S5-P15

Wikipediaから文化の広がりを理解する

村山太一
(大阪大学産業科学研究所)

背景

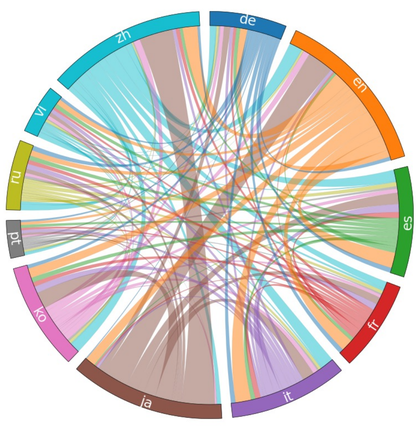
- WikipediaはOpen Knowledge Graphの一種
- Wikipediaには300を超える言語版があり、5,570万記事、2,040万のEntityが登録されており、毎月25万記事が作成
- Wikipediaの各記事はボランティアなどによって編集され、各言語の文化や人気が反映される

Question・問題提起

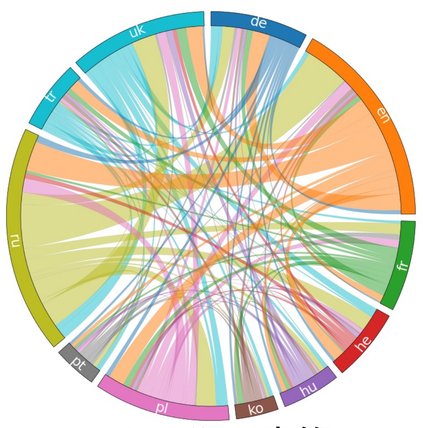
各言語のWikipediaの編集履歴やEntityの登録履歴を分析することで、
- どのような文化が広がっているのか？
- 言語を超えて広がる文化はどのような特徴があるのか？
- 文化が広がるためにキーとなる言語が存在するのか？
- コンテンツに含まれる要素の広がりはどのようなものか？

上記のことがわかると、文化と言語の関係が明らかになるとともに、商品やアイテムを世界に展開したい場合に必要となってくる要素が明らかになる

マクロの観点での取り組み



日本語の書籍
の広がりのパス



ロシア語の書籍
の広がりのパス

5年後に複数言語にページが作成されているか
各特徴のオッズ比

Genre
- Business: 6.23
- Crime Novel: 5.29
- encycropedia: 4.20
- Isekai (異世界物): 4.08

Language
- アイルランド語: 4.41
- カタルーニャ語: 3.32
- ラテン語: 2.92
- ハンガリー語: 2.63

- オフィシャルサイト: 1.70
- FreeBase登録: 1.60

ミクロの観点での取り組み

各言語のWikipediaの編集履歴を見ることで、コンテンツの各要素がどこから生まれ、どの要素が世界に広まっているかが理解できるのではないか？

Example

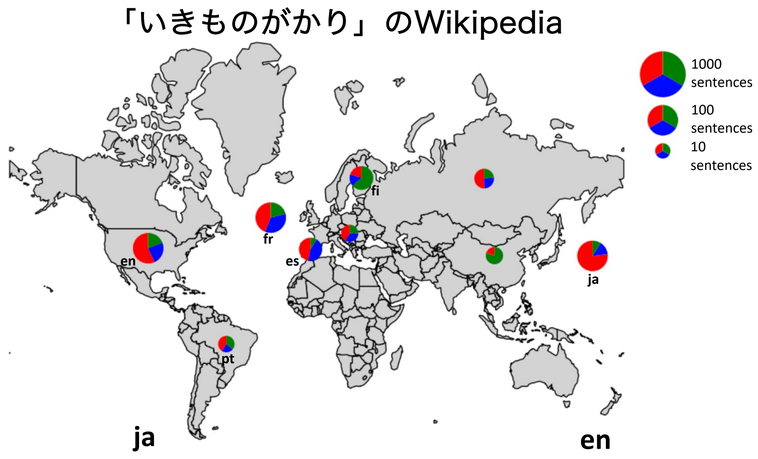
「ORANGE RANGEが2012年にレーベルが変更した情報」はどの言語から始まり、どの程度広まっているのか？

手法の概要

- 各言語のWikipediaで同じ意味の文を抽出し、どの言語版でその意味の文書が生まれたのか？どれくらい広がっているかについて示す
- 文の埋め込み表現の抽出にはMultilingual Sentence-BERTを採用
  (huggingfaceで提供のdistiluse-base-multilingual-cased-v2)
- 文のクラスタリングには**コサイン類似度**を採用
- 各言語のWikipediaに含まれる文を3つのクラスに分類

