

テキストマイニング入門（基礎編）

前田利之（まえだ としゆき）

maechan@hannan-u.ac.jp

阪南大学 経営情報学部

(2024.12.04)

テキストマイニング（背景）

- 近年、WWW/mail などにより非常に多くの文書が されている
- 文書が多くなると、自分が知りたいことを見つける材料は増えるが、本当に を見つけ出すことは難しくなる
- →多くの文書から自動的に特徴を抽出することで文書を整理したり、整理を通じて新たな 技術が重要となる

⇒

テキストマイニング（背景）

- 近年、WWW/mail などにより非常に多くの文書が蓄積されている
- 文書が多くなると、自分が知りたいことを見つける材料は増えるが、本当に [] を見つけ出すことは難しくなる
- →多くの文書から自動的に特徴を抽出することで文書を整理したり、整理を通じて新たな [] 技術が重要となる

⇒

[]

テキストマイニング（背景）

- 近年、WWW/mail などにより非常に多くの文書が蓄積されている
- 文書が多くなると、自分が知りたいことを見つける材料は増えるが、本当に欲しい情報のみを見つけ出すことは難しくなる
- →多くの文書から自動的に特徴を抽出することで文書を整理したり、整理を通じて新たな
技術が重要となる

⇒

テキストマイニング（背景）

- 近年、WWW/mail などにより非常に多くの文書が **蓄積** されている
- 文書が多くなると、自分が知りたいことを見つける材料は増えるが、本当に **欲しい情報のみ** を見つけ出すことは難しくなる
- →多くの文書から自動的に特徴を抽出することで文書を整理したり、整理を通じて新たな **知見を見出す** 技術が重要となる

⇒

テキストマイニング（背景）

- 近年、WWW/mail などにより非常に多くの文書が蓄積されている
 - 文書が多くなると、自分が知りたいことを見つける材料は増えるが、本当に欲しい情報のみを見つけ出すことは難しくなる
 - →多くの文書から自動的に特徴を抽出することで文書を整理したり、整理を通じて新たな知見を見出す技術が重要となる
- ⇒ テキストマイニング

テキストマイニング（続き）

【参考】言語処理のステップ

- :文を単語に分解する（後述）
- :文を構成する全ての単語間の構造（構文木）を作る
- :語の意味を利用して（複数の可能性のあるなかで1つの）構文木を選択する

テキストマイニング（続き）

【参考】言語処理のステップ

- **形態素解析**: 文を単語に分解する（後述）
- : 文を構成する全ての単語間の構造（構文木）を作る
- : 語の意味を利用して（複数の可能性のあるなかで1つの）構文木を選択する

テキストマイニング（続き）

【参考】言語処理のステップ

- **形態素解析**: 文を単語に分解する（後述）
- **構文解析**: 文を構成する全ての単語間の構造（構文木）を作る
- : 語の意味を利用して（複数の可能性のあるなかで1つの）構文木を選択する

テキストマイニング（続き）

【参考】言語処理のステップ

- **形態素解析**: 文を単語に分解する（後述）
- **構文解析**: 文を構成する全ての単語間の構造（構文木）を作る
- **意味解析**: 語の意味を利用して（複数の可能性のあるなかで1つの）構文木を選択する

テキストマイニング（続き）

【参考】言語処理のステップ

■ 意味解析の例：「高い山と海」

- 「高い山と海」は【【高い山】と海】と【高い【山と海】】の可能性がある
- 「高い海」は意味としておかしい（【温度が高い】は、ありうるが…）
- よって、前者が正しいとする

テキストマイニング（続き）

【参考】言語処理のステップ（続き）

- :複数の文にまたがる構文木作成と意味解析を行なう
- …ここまでの解析で文章の意味を理解した上で、
、などを行なう
 - Siri などに関しては、さらに前段階として音声認識が伴う（これがまた大変）

