くらしの中に
総務省
MIC Ministry of Internal Affairs and Communications

書類番号
1

人工知能を活用した衣料品ネット販売価格のデータ分析業務検討状況

令和４年６月１日

総務省統計局
物価統計室

Statistics Japan

目次

１．分析業務の概要

1　分析業務の概要

- ウェブスクレイピングにより膨大なデータ収集が可能となったが、種類（品質）の異なる商品が多数混在
- 更なる活用拡大に向け、商品を自動分類する**AIの開発・実用化に向け研究**に着手（2019年度～）

ウェブスクレイピングによる商品情報収集

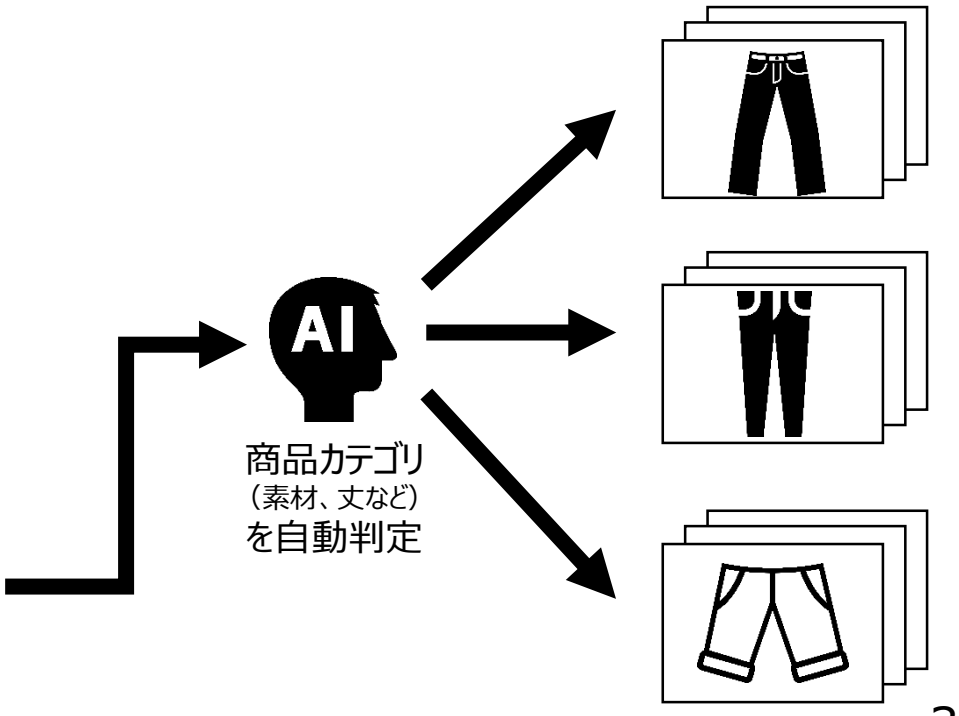
▶ 特に衣料品などの場合、商品の同一性・同質性の判別が困難なため、このままでは統計処理ができない

商品名	商品コード	価格	説明文
コーデュロイパンツ	なし	4,798	･･･コーデュロイパンツ細かな畝のコーデュロイ生地を使ったパンツ。トレンドを意識したシルエットがキュートです♪ウエストが総ゴムなので、着脱も着心地もよいのがうれしい特徴☆ ･･･
軽いデニム生地がかっこいいクロップド丈タックパンツ	なし	3,299	･･･ゆったりシルエットのクロップド丈タックパンツブルー100-130前身と後身の濃淡配色が目を引きます。シルエットも抜群です♪クロップド丈で足首出して元気に穿けます。軽いデニム生地がかっこいいです。･･･
･･･	･･･	･･･	･･･

AI（機械学習）による自動判定

▶ 商品の説明から、同質商品にカテゴリ分けするAIを試験的に開発

▶ AIアルゴリズムの改良、対象商品の拡大など、将来的な実用化に向け研究中



1　分析業務の概要（続き）

- 価格データが掲載されているサイトには、販売店舗を持つ個別企業のサイトや、多数の出展者の商品を扱うEC事業者のサイトが存在
- EC事業者のサイトでは、幅広い商品の情報が掲載されているものの、商品説明（商品名、素材など）や表記方法などの質・量が出展者によってバラバラな傾向

