テクニカルサンプルプロジェクト

テキストのクラスタリング

k-means 二項ソフトクラスタリング



株式会社 NTTデータ数理システム

NTT Data © 2025 NTT DATA Mathematical Systems Inc.

このプロジェクト について

こんな方におすすめします

・ テキストデータや属性データを利用してテキストをグループ分け、話題を抽出したい方

・ テキストデータを利用した機械学習を行いたい方

何をするプロジェクト?

このプロジェクトでは、いわゆる「教師なし学習」であるクラスタリ ングという手法を用いてテキストデータをクラスタリング(=グルー プ分け)する一連の流れを紹介します。

ここでは、クラスタリングとして有名な、k-means法と二項ソフト クラスタリングの2手法をご説明します。

この流れは、テキストデータを利用した機械学習の一般的なフ ローであり、これを応用することで様々な機械学習手法をテキ ストデータでも扱うことができます。



プロジェクトの解説

© 2025 NTT DATA Mathematical Systems Inc.

プロジェクトを構成する要素

本プロジェクトは大きく分けて以下の3つの要素に分けられます。

本サンプルプロジェクトでは、文章の分割の粒度を変えてクラスタリングを行っています。「自動連結あり」の結果は、「形態素・構文解析+」 アイコンにて分割の粒度を大きく設定したフロー、「自動連結なし」の結果は、分割の粒度を小さく設定したフローとなっています。

次ページからは「自動連結あり」の結果のフローについて、各要素を構成するアイコンの中身について説明します。

<complex-block></complex-block>								
Image: Control in the control in th	対象データ							「自動連結あり」の結果
<complex-block> Image: State Sta</complex-block>		▲ 1.データについて ▲	3. テキストの単語分割	4. 対象単語の抽出		6. k-means法	7. 単語使用状況の確認	
With With With With With With With With		本分析では、ホームバーカリー製造賞者でパールパンのの 実営売買い気品のロニテータを取って、製作して、サイ トから取職したデータを発達して作成したもの。す	「形態素・構文解析+」 アイコン	「語句のフィルタリング」アイコン) 「k-meansi造」アイコン	ロドンログイコンを利用して、クラスタリングの緒 果、各クラスタの学品使用の利用の可能にします。	
・ ・		列の構成	デキストデータを文簡単位に分割し、単語や 品は、体り気け汚動を抽出します。	分析に利用する単語を加出します。	5. 文章のベクトル化	▲ Рездес, Малниянет ▼	 回「編訂 アイコン 回「施御」アイコン 	
Notes Notes <t< td=""><td></td><td>テーブルエル下の州から時点されます。</td><td>分析の対象は「トロレビューデータニホ」デー 分の「レビュー」列です。</td><td>対かり自己構成のひつち、対称に利用する学校 をフィルタリングを 行は高尚もしくは出現開発です。</td><td>internate、2歳シブトクラスタリングのそれぞれの生活で分析を行うため、それぞれ文章をベクトル(抽画表明)に変更します。</td><td></td><td>(3)「集計」アイコンによる操作</td><td></td></t<>		テーブルエル下の州から時点されます。	分析の対象は「トロレビューデータニホ」デー 分の「レビュー」列です。	対かり自己構成のひつち、対称に利用する学校 をフィルタリングを 行は高尚もしくは出現開発です。	internate、2歳シブトクラスタリングのそれぞれの生活で分析を行うため、それぞれ文章をベクトル(抽画表明)に変更します。		(3)「集計」アイコンによる操作	
image: state of the stat		916 MB			Bernans - 18 文明ベクトル化		クラスターごとに、どのような単価がどの保険使われ ているかを集計します。	
市 市		D レビューごとに対号されるD 3時度 レビューの3時時時(1920年)			・ 10 Hymon songr 2 項ンフトクラスタリング			
		2588名 レビュー対称の2588名や2117			 日文章ベクト5化 			
		性別 レビュワー信頼:「労性」も以くま 「女性」			(a) 「文意ベクトル化」アイコンによる操作		°= " °++ "	
Image: State St		年代 レビュワーが相: 10歳5歳 0年代			lemman法で分析を行うために、マトリックス形式(勝特データ)で出力しま す。			
・ 185-9C.0VC ・ 150-95-95-90-95-000 ・ 186-9C-0XC ・ 150-95-95-90-95-000 ・ 100-100-100-100-100-100-100-100-100 ・ 150-95-95-90-95-000 ・ 100-100-100-100-100-100-100-100 ・ 150-95-95-90-95-000 ・ 100-100-100-100-100-100-100-100 ・ 150-95-95-95-95-95-95-95-95-95-95-95-95-95-		第入価格 レビュー対象の数量の価格 書き込み			(b)「Python script」アイコンによる操作	8. 二項ソフトクラスタリン	2 NU192813	
Image: State St		1 1771-194360418704 A				<u>d</u>		
image: spectrum image: spectrum image: spectrum image: spectrum <tr< td=""><td></td><td>i ji</td><td></td><td></td><td>i !</td><td>「二項ソフトクラスタリング」ア イコン</td><td>9. クラスター情報の整理</td><td></td></tr<>		i ji			i !	「二項ソフトクラスタリング」ア イコン	9. クラスター情報の整理	
1 ##						•	以下につめアイコンを利用して、クラスタリングの結果を確認します。	
 1. ## → #100 -		1000 (m) 1000 (m) 1000 (m)	1		į į		名レビューの所属クラスタの構成	
2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 885-94:001 1 2. 95:001 1 2. 95:001 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1 3. 95:016 1								
************************************		2.辞書データについて ^					 ※10005249へのRigitie#00662 ※11「マトリックス化」アイコン 	
INUCAS-9-3-9-9-800 mm Inucasion I		データの前部用の際に、参数の性語に分割されている体語を、 一つの単語としてまとめたい場合に設定します。					(a) 「行選択」アイコンの操作	
**		HBレビューデータ_ユーザー辞書の列の構成					レビューごとにクラスタへの所属等キが一番大きなクラスタを一急に決定しま	
** * ** ** <t< td=""><td></td><td>本データの列換品は以下の通りです。</td><td></td><td>_</td><td></td><td></td><td>小」「利用計算書」マイコンの場合</td><td></td></t<>		本データの列換品は以下の通りです。		_			小」「利用計算書」マイコンの場合	
Image: Walk New Willing Image: Walk New Willing<		NE 148					(U) 7時128史) アコゴンUN#1F マ	
		表記 単語として抽出したい文字列				1		「自動連結なし」の結果
3.742Nの単語の目 (FREA: HEXERF: 1742) (HEXERF: 12742) (HEXERF: 1			SALES BER	70%880226 693-0(6	\$8679.66.7 101-02(0) Print wing (10) Roma(1)	k-man(Z(1)		
アキスト前処理 クラスタリングと結果の可視化 4		L;;	2 ニナフトの単語分割				_,¶#=	
「####: #XX#: ! Y<2> !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!			5. デキストの単語方割		XH()	US40	- PARSER/O	
			「形態素・構文解析+」アイコン			1		
テキスト前処理 クラスタリングと結果の可視化 4			以降のフローは上記「分かち書き(自動連結あり ローと知一です。	7010				
テキスト前処理 クラスタリングと結果の可視化 4		L		-	d i			
			テキス	ト前処理 🖡		- クラスタリ	シクと結果の可視化	
								- 4

