データサイエンス『実践コース』

数理工学PBL

自然言語処理と深層学習

Day 1-1：自然言語処理の基礎

東北大学　小池 敦

このPBLの内容

- 自然言語処理
- 深層学習
- 深層学習を用いた自然言語処理

基本的な技術を理解し，実際にプログラミングできるようになることを目指す

自然言語処理とは？

人間が使用する言語に対する処理
（コンピュータで処理する）

- タスク例
  - 文書分類：カテゴリー（政治，スポーツ等）を推定
  - 情報抽出，文書要約
  - 機械翻訳
  - 質問応答，対話システム

日程（1日目）
2/18(土)10:00-12:00, 13:00-16:00

- 自然言語処理と深層学習の基礎
  - Python速習，基本的な文字列処理
  - Webからのテキスト取得と前処理
  - 日本語に対する自然言語処理
  - 深層学習の基礎（仕組みと進展，Kerasの基礎）

日程（2日目）
2/19(日) 10:00-12:00, 13:00-16:00

- 深層学習による自然言語処理
  - 分散表現
  - アテンション
  - **大規模言語モデル（応用と基礎技術）**
  - Hugging Faceライブラリによる高度自然言語処理

自然言語処理の基礎（1日目前半）

- 自然言語処理（講義）
  - 文字列の符号化
  - 自然言語の数学的表現
- 実習
  - Pythonの基礎
  - 文字列処理
  - 正規表現
  - Webからの文書取得と前処理

深層学習の基礎（1日目後半）

- 深層学習（講義）
  - 教師あり機械学習
  - 深層学習（基礎技術，実装の流れ）
- 実習
  - 深層学習フレームワークKerasの基本的な使い方

文字列の符号化

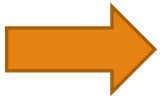
データの符号化

- 様々なデータをビット列に変換する

符号化

0101
1100
1110
1010

符号化



1101
1000
1010
1011

テキストの符号化

- 文字コード（変換表）を用いて
テキストから2進数に変換する
- 主な文字コード
  - ASCIIコード (US-ASCII)：英語用
    - 多くの文字コードがASCIIコードから拡張されている
    - 1文字につき1バイト（7bit）
（1bitはパリティビット（エラーチェック用））
  - UTF-8：ユニコード（世界の全言語をカバー ）
    - 1文字につき1バイトから4バイト
  - Shift_JIS：旧Windowsで使用されていた日本語向けコード
  - JISコード（ISO-2022-JP）：日本語用（メール等）

ASCIIコード

下位4ビット↓ \ 上位3ビット→	0	1	2	3	4	5	6	7
0	NUL	DLE	(SP)	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAC	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

1文字につき7ビット

Nの文字コード
16進数：0x4E
2進数：1001110

Lの文字コード
16進数：0x4C
2進数：1001100

Pの文字コード
16進数：0x50
2進数：1010000

武蔵システム 文字コード
http://opentype.jp/charactercode.htm

ASCIIコード

普通の文字以外に制御コードがある

代表的な制御コード

16進数	コード名	名称	説明
00	NUL	NULL	ヌル（無視する文字）
08	BS	Back Space	後退
09	HT	Horizontal Tabulation	水平タブ
0A	LF	Line Feed	改行
0C	FF	Form Feed	改ページ
0D	CR	Carriage Return	復帰
1B	ESC	Escape	エスケープ（拡張）