⇒ **同品質の商品の価格追跡が必要**なCPIにおいては、**ウェブスクレイピング活用は難易度が高い**

商品名称にPR情報等が混在

特定の商品名・コードが存在しない

商品の品質に関する情報の表記が不統一で情報量が膨大

衣料品ECサイトの掲載情報の例（背広服）

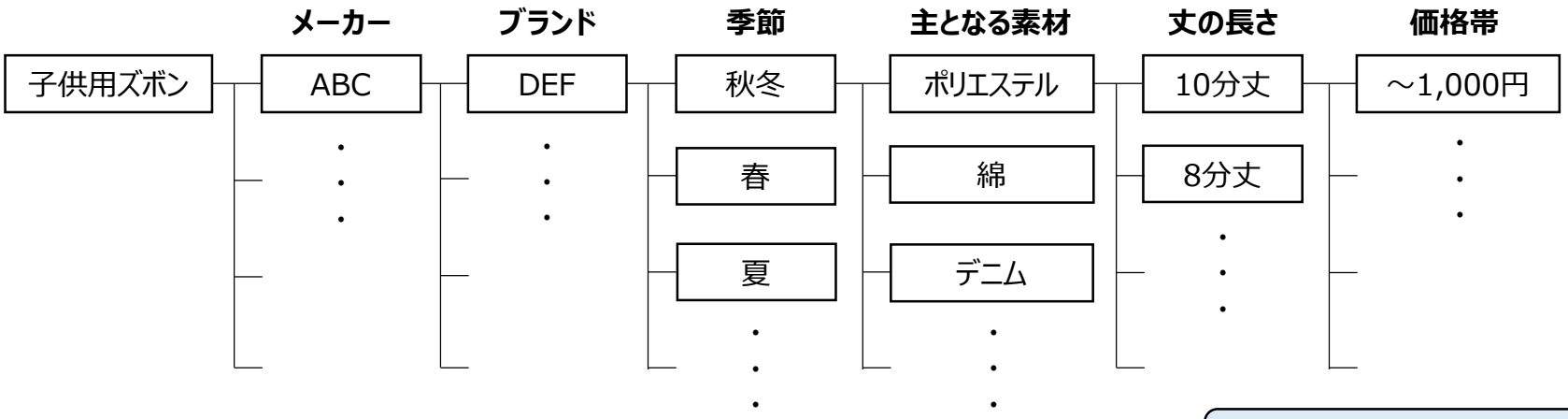
商品名	商品コード	出展者	商品説明の一部	商品説明の文字数
2つボタンスーツ ピンストライプ	–	A	･･･（略）･･･【生地】 Biella Finish イタリア技術者の指導を受け、原料選定から紡績、織布、仕上げまでの加工に、イタリアのノウハウを導入して作り上げた生地。高度な技術を具現化することで、イタリア製生地に遜色のない上質な生地に仕上がっています･･･（略）･･･【仕様：ジャケット】2つボタン／本切羽／センターベント／胸ポケット／腰ポケット×2／総裏仕立て／内ポケット×2【仕様：パンツ】ノータック／スリムフィットテーパード／エキストラローライズ／サイドポケット×2／ヒップポケット×2／ウォッチポケット／前ヒザのみ裏地あり【季節】秋冬モデル【洗濯表示】ドライオンリー ●体型 YA（Drop8：スリム） A （Drop6：普通）AB（Drop4：ややがっしり） ●身長【3】160cm ･･･（略）･･･	約1,500文字
1着は持っていたい!ウエストのアジャスター付きがうれしいシングルフォーマルスーツ メンズ	–	B	ビジネスばかりでなく、礼服・フォーマルにも使える黒無地スーツはマストアイテム。 ウエストにはアジャスターを装備し、左右で±3cmずつ調整できるので、長時間の座り姿勢のときや、食後にお腹が張っているときなども快適。 ベーシックなスーツなので、年齢も関係なく着用できるのがうれしい。 ■表地:毛50%ポリエステル50%(織物・背抜き)裏地:ポリエステル100% ▼スラックスの裾は半仕上げになっています。ご家庭でお好みの寸法にお直しください。ご家庭で簡単にできる裾上げテープを販売しておりますのでご利用ください。 ●ウエストアジャスター機能付き ●ドライのみ(ジャケット・パンツ) ●ミャンマー製･･（略）･･･	約300文字
秋冬物ウォッシャブル2ボタンスーツ【ブラック / ストライプ織柄】	–	C	･･･（略）･･･ご家庭でお洗濯可能! 2ボタンスーツ (8250-03) Item Information Dady Costa 2ボタンスーツのご紹介です。ウォッシャブル生地を採用し、ご家庭でお洗濯が可能なスーツです。 スタイルは極端にラインを細く強調するのではなく、着心地を重視した2ボタンスーツで、細身のスーツが苦手なお客様にもご満足いただけるスーツになっています。 ブラック地のストライプ織柄でオーソドックスながら品の良さがあり、どのような場面でも対応可能で幅広くご着用可能なスーツです。ズボンの折り目が消えない耐久折り目加工や、貴重品を守るフラップ付き内ポケットなど、充実した仕様となっております。 お手頃な価格ですので、スーツを何着も必要なお客様や家計のやりくりで苦労している奥様にも喜んでいただける商品です。 Dady Costa Washable 2 Button Suit 8250-03 WEB Price（税抜き）19,000 Information Size List A4 A5 A6 A7 AB4 AB5 AB6 AB7 BB4 BB5 BB6 BB7 Cloth Information ブラック / ストライプ織柄 ウール55% ポリエステル45% 日本製生地 ブラックのバイアス織柄地に4mm間隔でトライプが入っています･･･（略）･･･	約2,000文字

商品の同一性・同質性の判別が困難

1　分析業務の概要（続き）

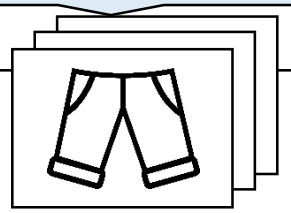
衣料品ネット販売データの製品区分などの格付けのため、AI（機械学習）の活用可能性について検証中
- 2019年度～20年度：　商品情報（文字情報）を対象に、最適な機械学習アルゴリズムを検証
- 2021年度～　　：　商品画像の併用可能性について検証、検証対象サイトの拡大

分類イメージ

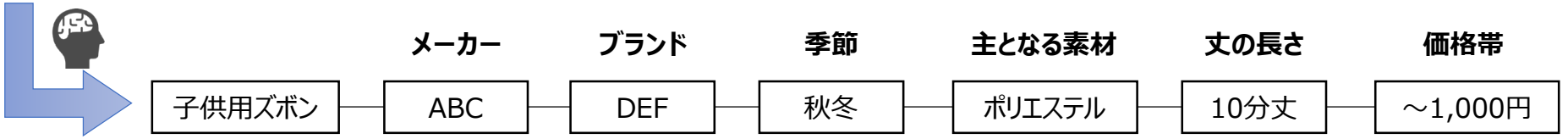


AI機械学習　イメージ

2021年度から、商品画像の併用可能性について検証中



商品名 ：**ABC** **長ズボン** キッズ 子供服 男の子 **2～5歳** **薄手のデニム** ジーンズ **青**
説明　：薄手なので**季節の変わり目**にぴったり！ キッズ　薄手のデニムジーンズ　長ズボン(２～５歳：**90･95･100･110**)
薄手のデニムが入荷しました。薄手なので、季節の変わり目や蚊の多い公園など**アクティブ**に過ごすときに最適！
半ズボンから長ズボンに変えるのを嫌がるお坊ちゃまにも。
【キッズ】【男の子】 素材 **ポリエステル100%** **サイズ 2T,3T,4T,5T**
注意点 ○洗濯機、乾燥機は弱で。○漂白剤不可。○**色落ち**する場合がございますので、最初の洗濯時は白いものとは分けてください。