© 2025 NTT DATA Mathematical Systems Inc.

プロジェクト解説 — 対象データ

1. HBレビューデータ.dft

ECサイトで様々なホームベーカリーに対してのレ ビューをまとめた、仮想の口コミデータです。MSIPの 上では、csv形式のデータをdft形式に変換し、シ ナリオ編集エリア上に配置して使用します。1行が1 レビューに対応します。

今回は口コミテキストの入ったレビュー列を利用しま す。データに含まれる列の詳細については、右の表 をご覧ください。

2. HBレビューデータ_ユーザー辞書.dft

既存の辞書にはないような、ユーザー独自の単語 を追加するためのデータです。テキストの分割処理 を行った結果、つながって出てきてほしい複合語が、 いくつかの単語として分かれて出てきてしまう場合な どに利用します。

列名	内容
ID	レビューごとに付与されるID
評価	レビューの評価列(5段階)
製品名	レビュー対象の製品名や型 番
性別	レビュワー情報 : 「男性」もし くは「女性」
年代	レビュワー情報 : 10歳刻み の年代
購入価格	レビュー対象の製品の価格
書き込み 日	レビューが投稿された日付
レビュー	レビュー内容 分析対象のテキスト列
	タ ユーザー辞書.dft-data 列数:2 行数:

品詞

Category

名詞 一般

形容詞 一般

表記 Category

1 パン焼き機

2 a lot of

No

1. データに	ついて			
本分析では、ホームペーカリー製品自宅でパンを作るための家庭 用電化製品の口コミデータを扱います。ある口コミサイトから収 集したデータを想定して作成したものです。				
列の構成				
テーブルは以下の	の列から構成されます。			
列名	内容			
ID	レビューごとに付与されるID			
評価	レビューの評価列(5段階)			
製品名	レビュー対象の製品名や型番			
性別	レビュワー情報:「男性」もしくは「女 性」			
年代	レビュワー情報:10歳刻みの年代			
購入価格	レビュー対象の製品の価格			
書き込み 日	レビューが投稿された日付			
	しビュー内容 分析対象のデキスト列			

1

2.

5

TextExtension

プロジェクト解説 — テキスト前処理

3. テキストの分割

テキストデータを分析する際、記載されている文章の長さや内容が統一され ていないため、テキストデータそのままでは分析を行うことができません。そこで、 「形態素・構文解析+」アイコンを利用して、テキストデータを単語単位に分 割します。さらに、単語の品詞や係り受け関係などの情報も抽出します。 分析の対象は「HBレビューデータ、dft」データの「レビュー」列です。

4. 対象単語の抽出

クラスタリングの対象とする単語を品詞と頻度の観点から絞り込みます。ここでは意味のある単語でベクトルを作成するために、品詞が「名詞」「動詞系」 「形容詞・形容動詞系」「副詞」の単語を取り出しています。更に、頻度の 大きい単語はクラスタリングに影響するため、上位5単語を除外し、ベクトル の次元数を調整するため、頻度上位100単語のみを取り出しています。



