2020 年度　卒業論文

大学ホームページから大学らしさの抽出と言語化

経営学部 経営学科

学籍番号 17161278　佐藤功基

目次

1 はじめに

少子化の影響により入学希望者総数が入学定員総数を下回る大学全入時代において、人気のある大学には受験生が集中し、人気のない大学では定員割れが起こり、存続の危機に立たされる。一方で、受験生は数多の大学の中から自身が受験する大学を選択する。このとき、参考にされる要素は設置している学部、専門領域、偏差値、キャンパスの立地、サークル、部活動などが挙げられる。これらの情報収集と各受験生の価値観や基準とが照らし合わさることで大学を選択するのが一般的である。しかし、私はそれに加えて、その大学を象徴するような言葉や文化、独自の制度など、オープンキャンパスや説明会のパンフレットに載っているが認知度の低い、もしくは載らずに漏れてしまったその大学の『大学らしさ』がより意識されるべきであると考える。

本研究では、大学が公式にコンテンツを公開しているホームページより取得可能なテキスト情報から、その大学の『大学らしさ』を抽出することを目標にしている。加えて、在学生、卒業生によるレビューデータについても分析を行うことで多角的な視点を取り入れる。加えて、2020 年に全世界を混乱に招いた新型コロナウィルス感染症による影響を、騒動以前と以後のデータを比較することで調査する。本研究が大学経営の観点から、主に広報部の『受験生や企業に対して何をアピールするか』という課題において、違和感のない、腹落ちするキーワードの獲得と今まで意識されなかったその大学に深く関連する特徴的な単語の発見の足掛かりになることを、それから、大学受験生の進学先の積極的な選択に寄与することを期待する。

2 研究対象の大学の選択

『大学のホームページのテキストデータから大学の特徴を抽出可能か』という観点から、まず、自身が在籍する大東文化大学とおおよそ同程度の規模であり、学部、学科という観点からバリエーションに富んだ 6 つの大学を選択する。5 章では、『大東文化大学』[1]、『東京大学』 [2]、『東京理科大学』[3]、『日本体育大学』[4]、『東京家政大学』[5]、『多摩美術大学』[6]の 6 つの大学の 2019 年 10 月時点のホームページと、7 章ではレビューを対象に分析する。大学へのレビューは、『大学スクールナビ』[7]という口コミサイトを利用する。また、8 章では、2020 年 8 月の同大学のホームページと比較し、新型コロナウィルスの出現以前と以降のホームページのテキストを比較する。

さらに、主に大学受験において大学はしばしば大学群というグループにされて表現される。大学群は、受験難易度や歴史、地域といった共通点から複数大学をまとめて名称をつけたものであることが多いが、同時に比較されることが多い大学同士でもあろう。ここでは、大学群の中から『大東亜帝国』に該当する 5 つの大学、『大東文化大学』、『東海大学』[8]、『亜細亜大学』[9]、『帝京大学』[10]、『国士舘大学』[11]について同様の分析を行う。

3 分析環境

分析に使用した PC スペックは、プロセッサが Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz　2.70 GHz、実装 RAM が 8GB である。OS は Winodows10 Home エディションの Windows Subsystem for Linux 機能を利用し、Ubuntu18.04 LTS に Python3.9.1 をインストール、各種サードパーティライブラリ(Pandas, Matplotlib 等)をインポートした上で分析する。

4 分析データの作成

4.1 大学ホームページからテキスト抽出

まず、各大学の Web サイトよりクローリングを行う。ダウンロード処理は再帰的に行うものとし、階層の深いファイルも取得する。取得したファイルから HTML ファイルのみを取得し、一つのテキストファイルにまとめる。次に、HTML ファイルに含まれる HTML タグと、JavaScript 部分は不要なので除去する。

4.2 形態素解析と複合名詞の抽出

4 章 1 節で取得したテキストを文ごとに MeCab を利用して形態素解析を行う。このとき、解析の結果から品詞が名詞になっている単語の原型のみを抽出するが、連続して出現

する名詞は複合名詞の可能性があるため、一度形態素として分解したのち、再度結合して一つの単語として扱う。さらに、連続する名詞を結合した結果が 4 語、5 語と長くなっている場合、不自然な結合となっている可能性が考えられる。したがって、『3 語以上連続する名詞は 2 語ずつに分解して複合名詞とみなす。』というルールに基づいて処理を実施する。しかし、十分に意味が通じる複合名詞をつくるためには課題があることを理解されたい。本節の処理によって、単純な 1 語の名詞と 2 語ずつ結合された複合名詞とみなされた単語をスペース区切りでテキストファイルに出力する。

