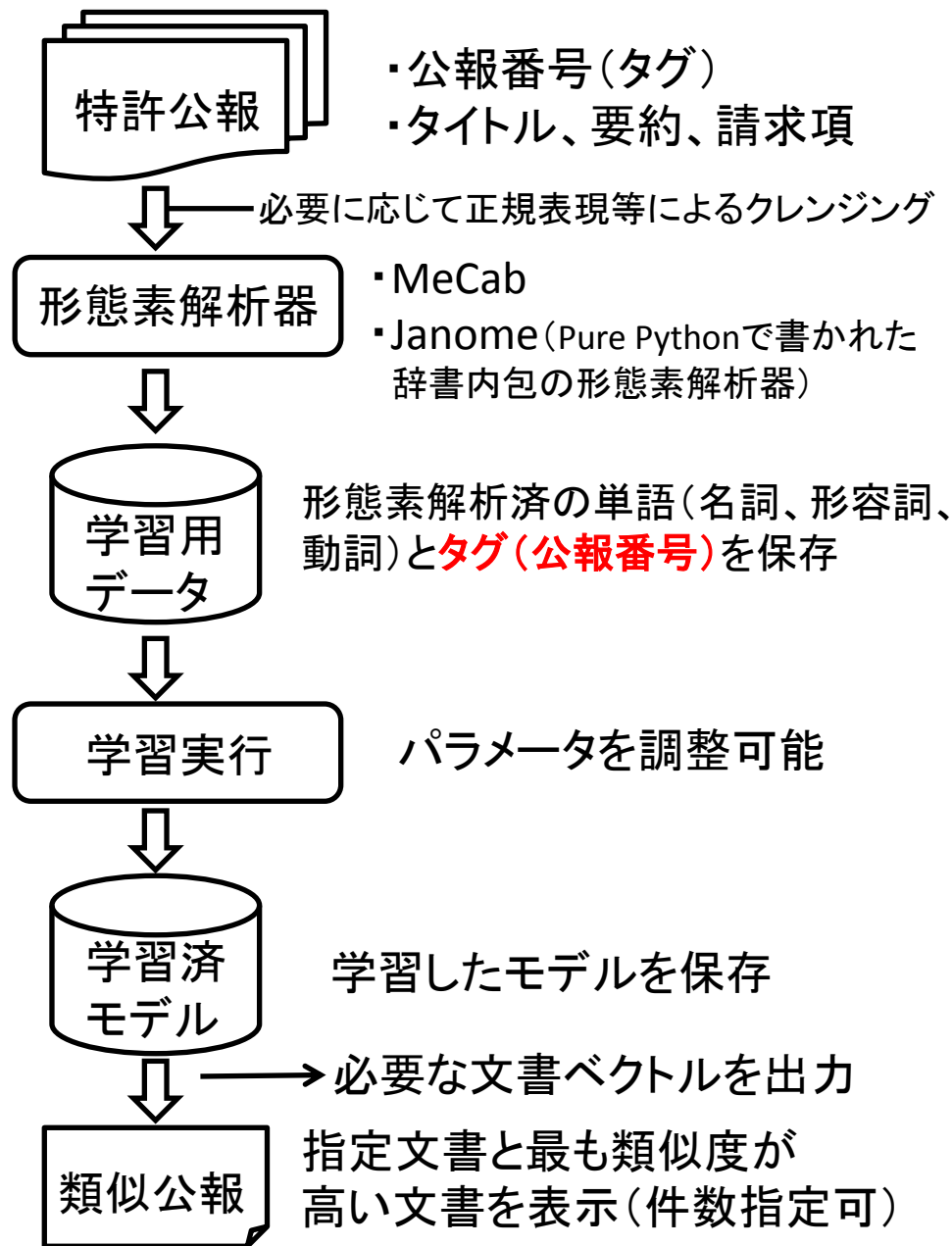


類似色が近くに配置

Pythonで自己組織化マップ NumPy版
https://qiita.com/T_Shinaji/items/609fe9aabd99c287b389

doc2vecによる文書のベクトル化処理の概要

YEARBOOK2018



改良ポイント (YEARBOOK2018)

- ① 公報を**文単位**に分解して**タグ付け**
- ② **実施例**追加
- ③ クエリ: 請求項1、構成要素a~g

タグ付け詳細

公報番号_記載部分:文番号

例:P2001-123456_c6

記載部分略号

T:タイトル

A:要約

C:請求項

E:実施例

構成要素分析(検索競技大会の模範解答例)

YEARBOOK2018

熱可塑性樹脂フィルム基材層、酸化ケイ素蒸着層、ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層が他の層を介して又は介さずにこの順に積層されてなることを特徴とするガスバリア性包装用フィルム。

正解例と解説:【問2】(1)構成要素分析

(1)調査依頼された請求項1に対して、検索すべき技術の構成要素(概念)を記述しなさい。

構成要素に対する重み付け※醜いアヒルの子の定理

記号	構成要素(概念)	重み1	重み2
a	熱可塑性樹脂フィルム基材層	10%	5%
b	酸化ケイ素蒸着層	20%	30%
c	ポリビニルアルコール系樹脂を含む塗膜層	10%	10%
d	塗膜層に粘土鉱物を含む	30%	30%
e	他の層を介してまたは介さずにこの順に積層	5%	1%
f	ガスバリア性	15%	19%
g	包装用フィルム	10%	5%

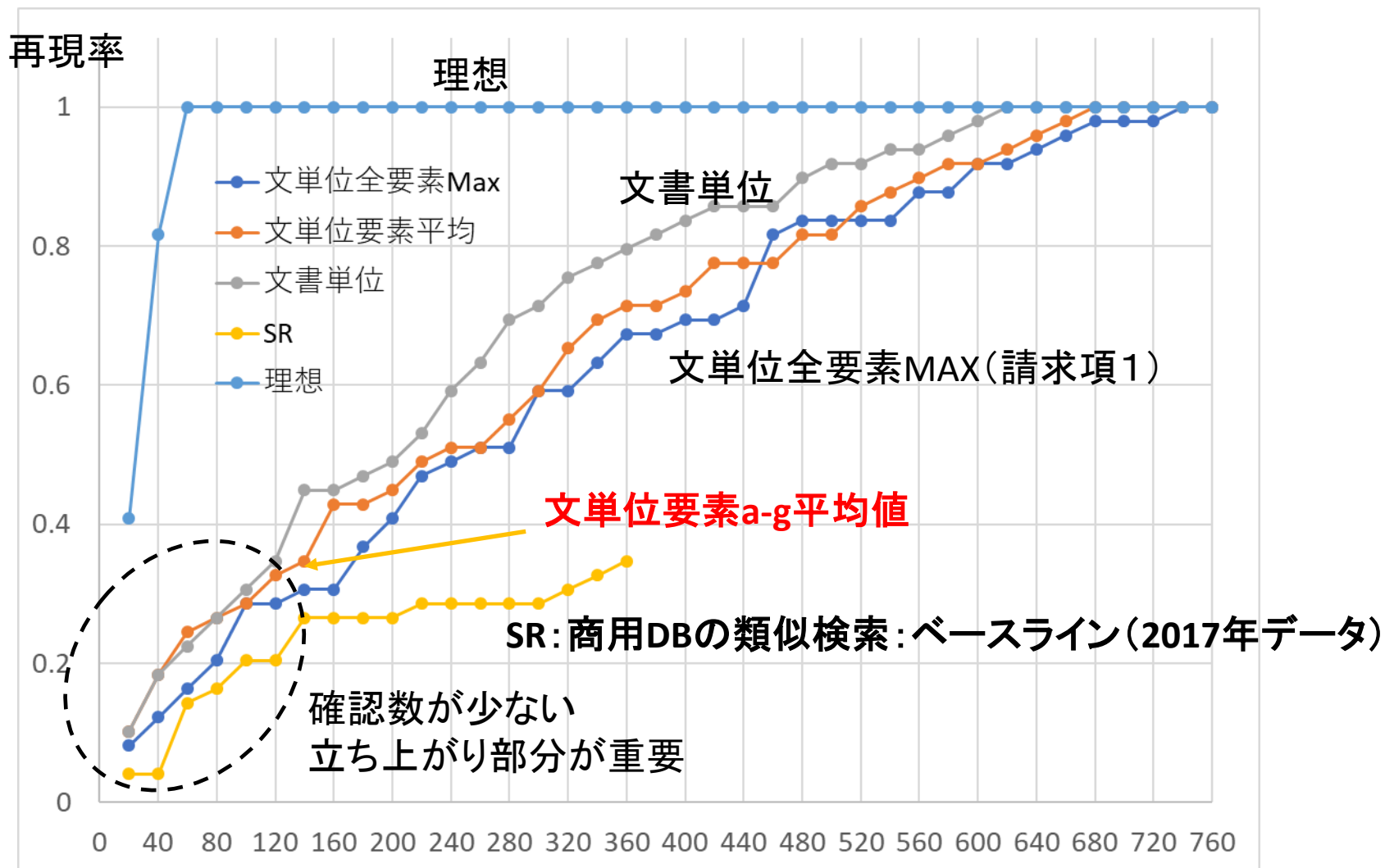
※構成要素の分け方は本例に限定しない

同じ重みだと
1/7=14.3%

図6. 構成要素分析(検索競技大会の模範解答例)