テキストマイニング（続き）

【参考】言語処理のステップ（続き）

- **文脈解析**: 複数の文にまたがる構文木作成と意味解析を行なう
- …ここまでの解析で文章の意味を理解した上で、
、などを行なう
 - Siri などに関しては、さらに前段階として音声認識が伴う（これがまた大変）

テキストマイニング（続き）

【参考】言語処理のステップ（続き）

- **文脈解析**: 複数の文にまたがる構文木作成と意味解析を行なう
- …ここまでの解析で文章の意味を理解した上で、**機械翻訳**、などを行なう
 - Siri などに関しては、さらに前段階として音声認識が伴う（これがまた大変）

テキストマイニング（続き）

【参考】言語処理のステップ（続き）

- **文脈解析**: 複数の文にまたがる構文木作成と意味解析を行なう
- …ここまでの解析で文章の意味を理解した上で、**機械翻訳**、**問い合わせ処理**などを行なう
 - Siri などに関しては、さらに前段階として音声認識が伴う（これがまた大変）

テキストマイニング（続き）

- 文を構成する最小の要素=
- 文を に分解し、個々の品詞を特定すること
=
 - 日本語の品詞@学校文法：動詞、形容詞、形容動詞、名詞（&代名詞）、連体詞、副詞、接続詞、感動詞、助動詞、助詞

テキストマイニング（続き）

- 文を構成する最小の要素= **形態素（単語）**
- 文を に分解し、個々の品詞を特定すること
=
 - 日本語の品詞@学校文法：動詞、形容詞、形容動詞、名詞（&代名詞）、連体詞、副詞、接続詞、感動詞、助動詞、助詞

テキストマイニング（続き）

- 文を構成する最小の要素= 形態素（単語）
- 文を 形態素 に分解し、個々の品詞を特定すること
=
- 日本語の品詞@学校文法：動詞、形容詞、形容動詞、名詞（&代名詞）、連体詞、副詞、接続詞、感動詞、助動詞、助詞

テキストマイニング（続き）

- 文を構成する最小の要素= **形態素（単語）**
- 文を **形態素** に分解し、個々の品詞を特定すること
= **形態素解析**
 - 日本語の品詞@学校文法：動詞、形容詞、形容動詞、名詞（&代名詞）、連体詞、副詞、接続詞、感動詞、助動詞、助詞

形態素解析（おまけ）

- 「日本語の品詞@学校文法」とあるように、「学校文法」じゃない【文法】もある
 - 山田文法
 - 松下文法
 - 橋本文法 …学校文法の規範とされる
 - 時枝文法 …前田的には、この中で一番馴染みがある

形態素解析（おまけ）

- そもそも、文法があって言語がある
- 言語を分析して、文法という
している
- →いろいろな文法があって当然（?!）

形態素解析（おまけ）

- そもそも、文法があって言語があるわけではない！
- 言語を分析して、文法という している
- →いろいろな文法があって当然（?!）

形態素解析（おまけ）

- そもそも、文法があって言語がある
わけではない！
- 言語を分析して、文法という
概念を構築している
- →いろいろな文法があって当然（?!）

形態素解析（おまけ）

- 子供は文法など考えずに喋ってる←言語を取り扱っている
- 非文（文法的に正しくない文）でも実際に運用されることもある＝これで正しい、とも言える
- 初学者にとっては文法は になるが、実際に使いこなせるようになるためには のが大事

形態素解析（おまけ）

- 子供は文法など考えずに喋ってる←言語を取り扱っている
- 非文（文法的に正しくない文）でも実際に運用されることもある＝これで正しい、とも言える
- 初学者にとっては文法は **理解の助け** になるが、実際に使いこなせるようになるためには のが大事

形態素解析（おまけ）

- 子供は文法など考えずに喋ってる←言語を取り扱っている
- 非文（文法的に正しくない文）でも実際に運用されることもある＝これで正しい、とも言える
- 初学者にとっては文法は **理解の助け** になるが、実際に使いこなせるようになるためには **経験値を高める** のが大事