4.3 単語のカウントとフィルタリング

次に 4 章 2 節で得られた名詞の出現回数を求める。大学間で比較するために、取得できた HTML ファイル数を単語の出現回数の分母において標準化を行う。出現した名詞は重複を除き 10 万件程度である。しかし、出現頻度が著しく高い単語には『方』や『こと』など一般的な単語が含まれる。一方で、出現頻度が著しく低い単語は、人の名前等、大学との関連が薄い単語が大部分を占める。したがって、それによって大学らしさを表現することは不自然である。ここで、単語のフィルタリングを実施する。分析の対象とする単語は、出現頻度の合計が全大学で 0.1 以上かつ 5 より少ない単語に限定する。上記閾値は最終的に分析対象単語が 1500 単語になるように設定したものである。

4.4 各大学における単語の大学らしさを計算

大学らしさを求めるにあたって、単語 t における大学 u らしさは次式にて求める。

$$\textit{単語}\mathbf{t}\textit{における大学}\mathbf{u}\textit{らしさ} = \frac{\text{大学 u の単語 t の出現回数}}{\text{u 以外の大学の単語 t の出現回数の平均}}$$

単語 t の出現頻度を大学 u とその他の大学とで比較する。出現頻度が他の大学より相対的に多く出現する単語をその大学を象徴する、特徴的な単語とみなす。

5 大東文化大学、東京大学、東京理科大学、日本体育大学、東京家政大学、多摩美術大学

5.1 ワードクラウド

大学ごとに抽出した大学を特徴づける単語を可視化するためにワードクラウドを作成する。ワードクラウドは、出現頻度の値が大きくなればなるほどフォントサイズを大きく表示する。4 章 4 節の結果を大学ごとに値の降順に並び替えた後、それぞれ先頭から 100 件の単語を出力する。

図 1　2019 年 大東文化大学のワードクラウド

図 1 は、6 つの大学のうち大東文化大学のワードクラウドである。『書道学科』や『社会学部』は、他大学にはない特別な学部であることがわかる。『学生手帳』、『青桐会』、『フォトアルバム』は大学特有の取り組みで十分に特徴的な単語といえる。他にも『学部案内』『海外同窓会』、『就職実績』、『受験生サイト』という単語が他の大学より多く出現していることがわかる。これらは、独自の取り組みとは必ずしもいえないが出現頻度が相対的に高くなっていることに注目したい。

図 2　2019 年 東京大学のワードクラウド

図 2 は東京大学のワードクラウドである。『総長室』『教員検索』といった単語が出現したことは、それぞれ、『大学総長の発言が非常に注目されている。』、『産学連携、取材などの目的で教員がよく検索されている。』といったことを反映している可能性がある。他にも、『研究倫理』、『生命倫理』、『帰国子女』という単語が比較的多く出現する。また、『東京大学基金』は名前の通り東京大学独自の取り組みである。

図 3　2019 年 東京理科大学のワードクラウド

図 3 は東京理科大学のワードクラウドである。『経営工学科』、『改善活動』と自身の大学名が大きく表示されている。『科学フォーラム』、『坊っちゃん選書』はそれぞれ、東京理科大学が発刊している科学教養雑誌と、同大学が発刊している中高生向けに最先端技術を紹介する書籍シリーズである。他には、『サイエンス道場』、『こうよう会』、『システムデザイン』といった単語が特徴的である。

図 4　2019 年 日本体育大学のワードクラウド

図 4 は日本体育大学のワードクラウドである。『和泉寮』、『体育学部』の出現回数が多いことがわかる。『体育研究所』、『健康管理』、『トレーニングセンター』、『スポーツ局』、『団体優勝』、『合宿寮』、『決勝戦』、『オリンピック』等、スポーツを連想させる単語が多数出力されている。学生生活に注目すると『アルバイト案内』などがある。また、『シャトルバス』という単語は比較している 6 つの大学限定で特有の単語である可能性が高い。

図 5　2019 年 東京家政大学のワードクラウド

図 5 は東京家政大学のワードクラウドである。『学長便り』、『狭山市』という単語が大きく表示されている。唯一の女子大学であるが、特徴的な単語としては『放課後デイサービス』『デイサービスつくし』、『挨拶運動』、『イングリッシュラウンジ』、『節電』といった他大学には存在しない施設や、より積極的に力を入れている取り組みが出現されている可能性が高い。