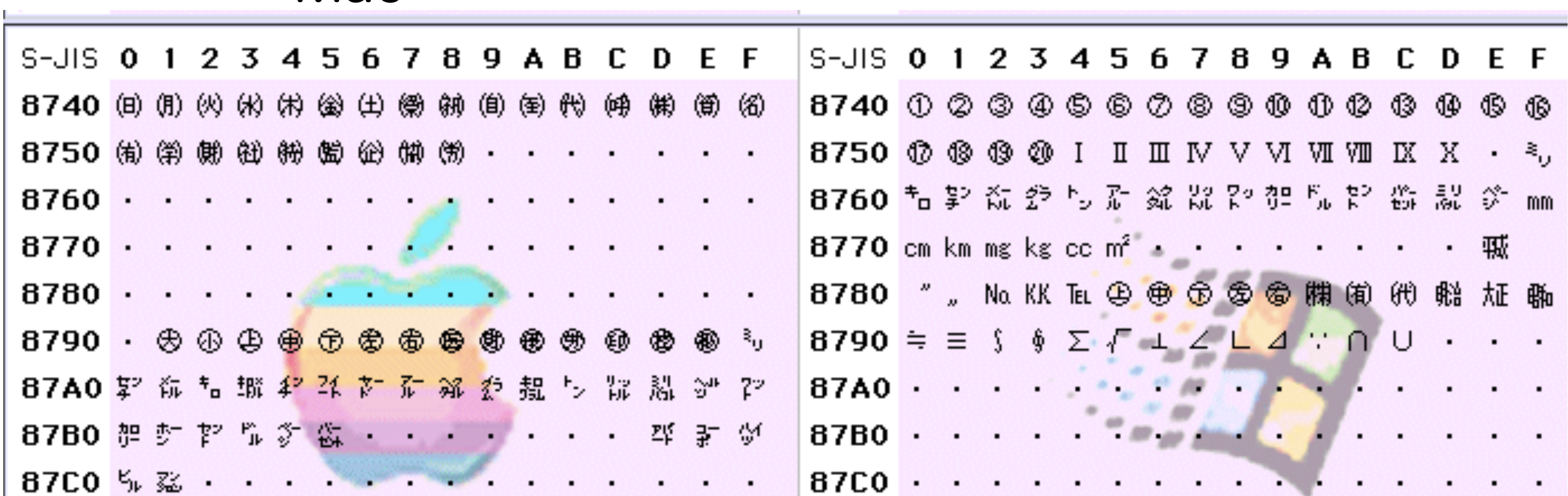
武蔵システム 文字コード
http://opentype.jp/charactercode.htm

日本語における機種依存文字

- Shift_JISやJISコードにおいて，コンピュータメーカが文字コードの独自拡張を行っている
  - 機器間の互換性がなく，文字化けが発生する（**機種依存文字**）
  - UTF-8を使えば問題なし
  - 現状はメールでの機種依存文字の使用を控えた方が良い
    - メールソフトは相手により文字コードを自動選択する
    - 代表的な機種依存文字：①②③ Ⅰ Ⅱ Ⅲ

日本語における機種依存文字

- 機種依存文字
  - WindowsとMac OS X 間での
    シフトJISコード非互換文字



Mac

S-JIS	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8740	㈰	㈪	㈫	㈬	㈭	㈮	㈯	㉀	㈷	㉂	㉃	㈹	㈺	㈱	㈾	㈴
8750	㈲	㈻	㈶	㈳	㈵	㈼	㈽	㈿	㈸	·	·	·	·	·	·	·
8760	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
8770	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	
8780	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
8790	·	(大)	(小)	㊤	㊥	㊦	㊧	㊨	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	㍉
87A0	㌢	㍍	㌔	[illegible]	[illegible]	[illegible]	[illegible]	㌃	[illegible]	[illegible]	[illegible]	㌧	[illegible]	[illegible]	[illegible]	[illegible]
87B0	㌍	[illegible]	㌣	㌦	㌻	㌫	·	·	·	·	·	·	·	[illegible]	[illegible]	[illegible]
87C0	[illegible]	[illegible]	·	·	·	·	·	·	·	·	·	·	·	·	·	·

Windows

S-JIS	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8740	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯
8750	⑰	⑱	⑲	⑳	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ	Ⅸ	Ⅹ	·	㍉
8760	㌔	㌢	㍍	㌘	㌧	㌃	㌶	㍑	㍗	㌍	㌦	㌣	㌫	㍊	㌻	㎜
8770	㎝	㎞	㎎	㎏	㏄	㎡	·	·	·	·	·	·	·	·	㍻	
8780	〝	〟	№	㏍	℡	㊤	㊥	㊦	㊧	㊨	㈱	㈲	㈹	㍾	㍽	㍼
8790	≒	≡	∫	∮	∑	√	⊥	∠	∟	⊿	∵	∩	∪	·	·	·
87A0	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
87B0	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
87C0	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·

文字化け　機種依存文字
http://www.geocities.co.jp/SiliconValley-SanJose/5148/moji/

自然言語の数学的表現

コーパス

自然言語に関するデータ（文書，音声等）を集めたもの

- 文書コーパスの例
  - Project Gutenberg（主に英語）
  - 青空文庫（日本語）
- 用途に応じて色々ある
  - 質疑応答：SquAD等
  - 対訳コーパス：日本語対訳データ 参照
  - その他，この辺も参照

トークン化

文を単語等の単位（トークン）に分割すること.