２．2021年度の研究内容と結果

２．１　研究概要

2021年度は、以下のとおり研究を行った。

○研究概要

衣料品の通販サイトから文字情報及び画像情報を取集し、人工知能（AI）を使用して、品目の分類を行い、より精度の高いモデルの構築や消費者物価指数への影響を研究する。

○分析対象品目（３商品、６品目）

ワンピース（春夏物）、ワンピース（秋冬物）、スラックス（ブルージーンズ）、 スラックス（秋冬物）、
子供用ズボン（春夏物）、子供用ズボン（秋冬物）

○データ取集対象サイト

衣料品専門店（３サイト）⇒モール型販売サイト･･･１サイト、自社商品販売型サイト･･･２サイト

○研究内容

１．文字情報によるＡＩ分析手法の研究

商品ページに記載されている商品説明等の文字情報について、AIを用いた各品目への分類の精度を検証する。

２．画像情報によるＡＩ分析手法の研究

商品ページに記載されている画像情報について、AIを用いた各品目への分類の精度を検証する。

３．文字情報と画像情報を併用した手法の研究

文字情報と画像情報を併用し、AIを用いた各品目への分類の精度を検証する。

４．価格指数の作成及び比較

分析によって得られた分類結果から試験的に価格指数を作成する。また公表値と比較分析を行う。

２．２　文字情報を利用した研究

文字情報を利用した研究は、以下の内容について行った。

（１）正解ラベルの付与

分析対象品目ごとにあらかじめカテゴリを作成し、そのカテゴリごとにラベルを設定した。

サイトから取集した分析対象品目の文字情報、全1万件に対して、人手による作業でラベルを付与し、正解データを作成した。

例）商品：　ワンピース、スラックス及び子供用ズボン

カテゴリ：　季節、丈の長さ…

ラベル：　季節→春夏物、秋冬物、通年　　丈の長さ→フルレングス、7分丈…

（２）前処理

① データクリーニング

取集した文字情報の半角全角及び大文字小文字の統一のみを行った。

（省略語の統一等の処理は行っていない。）

② 形態素解析

文字情報を「名詞」「動詞」「助動詞」などの形態素に分割を行った。形態素解析にはMecabを用いた。

③ 形態素のベクトル化

TF-IDFを用い、分割した形態素のベクトル化を行った。TF-IDFとは「ある単語の文章中における出現頻度の値」（TF）と「ある単語の全文章中での希少性の値」（IDF）を乗じた値であり、文章内における単語の重要度を表す。

２．２　文字情報を利用した研究（続き）

（３）モデルの作成

教師データを学習させるアルゴリズムには、以下の２種類を用いた。

① ロジスティック回帰

いくつかの要因（説明変数）から二値の結果（目的変数）が起こる確率を0~1で説明・予測する手法である。この手法を用いて、ラベルごとに起こる確率を求め、最も確率が高いラベルに分類することにより、多値分類を行う。

② LightGBM

LightGBMは、勾配ブースティング決定木の手法の一つである。類似の手法としてXGBoostなどが存在するが、LightGBMはそれらと比較し早く学習を行える特徴がある。

勾配ブースティング決定木は「勾配降下法(Gradient)」、「アンサンブル(Boosting)」、「決定木(DecisionTree)」を組み合わせた手法である。まずはカテゴリを当てる「決定木」を作成、次にその誤差を当てる「決定木」を作成…を繰り返し、それらを組み合わせ誤差を修正してモデルを作成する。

（４）精度検証

作成したモデルを用いて検証データ及び教師データのカテゴリごとの分類を行い、付与された正解データとの比較検証を以下の手順で行った。

① 分類された結果と正解データが一致したか（多値分類で一致したか）の検証（２種類のモデル）
② 分類された結果が特定の条件に一致したか（二値分類で一致したか）の検証（１種類のモデル）

２．２　文字情報を利用した研究（続き）

（５）検証結果

① 多値分類による検証

多値分類の検証では、(1)及び(2)の作業を行った1万件のデータを8,000件と2,000件にランダムに分割した。そのうち、8,000件を教師データとして２種類のアルゴリズムに学習させ、モデルを作成した。2,000件を検証データとして、教師データと共にモデルを用いて分類を行った。

多値分類の結果を２種類のモデルごとに比較すると、下表のとおり、LightGBMがロジスティック回帰よりも、若干精度が高かったが、大きな差は生じなかった。このため、解釈性の高さを考慮して、この後の二値分類では、ロジスティック回帰を採用する。

多値分類結果（スラックスを対象）

カテゴリ　（（）内は当該カテゴリのラベル）	ロジスティック回帰		LightGBM	
	検証データ精度	教師データ精度	検証データ精度	教師データ精度
種類（デニム、スラックス、カーゴ等）	0.779	0.830	0.774	0.868
季節（春夏、秋冬等）	0.826	0.845	0.878	0.995
丈（フルレングス、ショート等）	0.897	0.947	0.961	0.993
シルエット（スタンダード、ワイド、スキニー等）	0.923	0.948	0.955	0.993
素材（綿、毛、化繊等）	0.893	0.911	0.918	0.976
セール（セール、定価）	0.982	0.995	0.989	1.000
柄（無地、ライン、花柄等）	0.859	0.911	0.893	0.987

精度の計算方法・・・データ総数に占める、多値分類の結果と正解ラベルが一致したデータ数の比率

2．2　文字情報を利用した研究（続き）

(５) 検証結果（続き）

② ロジスティック回帰を用いた二値分類による検証

二値分類の検証では、(1)及び(2)の作業を行った1万件のデータから、4,000件をランダムで抽出した。そのうち、2,000件を教師データとして、新規でロジスティック回帰のアルゴリズムに学習させ、モデルを作成した。残りの2,000件を検証データとして、教師データと共にモデルを用いて分類を行った。

二値分類の結果をみると、下表のとおり、最低でも83%、最も高いものでは99%の精度となった。

ロジスティック回帰を用いた二値分類結果

品目	カテゴリ	条件	検証データ精度	教師データ精度
ワンピース 5161	素材	綿100%、綿50%以上の混紡	0.895	0.920
	季節	春、夏、通年	0.904	0.918
ワンピース 5163	素材	化繊	0.874	0.927
	季節	秋、冬、通年	0.867	0.877
スラックス 5179	種類	デニムパンツ（青）	0.832	0.864
	長さ	フルレングス	0.931	0.963
	素材	綿100%	0.952	0.966
スラックス 5181	長さ	フルレングス	0.931	0.963
	素材	毛100%、毛50%以上の混紡	0.987	0.990
	季節	春、秋、冬、通年	0.981	0.972
子供用ズボン 5191	対象	女児、乳児を除く	0.882	0.906
	長さ	ショート、ハーフ	0.976	0.981
	素材	綿100%、綿50%以上の混紡	0.882	0.914
	季節	春、夏、通年	0.945	0.953
子供用ズボン 5196	対象	女児、乳児を除く	0.882	0.906
	長さ	フルレングス	0.939	0.952
	素材	化繊、綿100%、綿50%以上の混紡	0.849	0.929
	季節	秋、冬、通年	0.928	0.928