プロジェクト解説 — テキスト前処理

5. 文章のベクトル化

文章に現れる単語をもとに、文章を数値で表したベクトル表現を獲得します。 テキストデータを機械学習で用いるためによく使われる手段です。

ここでは、BoW (Bag of Words) を用いて、「どの単語がどのくらい出現しているか」という数値のベクトル表現を獲得します。出現しているかどうかのみに着目する場合には、単語の頻度ではなく、有無を表す「0/1」の数値表現を利用します。

k-means法の入力として、(a)横持データであるマトリックス形式のベクトル表 現を獲得したのち、単語が出現するかのみに着目するため、(b)「Python Script」アイコンを用いて 0/1 のベクトル表現を作成しています。

また二項ソフトクラスタリングの入力として、(c)縦持データであるリスト形式のベ クトル表現を作成しています。



プロジェクト解説 ― クラスタリングと結果の可視化

6. k-means法

どのような単語が出現しているかをもとに、文章をクラスタリングします。

ここではクラスター数を3に設定し、1行のテキストデータを1件とし てクラスタリングしているため、レビューの1件ずつがクラスター1~3 のいずれか一つのみに所属します。



7. 単語使用状況の確認

各クラスターがどのような性質をもつか、どのような話題が多いかを 確認します。

(a)クラスターごとに、どのような単語がどの程度使われているかを 集計し、(b)集計結果を転置し折れ線グラフで可視化します。

プロジェクト解説 ― クラスタリングと結果の可視化

8. 二項ソフトクラスタリング

レビュ―データとレビュー全体で利用されている単語のクラスタリングを同時に行います。各レビューが所属するクラスターを把握するとともに、単語のクラスタリング結果をもとに各クラスターがどのような話題を持つかを見る ことができます。

二項ソフトクラスタリングは、レビューデータや単語がそれぞれのクラスター に所属する確率を算出します。レビューデータが複数の話題を持っている、 ひとつの単語が複数の話題で語られるということを見ることができます。

9. クラスター情報の整理

クラスタリングの結果を確認します。

極端に偏りがないか、などを確認するために、(a)各レビューの所属確率 が一番高いクラスターを抽出し (b)各クラスターに所属するデータ件数を 可視化します。

単語のクラスターへの所属確率を確認します。 二項ソフトクラスタリングの 結果はリスト形式で出力されるため、確認しやすくするために、 (c)テーブ

ル形式に整形します。

© 2025 NTT DATA Mathematical Systems Inc.



アウトプットの説明

© 2025 NTT DATA Mathematical Systems Inc.

アウトプット – k-means法

「k-means法」アイコンの結果を確認します。

resultテーブルには、cluster_id、元データの順で列が並びます。「cluster_id」列が、各行が所属するクラスターの値です。

cluster_infoテーブルでは、id, size, 単語, residual列があります。id が各クラスターを表し、クラスターごとにsize(データ件数)とクラスター 中心となる単語(各次元)の値を確認できます。residualは中心値との残差の絶対値の総和を表します。 クラスターごとのデータ件数を円グラフで可視化することも有効です。

eans法-I	result 列数: 102	行数: 361			
No.	cluster_id Category		RowID Integer	焼き立て _{Integer}	食べる Integer
1	1		1	1	1
2	2		2	0	1
3	3		3	0	0
4	2		4	0	1
5	2		5	0	0
6	1		6	0	0
7	2		7	0	0
8	3		8	0	1
9	3		9	1	1
10	2		10	0	1

k-means法-cluster_info 歹	J数: 103 行数: 3			
No. id Category		SiZe Integer	焼き立て _{Float}	食べる _{Float}
1 1		62	0.080645	0.161290
2 2		96	0.042036	0.354742
3 3		203	0.078654	0.231190



アウトプット - k-means法(単語使用状況)

各クラスターに割り当てられた文章ごとに、どのような単語がどの程度使われているかを集計し、折れ線グラフで可視化します。 それぞれのクラスターにおける単語の影響度合いを見ることができます。値の大きいものほど、どのクラスターに対する影響が大きい単語とみ なします。

cluster3(緑色)は特に突出した値の単語はなさそうです。複数の話題を含むために各単語の影響度合いがならされていることが考えられます。cluster3の結果から、クラスター数を増やすことでより話題を分割できそうなことが考えられます。cluster2(橙色)は、「買う」という単語が最も影響が強いことが分かります。



TextExtension

アウトプット - 二項ソフトクラスタリング

二項ソフトクラスタリングの結果には複数のテーブルが表示されます。pZXテーブルでは、X(レビュー)のZ(クラスター)への所属確率を 確認します。RowID 1 のレビューはz=2クラスターに43%、z=1クラスターに33%、z=3クラスターに24%の割合で所属していることが分かり ます。pYZテーブルでは、Z(クラスター)ごとのY(単語)の所属確率を見るため、クラスターの特色を表す単語を確認することができま す。z=1クラスターでは「焼ける」「美味しい」などの一般的な単語の他、「音」や「静か」が上位に現れるため、音に関する話題を持つクラス ターと考えられます。またz=3クラスターは、「お餅」や「米粉パン」などパン以外メニューの話題を持つと考えられます。