「文」単位での類似度計算による再現率曲線

YEARBOOK2018



立ち上がり部(スコア上位)では「**文単位要素a-g平均値**」が
再現率上は差はないが**根拠箇所が参照**できるので有効

確認数

発明の構成要素の重み付け検討

YEARBOOK2018

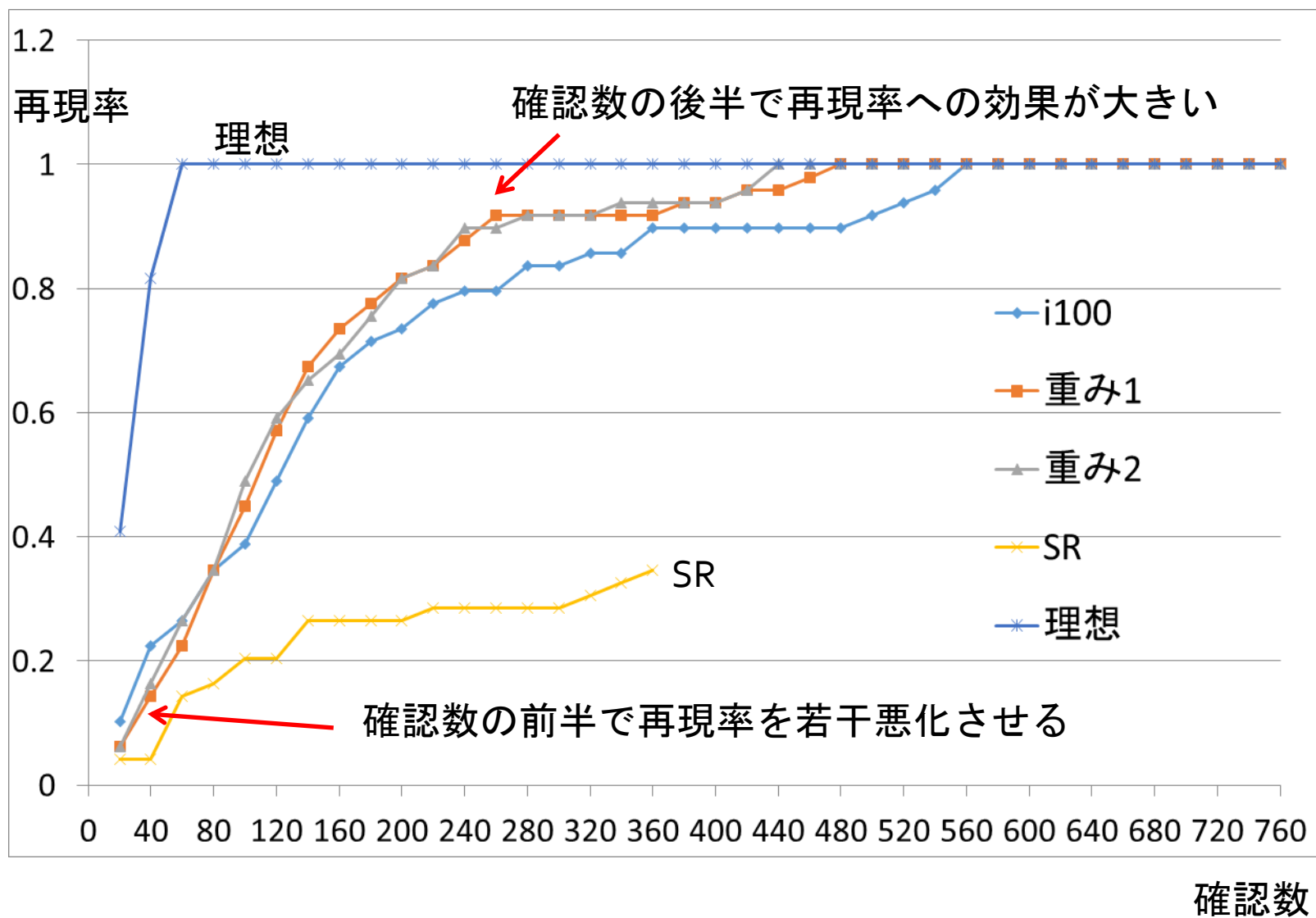


図9. 発明の構成要素の重み付け検討

文の分節とクエリ拡張の影響(クエリ)

実施例を請求項2、請求項3としてクレーム化してクエリ拡張

YEARBOOK2018

PatNo	TACE
P0_T1	ガスバリア性包装用フィルム。
P0_A1	ポリプロピレン、ポリエチレンテレフタレート、ナイロンなどの熱可塑性樹脂からなるフィルムは、透明性、耐熱性を有するため様々な用途に広く用いられている。
P0_A2	しかし酸素や水蒸気バリア性能が求められる用途、例えば鮮度が求められる食品のパッケージ用途には適さない。
P0_A3	そのため、従来から熱可塑性樹脂フィルムとアルミニウム箔とを積層したフィルムが食品用のパッケージとして用いられてきた。
P0_A4	しかしアルミニウム箔を積層したフィルムは、ガスバリア性能は優れる一方で、フィルムの向こう側が視認不能となる上、金属探知機の使用ができなくなるという問題がある。
P0_A5	これらの問題を解決するフィルムとして、熱可塑性樹脂フィルムに酸化ケイ素等の無機酸化物を蒸着したものが開発されているが、そのガスバリア性能は鮮度が求められる食品の保存用途としては十分でなかった。
P0_A6	そこで、酸化ケイ素蒸着層の上にポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層を設けることで、これらの問題を解決したガスバリア性包装用フィルムの発明に至った。
P0_C1	熱可塑性樹脂フィルム基材層、酸化ケイ素蒸着層、ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層が他の層を介して又は介さずにこの順に積層されてなることを特徴とするガスバリア性包装用フィルム。
P0_C2	熱可塑性樹脂がポリプロピレン、ポリエチレンテレフタレート、ナイロンから選ばれた請求項1記載のガスバリア性包装用フィルム。
P0_C3	粘土鉱物がカオリナイト、ディッカイト、ナクライト、ハロイサイト、アンチゴライト、クリソタイル、ヘクトライト、パイロフィライト、モンモリロナイト、白雲母、マーガライト、タルク、パーミキュライト、金雲母、ザンソフィライト、緑泥石から選ばれた請求項1記載のガスバリア性包装用フィルム。
P0_E1	ポリビニルアルコール水溶液に、モンモリロナイトを加え60℃で75分間攪拌した。
P0_E2	その後、さらに2-プロパノールを添加し、その混合液を室温まで冷却して塗工液を得た。
P0_E3	熱可塑性フィルム基材として厚さ15μ mのポリエチレンテレフタレートフィルムを用い、この一方の面上に酸化ケイ素を蒸着した。
P0_E4	蒸着層の上に塗工液をグラビアコート法により形成し、ガスバリア性包装用フィルムを得た。
P0a_C1	熱可塑性樹脂がポリプロピレン、ポリエチレンテレフタレート、ナイロンから選ばれた熱可塑性樹脂フィルム基材層。
P0b_C1	酸化ケイ素蒸着層。
P0c_C1	ポリビニルアルコール系樹脂を含む塗膜層。
P0d_C1	粘土鉱物がカオリナイト、ディッカイト、ナクライト、ハロイサイト、アンチゴライト、クリソタイル、ヘクトライト、パイロフィライト、モンモリロナイト、白雲母、マーガライト、タルク、パーミキュライト、金雲母、ザンソフィライト、緑泥石から選ばれた粘土鉱物を含む塗膜層。
P0e_C1	他の層を介してまたは介さずにこの順に積層。
P0f_C1	ガスバリア性。
P0g_C1	包装用フィルム。

クエリ拡張

記載部分略号

T: タイトル

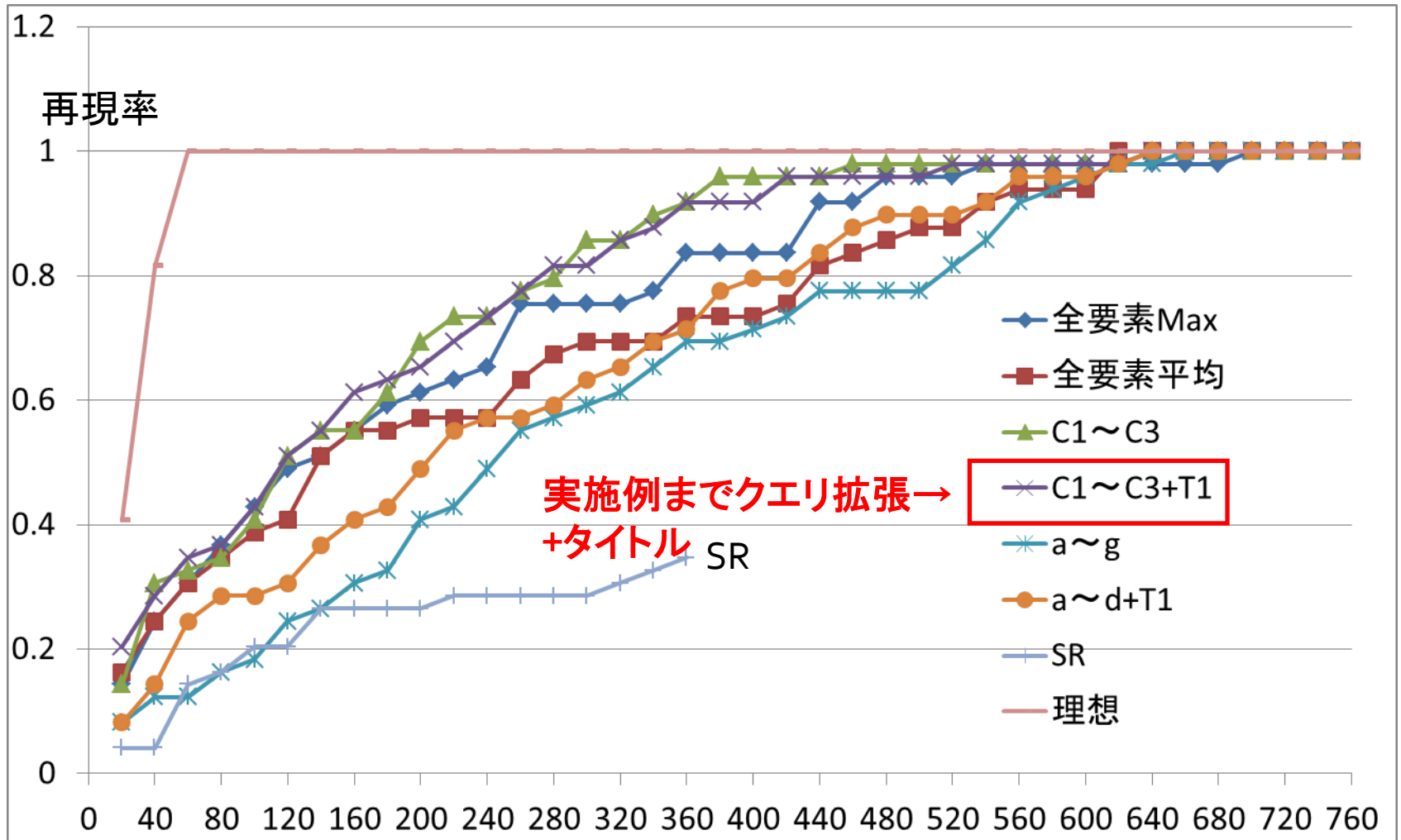
A: 要約

C: 請求項

E: 実施例

図10. 分の文節とクエリ拡張の影響(クエリ)

文の分節とクエリ拡張の影響(結果)



実施例までクエリ拡張+タイトルがスタート時点の再現率も良い

確認数

図 1 1. 文の文節とクエリ拡張の影響 (結果)

発明の構成要素毎の根拠箇所(文)抽出結果

各構成要素の最大類似度「文」の平均値で**順位2位** P1998-076325

正解公報

構成要素	記載部	類似度	該当文	適合
a	E94	0.728	さらに、これらの 熱可塑性樹脂基材 は、透明であることが好ましい。	○
b	E99	0.595	金属及び／または 金属酸化物 は特に限定されないが、アルミニウム、 ケイ素 、亜鉛、マグネシウムなどの金属及び／または 金属酸化物 であることが好ましい。	○
c	E55	0.523	さらに、本発明では塗膜中に架橋剤を含んでいてもよい。	×
d	E125	0.489	塗膜 の構成成分を含んだ塗剤は、溶媒に 無機板状粒子 が均一に分散もしくは膨潤しかつ水溶性または水分散性ポリマーが均一に溶解もしくは分散した溶液が好ましい。	○
e	E140	0.511	フィルム走行装置を具備した真空蒸着装置内にフィルムをセットし、冷却ドラムを介して走行させる。	×
f	E217	0.714	ガスバリア性 に特に優れるフィルムが得られた。	○
g	T1	0.633	ガスバリア フィルム 及び 包装材料	○

構成要素 **平均値: 0.599**

a: 熱可塑性樹脂フィルム基材層

b: 酸化ケイ素蒸着層

c: ポリビニールアルコール系樹脂を含む塗膜層

d: 塗膜層に粘土鉱物を含む

e: 他の層を介してまたは介さずにこの順に積層

f: ガスバリア性

g: 包装用フィルム

記載部分略号

T: タイトル

A: 要約

C: 請求項

E: 実施例



・現在人が判定

・機械で判定を
検討

教師あり分類アルゴリズムの適合判定への応用検討

スパムメール識別の正解率の箱ひげ図

隠れ層5層のDeep Learning

隠れ層(250,500,100,50,20)h2o

分類器

Dataset:Rパッケージkernlab
英文メール4601通
(スパム1813、非スパム2788)
57個の特徴項目
(単語、数値、記号)
10フォールド交差確認法