機械学習について（「歴史」から再掲）

☆人が英語を勉強すること（語学）に当てはめてみると…

- **初学者**：文法を学び単語を一所懸命覚えて、日本語との対応を考えながら読み書きする

⇒ 文法＝構造的な知識（ルール）をつめこむ

⇒ 処理速度的に限界がある

（第2次AIブームでやったこと）

- **使える英語**：文法とか考えず、とにかく英語とその意味との対応をどんどん蓄積する

- ここでの「意味」は日本語（文字情報）に限らず、海外で（挨拶返し等）状況を感じとることも含む ⇒



がやっていること

機械学習について（「歴史」から再掲）

☆人が英語を勉強すること（語学）に当てはめてみると…

- **初学者**：文法を学び単語を一所懸命覚えて、日本語との対応を考えながら読み書きする
⇒ 文法＝構造的な知識（ルール）をつめこむ
⇒ 処理速度的に限界がある
(第2次AIブームでやったこと)
- **使える英語**：文法とか考えず、とにかく英語とその意味との対応をどんどん蓄積する
 - ここの「意味」は日本語（文字情報）に限らず、海外で（挨拶返し等）状況を感じとることも含む ⇒
機械学習がやっていること

機械学習について (cont.)

- “This is a pen.” “That is a pencil.” ... などの文例を、意味と対応付けて大量に覚えると…
- 直接覚えていない「これは鉛筆です」を “This is a pencil.” と訳すことが出来る
(文法とか品詞情報を学習することなく、組み合わせまで理解している)
- ちなみに、このようなことはディープラーニングの一種である  で実現できる
(詳細略)

機械学習について (cont.)

- “This is a pen.” “That is a pencil.” ... などの文例を、意味と対応付けて大量に覚えると…
- 直接覚えていない「これは鉛筆です」を “This is a pencil.” と訳すことが出来る
(文法とか品詞情報を学習することなく、組み合わせまで理解している)
- ちなみに、このようなことはディープラーニングの一種である **RNN,LSTM 等** で実現できる
(詳細略)

形態素解析

- “This is a pen.” という英文の場合は；

- “This”:

- “is”:

- “a”:

- “pen”:

- “.”:

- もともと分かち書きされてる（単語と単語の間にスペースがある）ので、比較的簡単

形態素解析

- “This is a pen.” という英文の場合は；

- “This”: (指示) 代名詞

- “is”:

- “a”:

- “pen”:

- “. ”:

- もともと分かち書きされてる（単語と単語の間にスペースがある）ので、比較的簡単

形態素解析

- “This is a pen.” という英文の場合は；

- “This”: (指示) 代名詞

- “is”: (be) 動詞

- “a”:

- “pen”:

- “.”:

- もともと分かち書きされてる（単語と単語の間にスペースがある）ので、比較的簡単

形態素解析

- “This is a pen.” という英文の場合は；

- “This”: (指示) 代名詞

- “is”: (be) 動詞

- “a”: 冠詞

- “pen”:

- “.”:

- もともと分かち書きされてる（単語と単語の間にスペースがある）ので、比較的簡単

形態素解析

- “This is a pen.” という英文の場合は；

- “This”: (指示) 代名詞

- “is”: (be) 動詞

- “a”: 冠詞

- “pen”: (普通) 名詞

- “.”:

- もともと分かち書きされてる（単語と単語の間にスペースがある）ので、比較的簡単

形態素解析

- “This is a pen.” という英文の場合は；

- “This”: (指示) 代名詞

- “is”: (be) 動詞

- “a”: 冠詞

- “pen”: (普通) 名詞

- “.”: ピリオド【句点】

- もともと分かち書きされてる（単語と単語の間にスペースがある）ので、比較的簡単

形態素解析（続き）

- ちなみに、これを構文解析すると；

- “This”:

- “is”:

- “a” “pen”:

- ⇒

形態素解析（続き）

- ちなみに、これを構文解析すると；

- “This”: 主語 (Subject)

- “is”:

- “a” “pen”:

- ⇒

形態素解析（続き）

- ちなみに、これを構文解析すると；

- “This”: 主語 (Subject)

- “is”: (be) 動詞 (Verb)

- “a” “pen”:

- ⇒

形態素解析（続き）

- ちなみに、これを構文解析すると；

- “This”: 主語 (Subject)

- “is”: (be) 動詞 (Verb)

- “a” “pen”: 補語 (Complement)

- ⇒

形態素解析（続き）

- ちなみに、これを構文解析すると；

- “This”: 主語 (Subject)

- “is”: (be) 動詞 (Verb)

- “a” “pen”: 補語 (Complement)

- ⇒ SVC 構文

形態素解析（続き）

参考：英語の5文型

- [He walks.]
- [This is a pen.]
- [He has a pen]
- (O : Object (目的語))
- [He gives her a present.]
- [He makes her happy.]

形態素解析（続き）

参考：英語の 5 文型

- **SV** [He walks.]
- [This is a pen.]
- [He has a pen]
(O : Object (目的語))
- [He gives her a present.]
- [He makes her happy.]

形態素解析（続き）

参考：英語の5文型

- **SV** [He walks.]
- **SVC** [This is a pen.]
- [He has a pen]
(O : Object (目的語))
- [He gives her a present.]
- [He makes her happy.]

形態素解析（続き）

参考：英語の5文型

- **SV** [He walks.]
- **SVC** [This is a pen.]
- **SVO** [He has a pen]
(O : Object (目的語))
- [He gives her a present.]
- [He makes her happy.]

形態素解析（続き）

参考：英語の5文型

- **SV** [He walks.]
- **SVC** [This is a pen.]
- **SVO** [He has a pen]
(O : Object (目的語))
- **SVOO** [He gives her a present.]
- [He makes her happy.]

形態素解析（続き）

参考：英語の5文型

- **SV** [He walks.]
- **SVC** [This is a pen.]
- **SVO** [He has a pen]
(O : Object (目的語))
- **SVOO** [He gives her a present.]
- **SVOC** [He makes her happy.]

形態素解析（続き）

- 「これはペンです。」 という日本語文の場合は；
 - “これ”:
 - “は”:
 - “ペン”:
 - “です”:
 - “。”:
- もともと分かち書きされていない（単語と単語の間にスペースがない）ので、英語に比べると難しい

形態素解析（続き）

- 「これはペンです。」 という日本語文の場合は；
 - “これ”： (指示) 代名詞
 - “は”：
 - “ペン”：
 - “です”：
 - “。”：
- もともと分かち書きされていない（単語と単語の間にスペースがない）ので、英語に比べると難しい

形態素解析（続き）

- 「これはペンです。」 という日本語文の場合は；
 - “これ”： (指示) 代名詞
 - “は”： (格) 助詞
 - “ペン”：
 - “です”：
 - “。”：
- もともと分かち書きされていない（単語と単語の間にスペースがない）ので、英語に比べると難しい

形態素解析（続き）

- 「これはペンです。」 という日本語文の場合は；
 - “これ”： (指示) 代名詞
 - “は”： (格) 助詞
 - “ペン”： (普通) 名詞
 - “です”：
 - “。”：
- もともと分かち書きされていない（単語と単語の間にスペースがない）ので、英語に比べると難しい

形態素解析（続き）

- 「これはペンです。」 という日本語文の場合は；
 - “これ”： (指示) 代名詞
 - “は”： (格) 助詞
 - “ペン”： (普通) 名詞
 - “です”： 助動詞
 - “。”：
- もともと分かち書きされていない（単語と単語の間にスペースがない）ので、英語に比べると難しい

形態素解析（続き）

- 「これはペンです。」 という日本語文の場合は；
 - “これ”：（指示）代名詞
 - “は”：（格）助詞
 - “ペン”：（普通）名詞
 - “です”：助動詞
 - “。”：記号【句点】
- もともと分かち書きされていない（単語と単語の間にスペースがない）ので、英語に比べると難しい

形態素解析（続き）

日本語文の形態素解析の困難性とは？

- まず ことから始める
- に決まるとは限らない
- 例：「経営情報学部」は
 - 「経営／情報／学部」
 - 「経営情報／学部」
 - 「経営情報学／部」
 - ...