		4				
No.	X Category		Z Integer	pZX Float	rank Integer	
1	1		2	0.429060	各クラスタ	(7)への所属確率
2	1		1	0.332568	ちफ=刃□⇒	(<i>L)</i> 、。)////////////////////////////////////
3	1		3	0.238372	で唯誌しま	. 9 °
4	10		3	0.757824	1	
5	10		1	0.242176	2	
6	10		2	0.000000	3	
						1

οYΖ	テーブル	z=1]			
No.	Y Category	I	Z	pYZ Float	rank Integer
1	音		1	0.071793	1
2	焼ける		1	0.057133	2
3	美味しい		1	0.041141	3
4	食べる		1	0.037988	4
5	簡単		1	0.035084	5
6	おいしい		1	0.032401	6
7	気になる		1	0.029272	7
8	静力		1	0.027919	8
9	大きい		1	0.024729	9
10	ちょっと		1	0.024038	10

No. Y Category	Z	pYZ Float	rank Integer
239 両足	3	0.009200	39
240 好き	3	0.008890	40
241 商品	3	0.008650	41
242 餅	3	0.008399	42
243 あまり	3	0.008114	43
244 バター	3	0.007640	44
245 言う	3	0.007546	45
246 普通	3	0.007310	46
247 もう	3	0.007261	47
248 すごい	3	0.006601	48
249 つく	3	0.006392	49
250 安い	3	0.006324	50
251 お餅	3	0.005971	51
252 考える	3	0.005840	52
253 今	3	0.005733	5

© 2025 NTT DATA Mathematical Systems Inc.

TextExtension

アウトプット – 二項ソフトクラスタリング(単語のクラスタリング結果)

TextExtension

各単語がどのクラスターにどのくらいの確率で所属しているかを把握するには pZY テーブルで確認します。

ただし、二項ソフトクラスタリングの結果では、リスト形式で表示されているため、各クラスターでまとめて確認したい場合には不便です。その ため、マトリックス形式に整形し、結果を確認することをお勧めいたします。

マトリックス化_pZY-resul	t 列数:4 行数:10	0			
No. Y Category		pZY.2	pZY.1 Float	pZY.3	
1 あと		1.000000	0.000000	0.000000	
2 あまり		0.130169	0.431374	0.438457	ション しょうしょう しょうしょう しんしょう しんしょ しんしょ
3 しいしい		0.000000	0.501736	0.498264	の程度所属しているかが
4 いう		0.904389	0.000000	0.095611 -	一覧で見やすくなります。
5 いる		0.000000	1.000000	0.000000	
6 NZNZ		0.000000	0.000000	1.000000	
7 うるさい		0.000000	1.000000	0.000000	
8 おいしい		0.292957	0.312449	0.394595	
9 おすすめ		0.000000	1.000000	0.000000	
10 お餅		0.381104	0.351876	0.267020	

アウトプット – 二項ソフトクラスタリング(データ数の確認)

二項ソフトクラスタリングの特徴として、各要素は複数のクラスタに所属しうるということがあります。一方でどれか一つのクラスターにのみ所属 するとみなしたい場面も起こりえます。そのようなときには、所属確率の一番大きいクラスターに所属するとみなして、所属先を一意に決め ることがあります。

二項ソフトクラスタリング結果のpZXにて、rank=1の行のみを抽出することで、各レビューと所属確率が一番大きいクラスターの行を抜き出すことができます。その結果を円グラフで可視化することで、各クラスターのサイズをおおよそ把握することが可能です。

rank=1 のみを

抽出します。

No.	X Category	Z Category	I	pZX Float	rank Integer
1	1	2		0.429060	1
2	10	3		0.757824	1
3	100	3		0.770846	1
4	101	3		1.000000	1
5	102	2		1.000000	1
6	103	1		0.701432	1
7	104	3		0.535244	1
8	105	2		0.524179	1
9	106	1		1.000000	1
10	107	2		0.429607	1

Zの値はカテゴリ

として扱います。



TextExtension

アイコンの設定 アイコンの入力設定や処理実行時の設定項目について

アイコン – 形態素・構文解析+_自動連結あり

インプット設定

テキストデータと辞書ファイルの設定を行います。

ここでは、分割処理の対象のテキスト列を含むデータを「table」、HB レビューデータ ユーザー辞書を「usrdic」 に指定します。

辞書はそれぞれ、類義語辞書を「syndic」、ユーザー辞書を 「usrdic」、分割辞書を「sepdic」に設定します。いずれの辞書も必 須ではありません。詳細は補足情報の『辞書ファイル』をご参照ください。

対象テキスト列

● テキスト列

分割処理の対象としたい列を指定します。1列のみの指定 が可能です。ここでは「レビュー」列を対象とします。

Input Matching (Controller				×
		table	syndic	usrdic	sepdic
HBレビューデータ.dft	data	•• @ ••			
HBレビューデータ	data			-0-	
* 複数可					

形態素・構文解析+	▶_自動連結あり		? _ ×
対象テキスト列			
テキスト列 レ	<u>_</u>	•	
※String型・Category	/型の列を選択		
言語の選択			
◉ 日本語 ☰			
○ 英語			
☑ 構文解析と自動連續	結を行う 🗮		
文章の区切りとみた	よす文字		
🔽 句点(。)	₩ 2 疑問符(?)	■ 🔽 感嘆符(!)	=
□ 空白	■□ 改行	≡	
その他			≡
並列処理数			
1	=		
✓ 原文参照のための:	 オブジェクトを出力する	=	
		-	4
		· · · · · · · · · · · · · · · · · · ·	