⇒ コンピュータでの扱いが容易になる

- トークン化の例
  - After night comes the day.
    ⇒ After / night / comes / the / day.
  - 明けない夜はない.
    ⇒ 明け / ない / 夜 / は / ない.

日本語ではスペースによる機械的な分割ができない

形態素解析

文を解析して形態素（最小単位）に分割すること.

⇒ トークン化にも使用される

文　　学校名に関するお話を聞く

文節　学校名に/関する/お話を/聞く

単語　学校名/に/関する/お話/を/聞く

形態素　学校/名/に/関する/お/話/を/聞く

日本語ではMeCab, Juman++，Janomeなどのライブラリがある

辞書

各単語（形態素）についての詳細情報.

ボキャブラリ（語彙）：管理する全単語の集合

例：MeCab IPA辞書

```
引き込む,762,762,7122,動詞,自立,*,*,五段・マ行,基本形,引き込む,ヒキコム,ヒキコム
引き込ま,764,764,7118,動詞,自立,*,*,五段・マ行,未然形,引き込む,ヒキコマ,ヒキコマ
引き込も,763,763,7122,動詞,自立,*,*,五段・マ行,未然ウ接続,引き込む,ヒキコモ,ヒキコモ
引き込み,767,767,7088,動詞,自立,*,*,五段・マ行,連用形,引き込む,ヒキコミ,ヒキコミ
引き込ん,766,766,7122,動詞,自立,*,*,五段・マ行,連用タ接続,引き込む,ヒキコン,ヒキコン
引き込め,760,760,7122,動詞,自立,*,*,五段・マ行,仮定形,引き込む,ヒキコメ,ヒキコメ
引き込め,765,765,7122,動詞,自立,*,*,五段・マ行,命令ｅ,引き込む,ヒキコメ,ヒキコメ
引き込みゃ,761,761,7122,動詞,自立,*,*,五段・マ行,仮定縮約１,引き込む,ヒキコミャ,ヒキコミャ
すきかえす,731,731,9279,動詞,自立,*,*,五段・サ行,基本形,すきかえす,スキカエス,スキカエス
すきかえさ,733,733,9279,動詞,自立,*,*,五段・サ行,未然形,すきかえす,スキカエサ,スキカエサ
すきかえそ,732,732,9279,動詞,自立,*,*,五段・サ行,未然ウ接続,すきかえす,スキカエソ,スキカエソ
すきかえし,735,735,9279,動詞,自立,*,*,五段・サ行,連用形,すきかえす,スキカエシ,スキカエシ
すきかえせ,729,729,9279,動詞,自立,*,*,五段・サ行,仮定形,すきかえす,スキカエセ,スキカエセ
すきかえせ,734,734,9279,動詞,自立,*,*,五段・サ行,命令ｅ,すきかえす,スキカエセ,スキカエセ
すきかえしゃ,730,730,9279,動詞,自立,*,*,五段・サ行,仮定縮約１,すきかえす,スキカエシャ,スキカエシャ
看取る,772,772,7150,動詞,自立,*,*,五段・ラ行,基本形,看取る,ミトル,ミトル
看取ら,780,780,7150,動詞,自立,*,*,五段・ラ行,未然形,看取る,ミトラ,ミトラ
```

MeCab: Yet Another Part-of-Speech and Morphological Analyzer
https://taku910.github.io/mecab/