形態素解析（続き）

日本語文の形態素解析の困難性とは？

- まず **区切りを見つける** ことから始める
- に決まるとは限らない
- 例：「経営情報学部」は
 - 「経営／情報／学部」
 - 「経営情報／学部」
 - 「経営情報学／部」
 - ...

形態素解析（続き）

日本語文の形態素解析の困難性とは？

- まず **区切りを見つける** ことから始める
- **一意（一通り）** に決まるとは限らない
- 例：「経営情報学部」は
 - 「経営／情報／学部」
 - 「経営情報／学部」
 - 「経営情報学／部」
 - ...

形態素解析（続き）

日本語文の形態素解析の困難性から…

- があるため、コンピュータに処理させても人手で行ってもいずれにせよ日本語形態素解析の結果は完全なものとみなすことはできない
- 現実的な対応策としては、コンピュータに処理を行わせ (=) その結果をできる限り人間が修正する
- その際ユーザーが辞書に項目を することができるので、辞書機能を充実させることで後の修正の手間が省けるようになる

形態素解析（続き）

日本語文の形態素解析の困難性から…

- **曖昧性**があるため、コンピュータに処理させても人手で行ってもいずれにせよ日本語形態素解析の結果は完全なものとみなすことはできない
- 現実的な対応策としては、コンピュータに処理を行わせ (=) その結果をできる限り人間が修正する
- その際ユーザーが辞書に項目を することができるので、辞書機能を充実させることで後の修正の手間が省けるようになる

形態素解析（続き）

日本語文の形態素解析の困難性から…

- **曖昧性**があるため、コンピュータに処理させても人手で行ってもいずれにせよ日本語形態素解析の結果は完全なものとはみなすことはできない
- 現実的な対応策としては、コンピュータに処理を行わせ (= **形態素解析器**) その結果をできる限り人間が修正する
- その際ユーザーが辞書に項目を することができるので、辞書機能を充実させることで後の修正の手間が省けるようになる

形態素解析（続き）

日本語文の形態素解析の困難性から…

- **曖昧性**があるため、コンピュータに処理させても人手で行ってもいずれにせよ日本語形態素解析の結果は完全なものとみなすことはできない
- 現実的な対応策としては、コンピュータに処理を行わせ (= **形態素解析器**) その結果をできる限り人間が修正する
- その際ユーザーが辞書に項目を **追加登録** することができるので、辞書機能を充実させることで後の修正の手間が省けるようになる

形態素解析器

- 形態素を解析するソフトウェア（プログラム & 辞書）
- 辞書の整備が重要（特に特殊な分野の文書解析の場合）
- , , ...
- R から利用できる を試してみる

形態素解析器

- 形態素を解析するソフトウェア（プログラム & 辞書）
- 辞書の整備が重要（特に特殊な分野の文書解析の場合）
- ChaSen(茶筌), , ...
- R から利用できる を試してみる

形態素解析器

- 形態素を解析するソフトウェア（プログラム & 辞書）
- 辞書の整備が重要（特に特殊な分野の文書解析の場合）
- ChaSen(茶筌), MeCab(和布蕪), ...
- R から利用できる を試してみる

形態素解析器

- 形態素を解析するソフトウェア（プログラム & 辞書）
- 辞書の整備が重要（特に特殊な分野の文書解析の場合）
- ChaSen(茶筌), MeCab(和布蕪), ...
- R から利用できる RMeCab を試してみる

