第 6 章

非構造化データ処理の基本

データを整理する際は表などを用いることが多いですが、最近では、データの取得技術の発展も手伝い、そのような形式をとらないデータであっても解析を行うことができるようになりました。例えば、非構造化データと言われる、文書や音声、画像などのようなデータです。

この章では、非構造化データの処理方法について理解していきましょう。

この章のゴール

非構造化データを用いた分析の基本手法とプロセスを理解する

非構造化データ処理の種類／R・Python を用いた非構造化データの実行

非構造化データと呼ばれるデータの主な種類には、テキストデータ*、音声データや画像データなどがあります。

*ここで述べるテキストデータとは、電子化された文書のことを指します。

最近では、SNS での書き込みやさまざまなメディアの情報の電子化に伴い、手軽にデータが得られるようになりました。これらのデータは、適切な処理を行うことでキーワードの抽出や文書の分類を行うことができます。

音声データの解析分野においては、音声入力や工場での異音検知など、社会における実用を目的に幅広い研究が行われています。

画像データの分野も盛り上がりを見せており、MRI や CT スキャンといった医療画像データについての診断支援や、天体観測で様々な特徴の測定支援に始まり、自動車の自動運転などの身近な技術にも応用されています。

非構造化データと特徴量

非構造化データは、もともとは表形式などのフォーマットを取らないデータでした。解析を行う際は、「最終的にデータからどのようなことを知りたいのか」といった目的を定めることが重要です。

解析の工程では、このような目的に応じた特徴を言い表す「数値」を作成します。非構造化データである音声や画像から算出された数値は、**特徴量**といいます。

音声データの場合は、周波数や振幅などが多く用いられています。解析者は、それらの特徴量をどのように導き出すかについて考える必要があります。

言語処理の基本とその実行方法

私たちが使用しているコンピュータは、色々な動作を論理的な信号処理によって行います。そして、その論理式がコンピュータたちにとっての言語といえます。この論理式は人為的に表現規則を決めて、解釈に齟齬が生まれないように調整されたものです。

一方で、私たちにとっての言語とは日本語や英語など、慣習や時代によって移ろうような、論理式に比べると表現の自由度が高いものです。そこで、コンピュータが用いる言語と区別をつけるため、私たちが普段用いる言語のことを「自然言語」と呼んでいます。

自然言語処理とは、私たちが普段用いる言語をコンピュータで処理する技術のことを言います。自然言語処理におけるポイントは、「いかに自然言語をコンピュータに理解させるか」です。

これから紹介する様々な手法は、頭の固いコンピュータに自然言語を理解させる努力の結果だと思うとわかり易いです。大まかには、文章を単語ごとに分け、構文を解析し、文章の意味を推測、という流れで進めます。

単語への分割

私たちが話す言語は、単語を文法規則にのっとって組み合わせることで出来ています。つまり、単語の意味と文法がわかれば理解できるのです。したがって自然言語処理の第一段階では、与えられた文章を単語に分割していく作業が必要です。特に、日本語のような単語間の区切れがない言語だとその作業は難しくなります。

このように自然言語の文章を言語上で意味を持つ最小単位に分けていく作業を「分かち書き」といいます。また、分割した最小単位のことを「形態素」といいます。文章を形態素に分割していくとき、「辞書」を用います。この辞書は私たちが普段用いるものとは少々異なり、各形態素の品詞や活用法などが載っているものです。文章を形態素に分解し、品詞を振り分ける作業のことをまとめて「形態素解析」と呼ぶことがあります。

構文解析

文章を形態素に分割できた後は、形態素間の関係を調べます。日本語だと、単語間の修飾・被修飾関係に注目して構文を読み解く「係り受け解析」が主流です。この解析では「どの単語がどの単語に係るか」という観点で文章を見ることにより、文章がただ単語を並べたものではなく、その単語の並びに意味があることをコンピュータに理解させます。