図 6　2019 年 多摩美術大学のワードクラウド

図 6 は、多摩美術大学のワードクラウドである。大学の助手の下で研究室の仕事や研究の補助をする『副手』と、『アニメーション』の単語が大きい。細部をみても、『アーティスト』、『スタジオ』、『版画』、『グラフィックデザイン』、『卒業制作』といった単語から美術系の大学らしさを抽出していることがわかる。

5.2　可視化とクラスタリング

4 章 4 節の結果を 2 次元のグラフにプロットする。本節では 6 つの大学を比較していることから、6 次元のデータを 2 次元に圧縮するために PCA(主成分分析)を使う。同時に、おおよその単語の分類を K 平均法によって可視化し、クラスタリングされた単語と色付けされた単語の分布の全体像を把握する。

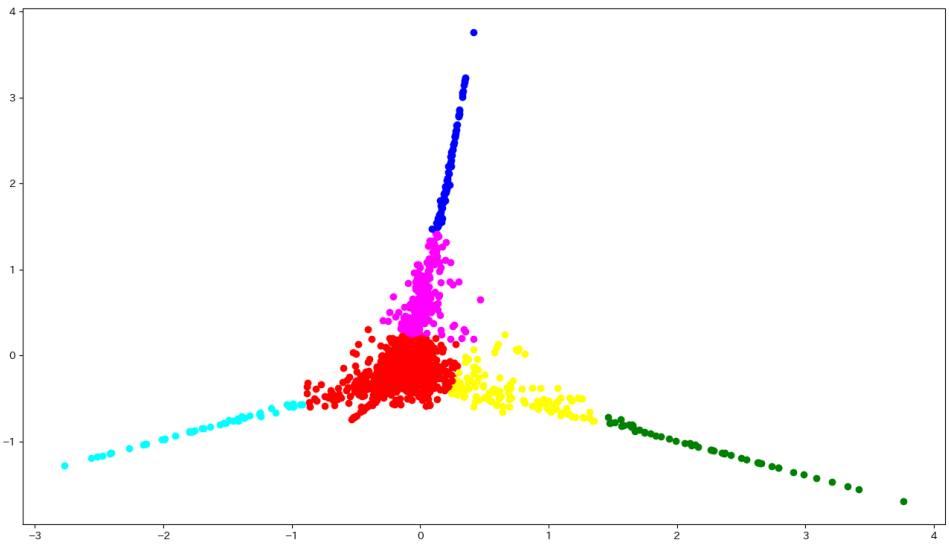

図 7　K 平均法によるクラスタリング

図 8 より青色部分には、『数学科』、『建築学科』、『土木工学科』、『薬学専攻』、『エネルギー技術』といった単語があり、東京理科大学に関連が強い単語が出力されている。同様に水色部分の『温水プール』、『こども園』、『栄養士』、『保育者』、『節電』といった単語は東京家政大学のワードクラウドに出力された単語が多いことがわかる。ピンク色部分は、『パスポート』、『健康診断』、『卒業研究』、『キャンパスマップ』、『海外留学』等、大学の一般的な単語が出力される。赤色部分は、『私』、『生活』、『日本語』等、大学関連の単語よりも更に一般的な単語と、『インターネット出願』、『研究テーマ』、『プレゼンテーション』等、ピンク色部分に出現されたような大学で一般的に使われる単語が含まれている。黄色部分は、『スポーツ』、『サークル活動』、『交通アクセス』等、大学で一般的に使われる単語が含まれているが、『フォトギャラリー』等、大東文化大学のワードクラウドに出現する単語も含まれている。

最後に、緑色部分は『シャトルバス』、『決勝戦』、『オリンピック』、『学生寮』等、日本体育大学のワードクラウドに出現している単語が出力されている。

東京理科大学、東京家政大学、日本体育大学はそれぞれの大学について独自性の強い単語が、青色、水色、緑色の部分にそれぞれ出力されている。一方で、大東文化大学、東京大学、多摩美術大学のワードクラウドの中で出現された単語は中央の赤色、ピンク色、黄色の中に大部分が含まれているといえる。