隠れ層1層のニューラルネットワークaveNNet

ロジスティック判別 Logit

ガウシアンカーネルSVM svm_Rad

多項式カーネルSVM svm_Poly

ランダムフォレスト RF

1番良い→ エイダブースト Boost

バギング Bagg

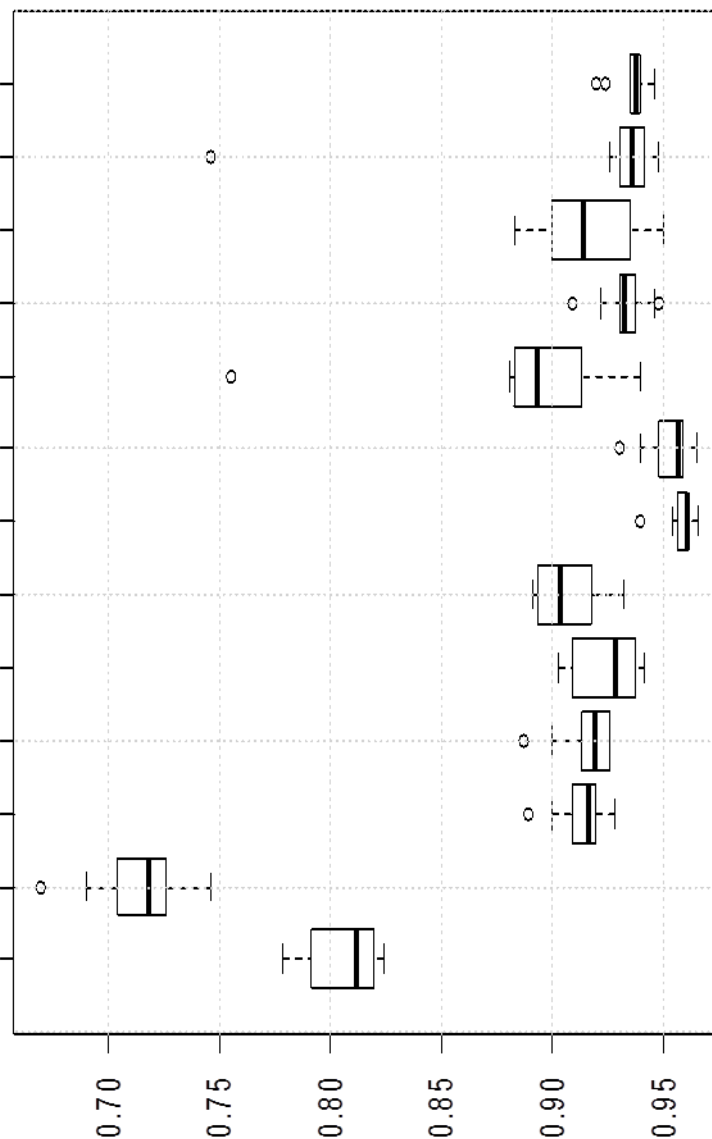
決定木C5.0 C50

CARTエントロピー rp_Entr

CART Gini係数 rp_Gini

ナイーブベイズ NBayes

k近傍法(k=7) k_NN



正解率

word2vecによる「粘土」の類似語抽出

YEARBOOK2017

分散表現ベクトル

word2vec「粘土」の類似語

順位	類似語	類似度
1	スメクタイト	0.774
4	サポナイト	0.646
5	ヘクト	0.637
7	スチーブン	0.630
8	ナイト	0.615
9	マイカ	0.614
11	モンモリロナイト	0.599
12	カオリ	0.597
14	タルク	0.587
16	ゼオライト	0.561
17	セリ	0.554

One hotベクトル

形態素 専門用語抽出

順位	頻度	専門用語	順位	頻度
555	26	スメクタイト	1655	7
2101	4	サポナイト	4655	2
2099	2	ヘクトライト	4656	2
2100	2	スチーブンサイト	4703	2
1448	4	カオリナイト	2669	4
1449	4	マイカ	3441	3
359	53	モンモリロナイト	246	52
1635	3	カオリナイト	2669	4
1446	4	タルク	2691	4
1175	7	ゼオライト	1652	7
2184	4	セリサイト	5112	2

黄色セルは形態素解析による分かち書きに失敗しているが類似語として上位に存在している

主な粘土鉱物(Wikipedia)

カオリナイト(高陵石)
スメクタイト
モンモリロン石(モンモリロナイト)
絹雲母(セリサイト)
イライト
海緑石(グローコナイト)
緑泥石(クロライト)
滑石(タルク)
沸石(ゼオライト)

<https://ja.wikipedia.org/wiki/粘土鉱物>

専門用語抽出(続き)

専門用語	順位	頻度
水素型スメクタイト	1657	7
合成スメクタイト	1979	6
スメクタイト族	3864	2
スメクタイト群粘土鉱物	4002	2
スメクタイト粘土鉱物	4740	2
合成マイカ	7890	1
カオリン	7203	1

word2vecを使用すると文脈に「粘土」の記載のない文からも具体的な粘土鉱物を学習しており検索クエリの拡張支援ツールとして有用である

自作PatAnalyzerによる「粘土」の文脈語集計

PatAnalyzer Ver.1.3.29

テキスト入力部

そこで、酸化ケイ素蒸着層の上にポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層を設けることで、これらの問題を解決したガスバリア性包装用フィルムの発明に至った。
 熱可塑性樹脂フィルム基材層、酸化ケイ素蒸着層、ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層が他の層を介して又は介さずにこの順に積層されてなることを特徴とするガスバリア性包装用フィルム。
 粘土鉱物がカオリナイト、ディッカイト、ナクライト、ハロイサイト、アンチゴライト、クリンタイト、ヘクトライト、パイロフィライト、モンモリロナイト、白雲母、マーガライト、タルク、パーミキュライト、金雲母、ザンソフィライト、緑泥石から選ばれた請求項1記載のガスバリア性包装用フィルム。
 粘土鉱物がカオリナイト、ディッカイト、ナクライト、ハロイサイト、アンチゴライト、クリンタイト、ヘクトライト、パイロフィライト、モンモリロナイト、白雲母、マーガライト、タルク、パーミキュライト、金雲母、ザンソフィライト、緑泥石から選ばれた粘土鉱物を含む塗膜層。
 市販品としては、スメクタイト系の粘土鉱物に属するサポナイト構造を有するスメクトンSA(クニミネ工業社製)、ソジウム型のモンモリロナイトであるクニピアーフ(クニミネ工業社製)、合成マイカPDM-5B(トピー工業社製)、精製された天然ベントナイトであるベンゲル(豊順洋行製)等を用いることができる。
 無機層状鉱物が天然の粘土の場合、結晶層間に、ナトリウムやカルシウムなどのカチオン性の無機イオンが

解析結果

分詞開始(中文)

そこで 接続詞,*,*,*,*,そこで,ソコデ,ソコデ
 、 記号,読点,*,*,*,*,ハハハ
 酸化 名詞,サ変接続,*,*,*,*,酸化,サンカ,サンカ
 ケイ素 名詞,一般,*,*,*,*,ケイ素,ケイソ,ケイソ
 蒸着 名詞,サ変接続,*,*,*,*,蒸着,ジョウチャク,ジョウ
 チャク
 層 名詞,接尾,一般,*,*,*,*,層,ソウ,ソー
 の 助詞,連体化,*,*,*,*,の,ノ,ノ
 上 名詞,非自立,副詞可能,*,*,*,*,上,ウエ,ウエ
 に 助詞,格助詞,一般,*,*,*,*,に,ニ,ニ
 ポリビニルアルコール 名詞,一般,*,*,*,*,
 系 名詞,接尾,一般,*,*,*,*,系,ケイ,ケイ
 樹脂 名詞,一般,*,*,*,*,樹脂,ジュシ,ジュシ

集計結果

構成	層	62
配位	面体構造	62
面体構造	構成	62
膨潤性	粘土化合物	36
ノントロナイト	サポナイト	34
-	配位	34
溶媒	膨潤性	33
水	配位	33
配位	吸収	33
パイ	ライト	32
層	対	31
ライト	ノントロナイト	31
対	結合	31

処理文数=480 KW抽出=6387 処理時間: 9200ms

Textファイル出力フォルダ

クエリ

粘土

正規表現

文字列サーチ

戻す

サーチ

文抽出

文末:改行

抽出

693 語 812ms

文字色

色設定

コピー

背景色

解析言語

中国語 日本語

Excel読込

一括処理

和布葉解析

形態素

専門用語

形態素+専門用語

分析用(文単位)

和布葉

分詞出力(類似率)

出力(類似率)

480文

Cabocha

トータル:480文

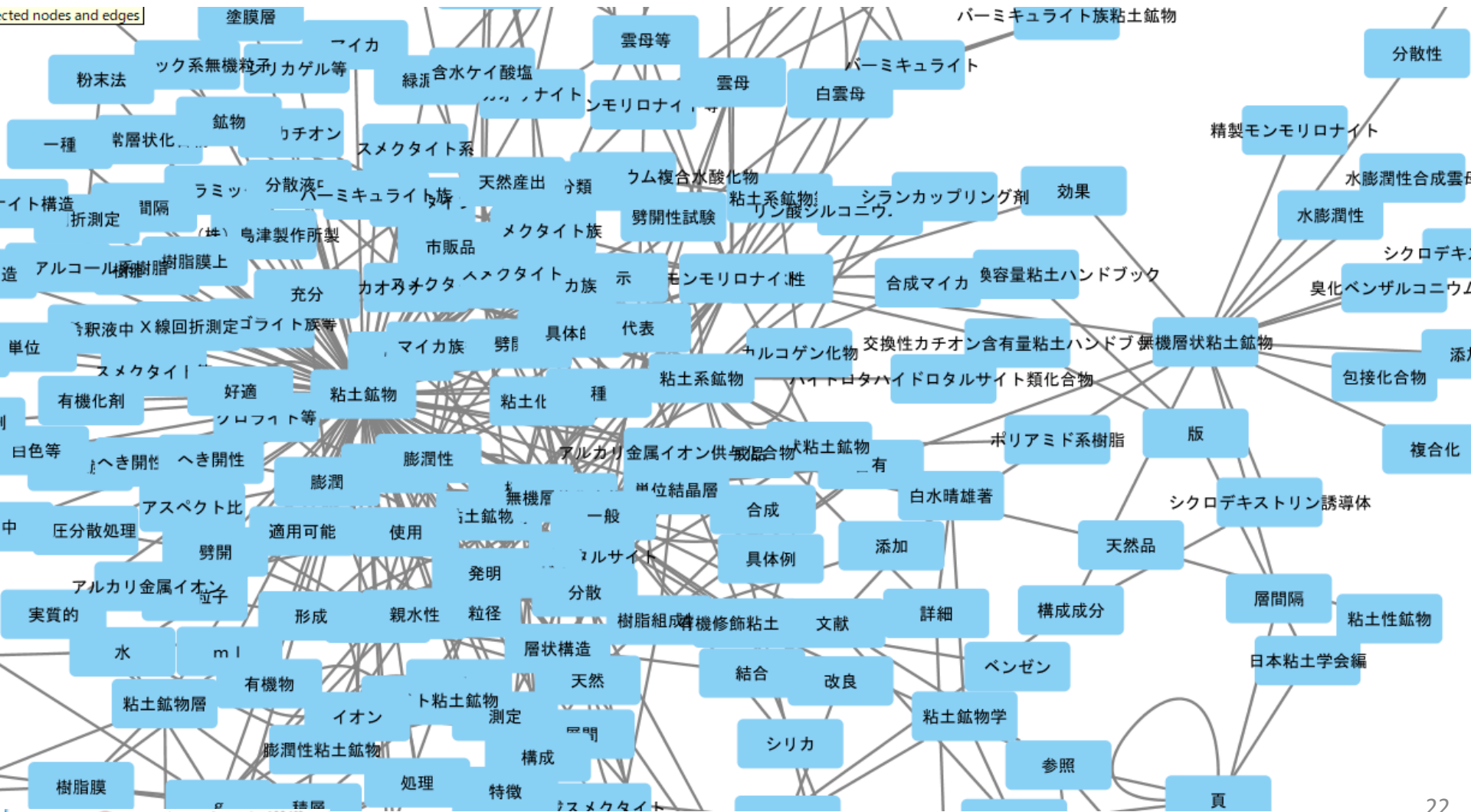
統計出力

参照

Excel2010対応

Cytoscapeによる「粘土」の文脈語のネットワーク分析

Cytoscapeはオープンソースのバイオインフォマティクスソフトウェアプラットフォームである。
特許公報の引用/被引用、単語/専門用語のノードとエッジによるネットワークグラフ可視化分析にも使用できる。