形態素解析（練習）

☆ 【練習】 R を立ち上げ、以下を実行する

```
library(RMeCab) # ライブラリの導入
res <- RMeCabC('私は勉強している。', 1)
# 形態素解析を実行、結果はリスト
res2 <- unlist(res)
# リストをベクトルに
res2[names(res2) == '名詞']
# 名詞のみ抽出
```

形態素解析（練習）

☆ 【練習の結果】

```
res2[names(res2)=='名詞']  
名詞    名詞  
"私"    "勉強"
```

RMeCab の関数

- RMeCabC(“文字列”,1) : オプションの 1 は動詞を原型になおす (オプションなしだと変形されたまま)
- RMeCabText(“ファイル名”) : ファイルの中の文を解析
- RMeCabFreq(“ファイル名”) : ファイルの中の文を解析した上で頻度表を作成
- docMatrix(“フォルダ名”,pos=c(“品詞”)) : 指定したフォルダの中にあるすべてのファイルを読み込みそれぞれに対して形態素の頻度表を作成する。オプションで品詞を指定できる

テキストマイニング（実践）

- 1 <https://puffin.hannan-u.ac.jp/lect/dats/tm1d.zip>
を **I:¥rdata** に保存する
- 2 そこで展開すると **I:¥rdata¥tm1dat** というフォルダができ、その中にいくつかテキストファイル（青空文庫から拝借した短編小説）が出来ているはず
 - **青空文庫**:著作権フリーの作品を集めたサイト

テキストマイニング（実践）

- 3** R を立ち上げ、以下を実行する（名詞や動詞では特徴が出にくいので、ここでは連体詞と副詞だけを対象としている）

```
library(RMeCab) # ライブラリを導入
rslt <- docMatrix('I:/rdata/tm1dat',
  pos=c("連体詞", "副詞")) # 本当は1行、↓はログ出力
file = tm1dat/gongitsune.txt
(中略)
if you remove these rows, run
result[rownames(result)!="[[LESS-THAN-1]]",]
result[rownames(result)!="[[TOTAL-TOKENS]]",]
```

テキストマイニング（実践）

- 4 ログの最後は「結果の行列は2行の情報が含まれているので、不要なら以下のようにして取ってね」と言ってるので、取る
- 5 そして念のためにデータ表を”hindo.txt” という名前でファイルに出力（保存）しておく

```
rslt1 <-rslt[rownames(rslt)
      != "[[LESS-THAN-1]]",] # 本当は1行
rslt2 <-rslt1[rownames(rslt1)
      != "[[TOTAL-TOKENS]]",] # 本当は1行
write.table(rslt2,"hindo.txt")
```

テキストマイニング（実践）

6 念のため "hindo.txt" を TeraPad かメモ帳で開き、中を確認する

- 1列目の単語を除くと7列なので7次元と見ることも出来るし、285行あって1行目はファイル名なので、284次元のベクトルの集まりとも見なせる

"gongitsune.txt" "hana.txt"

"hitofusano_budo.txt" "lemon.txt"

"miminashi_hoichi.txt" "rashomon.txt"

"torokko.txt" (この4行は本当は1行)

"あんな" 2 0 2 3 2 0 0

"いったん" 1 0 0 0 0 0 0

(... 以下略)

テキストマイニング（実践）

7 得られた頻度データをベクトルとみなして、多次元尺度法で分析する

```
c1 <- t(rs1t2) # 転置（縦横入替）  
c2 <- dist(c1) # ベクトルの距離計算  
c3 <- cmdscale(c2,eig=T) # MDS  
c4 <- c3$points # ポイントデータ
```