5.3 類似度分析

4 章の方法で求めた単語の大学らしさの指標を利用して各大学の類似度を計算する。本研究では、コサイン類似度を求めることで各大学における単語の大学らしさをベクトル空間モデルとして相互距離を計算し、類似している大学を定量的に計算することを検討する。現状の大学らしさの指標の中には、1 以下の少数からなる数値から、特定の大学で非常に大きな出現頻度を持つ単語には 100 といった数値が含まれており、過大となっている。この大きな幅のばらつきを過小評価したベクトルに整理するためにすべてのデータを自然対数に変換する。また、このとき値に 0 が含まれている場合、正しく対数変換できないためすべてのデータに定数 1 を足す。大学 a の単語ベクトルと大学 b の単語ベクトルのコサインは次式にて求める。

$$\cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}$$

6 つの大学を総当たりで単語ベクトル間のコサインを求める。

	大東大	東大	理科大	日体大	家政大	多摩美大
大東大	-	-	-	-	-	-
東大	0.508203	-	-	-	-	-
理科大	0.527565	0.460922	-	-	-	-
日体大	0.440884	0.320272	0.421618	-	-	-
家政大	0.436153	0.358356	0.365979	0.386239	-	-
多摩美大	0.521413	0.437949	0.449057	0.371553	0.370765	-

図 8　6 つの大学(大東文化大学、東京大学、東京理科大学、日本体育大学、東京家政大学、多摩美術大学)の単語ベクトルの cos 類似度

大東大	東大	理科大	日体大	家政大	多摩美大
0.486844	0.41714	0.445028	0.388113	0.383498	0.430147

図 9　他大学との cos 類似度の平均

類似度が大きい大学は上から『大東文化大学と東京理科大学』、『大東文化大学と多摩美術大学』、『大東文化大学と東京大学』である。逆に、類似度が低い大学は下から『東京大学と日本体育大学』、『東京大学と東京家政大学』、『東京理科大学と東京家政大学』である。文系総合大学である大東文化大学と理工系総合大学である東京理科大学の類似度が最も高いという結果は意外である。しかし、全体的には『東京理科大学』を例外に、前節の主成分分析と K 平均法において中心から外れた位置にある水色、緑色をそれぞれ意味して

いた『東京家政大学』、『日本体育大学』の cos 類似度の平均はどちらも低い値となっており、可視化の結果と大きく離れていない。『東京理科大学』に関しては、主成分分析と K 平均法による可視化において、大学との関連が高い単語としての水色部分と、『大東文化大学』、『東京大学』、『多摩美術大学』のように赤色、黄色、ピンク色部分といった中央部分の両方により多くの単語が出力されている可能性がある。

6 大東亜帝国

6.1 ワードクラウド

5 章では、大学のホームページからそれぞれの大学を特徴づける単語を分析した。本章では同様の処理を、意図的に選択した大学ではなく、大学群である『大東亜帝国』に含まれている 5 つの大学、『大東文化大学』、『東海大学』、『亜細亜大学』、『帝京大学』、『国士舘大学』を対象に行う。また、取得したホームページのテキストデータはすべての大学において 2020 年 8 月時点のものとする。

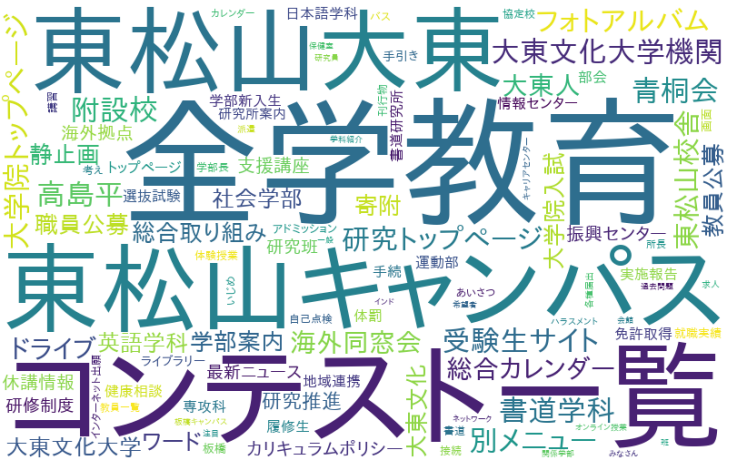

図 10　2020 年　大東文化大学のワードクラウド

図 10 は、2020 年の大東亜帝国のグループで比較したときの大東文化大学のワードクラウドである。『全学教育』、『コンテスト一覧』といった単語が大きく表示される。比較している大学データが変わっているが、『フォトアルバム』、『書道学科』、『青桐会』、『海外同窓会』といった単語は 5 章の結果同様に出力されている。これらは、大学特有の大学らし