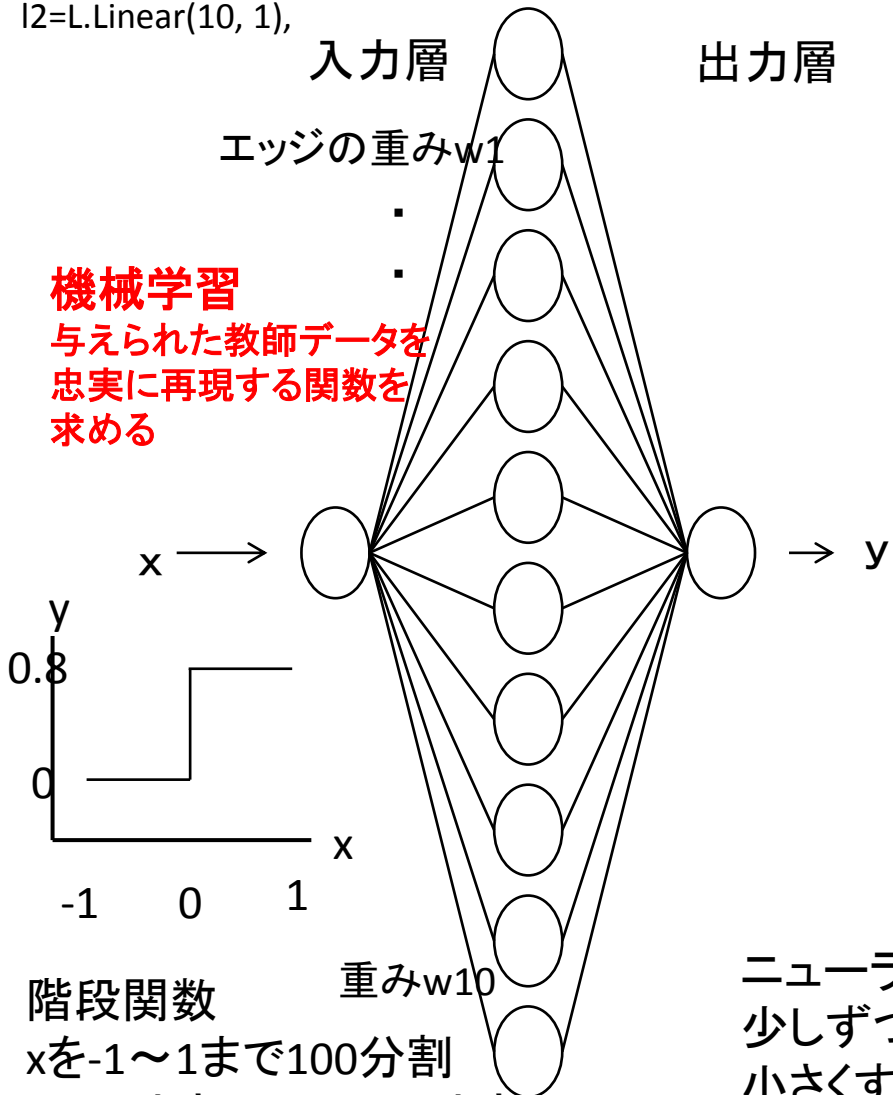
<https://cytoscape.org/>
Cytoscape (Wikipedia)
<https://ja.wikipedia.org/wiki/Cytoscape>

今津均.Cytoscapeによる特許情報のネットワーク解析とビジュアル化
https://www.jstage.jst.go.jp/article/johokanri/54/8/54_8_463/_pdf

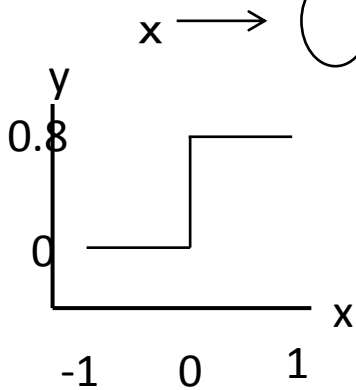
Chainerの多層パーセプトロンによる階段関数の学習

Chainerの層の定義
l1=L.Linear(1, 10),
l2=L.Linear(10, 1),

多層パーセプトロン(3層)
中間層(隠れ層)

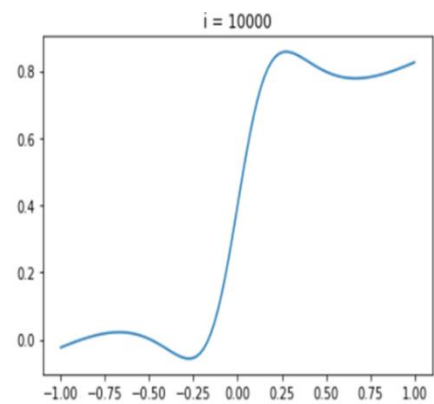


機械学習
与えられた教師データを
忠実に再現する関数を
求める

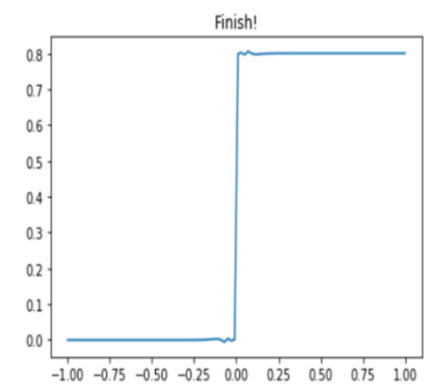


階段関数
xを-1~1まで100分割
x<0のときy=0,x>=0のときy=0.8

10000回「学習」

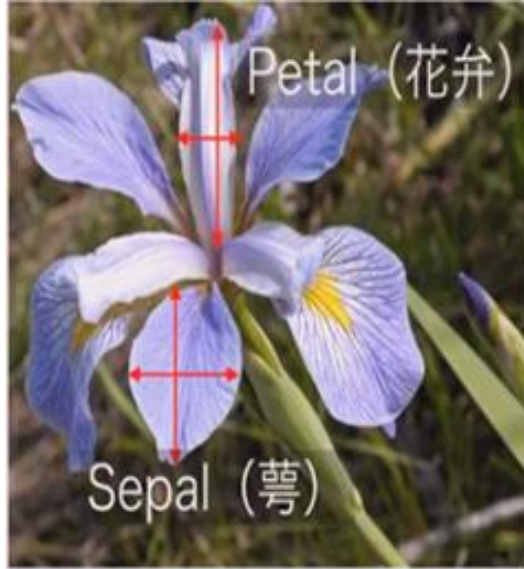


100000回「学習」



ニューラルネットワークの各エッジの重みを
少しずつ調整し、正解ラベルとの誤差を
小さくする事を、「学習」と呼ぶ

Iris (アヤメ) の品種分類 (Iris data set)



各花ごとに4つの測定値

- Sepal length がく片の長さ
- Sepal width がく片の幅
- Petal length 花弁の長さ
- Petal width 花弁の幅