単語（トークン）の表現

- 単語ID
  - すべて単語（ボキャブラリ）の
    それぞれにIDをつける

  - 例

    私　　⇒ 1

    は　　⇒ 2

    眠い　⇒ 3

    ・・・

単語（トークン）の表現

- One-hot ベクトル
  - 単語のベクトル表現
  - 各単語にひとつの次元を割り当てる

  - 例

		私	は	眠い		
私	⇒	1	0	0	0	・・・	0
は	⇒	0	1	0	0	・・・	0
眠い	⇒	0	0	1	0	・・・	0

・・・

未知語

ボキャブラリの数を一定数に制限し，
そこに含まれない単語は未知語 <UNK> にする．
⇒ One-hotベクトルの次元が小さくなる

パディング

文の単語数を増やしたいとき
（文の単語数を一定数にする場合など），
末尾にパディング文字 <PAD> をつける

- 6単語からなる文にする場合：

  私 / は / 眠い / <PAD> / <PAD> / <PAD>

サブワード

低頻度の単語は，文字（もしくは部分文字列）に分割する．

⇒ 未知語 <UNK> を使う必要がなくなる．

- 例： 京都 / の / 相 / 国 / 寺

  asinine situation ⇒ as/in/in/e situation

Sentencepieceライブラリ
⇒ 文から直接サブサードに分割

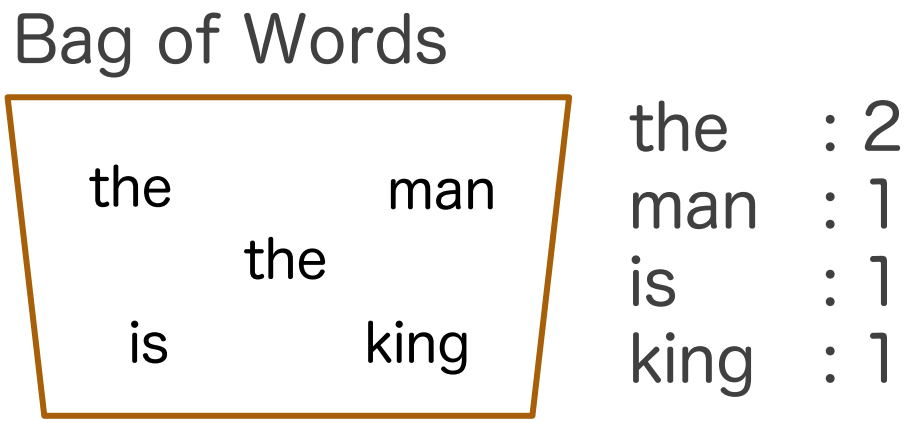
Sentencepiece：ニューラル言語処理向けトークナイザ
https://qiita.com/taku910/items/7e52f1e58d0ea6e7859c

文書の表現

- Bag of Words (BoW)
  - 文書を単語に分解し、各単語の出現回数のみを保持
  - 単語の出現順序の情報は捨てられることになる
  - トピックモデルなどの文書分類で使われる