精度の計算方法・・・データ総数に占める、二値分類の結果と正解ラベルが一致したデータ数の比率

２．３　画像情報を利用した研究

画像情報を利用した研究は、以下の内容について行った。

（１）正解ラベルの付与

サイト上の子供用ズボンの画像情報全276,925枚を、商品番号をキーとして、文字情報の研究においてラベルを付与した文字情報と紐づけることにより、正解データのラベルを付与した。

（２）モデルの作成

画像情報の分類には、以下の２種類のアルゴリズム（学習前モデル）を用いた。さらに、(1)の作業を行った276,925枚の画像情報をランダムに分割し、8割(221,540枚)を教師データとして、当該アルゴリズムに学習させ、モデル（学習後モデル）を作成した。また、残りの2割(55,385枚)を検証データとした。

① ResNet

Microsoftによって開発されたアルゴリズムで、それ以前に提案されていたアルゴリズムよりもはるかに層を深くできるように設計されている。

② EfficientNet

2019年にGoogleから発表されたアルゴリズムで、従来よりも少ないパラメータでImageNetにおける当時の最高精度を出した。ResNetと共に画像認識の分野における代表的なアルゴリズムのひとつとなっている。

２．３　画像情報を利用した研究（続き）

（３）精度検証

学習前モデル及び学習後モデルを用いて検証データの分類を行い、正解データとの比較検証を行った。

正解率は、下表のとおり、学習前モデルを用いると約50%にとどまるが、学習後モデルを用いると、ResNetは90.9%、EfficientNetは91.4%といずれも90%を超え、EfficientNetがわずかに上回る結果となった。

他方で、学習後モデルによる演算時間は、ResNetの2時間53分に対し、EfficientNetは9時間51分となった。以上を踏まえると、演算時間を抑えつつもEfficientNetとほぼ同等の精度を出すことのできるResNetの方が有用と考えられる。

○子供用ズボン　カテゴリ「種類」　ラベル「スラックス」分類結果

	検証データ数	学習前		学習後		演算時間
		正解数	正解率	正解数	正解率	
ResNet	55,385	28,634	51.7%	50,355	90.9%	2時間53分
EfficientNet		27,471	49.6%	50,648	91.4%	9時間51分

※正解数＝ラベルが正しく付与されたデータ数（正しく対象外となったデータを含む）

※正解率＝正解数／検証用データ数

２．４　文字情報及び画像情報を併用した研究

文字情報、画像情報のそれぞれを利用した研究で得られた特徴量を変数としてロジスティック回帰を用いたモデルを作成した。このモデルを用いて、文字情報及び画像情報の組合せを検証データとしてカテゴリごとの分類を行い、付与された正解データとの比較検証を行うとともに、文字情報又は画像情報のみの場合と精度の比較を行った。

なお、画像情報は、１つの商品につき複数存在する場合、当該画像情報の特徴量を平均化したものを当該商品の値として使用した。

文字情報と画像情報の合成イメージ

$$a + b_1 x_1 + b_2 x_2 + c_1 y_1 + c_2 y_2 \sim c_{499} y_{499} + c_{500} y_{500}$$

既存の自然言語処理モデルに画像モデルから得られたパラメータ等も投入する。

お客様の声から生まれたフレアジーンズに新色が登場。ハイウェストときれいなフレアではくだけで脚長、美脚に見える。ウエストラインを高い位置に設定し、裾にかけてゆるやかに広がるセミフレア。

画像1　画像2　…　画像9

TFIDF結果を出力

シルエット	アイテム	ウエスト	場合	ブランド	デニム	パンツ	～	デザイン
6,199	0	0	0	172,201	0	0		0

画像処理中間層の結果を平均する

0	1	2	～	511
0.070698	0.530883	0.453604		0.01586

あるカテゴリが指定のラベルである確率97.5%

テキスト結果と画像結果を用いたロジスティック回帰モデルで確率を計算

２．４　文字情報及び画像情報を併用した研究（続き）

文字情報及び画像情報を併用した分類結果は、一般に、文字情報又は画像情報のみを利用した結果に比べ、精度が高い結果となった。特に、検証データについては、全18カテゴリで文字情報のみを利用した結果に比べ、精度が高い結果となった。

			検証データ精度			教師データ精度		
品目	カテゴリ	条件	文字	画像	両方	文字	画像	両方
ワンピース5161	素材	綿100%、綿50%以上の混紡	0.895	0.943	0.948	0.920	0.925	0.958
	季節	春、夏、通年	0.904	0.969	0.985	0.918	0.957	0.990
ワンピース5163	素材	化繊	0.874	0.932	0.975	0.927	0.912	0.990
	季節	秋、冬、通年	0.867	0.962	0.971	0.877	0.945	0.981
スラックス5179	種類	デニムパンツ（青）	0.832	0.899	0.924	0.864	0.880	0.945
	長さ	フルレングス	0.931	0.882	0.953	0.963	0.848	0.972
	素材	綿100%	0.952	0.920	0.978	0.966	0.902	0.984
スラックス5181	長さ	フルレングス	0.931	0.882	0.953	0.963	0.848	0.972
	素材	毛100%、毛50%以上の混紡	0.987	0.992	0.997	0.990	0.991	0.997
	季節	春、秋、冬、通年	0.981	0.963	0.984	0.972	0.956	0.984
子供用ズボン5191	対象	女児、乳児を除く	0.882	0.952	0.956	0.906	0.935	0.989
	長さ	ショート、ハーフ	0.976	0.965	0.981	0.981	0.951	0.960
	素材	綿100%、綿50%以上の混紡	0.882	0.913	0.900	0.914	0.849	0.940
	季節	春、夏、通年	0.945	0.963	0.986	0.953	0.956	0.989
子供用ズボン5196	対象	女児、乳児を除く	0.882	0.952	0.956	0.906	0.935	0.960
	長さ	フルレングス	0.939	0.921	0.949	0.952	0.902	0.947
	素材	化繊、綿100%、綿50%以上の混紡	0.849	0.914	0.912	0.929	0.896	0.927
	季節	秋、冬、通年	0.928	0.962	0.991	0.928	0.958	0.985

