トピックも日ごとに変わるネットかな　＜トピック抽出＞

2020 年 11 月下旬　白田由香利

街がクリスマスの準備で賑わってきた．この時期，オクスフォードの市街地に出るクリスマス・マーケットに出かけて行って，プレゼントやオーナメントを買いたいなぁと強く思う．仕事があるので，東京を離れるわけにはいかないが，今年は行けないとなると，さらに気持ちは募る．2006 年，大学の壁に聖歌隊のミサのチラシが貼ってあった．場所はニュー・カレッジ内の教会で，ヘンデルのメサイアだ．このポスターに使われていた「天使の歌」というウィリアム・アドルフ・ブグロー作の宗教画があまりに美しくて驚いた．当日聞きに行ってしまったほどだ．この絵がパブリックドメイン作品であることは，後で知った．だから，ポスターに使われていて問題なかったのだ．「天使の歌」の絵では，女性の姿の天使がバイオリンやリュートを奏でながら，赤ちゃんのキリスト様をのぞき込んでいる．天使が楽器を演奏するというテーマ，世界で一番美しいものの一つだと思う．鬼に金棒，虎に翼．ロックダウンの私のところにも，シベリア大ヤマネコが天使になって，リュートを手に手に陣中見舞いに来てくれないものか，と思う．猫科の動物はどれも好きだが，とりわけオオヤマネコ系の野生ネコは大好きだ．太い前足をモフモフしながら天上の音楽を聴けたらどんなにか嬉しいことだろう．リュートに関しては，オクスフォードで何枚もリュートの CD を買ってきた．バイオリンに比べると眠くなるような曲が多いが，今でもよく聞いて楽しんでいる．学生の皆さんも，この状況下で勉強がつらくなったら，私のように，何でもいいので楽しいことを思って期末までラストスパートしてください．

さて，機械学習のアプリケーション分野のひとつに自然言語処理がある．人間が書いた大量の文書を入力として，形態素解析をして，トピック抽出する，というアプローチが典型的な処理のひとつである．その概要を説明する．

ID	DATE	TEXT
1	2020/11/17	学生センター学生課輔仁会大学支部所属団体（公認団体）及び任意団体　各位　標記の件について、年末年始の各事務室閉室に伴い、大学におい
2	2020/10/20	学生の皆さんへ学生センター学生課新型コロナウイルス感染症に伴う課外活動への対応について（10月20日以降）　課外活動の再開については、
3	2020/10/7	第71回四大学運動競技大会について、新型コロナウイルス感染症の影響により、全ての競技（正式種目、一般種目、教職員種目）が中止となりま
4	2020/10/7	本年度の大学祭（桜凛祭）は、「第51回桜凛祭特設ページ」をWeb上に設け、オンライン配信により開催します。　配信日時：11月２日・３日
5	2020/8/31	学生の皆さんへ緊急事態宣言解除及び東京都による休業要請緩和後の授業等の取扱いについて(第３報)　７月２０日に標記の第２報において、第
6	2020/7/20	学生の皆さんへ緊急事態宣言解除及び東京都による休業要請緩和後の授業等の取扱いについて(第２報)　８月の補講期間及び９月の集中講義期間
7	2020/7/7	学生の皆さんへ 学生センター学生課 新型コロナウイルス感染症に伴う課外活動への対応について（８月１日以降）　課外活動の再開について
8	2020/4/20	1令和２（2020）年４月 20 日学習院大学長遠隔授業実施のガイドライン＜本ガイドラインの構成＞1 背景及び目的2 定義（1）オンデマンド授
9	2020/9/3	1 / 4授業実施方法に関する大切なお知らせ（第２学期）0. はじめに新型コロナウイルス感染症の感染拡大を防止するため、学習院大学では、原則

データは，当初，学生の皆さんに WEB アンケートに書き込んでもらったデータを使おうと思ったが，データ量が少なくて，まっとうな分析ができなかった．そこで，学習院大学 WEB のコロナに関するお知らせを使った（上記参照）．9 個のお知らせと，その内容のテキストデータである．さて，どのようなことが書いてあるのでしょうか？