品種



0: Setosa



1: Versicolor

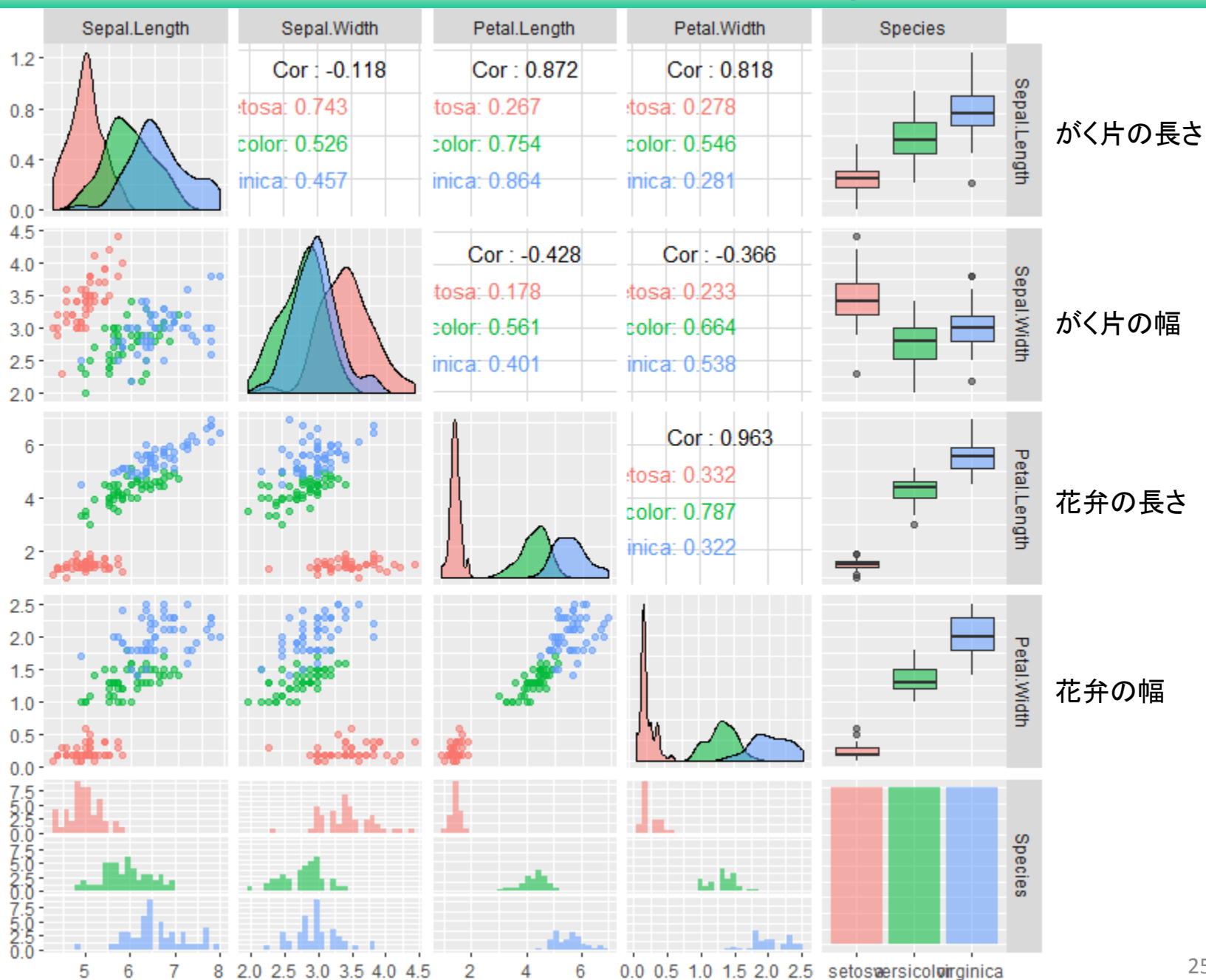


2: Versinica

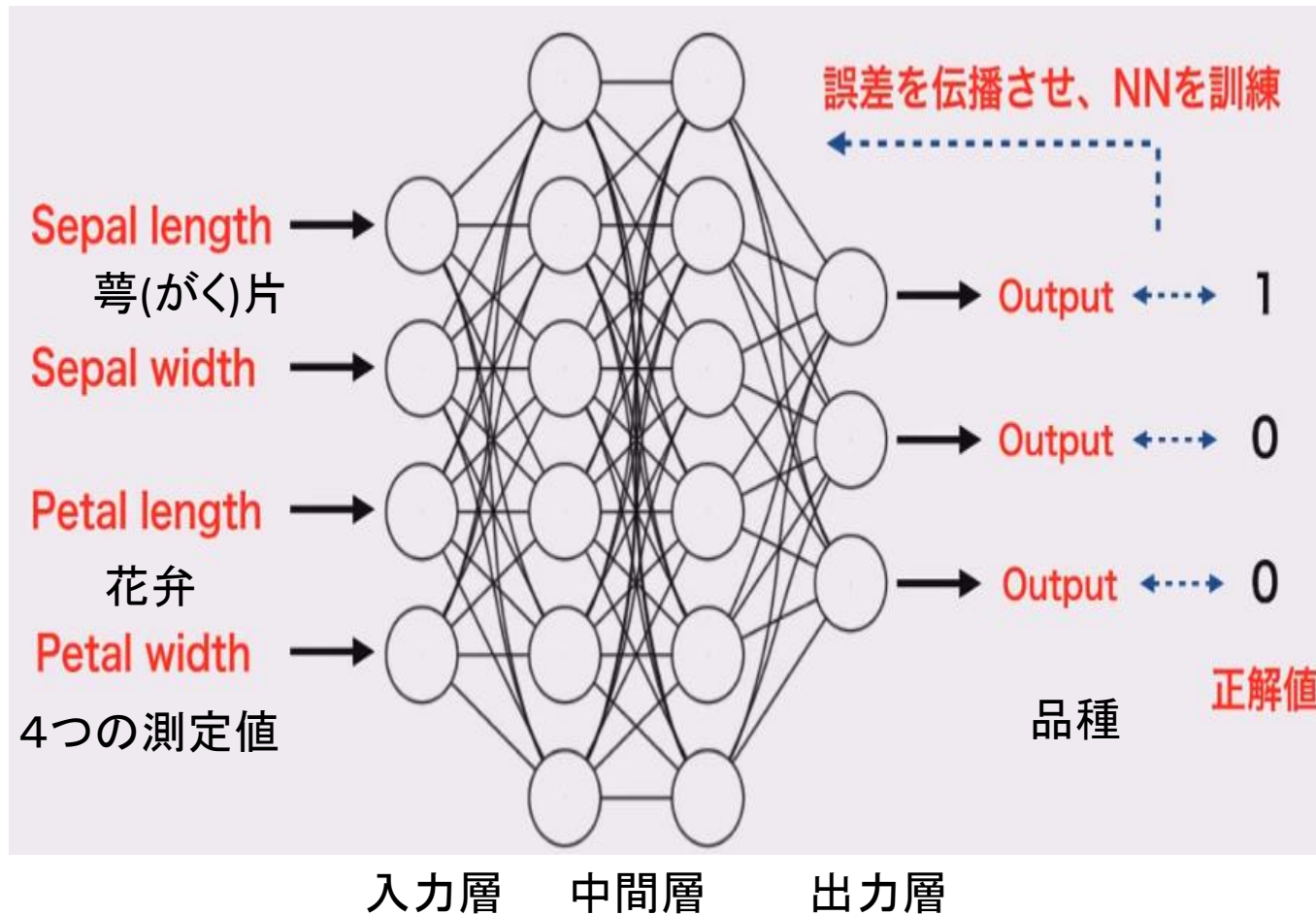
各花ごとに4つの測定値+3品種(教師データ) → 全部で150組

図14. Iris (アヤメ) の品種分類 (Iris data set)

ggplot2によるIris Data **setosa**/**versicolor**/**virginica**の可視化



ニューラルネットワークの訓練 (Chainer)



最低限の実装によるディープラーニング

- ・入力層には4つ、出力層には3つのニューロン
- ・4層からなるニューラルネットワークを訓練
- ・訓練済のニューラルネットワークによる品種分類

訓練済モデルによる品種分類実行例
Correct: 72 Total: 75 Accuracy: 96.0 %

図15. Chainerによるニューラルネットワークの訓練と分類実行

NTTデータ数理システム **Deep Learner**の データタイプ・学習別処理内容と特色

	教師あり学習	教師なし学習
テーブル	分類分析・回帰分析	次元圧縮
時系列	系列を考慮した 分類分析・回帰分析 例) 時系列センサーデータ等	可変長の系列データから 固定長の次元圧縮表現を獲得
テキスト	テキストの分類分析	
特色	目的変数は数値、カテゴリを 問わず複数指定可能	次元圧縮により得た表現を VMSの他のアイコンで使用可能 例) クラスタリング、可視化等

Deep Learner

<http://www.msi.co.jp/deeplearner/>

Deep Learner は、ディープラーニング (Deep Learning, 深層学習) のモデルを
プログラミング不要で対話的に設計し、実行するためのモジュール

Deep Learnerの変数設定

ドラッグ & ドロップ

用途

- 予測
- 次元圧縮

データ形式

- テーブル
- 時系列
- テキスト

	列名	列属性	目的変数	説明変数
1	irisNo	整数	<input type="checkbox"/>	<input type="checkbox"/>
2	sepal length	実数	<input type="checkbox"/>	<input checked="" type="checkbox"/>
3	sepal width	実数	<input type="checkbox"/>	<input checked="" type="checkbox"/>
4	petal length	実数	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	petal width	実数	<input type="checkbox"/>	<input checked="" type="checkbox"/>
6	Category	カテゴリ	<input checked="" type="checkbox"/>	<input type="checkbox"/>

**Category(目的変数)の列属性に注意！
数値:回帰、テキスト:カテゴリ判別**

<< 戻る 次へ >> Cancel

ニューラルネットワークのモデルデザイン

Deep Learner

Deep Learning

モデルデザイン

出力層

全結合層2

全結合層1

入力層

全結合層パラメーター一覧

追加 削除

名前	出力次元数	活性化関数	Dropout Ratio
全結合層2	6	ReLU	0.0
全結合層1	6	ReLU	0.0

<< 戻る 次へ >> Cancel

Deep Learning Help

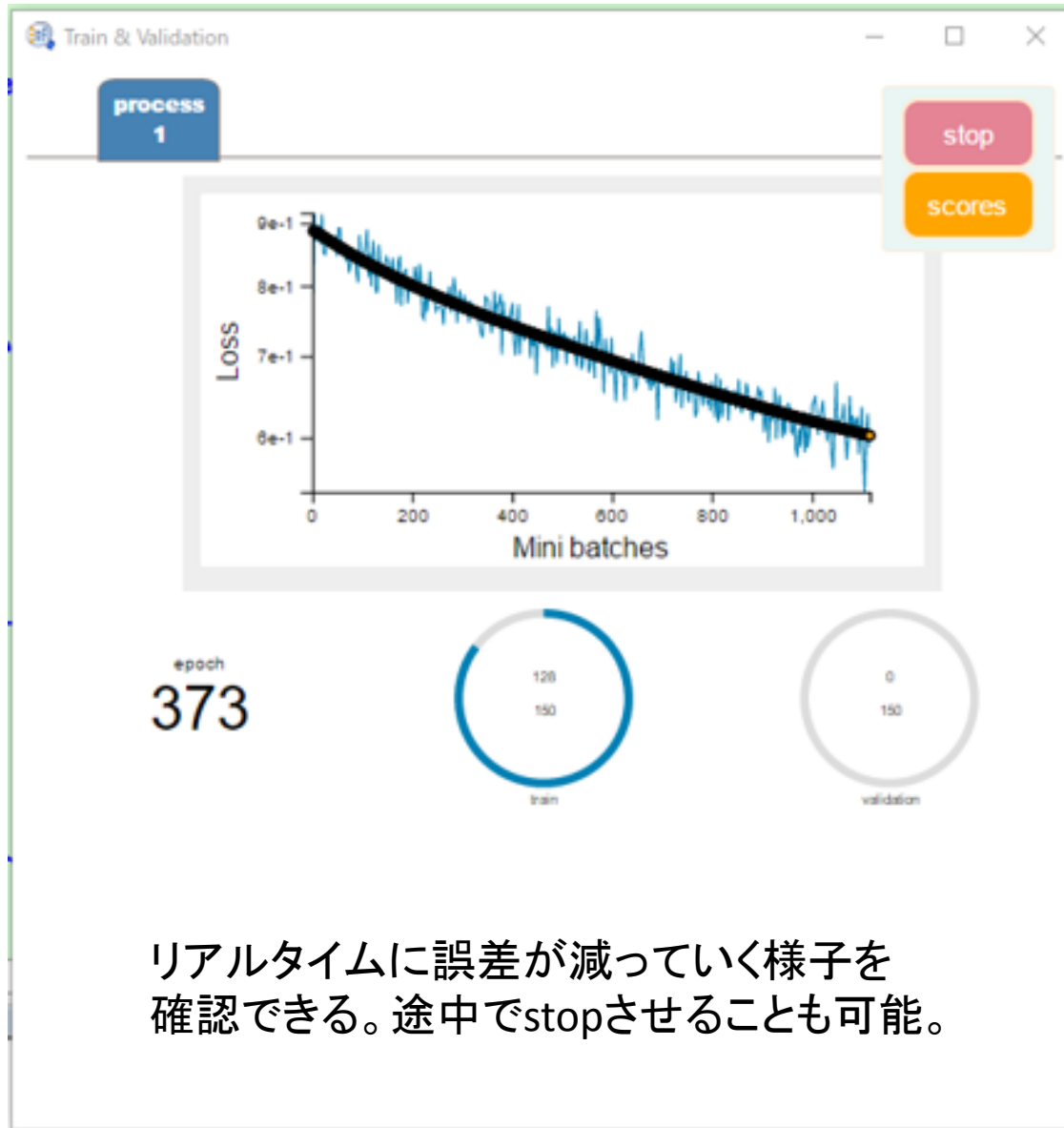
学習設定

学習設定	
ミニバッチサイズ:	64
SGD設定:	Adam
学習率:	0.0001
エポック数:	1000

Model Optimizer設定	
手法:	ランダム
指標:	Loss
最大探索数:	100
使用プロセス数:	1

環境設定	
GPU使用:	<input checked="" type="checkbox"/>
乱数の初期値:	<input checked="" type="radio"/> 自動
	<input type="radio"/> 手動 <input type="text" value="0"/>

<< 戻る 実行開始 Cancel



コンテンツ 🔍

result (150行/10列)

result

score

	No	sepal le...	sepal w...	petal le...	petal w...	Category	Category.予測	Category.c0	Category.c1	Category.c2
1	1	5.10	3.50	1.40	0.20	c0	c0	0.83	0.15	0.02
2	2	4.90	3.00	1.40	0.20	c0	c0	0.81	0.17	0.02
3	3	4.70	3.20	1.30	0.20	c0	c0	0.83	0.15	0.02
4	4	4.60	3.10	1.50	0.20	c0	c0	0.83	0.15	0.02
5	5	5.00	3.60	1.40	0.20	c0	c0	0.83	0.15	0.02
6	6	5.40	3.90	1.70	0.40	c0	c0	0.83	0.15	0.01
7	7	4.60	3.40	1.40	0.30	c0	c0	0.82	0.16	0.02
8	8	5.00	3.40	1.50	0.20	c0	c0	0.83	0.15	0.02
9	9	4.40	2.90	1.40	0.20	c0	c0	0.82	0.15	0.02
10	10	4.90	3.10	1.50	0.10	c0	c0	0.83	0.16	0.02

result (150行/10列)

