

講義コード	1200018402
講義名	知識情報学
開講期	2024年度後期
⑤単位数	2単位
②授業形態	講義
⑥担当教員	田近 一郎
ナンバリングコード	22M2A6202、19M4A6203

③科目概要	情報インフラの普及・拡大にともなうデータの爆発的な増加・多様化に対し、大規模データに潜むパターンを発見し、知識として活用する知識発見・機械学習・データマイニング・情報検索、自然言語処理等のデータサイエンス技術が重要度を高めている。本講義では、データベースや統計解析・機械学習・人工知能の分野で発展したこれらの技術の理論と応用について解説するとともにPythonプログラミングのためのクラウド開発環境Google Colab上でのPC演習を取り入れ理解を深める。
①達成目標	データの性質と目的に応じて、どのデータマイニング／機械学習手法を適用すればよいかを把握できる。 「基本情報技術者試験」で出題されるデータマイニング／機械学習関連分野の問題を理解できるようになる。
履修に必要な予備知識や技能（関連科目等）	より深く理解するには「アルゴリズムとデータ構造」の履修が望ましい。関連科目として「統計学」「データサイエンス」「データベース」がある。
学位授与方針との関連	1. 情報メディアに関して、基礎的な情報処理技術、ソフトウェアについての知識、また、資格試験の学修内容等、基礎的な知識・技能を身につけている 2. 専修コースにおいて、情報理論、映像史、音響、色彩など、必要な知識・教養を身につけている。 3. 専修コースにおいて、プログラミング、映像制作、PA、グラフィックデザインなど、必要な技術・技能を身につけている。
④授業計画、授業外学習の内容及び必要な時間	
第1回	ビッグデータを扱うデータ駆動型社会（Society5.0）におけるデータ分析や問題解決のための仮説検証・知識発見の考え方をビッグデータ活用事例を紹介しながら概説する。またそれらの分析に不可欠なデータマイニング、機械学習、情報検索および自然言語処理などの各手法についても概説し、本講義の範囲・目的、「基本情報技術者試験」関連科目とのかかわりを説明する。講義の進め方や成績評価方法等も説明する。講義前、ビジネスをはじめとするビッグデータ活用事例におけるデータマイニング、機械学習の位置づけを確認しておく（120分）。講義後、「基本情報技術者試験」の午前の過去問題を調査しデータマイニング分野が出題されることを確認する。また機械学習の学習過程と予測過程に関するWebClass課題を解く（120分）。
第2回	データの爆発的な増加・多様化の問題への新しい取り組みとして、大規模データから従来の統計解析では発見できないパターン・法則性を取り出せるデータマイニングの概念をその時代背景、発展の歴史、適用例とともに概説する。またデータマイニング用アプリケーションを紹介し、それらのバージョンを最新に保つ必要性などの情報セキュリティについても触れる。講義前、キーワード「データマイニング」で検索し、データマイニングの具体例を複数下調べしておく（120分）。講義後、調べておいた具体例の目的、出力されるパターンや法則性をまとめておく。またデータマイニング／機械学習の分析目的に関するWebClass課題を解く（120分）。
第3回	各種のデータマイニング手法について紹介する。具体的には、相関分析、クラスタリング、リンク分析、決定木分析、ニューラルネットワーク、Webマイニング等について、対象とするデータの性質、データ分析の目的、データ分析の進め方を概説する。またデータクレンジング処理などの前処理・マイニングアルゴリズム処理・出力の人による解釈検証などの分析手順を概説する。講義前、データマイニング／機械学習の各種手法を下調べしておく（120分）。講義後、手法ごとに処理対象とするデータの性質、分析目的、分析手順、出力されるパターンや法則性の表現形式をまとめておく（120分）。
第4回	小売業などビジネスで使われるデータマイニング手法の1つで、大規模販売データを集計してRDBを構成し、そこから購買パターンを表す相関ルール（if then ルール）を発見するマーケットバスケット解析（アソシエーション分析）について概説する。講義前、キーワード「マーケットバスケット解析」「相関ルール」等で検索し、アケットバスケット解析の具体的な応用例を下調べしておく（120分）。講義後は、マーケットバスケット解析の概念を振り返り、分析結果を相関ルールとして出力することの意味を理解しておく。また具体的な購買データから得られる相関ルールの当てはまりの良さに関するWebClass課題を解く（120分）。
第5回	マーケットバスケット解析のための各種アルゴリズムを紹介する。アルゴリズムの種類によりデータ処理の効率が著しく異なることにもふれる。また小規模なデータを用いて集合のベン図を用いた演算によりマーケットバスケット解析が可能であることを確かめる。講義前、マーケットバスケット解析の概念について復習しておく（120分）。講義後、マーケットバスケット解析の簡単な具体例を確認し小規模なデータを用いて集合の演算によりマーケットバスケット解析が可能であることを確かめる。またキーワード「マーケットバスケット解析」「アプリアリアルゴリズム」で検索し、アルゴリズムの発明が社会に与える影響力を確認しておく。またアプリアリアルゴリズムの動作をシミュレートして得られる相関ルールの集合を求めるWebClass課題を解く（120分）。