総件数（文字情報（検証データ・学習データともに）・・・ 1品目あたり2,000件、画像・・・文字情報に紐付く画像すべて）

２．５　価格分析及び指数精度の検証

文字情報及び画像情報を併用した分類結果を用いて、以下の要領にて価格指数を作成し、消費者物価指数の公表値と比較を行った。

○品質調整
・同一特性の商品群内で平均価格との差が標準偏差の3倍を超える商品をノイズデータとして除外

○計算手順
①以下の表のとおり分類を行い、企業毎に平均価格を算出
②各企業の平均価格に対し、各企業の年間売り上げをウエイトとして加重平均
③2021年8月～2022年2月の平均価格を100として、指数を計算

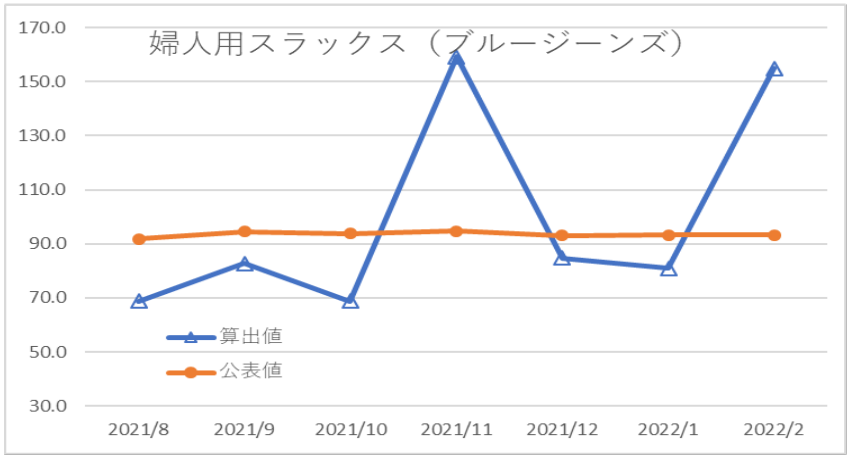
カテゴリ	ワンピース　5161 （春夏物）	ワンピース　5163 （秋冬物）	婦人用スラックス　5179 （ブルージーンズ）	婦人用スラックス　5181 （秋冬物）	子供用ズボン　5191 （春夏物）	子供用ズボン　5196 （秋冬物）
種類	すべて	すべて	デニムパンツ（青）	すべて	すべて	すべて
丈の長さ	－	－	フルレングス	フルレングス	ショート、ハーフ	フルレングス
対象	－	－	－	－	女児、乳児除く	女児、乳児除く
素材	綿100%、 綿50%以上の混紡	化繊	綿100%	毛100%、 毛50%以上の混紡	綿100%、 綿50%以上の混紡	化繊、綿100%、 綿50%以上の混紡
季節	春、夏、通年	秋、冬、通年	すべて	春、秋、冬、通年	春、夏、通年	秋、冬、通年

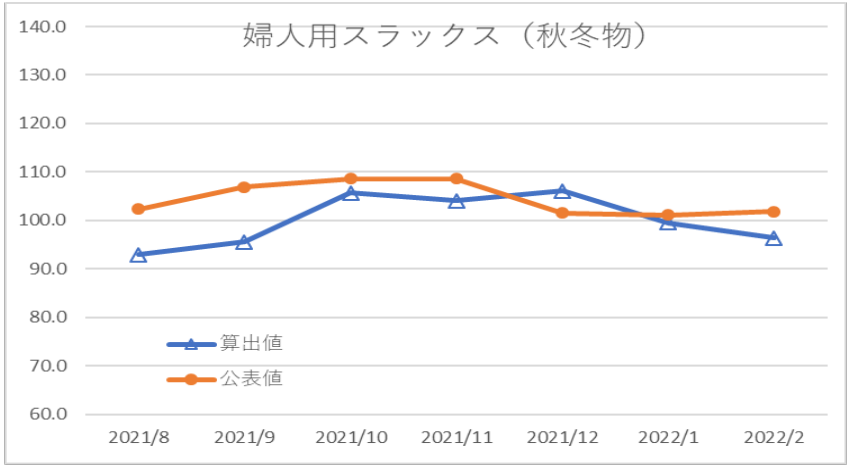
"－"は品目に設定していないカテゴリ

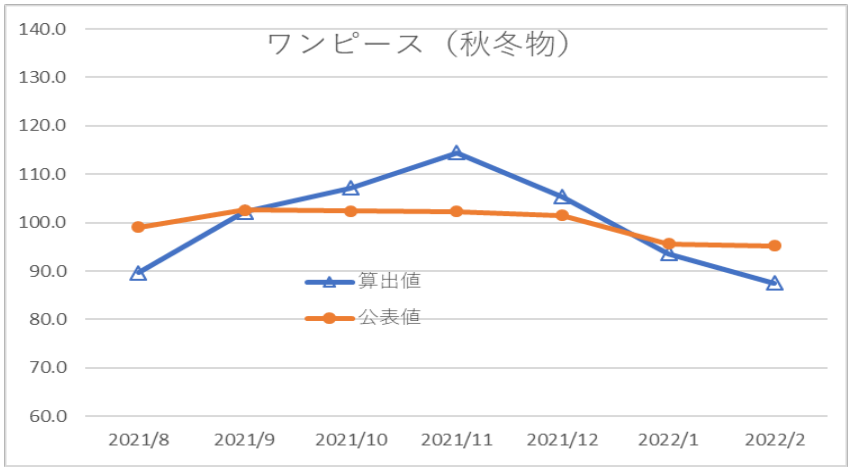
４．価格分析及び指数精度の検証

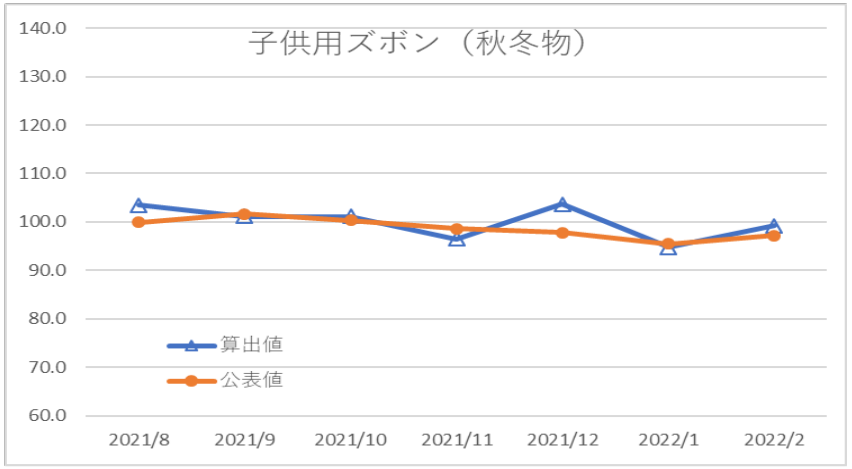
試算した指数の比較は以下のとおりとなった。

婦人用スラックス（ブルージーンズ）について、11月及び２月の価格指数は、通常の価格変動だけでは説明できない動きをしていることから、実用化に向けて、ノイズデータの排除など、精度向上に資するデータ処理が必要。