アイコン - 語句のフィルタリング

品詞フィルタ

よく利用される品詞セットは「デフォルト品詞セット」として設定され ています。名詞/動詞系/形容詞・形容動詞系/副詞の選択が 可能です。詳細に設定する場合には「オリジナル設定」を選択し、 利用する品詞を個別に指定します。

頻度フィルタ

● 対象列

頻度を指定したい単語列を指定します。

● 上位N単語を除外する

頻度上位から指定した数の単語を除外します。ここでは 初期設定の「5」を指定します。

● 上位N単語を抽出する

頻度上位から指定した数の単語を抽出します。ここでは 「100」を指定します。

語句のフィルタリング ? _ X
☑ 品詞フィルタを設定する
● デフォルト品詞セット ○ オリジナル設定
抽出する品詞
☑ 名詞
☑ 形容詞・形容動詞系 ☑ 副詞
☑ 頻度フィルタを設定する
対象列 replaced ▼ ····
※String型・Category型の列を選択
□ 最低頻度を設定 2
□ 最高頻度を設定 100
✓ 上位N単語を除外する 5
✓ 上位N単語を抽出する 100
□ 文字列フィルタを設定する
□ 文字数フィルタを設定する
実行して閉じる ▼ 保存

アイコン - 文章ベクトル化_マトリックス・文章ベクトル化_リスト①

変数選択

● 単語列

ベクトル化の対象となる単語列を指定します。ここでは置換語列である「replaced」列を選択します。

● キー列

ベクトルを生成するキー列を指定します。「形態素・構文 解析+」アイコンの結果を利用する場合、以下の列を選 択します。

- 1行(1セル)単位のベクトル化: RowID
- 1文単位のベクトル化: RowID, SntID
 ここでは、1行単位でベクトル化を行うため、「RowID」列
 を選択します。

文章ベクトル	ヒ_マトリックス		? _ ×
変数選択			
列名		列型	単語列 キー列
RowID		整数	
SntID		整数	
TokenID		整数	
form		文字列	
lemma		カテゴリ	
replaced		カテゴリ	✓ □ -
モデルの設定			
モデル	● BoW		\equiv
	⊖ tf-idf		
	○ SWEM		
SWEMの設定	:		
計算方法	◉ 平均 🔘 最大		=
乱数シード	0		生成 🗮
出力形式			
◉ マトリックス	₹形式 ○ リスト形式		≡
			実行 ▼ 保存

アイコン - 文章ベクトル化_マトリックス・文章ベクトル化_リスト②

モデルの選択

ベクトル表現のモデルを選択します。モデルの種類は、単語の出現 状況から文章データをベクトル化する手法として、

• BoW (Bag of Words)

tf-idf (Term Frequency-Inverse Document Frequency)
 単語の埋め込み表現を利用してベクトル化する手法として、

SWEM (Simple Word-Embedding-based Methods)
 があります。詳細はマニュアルをご参照ください。
 ここでは「BoW」を選択します。

出力形式

ベクトル化したデータの出力形式を指定します。マトリックス形式は キー列で指定した単位1行ごとにベクトル表現を出力します。リスト 形式は、キー列・単語・値の組を出力します。

k-meansはマトリックス形式、二項ソフトクラスタリングはリスト形 式の出力を利用します。

文章ベクトル	L マトリックス		2 _×
変数選択			
列名		列型	単語列 キー列
RowID		整数	
SntID		整数	
TokenID		整数	
form		文字列	
lemma		カテゴリ	
replaced		カテゴリ	-
モデルの設定			
モデル	● BoW		=
	◯ tf-idf		
	○ SWEM		
SWEMの設定			
計算方法	◉ 平均 🔘 最大		=
乱数シード	0		生成 🗮
出刀形式			
◎ マトリックス	ス形式 ○ リスト形式		=
			実行 ▼ 保存

アイコン - Python script(0,1)表の作成

Python script

BoWでベクトル化したデータをもとに、出現有無の情報に置き換え たベクトル表現を獲得します。

【スクリプト内で行っていること】

- (MSIP) DataFrame を pandas.DataFrame に変換 する
- ベクトル表現のテーブルにおいて、単語のキー列ごとの出現頻度である各セルの値に対して、以下の置換を行う
 - ・ 値 ≥ 1 … 新しい値:1 (=単語が出現している)
- pandas.DataFrame を (MSIP) DataFrame に変換す

る

Python script_(0,1)表の作成 ? _ ×
入力設定 🗸
出力設定 🗸
メタパラメータ参照設定 >
<pre>1 from msi.common.dataframe import DataFrame, cbind, rbind, merge, 2 from msi.common.dataframe.params import Axis, Merge, DType, Agg 3 from msi.common.dataframe.special_values import Na, Error, Negat 4 from msi.common.dataframe import pandas_to_dataframe 6 key = ['RowID'] 8 table_pd = table[len(key):table.ncol()].to_pandas() 10 result_pd = table_pd.mask(table_pd>=1,1) 11 result = cbind(table[key],pandas_to_dataframe(result_pd)))</pre>
<
→ 動作確認用インタプリタ → ×
egativeInf, PositiveInf;
In [2]:
実行 ← 保存