テキストマイニング（実践）

6 c2 (距離データ) を確認

```
gongitsune.txt hana.txt hitofusano_budo.txt lemon.txt
hana.txt                28.28427
hitofusano_budo.txt     24.97999 31.55947
lemon.txt               25.39685 29.98333                22.15852
miminashi_hoichi.txt   41.67733 33.71943                37.64306 30.29851
rashomon.txt           44.31704 29.22328                42.89522 37.93415
torokko.txt            22.40536 27.67671                21.63331 19.62142
miminashi_hoichi.txt rashomon.txt
hana.txt
hitofusano_budo.txt
lemon.txt
miminashi_hoichi.txt
rashomon.txt           26.77686
torokko.txt           34.91418 38.72983
```

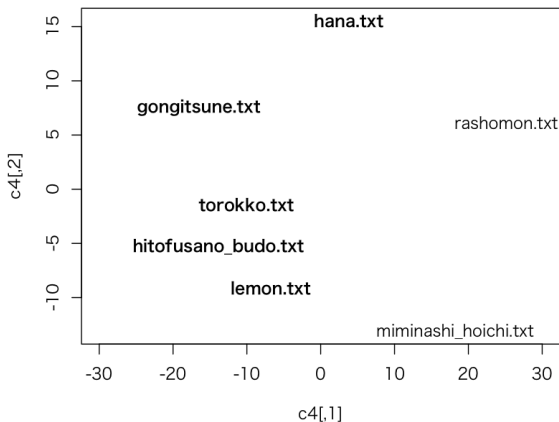
テキストマイニング（実践）

- 8 得られた頻度データを k-means で 2 分して MDS の結果とともにグラフ化する（グループでフォントをかえて表示）

```
c5 <- kmeans(c2, 2)
c6 <- c5$cluster
plot(c4, xlim=c(-30, 30), t="n")
text(c4, rownames(c1), font=c6)
```

テキストマイニング（実践）

☆ 出力結果



テキストマイニング（実践）

☆ 図の解釈（読み解き方は研究者次第）

- 「羅生門」・「耳無芳一の話」と、それ以外が別グループとなっている
- 「トロッコ」・「一房の葡萄」・「檸檬」は互いにかなり近い
- …（以下略）

おわりに

- テキストマイニング＝
形態素解析 → 量的データ作成 → データ分析
- 形態素解析の が大事な場合も多い
(WWW ページの場合、タグなど不要な情報を除いてプレーンテキストにする、など)
- 形態素の品詞によっては項目のほとんどが0になることが多い＝ が問題と
なることがある
- データがそろっていれば分析手法はなんらかの結果を出す ⇒ 結果の

おわりに

- テキストマイニング＝
形態素解析 → 量的データ作成 → データ分析
- 形態素解析の **前処理** が大事な場合も多い
(WWW ページの場合、タグなど不要な情報を除いてプレーンテキストにする、など)
- 形態素の品詞によっては項目のほとんどが0になることが多い＝ が問題となることがある
- データがそろっていれば分析手法はなんらかの結果を出す ⇒ 結果の

おわりに

- テキストマイニング＝
形態素解析 → 量的データ作成 → データ分析
- 形態素解析の **前処理** が大事な場合も多い
(WWW ページの場合、タグなど不要な情報を除いてプレーンテキストにする、など)
- 形態素の品詞によっては項目のほとんどが0になることが多い＝ **疎 (sparseness)** が問題となることがある
- データがそろっていれば分析手法はなんらかの結果を出す ⇒ 結果の

おわりに

- テキストマイニング＝
形態素解析 → 量的データ作成 → データ分析
- 形態素解析の **前処理** が大事な場合も多い
(WWW ページの場合、タグなど不要な情報を除いてプレーンテキストにする、など)
- 形態素の品詞によっては項目のほとんどが0になることが多い＝ **疎 (sparseness)** が問題となることがある
- データがそろっていれば分析手法はなんらかの結果を出す ⇒ 結果の **検証が重要**

おしまい

質問・コメント等あれば、何でもお気軽に
dats@e-chan.jp に
メールください！