※ワンピース（春夏物）、子供用ズボン（春夏物）については、価格調査対象外期間であるため、指数の試算は行っていない。

３．次回へ向けての課題
及び2022年度の研究内容

３．１　次回へ向けての課題

以上の研究結果から、文字情報及び画像情報を併用することによって、一定の分類精度を確保することができた。一方で、以下のような課題が見つかった。

１．指数作成の実務上の課題

画像情報は文字情報と比べて容量が桁違いであり、以下のような支障が出た。

・容量の制限により大容量転送システムが使用できず、外付けSSDにデータを格納し、セキュリティ対策を施した形で郵送を行ったが、郵送の手続き等を含めデータの送受に数日を要することになった。

・ウェブスクレイピングデータの受領から指数の作成までの作業時間を現状のウェブスクレイピング品目と比較すると、以下のとおりとなる。レコード数の増加する本番運用を考慮すると、許容範囲内に収まるか懸念が残る。

ウェブスクレイピング品目の演算時間の比較

ウェブスクレイピング採用済の品目		
	データ数	演算時間
航空運賃	約250万	約15分
宿泊料	約150万	約10分

今回検証の品目		
	データ数	演算時間
衣料品	文字データ　約1万	約５分
衣料品	文字データ　約1万 画像データ　約2000	約1時間30分

本番運用では、文字データ約100万とそれに付随する画像データを見込む

２．作成した指数の結果精度上の課題

・今年度は、衣料品販売のシェアを重視しサイトを選定し、一定程度の商品数を確保できたが、必ずしも、市場における充分なシェアを確保できていない。

・「婦人用スラックス」など通常の価格変動だけでは説明できない動きをしている指数について、実用化に向けて、ノイズデータの排除など精度向上に資するデータ処理が必要。

3．2　2022年度の研究内容（案）

2021年度の研究結果を踏まえ、2022年度は以下の内容について研究する。

１．実運用への適用の可能性の研究

○画像情報の運用可能性の研究

取集する画像数を、2021年度は1商品あたり10画像から、2022年度は半分程度に削減し、精度維持の施策として、以下の点について検証を行う。

・画像点数を増減（１枚～５枚）させ、精度への影響を確認する。
・取集した画像を向きを変える、角度を変える等により枚数を増大させ、精度を検証する。
・演算時間の短縮効果を検証する。

２．精度向上に関する研究

○取集対象サイトの拡大

本番運用を見据え、商品をより多く取集できるように取集対象企業を拡大する。（３企業→７企業）

拡大対象は、衣料品販売売り上げや小売物価統計調査の調査店舗等を考慮し決定する。

○より効果的なデータ処理の研究

分類結果や価格指数の精度向上のため、より有効なデータの処理を行う。

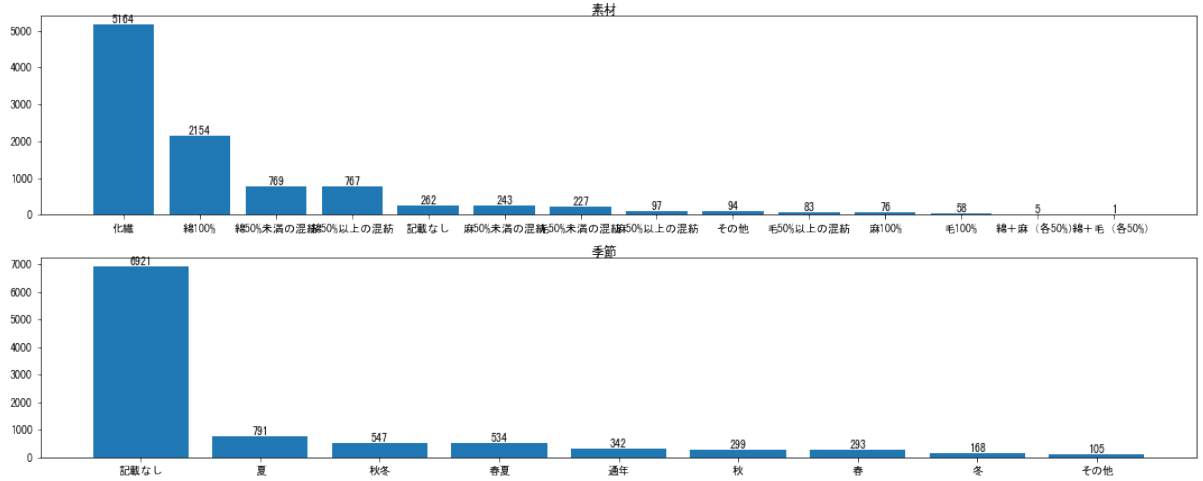
・作成した価格指数と消費者物価指数の比較
・比較時に観察された差の要因分析及び対応方法の検討・研究