アイコン - k-means法

変数選択

● 説明変数

クラスタリングのもととなる変数を指定します。ここではRowIDを除く 単語列をすべて指定します。

基本設定

● 距離計算方法

クラスターの中心と各要素のベクトルの距離を定義します。単語を 扱う場合、ユークリッド距離の他、コサイン距離などもよく利用します。

● クラスター数

いくつのクラスターに分けるかという値を指定します。ここでは「3」に 指定します。

初期クラスターの設定方法

初期クラスターの設定方法を指定します。ここでは KMeans++を選択します。 初期のクラスターを離れた位置に定める手法で、効率よくクラスタリングを行い ます。

k-means法			? _ ×
変数選択			
列名		列型	説明変数
RowID		整数	▲
焼き立て		整数	
食べる		整数	\checkmark
焼く		整数	\checkmark
選ぶ		整数	
自分		整数	-
🗹 入力データを	出力に含める		
基本設定			
距離計算方法	ユークリッド距離		_
クラスター数	3		=
繰り返し最大数	100		=
□ 規格化オプシ	'3>		=
初期クラスター	の設定方法		
○ ランダム			=
● KMeans++			
乱数シード	0		生成 🚍
		実行	▼保存

グラフ – 円グラフ(k-means法)

グラフの種類

作成するグラフの種類を指定します。ここでは「その他」の「円グラフ」を選択します。

データの列

● データ選択

円グラフを作成するデータを選択します。分析結果テーブルからグラ フを作成する場合は自動的に入力されています。

● カテゴリ

円グラフで表現したい列を指定します。ここでは「cluster_id」を選択します。

	を選択してください		() ×
折れ線			_
散布図			
	MSIP		
ヒストグラム			
その他			
BACK	NEXT	ОК	CANCEL
212. データの列をき	曜択してください フ		⊘ ×
212. データの列を3	確択してください フ	 ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	 ※ ご動して選択することもできます
222. データの列を3 アクタン アータ選択	羅択してください フ k-means法 - result	・・・・ で選択タイアログをJ	 () × () () () () () () () () () () () () () (
22. データの列をi アクラコ データ選択 カテゴリ	壁択してください フ k-means法 - result ∂2須一 cluster_id	・・・・で選択ダイアログをJ	 ② × 2200して選択することもできます ・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
22. データの列を込 アータ選択 カテゴリ オブション設定 (3 軒	壁択してください フ k-means法 - result ℓ cluster_id ŧ)	●■で選択タイアログを	() ×
222. データの列を注 データ選択 カテゴリ オブション設定 (3 和	壁択してください フ k-means法 - result ∂② cluster_id ŧ)	・・・・ で選択ダイアログを1	 > × ۵.2.4.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0
22. データの列を込 アータ選択 カテゴリ オブション設定 (3 軒	壁沢してください フ k-means法 - result cluster_id ♠)	で選択タイアログを	 ۲ ۲
22. データの列を注 アータ選択 カテゴリ オブション殺定 (3 新	壁択してください フ k-means法 - result ∂②須── cluster_id ♠)	・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	 > × > • • • • • • • • • • • • • • • • • •
22. データの列を込 アータ選択 カテゴリ オブション設定 (3 軒	壁沢してください フ k-means法 - result ℓ ^{20月} cluster_id	・・・・で選択ダイアログを	 > × ۲ ۲
22. データの列を込 アータ選択 カテゴリ オブション設定 (3 軒	壁択してください フ k-means法 - result ℓ/// cluster_id ♠)	・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	 > × > • • • • • • • • • • • • • • • • • •
22. データの列を込 アータ選択 カテゴリ オブション設定 (3 和	壁沢してください フ k-means法 - result	で選択ダイアログを	 > × > • • • • • • • • • • • • • • • • • •

アイコン - 集計_単語使用状況

インプット設定

集計対象のテーブルを指定します。k-means法の「result」テーブル が対象です。

対象テキスト列

● 集計項目

集計対象列と集計方法を設定します。ここでは、単語列を 集計対象とし、平均を算出します。

● キー列

指定したキー列ごとに、集計項目設定した集計が行われま す。ここではクラスターごとに単語の平均を算出するため、 「cluster_id」を指定します。

Input Matching (×	
		table	
\	result	•• @ ••	
k-means/g	cluster_info		
★ 複数可			

集計_単語使用状況 ? _ X							
集計項目							
列名	列型	項目数	合計	平均	最小値		
RowID	integer				•		
焼き立て	integer			\checkmark			
食べる	integer			\checkmark			
焼く	integer			\checkmark			
選ぶ	integer			\checkmark			
古 八 4					· · ·		
□ 集計項目を	マトリックス形式	式で出力する	=				
キー列					_		
🗹 キー列の4	件数 ☰						
□ キー列の4	牛数割合 📃						
集計≠−	集計キー cluster_id ② v v						
オプション							
□ 重み付けをする =							
重み列			Ŧ		-		
				実行	保存		