さを表現する単語である可能性が高い。

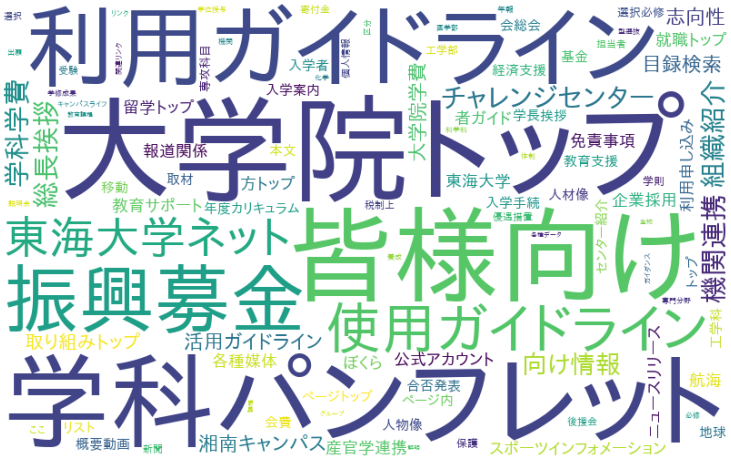

図 11　2020 年　東海大学のワードクラウド

図 11 は、東海大学のワードクラウドである。『大学院トップ』、『学科パンフレット』、『利用ガイドライン』等の単語が大きく出力される。これらは、ヘッダーやフッターに含まれたため出現頻度が高くなった可能性がある。特徴的な単語を挙げるとしたら『振興募金』、『東海大学ネット』、『チャレンジセンター』、『産官学連携』、『スポーツインフォメーション』等といった単語が出力されており、同大学の取り組みや方針に関連している。しかし、5 章と比較すると大学独自の特徴的な複合名詞よりも、やや一般的な単語の出力が増えている。

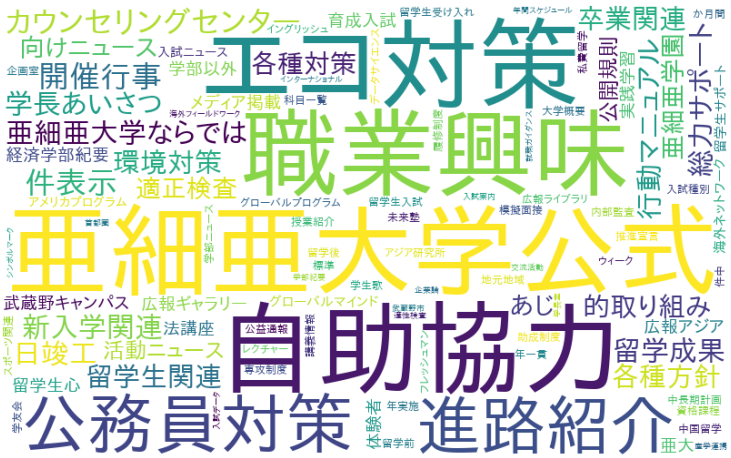

図 12　2020 年　亜細亜大学のワードクラウド

図 12 は、亜細亜大学のワードクラウドである。特徴的な単語としては『エコ対策』、『職業興味』、『自助協力』、『公務員対策』、『進路紹介』、『カウンセリングセンター』が挙

がる。これらの単語が出力しているからといって、『亜細亜大学が他大学より学生の進路へのアプローチに特徴がある、より積極的である。』、『環境問題への貢献度が優れている』と断定することはできない。また、どういった文脈でこれらの単語が出現しているかも分からない。しかし、出現した単語はどれも亜細亜大学以外の大学に出力されていても違和感がないような一般的な単語であり、その相対的な出現頻度が高いということは、亜細亜大学においては少なくともウェブページ上での進路指導に関する情報は充実している可能性がある。

図 13　2020 年 帝京大学のワードクラウド

図 13 は帝京大学のワードクラウドである。『サポート資格』、『連携スポーツ』、『所有校』という単語が大きく目立つ。『帝京大学的』という単語が出力されているが、本研究の趣旨はまさにその性質を言語化することである。『グループ帝京大学』は、『帝京平成大学』、『帝京科学大学』等、その他国内外の短期大学や専門学校、高等学校、小中学校から、病院にいたるグループのことである。『問い合わせ言語』といった単語も特徴的であり、コンピュータ言語についての記述であろう。