参考資料

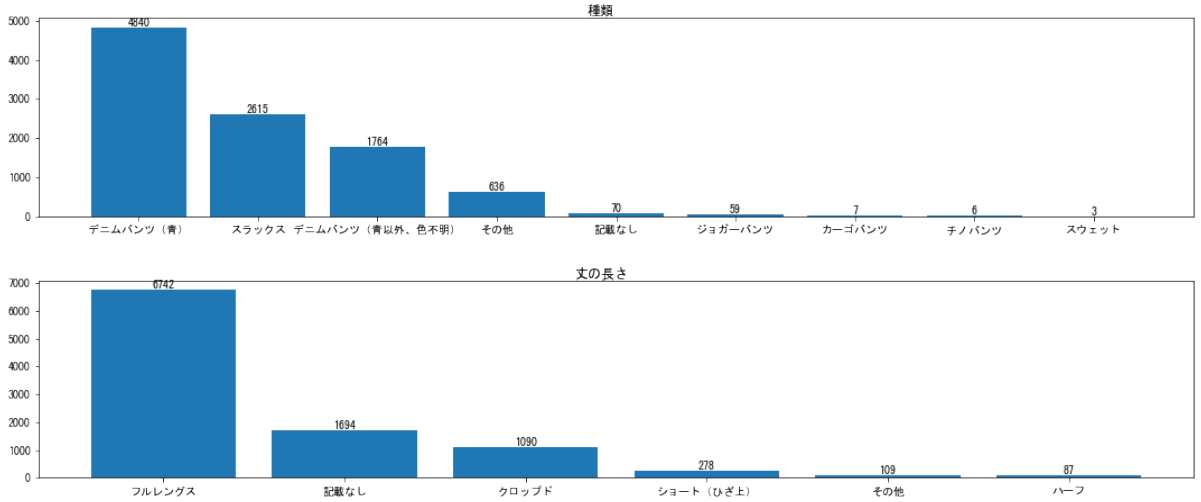
（参考）文字情報の分布状況等

・正解ラベルの分布状況

・ワンピース



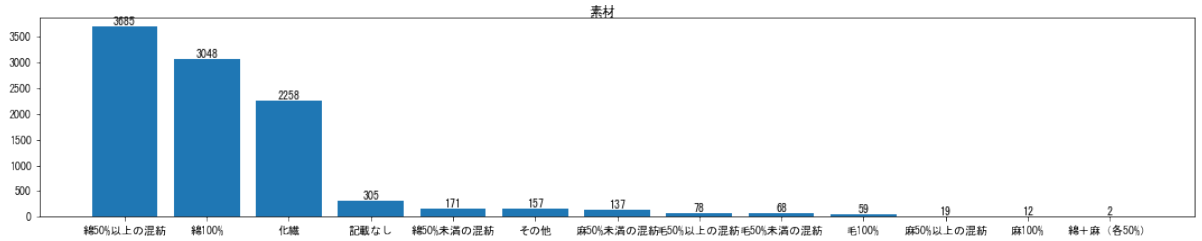
・スラックス

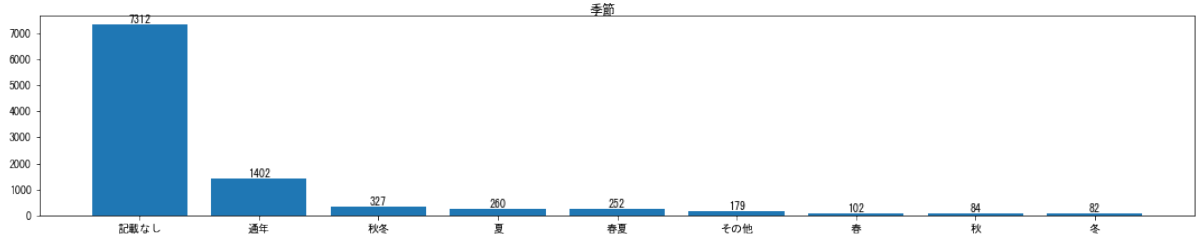


（参考）文字情報の分布状況等（続き）

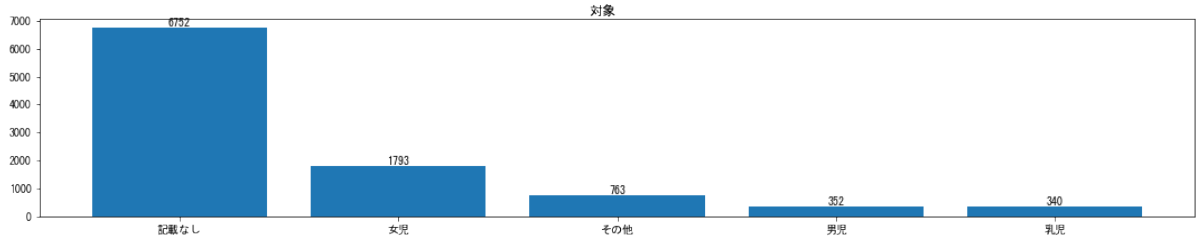
・正解ラベルの分布状況（続き）

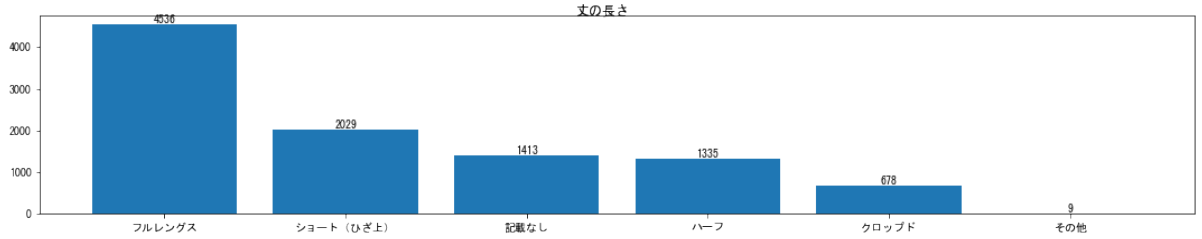
・スラックス（続き）





・子供用ズボン

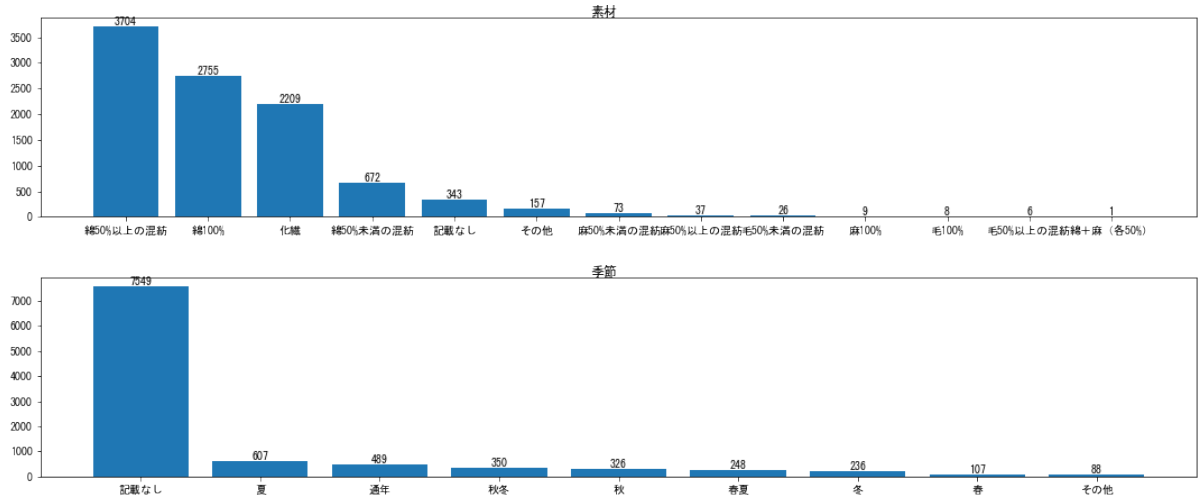




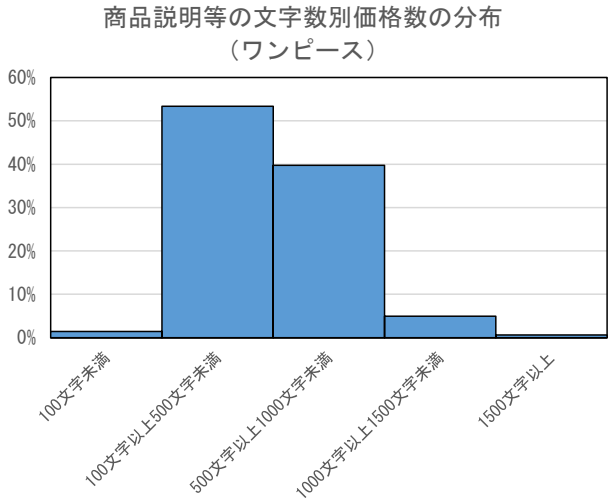
（参考）文字情報の分布状況等（続き）

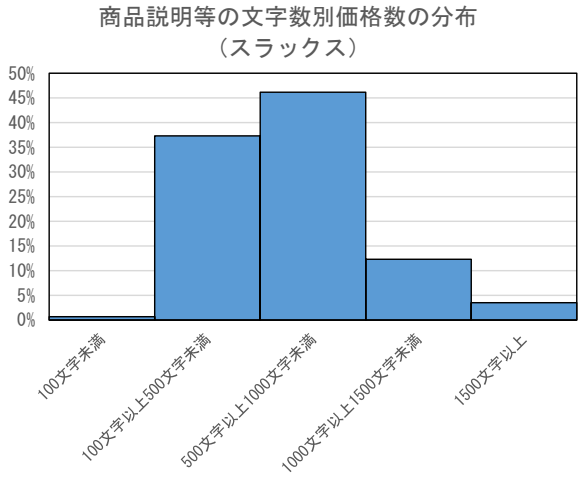
・正解ラベルの分布状況（続き）

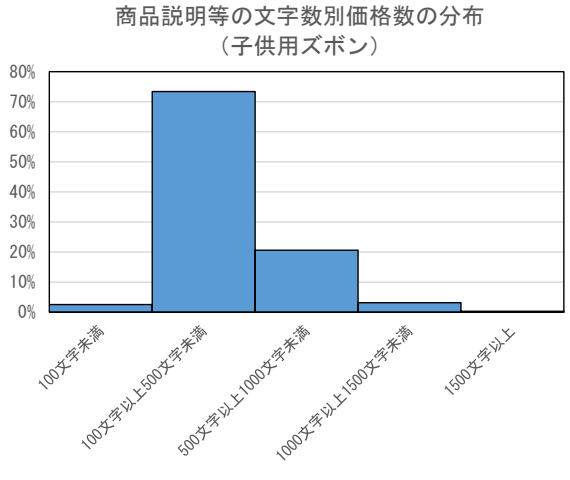
・子供用ズボン（続き）



・商品説明等の文字数の分布状況







（参考）形態素解析について

・形態素解析

文章を「名詞」「動詞」「助動詞」などの形態素に分割を行う。形態素解析にはMecabを用いた。
（形態素解析の例）

きれいなシルエットにこだわったデニムです

きれい / な / シルエット / に / こだわっ / た / デニム / です
（形状詞） （助動詞） （名詞） （助詞） （動詞） （助動詞） （名詞） （助動詞）

・形態素のベクトル化