形態素解析にかけると，名詞，動詞(その原形)，副詞などの形態素に分けて，出力してくれる．形態素解析エンジンとしては MeCab（めかぶ）というフリーソフトが広く使われている．これは京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発されたオープンソースである．私が初めて使ったのは今から約 20 年前だった．このようにすごいことができるソフトがフリーで使えるなんて，なんとすごいことだろう，と感激した．日曜日に大

雪の中を興奮しながら，分析の続きをしに研究室に行ったことを今でも覚えている．人間，夢中になっているときは，雪が降ろうが槍が降ろうが，一切気にならない．

ここでは，名詞のみを使って（つまり，それ以外は捨てる），名詞―名詞と２つつながる複合語だけにした．例えば，次の表にあるように，対面―授業，ウイルスー感染，などである．２つつながった複合語をバイグラムと呼ぶ．長く，トピック抽出の研究をしているが，名詞―名詞のバイグラムが最もよい結果を生んでくれる．動詞を使っても，名詞のバイグラムほど良い結果はでない，ことが多い．

#	1:管理システム-対面授業-配信授業-オンデマンド授業-会議システム	#	2:コロナウイルス-ウイルス感染-履修登録-授業実施-課外活動	#	3:遠隔授業-同時配信-学習管理-公衆送信-文部科学	#	4:著作権-設置基準-対面形式-授業等-感染症
対面授業	13	コロナウイルス	14	遠隔授業	36	著作権	12
管理システム	13	ウイルス感染	14	同時配信	13	設置基準	11
配信授業	12	課外活動	11	学習管理	13	対面形式	8
オンデマンド授業	9	授業実施	11	公衆送信	8	授業等	7
会議システム	7	履修登録	11	文部科学	7	感染症	6
一部改正	5	大学設置	10	科学省	7	Web会議	5
メディア利用	5	新型コロナ	9	当該授業	5	桜凛	4
開設部門	4	感染症	9	改正著作	5	凛祭	4
権者	4	権法	6	授業開始	5	東京都	4
修正期間	4	全面的	5	著作物	4	遠隔授業	4
成績評価	3	単位数	5	登録修正	4	入構許可	4
機会確保	3	対面授業	4	担当教員	3	補償金	4
補助員	3	学生課	4	教室等	3	授業受講	3
大学履修	3	任意団体	4	科目開設	3	支部所属	3
指定管理	3	令和	4	履修者	3	新型コロナ	3
管理団体	3	10月	4	授業内容	3	遠征等	3
施行等	3	日以降	4	設問解答	3	51回	3

トピック抽出の手法も多種多様であるが，最も古くから使われている標準的な手法のひとつである潜在的意味解析（Latent Semantic Analysis，LSA）を使った．LSA の結果から，主な４つのトピックを選んで，その内訳を上記の表に示した．お知らせの内容が類似しているので，４つのトピックも似たような内容であるが，若干異なる．