第6回	クラウド開発環境Google Colab上でPythonによるコンピュータ演習をおこなう。データから相関ルールを抽出するマーケットバスケット解析の過程を体験してもらい、データマイニングのプロセスの理解を促す。講義前にマーケットバスケット解析の簡単な具体例について復習しておく（120分）。講義後、PC演習で実施した、データから相関ルールを抽出する過程をトレースし習熟しておく。また具体的に商品を指定し、その商品を含む有益な相関ルールを探索的に調べ、ミニ演習課題の課題として提出する（120分）。
第7回	データマイニング手法の1つであり、データをそのカテゴリに応じて分類するために使われる決定木分析について説明する。一般的な機械学習の枠組みとして、分類器の生成に使われる学習データ、分類器の性能評価に用いるテストデータなどの概念についてまとめる。またデータから決定木を生成する学習アルゴリズムを解説する。講義前、キーワード「決定木」「学習」等で検索し、決定木分析の具体的な応用例を下調べしておく（120分）。講義後、決定木分析の概念を振り返り、分析結果を決定木として出力することの意味を理解しておく。また具体的な決定木を利用してデータの分類をおこなうWebClass課題を解く（120分）。
第8回	データから決定木を生成する学習アルゴリズムを解説する。学習アルゴリズムが生成した決定木のサイズとその分類性能との関係にふれ、場合により決定木の枝刈りが必要になることにも言及する。講義前、決定木分析の概念について復習しておく（120分）。講義後、決定木学習アルゴリズムの学習過程を振り返り、自分なりにまとめておく。また決定木学習アルゴリズムの学習過程に欠かせない木の成長に必要なパラメータ調整に関するWebClass課題を解く（120分）。
第9回	クラウド開発環境Google Colab上でPythonによるコンピュータ演習をおこなう。実際のデータを用いて学習プログラムを実行し、生成された決定木を観察する。また決定木解析の結果得られた3次元散布図を観察する。データの品質に応じて生成される決定木の性能に差が生じることにも注目する。また決定木解析で得られた変数を用いて3つのカテゴリからなるデータの3D散布図を作成するとデータが3つのクラスに適切に分離できていることを確かめる。講義前、決定木学習アルゴリズムについて復習しておく（120分）。講義後、PC演習で実施した決定木分析に関して、分析全体における決定木学習の位置づけと、それ以外の前処理等の作業手順について確認しておく。また得られた決定木をミニ演習課題として提出する（120分）。
第10回	大規模データの検索で用いられる全文検索のしくみをWEB検索エンジンを例に概説する。ワープロ文書のような小規模データで用いられる逐次検索についてもふれる。全文検索の一種である内容型検索では、文書に出現する単語集合をベクトルで表現し、ベクトルのコサイン類似度（内積）で文書間の類似度が計算できることを解説する。講義前、キーワード「情報検索」「全文検索」で検索し、情報検索の具体的な応用例を複数下調べしておく（120分）。講義後、全文検索と逐次検索の違いについて確認しておく。文書の類似度計算に関する例題のWebClass課題を解く（120分）。

第11回	WEB検索エンジンのコア技術である、超大規模データに対して用いられるインデックスサーチについて、そのしくみを詳しく解説する。インデックスデータベース作成のためのWebクローラによるWebデータ収集、Webデータのインデックスデータベースへの変換、インデックスのソート（並び替え）、さらに検索時のインデックスデータベースに対する二分探索法を含むインデックスサーチ（探索）の手順に具体例を通して習熟する。講義前、キーワード「インデックスサーチ」で検索し、どのように情報検索エンジンが構成されているかを下調べしておく（120分）。講義後は、インデックスデータベース生成とインデックスサーチの手順を自分なりにまとめておく（120分）また簡単なインデックスサーチの仕組みに関するWebClass課題を解く。			
第12回	大量の検索結果をランキングすることの重要性を説明し、その具体例としてページランクアルゴリズムについて解説する。クラウド開発環境Google Colab上でPythonによるコンピュータ演習を通して、ページランクアルゴリズムの特徴を把握する。講義前、キーワード「Google」「情報検索」等で検索し、検索エンジンGoogleの概略を下調べしておく（120分）。講義後、ページランクのアイデアを振り返り自分なりにまとめておく。また各種のインターネットのネットワーク構造に対してページランクの数値を求めるWebClass課題を解く（120分）。			
第13回	テキストデータ（文章、SNSデータ）からテキスト分類、文書要約をはじめとする自然言語処理をおこなうテキストマイニングアルゴリズムについて概説し自然言語処理の活用事例を紹介する。次に与えられたテキストを形態素解析アルゴリズムで単語分割し、日本語Word2Vecモデルと呼ばれる日本語学習済みニューラルネットを利用してテキストに出現する各単語を意味に基づいて配置した散布図を作り、テキストの特徴を把握する自然言語処理PC演習をクラウド開発環境Google Colab上でおこなう。講義前、「自然言語処理」「形態素解析」等で検索し、これらの用語の背後にある自然言語処理分野の概略を下調べする（120分）。講義後、PC演習で実施した形態素解析などを含む前処理の作業手順を復習する。またインターネット上から選んだ複数のテキストをWord2Vecモデルを用いて可視化しその結果をミニ演習課題として提出する（120分）。			