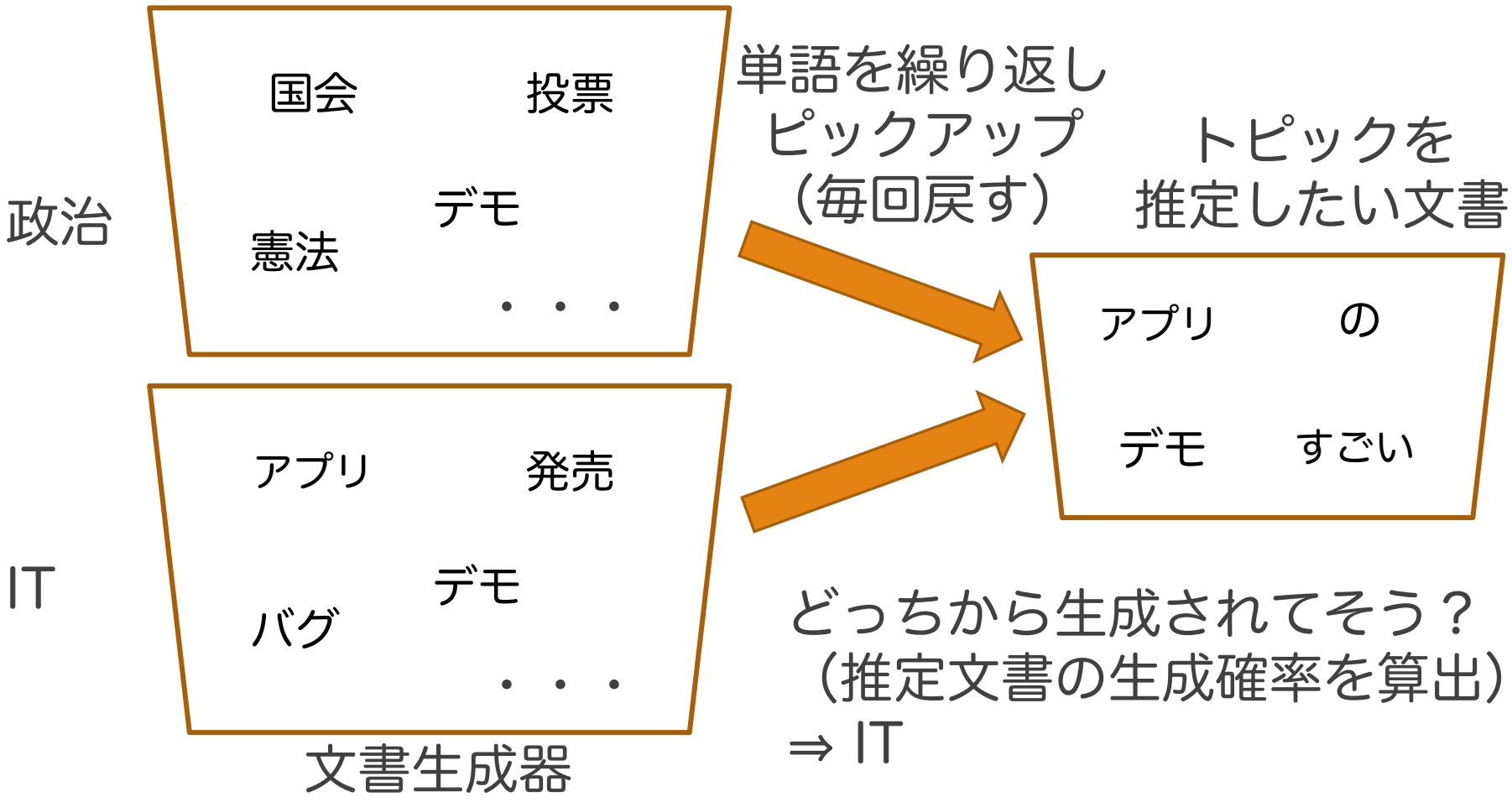
元の文書

The man is the king. ⇒



Bag of Wordsでは通常、大文字と小文字の違いは無視し、ピリオドなどの記号は捨てられる

BoWによるトピック分類（生成モデル）



文書の表現

出現順序を考慮した表現を使いたい！

- n-gram
  - 単語 n-gram：連続するn単語を組にして扱う
    - n = 1 の時ユニグラム，n = 2 の時バイグラムと呼ぶ
    - 例：「お好み焼き / を / 無性に / 食べたい」
      n = 2 ： (お好み焼き, を), (を, 無性に), (無性に, 食べたい)
      n = 3 ： (お好み焼き, を, 無性に), (を, 無性に, 食べたい)
  - 文字 n-gram：連続するn文字を組にして扱う
    - 例：「お好み焼き」
      n = 2 ： (お, 好), (好, み), (み, 焼), (焼, き)
      n = 3 ： (お, 好, み), (好, み, 焼), (み, 焼, き)

その他の前処理

- 文字種の統一：全角，半角，大文字，小文字
  - 1月３０日 ⇒ 1月30日（数字を全部半角に）
- 表記揺れの修正
  - 「午前も打合せ，午後も打ち合わせだ」
    ⇒ 「午前も打合せ，午後も打合せだ」
- 単語の原形化・ステミング：
  - 「食べろ」⇒「食べる」，「食べ（ステミング）」
- ストップワードの除去
  - 日本語の助詞や英語の冠詞等は「ストップワード」とし，BoW等に入れない