1:Versicolorの予測結果一部抜粋

	No	sepal le...	sepal w...	petal le...	petal w...	Category	Category.予測	Category.c0	Category.c1	Category.c2
51	51	7.00	3.20	4.70	1.40	c1	c1	0.03	0.68	0.29
52	52	6.40	3.20	4.50	1.50	c1	c1	0.02	0.57	0.41
53	53	6.90	3.10	4.90	1.50	c1	c1	0.02	0.62	0.36
54	54	5.50	2.30	4.00	1.30	c1	c1	0.04	0.56	0.40
55	55	6.50	2.80	4.60	1.50	c1	c1	0.02	0.59	0.39
56	56	5.70	2.80	4.50	1.30	c1	c1	0.02	0.54	0.44
57	57	6.30	3.30	4.70	1.60	c1	c2	0.01	0.49	0.50
58	58	4.90	2.40	3.30	1.00	c1	c1	0.14	0.57	0.29
59	59	6.60	2.90	4.60	1.30	c1	c1	0.03	0.67	0.30
60	60	5.20	2.70	3.90	1.40	c1	c2	0.03	0.46	0.50

result (150行/10列) 2:Versinicaの予測結果一部抜粋

	No	sepal le...	sepal w...	petal le...	petal w...	Category	Category.予測	Category.c0	Category.c1	Category.c2
101	101	6.30	3.30	6.00	2.50	c2	c2	0.19	0.12	0.69
102	102	5.80	2.70	5.10	1.90	c2	c2	0.22	0.19	0.60
103	103	7.10	3.00	5.90	2.10	c2	c2	0.20	0.14	0.66
104	104	6.30	2.90	5.60	1.80	c2	c2	0.20	0.16	0.64
105	105	6.50	3.00	5.80	2.20	c2	c2	0.19	0.13	0.68
106	106	7.60	3.00	6.60	2.10	c2	c2	0.16	0.09	0.75
107	107	4.90	2.50	4.50	1.70	c2	c2	0.24	0.23	0.53
108	108	7.30	2.90	6.30	1.80	c2	c2	0.18	0.12	0.70
109	109	6.70	2.50	5.80	1.80	c2	c2	0.19	0.13	0.67
110	110	7.20	3.60	6.10	2.50	c2	c2	0.19	0.12	0.69

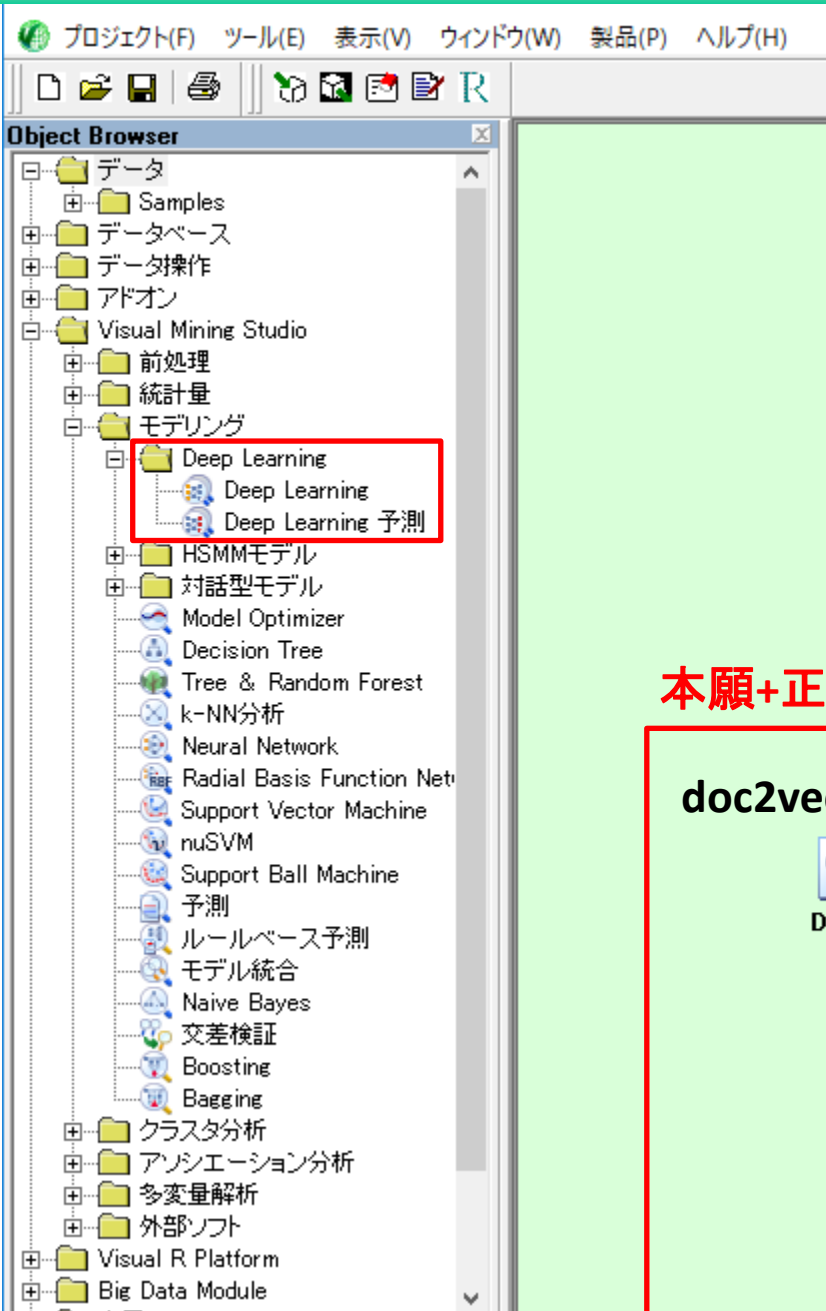
Iris (アヤメ) の品種分類 (Iris data set)

	c0 (予測)	c1 (予測)	c2 (予測)
c0	50	0	0
c1	0	47	3
c2	0	0	50

	誤答数	正答数	正答率
c0:Setosa	0	50	100.0%
c1:Versicolor	3	47	94.0%
c2:Versinica	0	50	100.0%
	3	147	98.0%

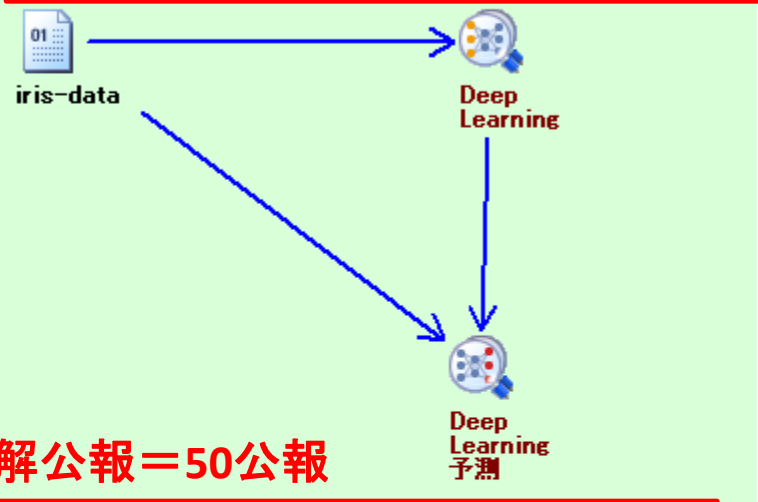
Deep Learningによる正解/ノイズの2値分類検討

YEARBOOK2018

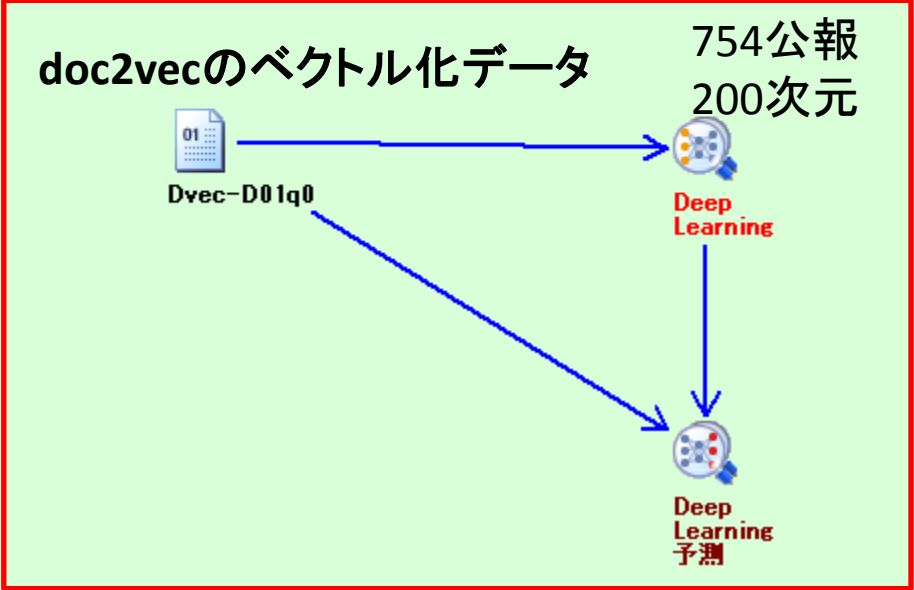


予備検討

	正答数	誤答数	正答率	精度	再現率
正解公報	30	20	60.0%	100.0%	60.0%
ノイズ公報	704	0	100.0%		
	734	20	97.3%		