図 14　2020 年 国士舘大学のワードクラウド

図 14 は国士舘大学のワードクラウドである。『国士舘オリンピック』をはじめ、『シンクロナイズドスイミング』、『新体操部』、『ホッケー』、『水球女子』、『ソフトボール部』、『バスケットボール』、『駅伝』、『スポーツ情報』、『スポーツマンシップ』等、スポーツ関連の単語が非常に多く含まれている。また、『広報犬』という単語があるが、国士舘大学では『こくしば』というキャラクターが大学の様子を紹介、応援し、広報に従事している。

6.2 類似度分析

大東亜帝国の 5 つの大学のデータを対象に 5.3 と同じ方法で各大学の単語ベクトルのコサイン類似度を計算する。

	大東大	東海大	亜大	帝京大	国士舘
大東大	-	-	-	-	-
東海大	0.494079	-	-	-	-
亜大	0.493396	0.383354	-	-	-
帝京大	0.541098	0.416401	0.350457	-	-
国士舘	0.598935	0.431616	0.396332	0.490027	-

図 15　5 つの大学(大東亜帝国)の単語ベクトルの cos 類似度

大東大	東海大	亜大	帝京大	国士舘
0.531877	0.431362	0.405885	0.449496	0.479228

図 16　他大学との cos 類似度の平均

最も類似度が高い大学の組み合わせは上から『大東文化大学と国士舘大学』、『大東文化大学と帝京大学』、『大東文化大学と東海大学』である。逆に、最も類似度が低い大学の組み合わせは下から『亜細亜大学と帝京大学』、『東海大学と亜細亜大学』、『亜細亜大学と国士舘大学』である。類似度の合計が高いのは『大東文化大学』であり、低いのは『亜細亜大学』ということになる。5 章でも同様だが、『大東文化大学』は他の大学との類似度が高い傾向にあり、大学として一般的なコンテンツを提供しているとみることもできる。一方で、『亜細亜大学』はより独自方向のベクトルとなっており出現する単語の性質が独特であることがわかる。

7 各大学への口コミ

本章では 5 章、6 章と同様の分析を 5 章で扱った 6 つの大学『大東文化大学』、『東京大学』、『東京理科大学』、『日本体育大学』、『東京家政大学』、『多摩美術大学』についてそれぞれのレビューデータを対象に行う。レビューデータは 2 章で紹介した、口コミサイト、『大学スクールナビ』から取得する。同サイトでは、大学について『満足しているポイント』、『不満に感じているポイント』、『おすすめの学部』、『通ってよかったか』という項目で口コミを記述する。本研究では、上記の 4 つの項目すべてを調査の対象とする。調査方法も、5 章、6 章とおおよそ同様であるが、単語の標準化に html ファイル数を利用していた部分をレビュー数に置き換える。しかし、大学によってはレビュー数が 3 つ程度しか存在しない場合があり、レビューデータが主観的になっているという課題が残っていることを理解されたい。

7.1 ワードクラウド

図 17　大東文化大学のレビューデータのワードクラウド

図 17 は、大東文化大学のワードクラウドである。レビューにおいて他大学より出現頻度が高い単語が出力される。『競技場』、『埼玉県』、『キャリアセンター』、『就職支援』といった単語が出力され、フォントサイズが大きい。出力される単語は、どういった文脈で出現しているかはわからないため、例えば、『就職支援』が手厚いのか、手薄なのかは分析していない。しかし全体的にスポーツについての書き込みが多い印象を受ける。

図 18　　東京大学のレビューデータのワードクラウド

図 18 は、東京大学のワードクラウドである。出現頻度が高いのは、『最先端』、『幻想』、『理系』、『知名度』、『文献』等である。また、3 年生になってから研究する専門分野を決定する『進振り制度』は、東京大学の特有の制度である。

図 19　東京理科大学のレビューデータのワードクラウド

図 19 は東京理科大学のワードクラウドである。『数学科』、『留年』という 2 語が特に大きく出力されている。東京理科大学は留年率が高いことで有名であるがこれも例に漏れず大学らしさである。

図 20　日本体育大学のレビューデータのワードクラウド

図 20 は日本体育大学のワードクラウドである。フォントサイズが大きい単語は、『マット運動』、『大盛り』、『消防士』、『ジャンプ』、『スリル』、『スポーツ選手』である。スポーツや体育関連、食事に関する単語が出力されている。