アイコン - 転置

転置設定

● 対象列

行と列の転置を行いたい対象の列を指定します。ここでは可 視化(折れ線グラフの作成)のため、各単語列名を値に 持つ単語列を作成します。

詳細設定

転置前の列名:

転置前の列名の扱いを指定します。「元の列名を転置後の 一列目とする」と指定することで、元の列名を1列目のデータ と指定します。

● 転置後の列名

転置後の列名を指定します。「列名となる列を選択する」を 指定することで、いずれかの1列の値を列名として扱います。



集計_単語使用状況 ? _ ×						
集計項目						
列名	列型	項目数	合計	平均	最小値	
RowID	integer				· ·	
焼き立て	integer			\checkmark		
食べる	integer			\checkmark		
焼く	integer			\checkmark		
選ぶ	integer			\checkmark		
古八				-		
 二 集計項目を キー列 	Eマトリックス形i	式で出力する	=			
☑ ≠−列の	件数 ☰					
□ キー列の	件数割合 🗮					
集計字- cluster_id ② v						
オプション						
□ 重み付けをする 🗮						
重み列			Ŧ	•••	-	
				実行	▼ 保存	

グラフ – 折れ線グラフ(転置)

グラフの種類

作成するグラフの種類を指定します。ここでは「折れ線」の「折れ線」グ ラフを選択します。

データの列

● データ選択

折れ線を作成するデータを選択します。分析結果テーブル からグラフを作成する場合は自動的に入力されています。

● X軸

折れ線グラフのX軸(横軸)を指定します。ここでは「単語」を選択します。

● Y軸

折れ線グラフのY軸(縦軸)を指定します。ここでは各クラ スターごとの単語の利用状況を描画するため「1」「2」「3」 を選択します。



アイコン - 二項ソフトクラスタリング

変数選択

● X列

クラスタリング対象の列を指定します。ここでは「RowID」列を選択します。

● Y列

X列と同時にクラスタリングしたい対象の列を指定します。ここでは「word」列を選択します。

● スコア列

X列とY列の共起度合いを表す列を指定します。ここでは、各列の単語の出現頻度である「value」列を選択します。

パラメータ設定

● 隠れ変数(Z)の数/クラスター数

いくつのクラスターに分けるかという値を指定します。ここでは「3」 に指定します。

二項ソフトクラスタリング	7				? _ ×
変数選択					
列名		列型	X列	Y列)	スコア列
RowID		整数	\checkmark		
word		カテゴリ		\checkmark	
value		整数			
					- 1
ハフメータ設定					
隠れ変数(Z)の数/クラスタ数	3				_
学習回数	10				=
繰り返し数	10				=
比較候補数	1				=
14					
推薦指定					
□ Xに対するYの推薦を行う					=
件数上位 10				C	
オプション指定					
□ 確率上位の隠れ変数のみ出	力				= -
_			_		
			Į	新 -	保存

グラフ – 円グラフ(二項ソフトクラスタリング)

グラフの種類

作成するグラフの種類を指定します。ここでは「その他」の「円グラフ」を 選択します。

データの列

● データ選択

円グラフを作成するデータを選択します。分析結果テーブル からグラフを作成する場合は自動的に入力されています。

● カテゴリ

円グラフで表現したい列を指定します。ここでは「Z」を選択します。

1/2. グラフの種類を選択し	てください		@ ×
			*
折れ線			
散布図			
	1SIP		
ヒストグラム			
その他			
			v
BACK	NEXT	ок	CANCEL
2/2. データの列を選択して	こください		
1000000000000000000000000000000000000			

2/2. データの列を	選択してください		@ ×
● 用グラ:	7		
		で選択ダイアログを起動	して選択することもできます
データ選択	列属性変更 - result		•
カテゴリ	必須 Z		• •••
オプション設定 (3 #	油)		~
			θ
			•
BACK	NEXT	ок	CANCEL

アイコン - 行選択_pZX_rank1

インプット設定

行選択を行いたい対象列を指定します。ここでは、レビューのクラスタリング結果を見るため、「pZX」テーブルを指定します。

対象列

行選択を行う条件の対象列を指定します。ここでは、各レビューの所属確率が一番大きいデータを抽出するため、「rank」列を指定します。

rank

「rank」列に対する条件を指定します。

● 演算子

条件の演算子を指定します。「==」を指定し、一致する条 件を抽出します。

●式

条件式を指定します。ここでは rank=1 の行のみを抽出す るため、「1」とします。



行選択_pZX_ra	ink1	? _ ×
対象列		
🗖 列名	列型	条件式
X	category	in []
🗌 z	integer	
D pZX	float	
🔽 rank	integer	== 1
	対象列条件式	◉ 論理積(かつ) ○ 論理和(または)
rank		
	数値列条件式	● 論理積(かつ) ○ 論理和(または)
演算子 式		
1		×
+ 1		
		•
入力補助		
列名	列型	関数名 説明
Х	category	table[columnir 列を取得します。タ^
Z	integer	table.get_valu 指定されたセルの(
pZX	float	table.nrow() 行数を取得します。
		実行 ▼ 保存