形態素のままでは機械学習の処理を行うことが難しいため、TF-IDFを用い形態素のベクトル化を行った。TF-IDFとは「ある単語の文章中における出現頻度」と「ある単語の全文章中での希少性」を掛け合わせた値である。これを用いることで文章内の単語の重要度を求めることができる。

TF-IDFの算出及び特徴語の抽出

TF-IDFは各文章中に含まれる各単語が「その文章内でどれくらい重要か」を表す尺度で、具体的には「（ある単語の文章中における出現頻度）×（ある単語の全文章中での希少性）」であらわす。
これにより文章内の単語の重要度を求めることができる。

TF-IDF= TF×IDF

TF = 単語iの文章dにおける出現回数/文章dにおける全単語の出現回数の和
IDF = log(総文章数/ある単語iを含む文章の数)＋ 1

（参考）形態素解析について（続き）

・形態素のベクトル化

TF-IDFの算出及び特徴語の抽出の具体例

（例）：以下のような商品AとBの説明文がある場合
A：「ストレッチデニムを使用した、冬にぴったりなデニムパンツです。」
B：「着回しに便利なデニムパンツです。」
↓上記を形態素解析
A：['ストレッチ' 'デニム' '使用し' '冬' 'ぴったり' 'デニム' 'パンツ']
B：['着回し' '便利' 'デニム' 'パンツ']

「デニム」「冬」という単語のTF値の求め方は、以下のとおりである。
TF(デニム,文章A) ＝ 2/7≒0.28
TF(冬,文章A) ＝ 1/7≒0.14

IDF値の求め方は、以下のとおりである。
IDF(デニム) =log(2/2)+1 =1
IDF(冬) =log(2/1)+1 ≒1.3

上記を掛け合わせた値（TFIDF）は、以下のとおり。
TFIDF（デニム,文章A）≒0.28 ×1 =0.28…
TFIDF（冬,文章A） ≒0.14 ×1.3=0.18…

上記のような計算を他の単語に対しても行うことで商品Aを表す語として、文章内に頻繁に出現する「デニム」や他の文章ではあまり見られない「冬」といった単語の抽出を行う。