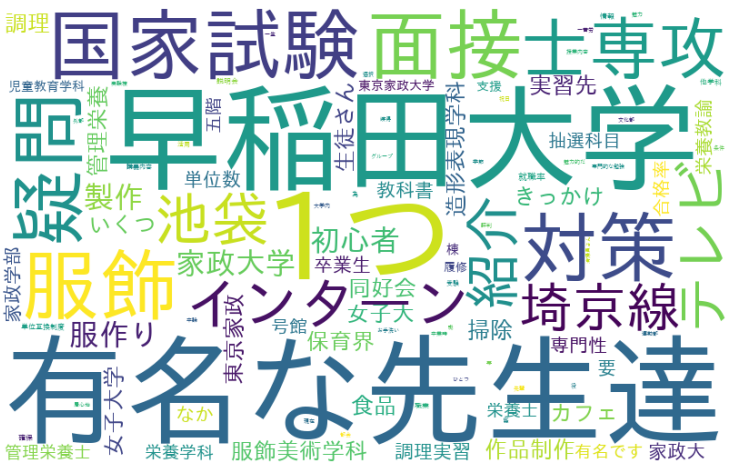

図 21　東京家政大学のレビューデータのワードクラウド

図 21 は東京家政大学のワードクラウドである。『国家試験』、『面接』、『服飾』、『早稲田大学』、『有名な先生達』といった単語のフォントサイズが大きい。東京家政大学は、他大学との単位互換制度があり、早稲田大学で受講した講義を卒業単位として認めているようだ。また、『国家試験』といった単語から国家試験の取得を目指す学生が比較的多い可能性がある。

図 22　多摩美術大学のレビューデータのワードクラウド

図 22 は、多摩美術大学のワードクラウドである。『時間管理』、『昼食』といったサイズの大きな単語や、『立地条件、40 分かかり、片道 2』という立地、キャンパスへのアクセス関連の単語が目立つ。就活と卒業制作を両立させるための『時間管理』は、特に美術系大学では不可欠な能力かもしれない。

7.2 類似度分析

	大東大	東大	理科大	日体大	家政大	多摩美大
大東大	-	-	-	-	-	-
東大	0.670079	-	-	-	-	-
理科大	0.72663	0.759591	-	-	-	-
日体大	0.506744	0.445596	0.484725	-	-	-
家政大	0.667037	0.673042	0.709795	0.487226	-	-
多摩美大	0.571558	0.562266	0.582249	0.399996	0.562318	-

図 23　6 つの大学(東京理科大学、東京大学、東京理科大学、日本体育大学、東京家政大学、多摩美術大学)のレビューデータの単語ベクトルの cos 類似度

大東大	東大	理科大	日体大	家政大	多摩美大
0.62841	0.622115	0.652598	0.464857	0.619884	0.535677

図 24　他大学との cos 類似度の平均

レビューデータに関しては、キャンパスやサークル、ゼミなど、学生生活に関連する記述がホームページより多くなっているため類似度の数値は高くなったと考えられる。最も類似度が高い組み合わせは上から『東京大学と東京理科大学』、『大東文化大学と東京理科大学』、『東京理科大学と東京家政大学』である。逆に、最も類似度が低い組み合わせは下から『日本体育大学と多摩美術大学』、『東京理科大学と日本体育大学』、『東京理科大学と日本体育大学』である。また、『日本体育大学』は、類似度の合計が際立って低い。レビューデータは、学生生活についてホームページよりも直接的に表現しており、学生自身が『どういった基準で大学を評価しているか』という学生の性質も結果に影響するだろう。

8 コロナ感染症による各大学ホームページへの影響

本章では、5、7 章で扱った 6 つの大学のうち、『大東文化大学』、『東京大学』、『日本体育大学』の 3 つの大学について新型コロナウィルス感染症の感染が拡大する以前の 2019 年 10 月時点と、拡大した後の 2020 年 8 月時点を対象に分析する。

4 章 2 節までの手順で名詞、または、2 つの形態素からなる複合名詞のデータを作成する。4 章 3 節と同様に、単語の出現頻度をカウントし、html ファイル数で標準化する。大

学 u の単語 t がコロナ騒動前から、コロナ騒動後にどう変化しているかを次式にて求める。

$$\text{コロナ騒動後の大学}u\text{の単語}t\text{の出現頻度の変化} = \frac{\text{2020 年 8 月の大学 u の単語 t の出現回数}}{\text{2019 年 10 月の大学 u の単語 t の出現回数}}$$

以上の計算をコロナ騒動前後で出現するすべての単語について行い、単語の相対的な出現頻度からベクトルを作成する。したがって、前年より出現頻度が高い単語のスカラの数値が高くなる。