第14回	テキストデータからの自然言語処理について、OpenAI社ChatGPTやGoogle社Geminiなどの大規模言語モデル（基盤モデル）の基礎になるTransformerの仕組みについて解説する。また、大規模言語モデルを前処理用プログラムとして組み込み、特定タスクのための小規模なニューラルネットを教師ラベル付きデータで学習させるファインチューニングの仕組みについて解説し、その適用事例として日本語BERTモデルと呼ばれる日本語学習済み深層ニューラルネットを前処理用に組み込んだ文書の感情分析手法を紹介する。次にクラウド開発環境Google Colab上でPython実装により模擬的なAIの開発環境を構築し、AIの学習の事例としてファインチューニングプログラムを利用するSNS文書データの感情分析のPC演習をおこなう。講義前、「生成AI」「自然言語処理」「テキストマイニング」「GPT」「BERT」で検索し、これらの用語の背後にあるビジネス分野での自然言語処理活用事例を下調べする（120分）。講義後、PC演習で実施した作業手順のうち、各文書をBERTに入力できるように「トークンベクトル」に加工するまでの処理の過程についてまとめて、トークン化の重要性を確認する（120分）。			
第15回	クラウド開発環境Google Colab上で引き続き、Python実装により模擬的なAIの開発環境を構築し、AIの学習の事例としてファインチューニングプログラムを利用するSNS文書データの感情分析のPC演習をおこなう。その後、講義期間中に出したミニ演習、ミニレポートの解説をおこなう。これまでに配布した授業プリントや補助資料を持参すること。欠席回の授業プリント等を把握した上ですべての授業プリントを整えること。講義前、これまでに出示されたWebClass課題レポートを見直し、理解不足の部分について把握しておく（120分）。講義後、講義内容を復習し理解不足の内容に関するWebClass課題レポートを解く。また感情分析の結果をミニ演習課題として提出する（120分）。			
講義進行方法、課題へのフィードバック方法		毎回配布する講義プリントへの記入とpdf補助資料、参考webサイトの紹介などで講義を行う。また、適宜講義内で紹介したアルゴリズムの処理手順に関する演習を実施する。講義内である程度まとまった概念群を解説した後、あるいはアルゴリズムの処理手順を解説した後、適宜理解度の確認と習熟のためのミニ演習やミニレポートを実施する。また、基本情報処理技術者試験などで実際に出題された問題もミニ演習の一部に組み込み、履修者に解いてもらう。提出されたサンプルコード等は動作に関して一定の品質を満たすまで修正作業を促す。		
アクティブラーニング		実習、フィールドワーク		
講義前・講義後の自主的学修活動への助言		講義前には学修のポイントやキーワードに従って、事前に下調べをしておき講義対象の把握に務めてもらいたい。また、講義後は、指示されたまとめや紹介した手法の処理手順のトレースなどをおこない処理手順等の習熟に努めてもらいたい。講義内容についての疑問点もそのままにせず教員に積極的に質問してもらいたい。		
⑦成績（達成度）評価方法・評価基準、割合		成績評価は、１．定期試験（５０％）、２．ミニ演習の達成状況等（２５％）、３．ミニレポートの提出・達成状況等（１５％）４．その他（１０％）に基づきおこなう。また、講義およびミニ演習やミニレポートの作業への取り組みの状況、講義前・講義後の自主的な学修活動に基づく積極的な質問等も評価する。		
教員の実務経験と授業科目との関連				
テキスト				
書名	著者	出版社	ISBN	備考
『なし』				
参考図書など				
書名	著者	出版社	ISBN	備考
『Googleを支える技術』	西田圭介	技術評論社	978-4774134321	2008
『つくりながら学ぶ! PyTorchによる発展ディープラーニング』	小川雄太郎	マイナビ出版	978-4839970253	2019
『ビジネス活用事例で学ぶデータサイエンス入門』	酒巻隆治他	SBCreative	978-4797376333	2014
『ゼロから作るDeep Learning』	斎藤康毅	オライリー・ジャパン	978-4873117584	2016
『大規模言語モデル入門』	山田 育矢， 鈴木 正敏， 山田 康輔， 李 凌寒	技術評論社	978-4297136338	2023
『LangChain完全入門 生成AIアプリケーション開発がはかどる大規模言語モデルの操り方』	田村悠	インプレス	978-4295017967	2023
参考URL				
NO	表示名	URL	説明	
1.	Colaboratoryへようこそ	https://colab.research.google.com/?hl=ja	プログラミング言語Pythonで基礎的プログラムからディープラーニング・人工知能プログラムまで作成できるクラウドプログラミングサイト	
2.	Qiita	https://qiita.com/	コンピュータ・ソフトウェア関連の技術に関する知識を記録・共有するための国内のコミュニティサービス	