意味解析

次はその文章がどのような意味を持つのか、どのような場面を表しているのかについてパソコンに理解させます。ここで用いるのがコーパスです。

コーパスとは、自然言語のさまざまな用例の集まりです。コーパスを用いると、慣用的な表現を学ぶことができます。例えば、「おいしいご飯、自転車で食べに行ってくる！」という文章があるとします。「おいしい」という形容詞は「ご飯」、「自転車」という名詞のどちらと結びつくのか、はたまた両方と結びつくのか、コンピュータにはわかりません。そこでコーパスを用いると、「おいしい」は「ご飯」と結びついて出てくることが多く、「自転車」とはあまり結びつかない、とわかります。その結果「おいしいご飯」が適切な解釈だ、と落ち着きます。つまり、コンピュータでは単語間の結びつきの強さによって意味を定義するのです。このような一連の作業によって、コンピュータは与えられた文章の意味を理解します。

文脈解析

これまでの作業で、一文に対して処理を行うことはできました。ただ、一文だけでなく複数の文章について解析をして、それらの文章間の関係を見たい場合も多いです。そのような時でも上記の処理を大幅に変えることはありません。しかし、複数の文章を扱う場合、一つ大きな問題があります。それは代名詞や指示詞の存在です。それらが一体何を指しているのか推定を行う必要があり、処理の難易度がぐっと上がります。それらを推定するための方法は現在も研究が進んでいるところです。推定を考えない手法だと、ある文書における単語の出現回数を用いて行う重要語の解析などがあります。

画像処理の基本とその実行方法

画像処理は、非構造化データの中でも特に幅広く様々な手法が用いられる分野です。
本解説では、画像をどのように扱い、どのような技術が話題となっているかについて説明します。

私たち人間は視覚を持っており、さまざまな画像から、そこに写っている動物や車、人物を見つけることは比較的容易に行うことができます。一方で、人間が直感的に行っているこの作業をコンピュータへ持ち込むことは高い技術を要します。 私たちは視覚情報を処理する際、長年の経験や焦点の絞りをはじめとする眼の自動制御など、多くの働きを無意識にこなしています。コンピュータでこれを再現するには、画像を格子状に切って、色や明るさを数値に変換したものを情報として取り込む必要があります。

ここで、「画像に鉛筆が含まれているか」という問題を解くことを考えます。問題を解くには、鉛筆とはどのような形状をしたものか、という定義を行わなければいけません。これはすごく難しい問題です。画像に写っているのは長い鉛筆か、短い鉛筆か、えんじ色の鉛筆か、緑色の鉛筆か、6 角か、丸型か、削られた方が上を向いているのか、削られていないのか、どんな鉛筆が来るのかわかりません。この世にある一つ一つの鉛筆について定義していては大変だということは想像が付きます。そこで、最近ではたくさんの鉛筆の画像を用意して、「これは鉛筆です」というようにコンピュータへ取り込みます。そうすることでコンピュータが勝手に「鉛筆とはどういう数値が並んだ物体なのか」というパターンを学べるようにします。ただ、画像をたくさん用意するのも大変です。そこで、手持ちの画像を白黒写真にしたり、少しぼかしたり、回転させたりしたものも取り込みます。なぜなら、鉛筆はコーティングの色が変わっても、少しぼやけていても、向きが変わっても鉛筆であることに変わりはないからです。そのように様々な鉛筆の画像を見せることで、鉛筆が持つより抽象的な特徴をコンピュータが捉えることができるのです。その特徴が定義できたら、あとはその特徴に当てはまるかどうかで問題が解けます。

鉛筆の例でみたように、画像処理においては一つの課題をクリアするのにも多くの労力が必要です。したがって、画像処理の分野では、かなり具体的な目的や動機を持った研究が多いです。ただ、その過程で入力する画像に白黒加工やぼかし、回転を施す操作は共通して必要だと言われています。それらの操作を簡単に行うアプリは多く作られています。その後は、各目的に対して個別に処理をしていきます。そのような特性をポジティブに捉えると、画像さえあれば様々な分野に応用が利くと言えます。実際に、応用範囲は地図アプリのストリートビューからスマホの写真加工アプリ、自動車の自動運転や生物医学分析、防犯システムの構築など多岐に渡ります。