トピック１：対面授業の代わりに配信授業やオンデマンド授業

トピック２：コロナウイルスによる課外活動や授業実施への影響

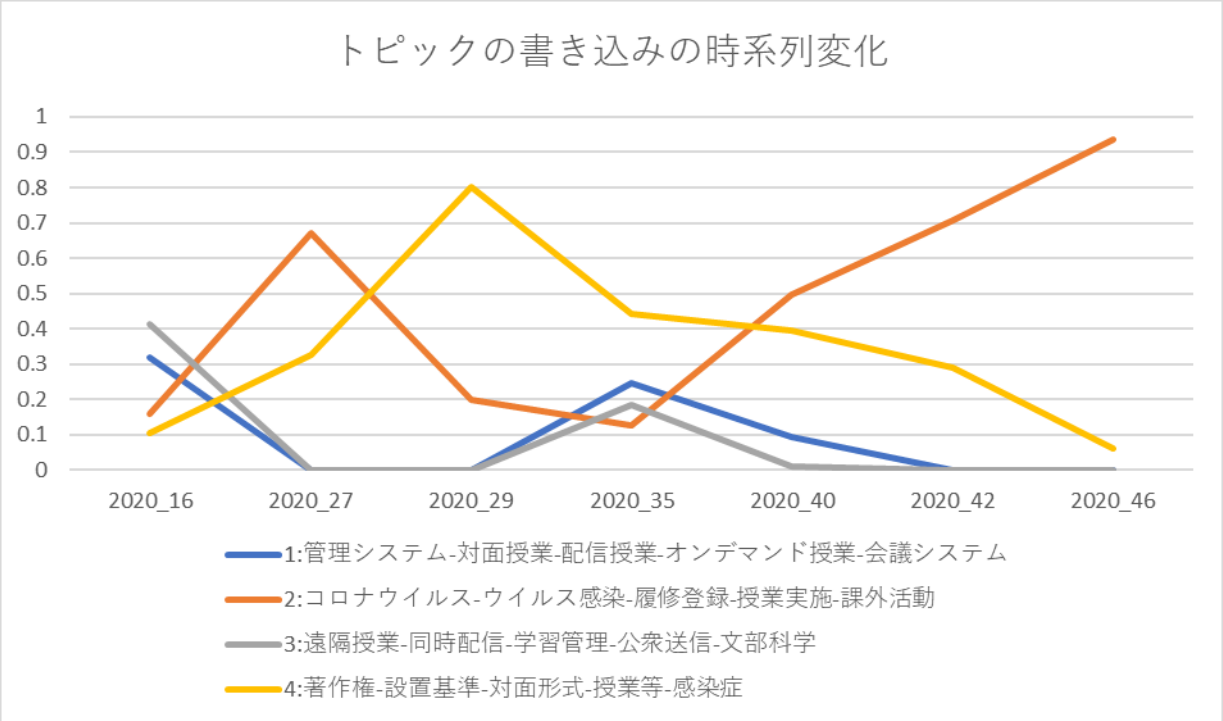
トピック３：遠隔授業について

トピック４：著作権について

LSA が教えてくれるのは，どのトピックではどの複合語が何回出現している，という回数のみである．それを見て，適切なトピック名を考えるのは，人間の役目である．上記のトピック名は苦し紛れにつけてみた．一般に入力テキストデータはもっと膨大にいれるべきであり，このデータでは十分な分析ができていないが，やり方の大筋だけ理解してほしい．大量のデータの場合，一般にトピックがきれいに分かれることが多くなる．

次にどのお知らせが，どのトピックの要素を多く含んでいるかを比率で表してみる．例えば，お知らせ＃１は９割，トピック２の話題について記していることが分かる．

# ID	DATE	TIME	1:管理システム-対面授業-配信授業-オンデマンド授業-会議システム	2:コロナウイルス-ウイルス感染-履修登録-授業実施-課外活動	3:遠隔授業-同時配信-学習管理-公衆送信-文部科学	4:著作権-設置基準-対面形式-授業等-感染症	
1	20201117 ,		0.0000	0.9375	0.0000	0.0625	1.0000
2	20201020 ,		0.0000	0.7083	0.0000	0.2917	1.0000
3	20201007 ,		0.0000	0.9762	0.0238	0.0000	1.0000
4	20201007 ,		0.1875	0.0208	0.0000	0.7917	1.0000
5	20200831 ,		0.0902	0.0328	0.0000	0.8770	1.0000
6	20200720 ,		0.0000	0.1985	0.0000	0.8015	1.0000
7	20200707 ,		0.0000	0.6724	0.0000	0.3276	1.0000
8	20200420 ,		0.3189	0.1608	0.4135	0.1068	1.0000
9	20200903 ,		0.4059	0.2206	0.3676	0.0059	1.0000



上図のように，その時期のトピックの比率の時系列変化を見ることもできる．上記では，データを週ごとに分けた．2000_16 というのは，西暦 2000 年の第 16 週を意味する．お知らせは時間的に飛んでいるので，何もない期間も多いが，トピック 2 は 11 月になっても勢いがあることが分かる．他方，トピック 4 は 11 月には話題としては下火になっている．

今は人が考えているが, トピック抽出で, 適切なタイトルを AI がつけてくれるようになるのがいつか, 楽しみである．

終わり

引用元：枝ぶりの日ごとに変わる芙蓉かな　　芭蕉