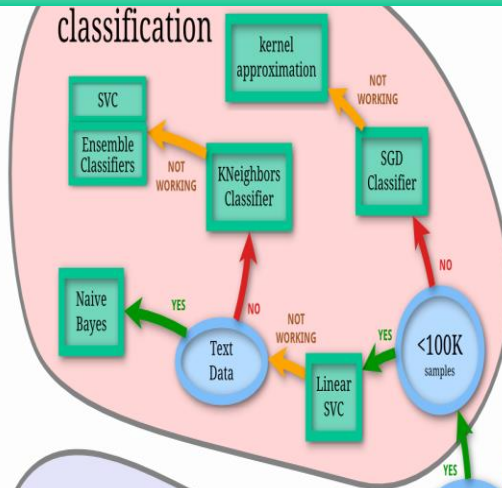
本願+正解公報=50公報



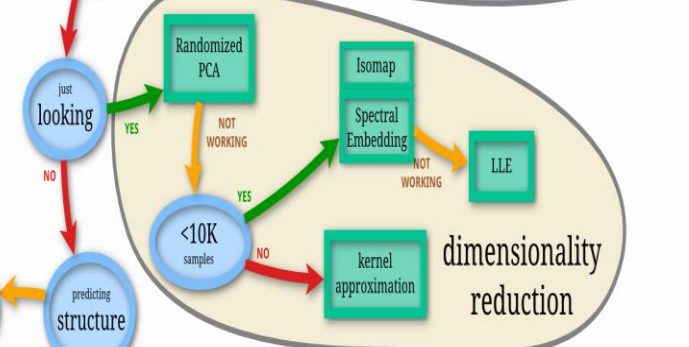
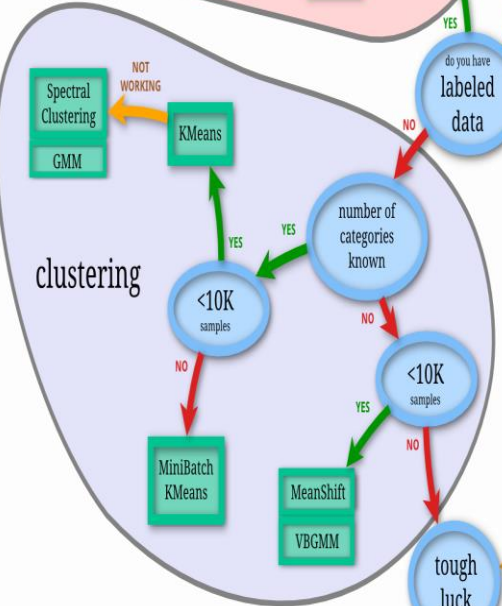
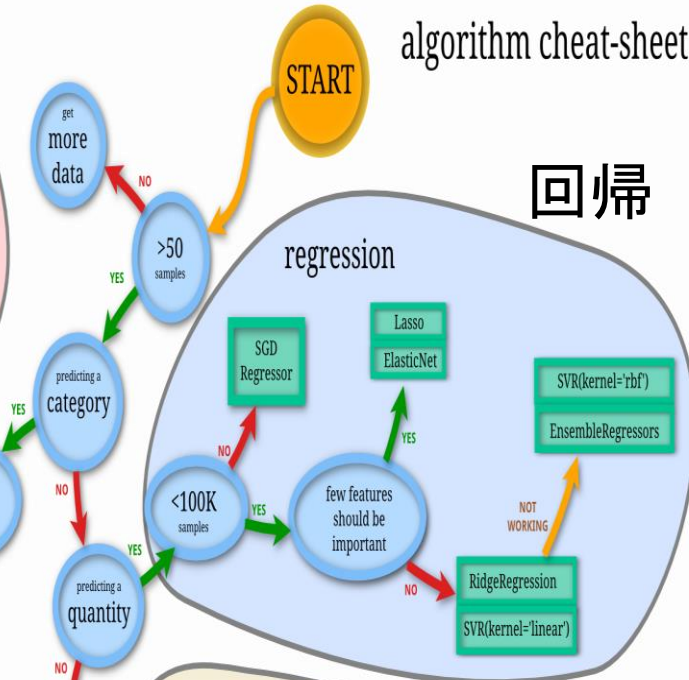
scikit-learn アルゴリズム早見表

scikit-learn
algorithm cheat-sheet

クラス分類
文書分類



回帰



クラスタリング

次元圧縮



http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Orbitのランドスケープマップ 2次元に次元圧縮

技術開発の
水先案内人
パイロット

自社技術
開発技術

他社技術

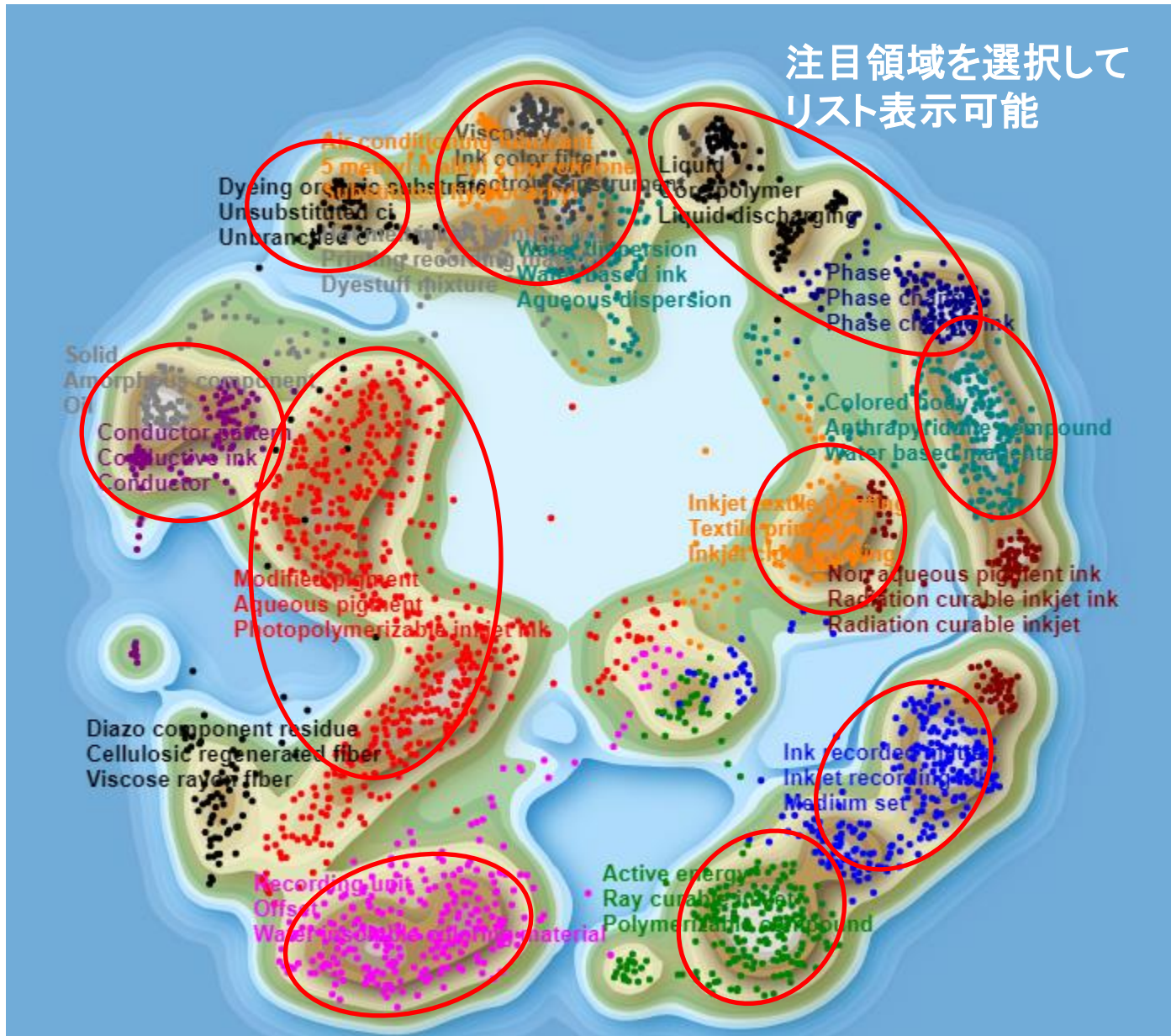


図19. Orbitのランドスケープマップ

地図の図法とその特徴

次元圧縮

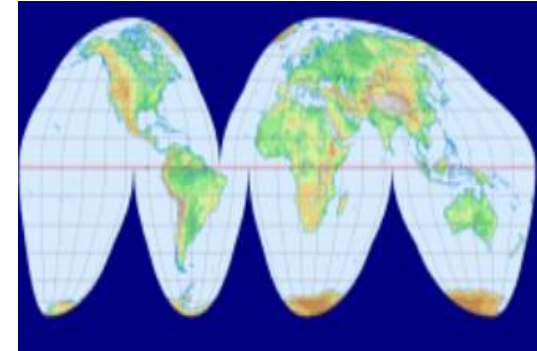
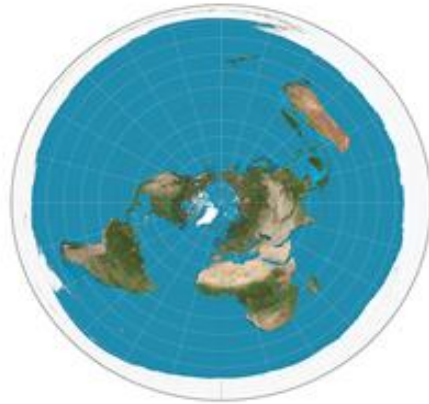
地球は**3次元**→**2次元**の地図に**次元圧縮**する各種方法とその特徴を理解して使用する

メルカトル図法

正距方位図法

メルワイデ図法

グート図法



円筒投影による図法
地球表面のすべての部分の**角度**が**正**しく表される
図上の2点を結ぶ直線は等角航路となる
羅針盤による航海に便利
海図に利用される

図の中心から他の1地点を結ぶ直線が図の中心からの**正しい方位**、**最短経路**を表し、図の中心からの距離を正しく求めることができる
飛行機の最短経路や方位を見るために使われる

地球を楕円形にして、極地方の**形のゆがみ**を少なくした図法
分布図に使用される

世界全体を通して**大陸部分の形の歪み**ができるだけ小さい地図を作るために、サンソン図法の地図とメルワイデ図法の地図を組み合わせ作成
面積が正しく表されている

地図投影法学習のための地図画像素材集

http://user.numazu-ct.ac.jp/~tsato/tsato/graphics/map_projection/

図20. 地図の図法とその特徴

コード	カテゴリ名	内容
SCL	school	学校
RLW	railway	鉄道 (交通関連)
FML	family	旧家
BLD	building	建造物
SNT	Shinto	神道
PNM	person name	人名
GNM	geographical name	地名
CLT	culture	伝統文化 (現代文化も含む)
ROD	road	道路
BDS	Buddhism	仏教
LTT	literature	文学
TTL	title	役職・称号
HST	history	歴史
SAT	shrines and temples	神社仏閣
EPR	emperor	天皇

Wikipedia日英京都関連文書対訳コーパス
<https://alaginrc.nict.go.jp/WikiCorpus/>

MXNetによる文書分類

トレーニング文書で教師データ
(左記カテゴリ)を学習させ、
テスト文書をカテゴリに分類
して正解数をカウントする。

- ・トレーニング文書: 9877記事
- ・テスト文書: 4234記事

Accuracy=0.799953

約80%正解

学習アルゴリズム
一次元CNN

Convolutional Neural Network

畳み込みニューラルネットワーク

Yoon Kim, Convolutional Neural Networks for Sentence Classification
<https://arxiv.org/abs/1408.5882>

ディープラーニング

フレームワーク: MXNet

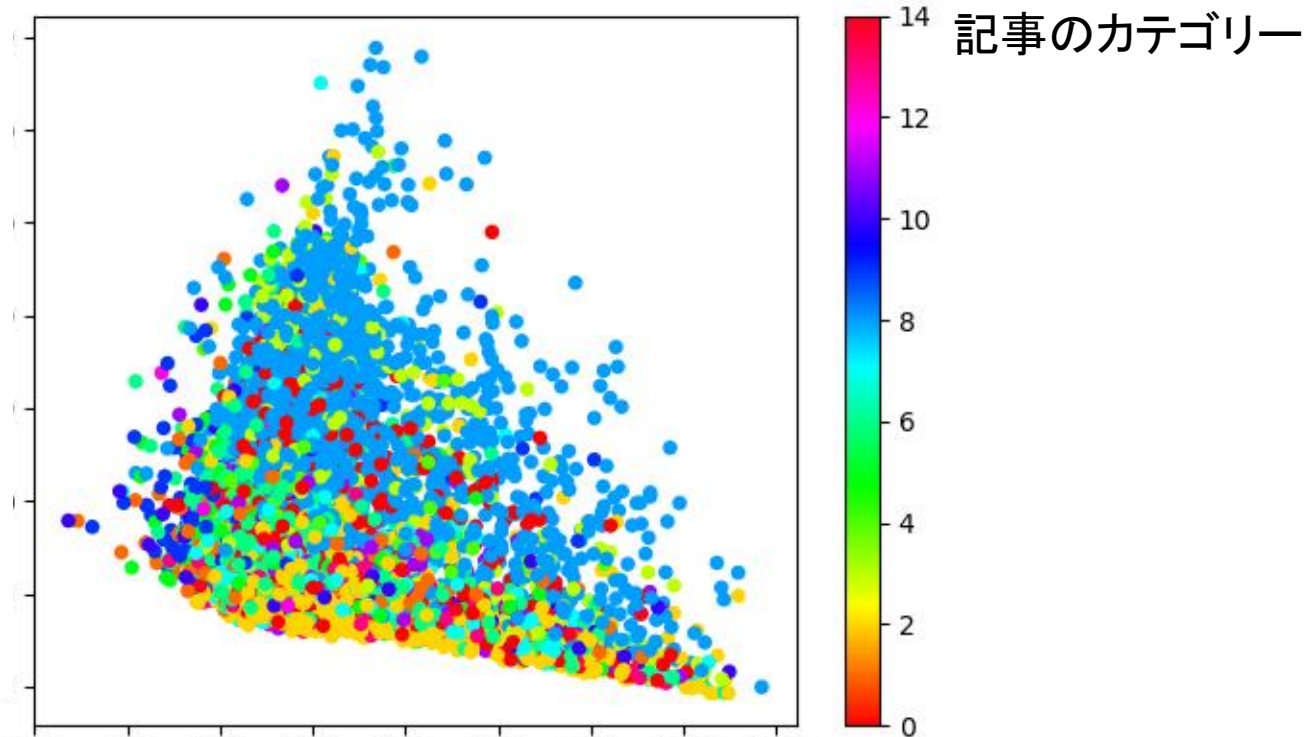
Apache MXNet <https://mxnet.apache.org/>