アイコン – 列属性変更

対象列

列属性を変更したい対象列を指定します。円グラフ作成のため「Z」列 (クラスターID列)のデータ型を変更します。

Ζ

「Z」列の変更の設定を行います。

● 新列名

新しい列名に変更できます。特に変更がなければそのまま 利用します。

● 新列型

新しい列型を指定します。円グラフの変数として指定するため「category」とします。

列属的	按 更				0 _ ×
対象列					
	列名		列型	新列名	新列型
	Х		category	Х	category
\checkmark	Z		integer	Z	category
	pZX		float	pZX	float
	rank		integer	rank	integer
Z					
新列名		Ζ			
新列型		categor	у		·
					宝行 _ 保在

アイコン - マトリックス化_pZY

インプット設定

単語のクラスターへの所属確率を確認するため、「pZY」テーブルを指定します。

キー列

マトリックスのキーとなる列を指定します。単語である「Y」列を選択します。

横展開列

キー列以降の列名になる列を指定します。ここでは「Z」列を選択します。

内容列

所属確率である「pZY」列を指定します。

Input Matching Controller			×
		table	*
	рZX		
	pZY	-0-	
	рХZ		
二項ソフトクラスタ	pYZ		

マトリックス化	_pZY		? _ ×			
マトリックス化設定						
≠—列	Y 😒	*)			
横展開列	Z	*				
内容列	pZY 🕲	•				
詳細設定						
□ 横展開列名の接頭辞 prefix_ ==						
キー列名と横展開列名の重複 🚃						
◉ 最後に出現した項目を採用 ○ エラーで停止						
欠損の補填						
□ 整数	0	=				
□ 実数	0	=				
□ 日付	2022/12/27	\equiv				
□ 日時	2022/12/27 11:01	=	-			
		実行	▼ 保存			

補足情報 技術的な情報や利用規約について

© 2025 NTT DATA Mathematical Systems Inc.

辞書ファイル

「形態素・構文解析+」アイコンを利用する際には、ユーザー辞書、類義語辞書、分割辞書を利用することができます。

ユーザー辞書

単語の切れ目を変える辞書です。主に、つながって出てきてほしい複合語が、いくつかの単語として分かれて出てきてしまう場合などに利用します。

分割辞書

単語の切れ目を変えるために用いる辞書です。「構文解析と自動連結を行う」 にチェックを入れて分かち書きする際に、登録した内容に応じて「連結しないよう に」します。

類義語辞書

キー列以降の列名になる列を指定します。ここでは「Z」列を選択します。類義語をまとめ上げるために用いる辞書です。表記ゆれのまとめ上げに有用です。

これらの辞書はテキストの分割処理の際、右図のような流れで用いられます。



k-means法

クラスターの「中心」の探索、および、クラスター対象の要素がどの重心に一番近いかを計算し所属するクラスターを決めるという操作を繰り返し行うことで、全ての要素をいずれかのクラスターに割り当てます。

一度、分類した後にそのグループ分けが本当に最適か計算し、最適でなければ少しだけ分類を変えてグループを分けなおす作業を繰り返し行います。そのため、Step1の最初の点が異なると、最終的な分類結果が少し異なることがあります。(初期値依存性)

● クラスター数2の場合



代表点が変化しなくなるまで繰り返す

© 2025 NTT DATA Mathematical Systems Inc.

二項ソフトクラスタリング

二項ソフトクラスタリングは、2つの項目 X と Y の組み合わせで表されるデータに対して、同時に発生している組み合わせとクラスターとして抽出する方法です。テキストデータにおいては、1行のテキストデータと単語に対して適用することで、特許テキストと技術用語やレビューと評価単語のクラスタリングを行うことができます。

二項ソフトクラスタリングは、隠れクラスター $Z = \{Z_1, ..., Z_k\}$ があると仮定し、行と単語の間に隠れクラスターが介在し、それらを通じて行から単語が発生するというモデルを考えます。(右図)

内部的な計算のロジックとしては、行と単語の二項目の組み合わせの行列データを3つの 行列に分解したうえで、右辺の3つの行列について、はじめにランダムな値を初期値として 設定し、その積がXYの共起行列であるP(X,Y)と等しくなるように更新していく、という学 習をしています。(下図)





本文書・プロジェクトファイルのご利用にあたって

本文書ならびにプロジェクトファイルは、(株)NTT データ数理システム (以下「弊社」)が開発・販売 する分析プラットフォーム MSIP および Alkano と TextExtension の機能についての情報提供として弊 社が作成を行ったものです。弊社による事前の許可なしに、本文書の再配布や引用の範囲を超える複製 といった行為、およびリバースエンジニアリングを禁じます。

本文書ならびにプロジェクトファイルのご利用に際して、ご利用者様および第三者に損害が発生したとしても、 弊社は責任を負わないものとします。

プロジェクトファイルは、その中に同梱されているデータを利用し、本文書内で解説している設定可能なパラ メータで動作させた場合についてのみ、弊社にて動作の検証を行っております。これを超えるような状況にお ける動作は保証いたしません。

本プロジェクトファイル は、MSIP1.10.0 および Alkano1.4.0、TextExtension1.2.0 にて動作確認 を行っております。

お問い合わせ:株式会社NTTデータ数理システム 営業部 WEB: https://www.msi.co.jp/solution/analytics/index.html

株式会社 NTTデータ数理システム

NTT Data © 2025 NTT DATA Mathematical Systems Inc.