8.1 ワードクラウドと類似度分析

図 25　大東文化大学のワードクラウド

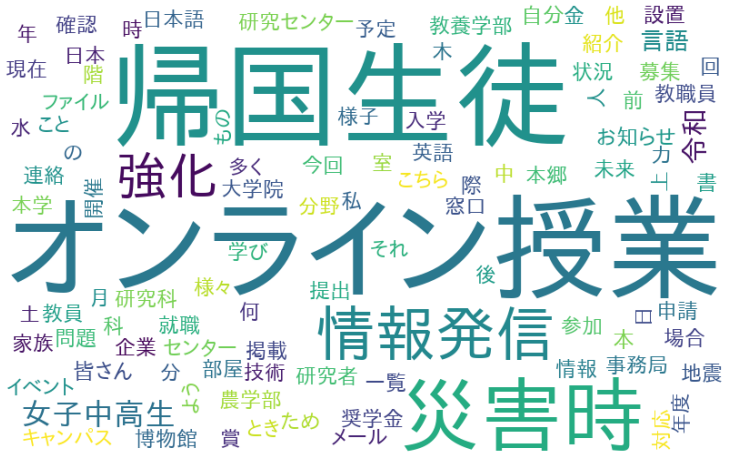

図 26　東京大学のワードクラウド

図 27　日本体育大学のワードクラウド

大東文化大学	0.979142202
東京大学	0.985486920
日本体育大学	0.997551926

図 28　　大東文化大学、東京大学、日本体育大学のコロナ騒動の前後のホームページの類似度

図 25、26、27 は、それぞれの大学のコロナ騒動後のワードクラウドである。『支援』、『サポート』、『各種届出』、『災害時』、『中止』、『オンライン授業』等の単語が感染症の影響を連想させる。また、『学部新入生』、『帰国生徒』等、一部の学生へのアナウンスが増加していることがわかる。

図 28 は、各大学のコロナ前後のホームページのコサイン類似度である。大学ホームページは、アーカイブとして記事を蓄積するという性質があるため類似度は高くなる。また、最も変化しているのは『大東文化大学』であるが、2020 年という節目であったため、ホームページのリニューアルの影響があった。

9 まとめ

本研究では、バリエーション豊かな合計 11 つの大学のホームページ、クチコミデータについて、『大学らしさ』を抽出するためにワードクラウドの作成、類似度分析、主成分分析と K 平均法による単語ベクトルの全体像の可視化とクラスタリングを行った。ホームページの単語の出現頻度という限られたリソースの中で、大学独自の制度や文化が言語化されているケースのいくつかを目の当たりにした。しかし、一言では理解しかねる概念を詳細に説明すれば自然と単語の出現頻度は高くなる。ホームページ上で重要な単語を強調したい場合には、フォントカラーを変えること、画像を作成すること等、ビジュアルとして目に留まるように工夫することの方が一般的なことかもしれない。『伝えたいことをページのどこに配置するか』という課題が重要であればあるほど、本研究から断定できることは限定的になってしまうだろう。

10 課題

4 章 2 節でも言及したが、形態素から意味が十分に通じる複合名詞を抽出する方法を探索することが課題として挙がる。7 章のレビューの分析に関しても、在学生、卒業生の本音である重要な章であったが十分なデータ数を集めることはできなかった。

また、ワードクラウドの作成、類似度分析もすべて 4 章 4 節の単語のフィルタリングの影響を大きく受ける。分析対象のデータを絞り込む際に、出現頻度が低い、もしくは高くなっているがフィルタリングで除外された、より特徴的な単語が存在している可能性があるため、精度の改善と評価の仕組みが必要だろう。

参考文献

[1] 大東文化大学（https://www.daito.ac.jp/）

[2] 東京大学（https://www.u-tokyo.ac.jp/）

[3] 東京理科大学（https://www.tus.ac.jp/）

[4] 日本体育大学（https://www.nittai.ac.jp/）

[5] 東京家政大学（https://www.tokyo-kasei.ac.jp/）

[6] 多摩美術大学（https://www.tamabi.ac.jp/）

[7] 大学スクールナビ（https://school-navi.org/university/）

[8] 東海大学（https://www.u-tokai.ac.jp/）

[9] 亜細亜大学（https://www.asia-u.ac.jp/）

[10] 帝京大学（https://www.teikyo-u.ac.jp/）

[11] 国士舘大学（https://www.kokushikan.ac.jp/）