SCDVによる文書のベクトル化散布図 次元圧縮

MXNet

score = 0.766572

scoreは小さい方がクラス毎に良くまとまっていることを示す



SCDV: Sparse Composite Document Vectors

全ての単語に対する単語ベクトル辞書を作成する (fastText)

全ての単語ベクトルを MinBatchKMeans によってクラスタリングする

各クラスターに属する単語のベクトルを加算して合成して文章ベクトルを生成する

fastText は Facebook が開発した単語のベクトル化とテキスト分類をサポートした機械学習ライブラリ

<https://dheeraj7596.github.io/SDV/>

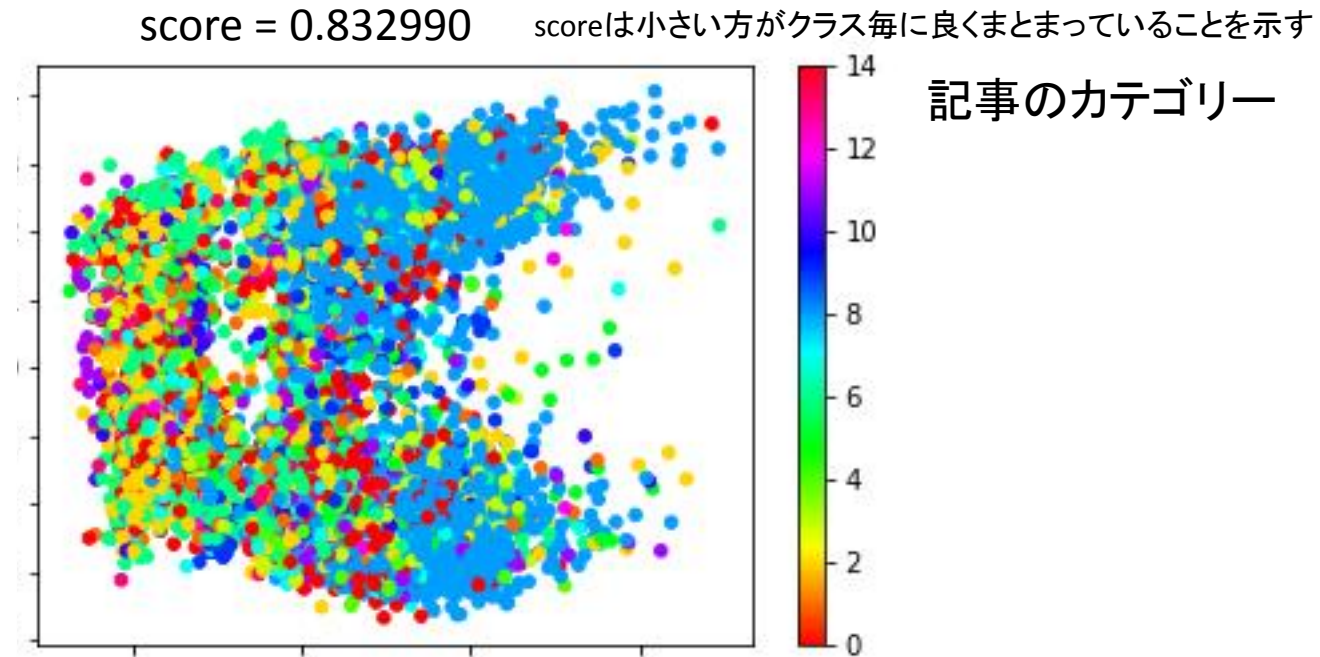
<https://arxiv.org/abs/1612.06778>

文書ベクトルをお手軽に高い精度で作れる SCDV って実際どうなのか日本語コーパスで実験した

<https://qiita.com/fufufukakaka/items/a7316273908a7c400868>

因子解析による文書のベクトル化散布図

次元圧縮
MXNet



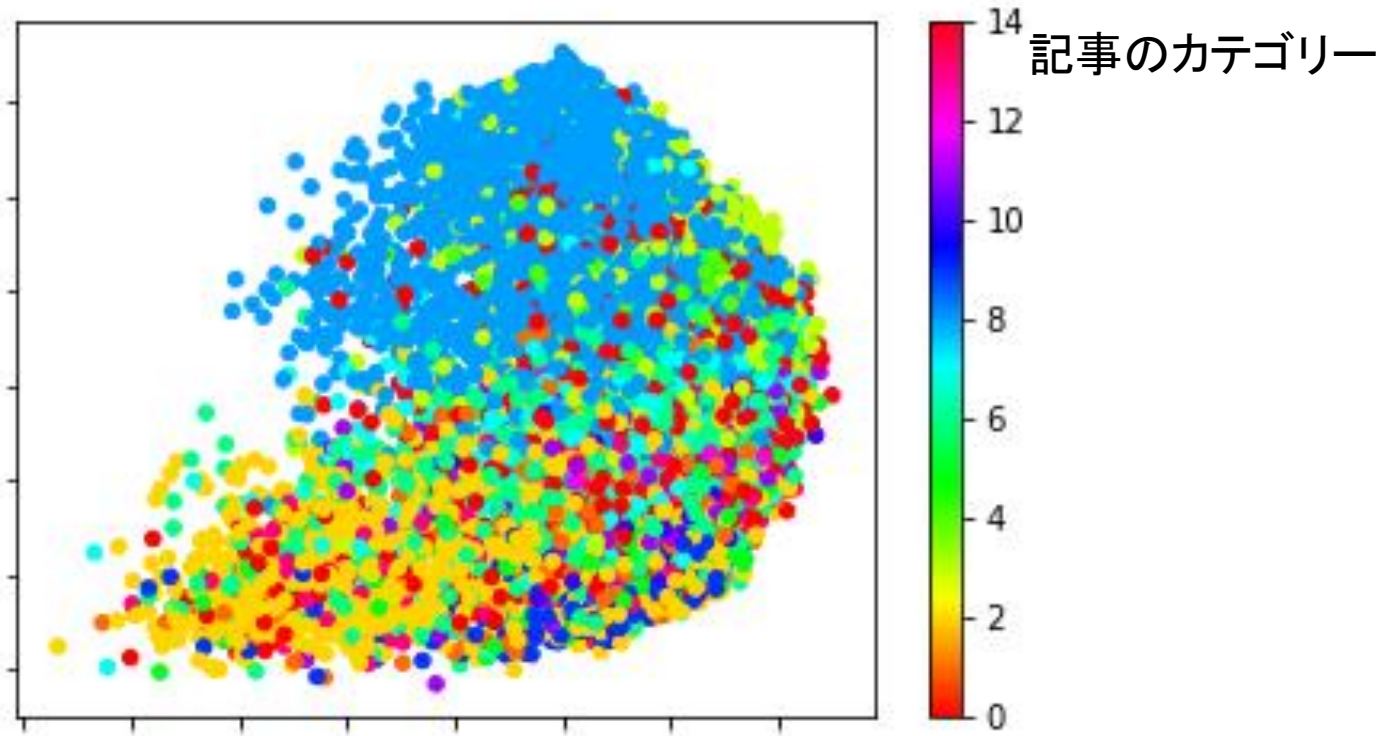
文書内に含まれているすべての単語ベクトルから因子成分を作成し
その因子を文書の意味合いを表すベクトルデータとする

RNNによる文書のベクトル化散布図 次元圧縮

MXNet

score = 0.575706

scoreは小さい方がクラス毎に良くまとまっていることを示す



RNN学習のエポック数: 10

Recurrent Neural Network(RNN)により直接文書データをベクトル化

計算時間: 4CPUで約1時間

まとめ

① 先行技術調査を念頭にdoc2vecによる文のベクトル化と発明の要素単位の類似文(根拠箇所)抽出、適合検討を行った。

② 動向調査を念頭に教師あり機械学習の1次元CNNによる文書分類と次元圧縮による公報の可視化検討を行った。

教師あり機械学習には良質な教師データの準備が重要である。

ディープラーニングの機械学習には大量の教師データが準備できるかで学習済モデルの性能が決まる。

調査目的に応じたアルゴリズムとデータの選択が重要である。

今後の予定

- ① 特許データを用いてベクトル化、根拠箇所抽出、適合判定を一気通貫で検討
- ② 動向調査、SDI調査用正解教師データ作成と文書分類、可視化検討
- ③ アテンションモデルの検討

参考文献

- 1) 安藤俊幸, 機械学習を用いた効率的な特許調査方法
ーディープラーニングの特許調査への適用に関する基礎検討ー
Japio YEAR BOOK 2018, 2018, p. 238-249.
http://www.japio.or.jp/00yearbook/files/2018book/18_3_05.pdf

- 2) 安藤俊幸, 機械学習を用いた効率的な特許調査方法
ーニューラルネットワークの特許調査への適用に関する基礎検討ー
Japio YEAR BOOK 2017, 2017, p. 230-240.
http://www.japio.or.jp/00yearbook/files/2017book/17_3_04.pdf

- 3) 安藤俊幸, テキストマイニングと機械学習による効率的な特許調査
数理システムユーザーコンファレンス2017(2017年11月2日)
http://www.msi.co.jp/userconf/2017/pdf/muc17_501_2.pdf

- 4) 安藤俊幸, 機械学習を用いた効率的な特許調査
アジア特許情報研究会における研究活動紹介
「特技懇」誌, 2018.11.26. no.291
https://tokugikon.smartcore.jp/tokugikon_shi

謝辞

免責

本報告は2018年の「アジア特許情報研究会」のワーキングの一環として報告するものである。本報の内容は筆者の私見であり所属機関の見解ではない。

謝辞

最後に機械学習の初心者である筆者を様々な形でサポートしていただいた研究会の研究活動で出会った多くの皆様に感謝申し上げます。

アジア特許情報研究会では研究員を募集しています！
アジア特許情報研究会 事務局 伊藤 徹男

私たちには研究したいテーマがたくさんありますが、研究メンバーが足りていません。アジア・新興国の知財情報、知財情報解析に関心がある方は、ぜひ事務局までご連絡ください。

Mail: patentsearch2006@yahoo.co.jp



アジア特許情報研究会 <http://www.geocities.jp/patentsearch2006/asia-research.html>