- その言語ではじめに生まれ、世界的に広がっている文
- その言語で生まれたが、世界に広がっていない文
- 他の言語で生まれ、輸入されてきた文

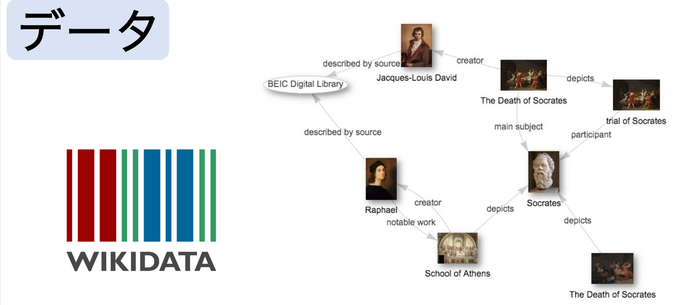
「いきものがかり」のWikipedia

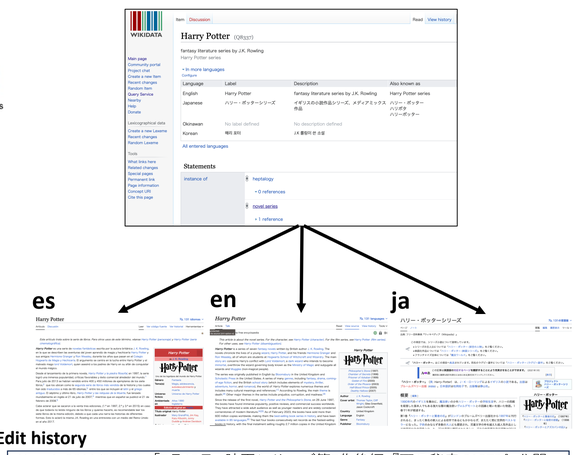


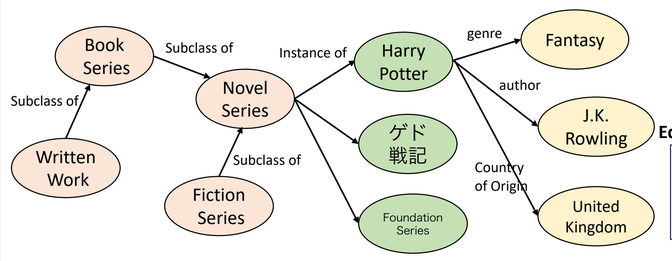
公式ファンクラブは「いきものがかりファンクラス1年2組」。
吉岡自身ももともと路上ライブに興味を持っていたこともあり、当時2人が相模大野駅前で行っていたライブに飛び入り参加したことがきっかけで、そのまま加入することとなった。
2005年3月26日に初のホールでのワンマンライブを厚木市文化会館小ホールで開催した。

He was born in Hamamatsu, Shizuoka Prefecture, but he later moved to the city Ebina, Kanagawa Prefecture, at an early age. It was announced that he registered his marriage on August 17, 2013, to a woman who has no connection to the media industry. He has released no further information about her identity.

データ





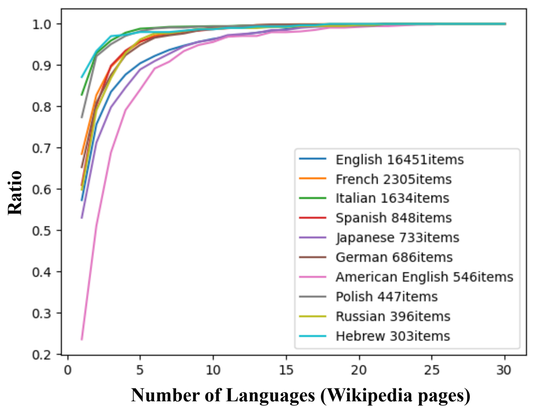


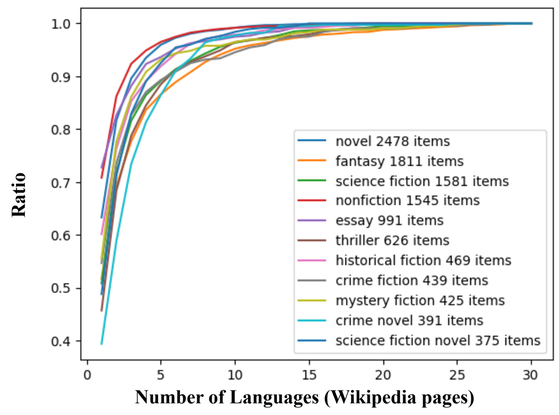
Edit history

User A	2011/06/24	「7月15日 - 映画シリーズ第7作後編『死の秘宝 PART2』公開。映画シリーズ完結。」
User B	2013/01/12	「日本語版文庫本（2013年2月14日発売）」
User C	2013/01/15.	「J・K・ローリングが『ハリー・ポッター』シリーズの完成後に執筆した、物語上に存在する童話集。」

- Wikidataは階層的なオントロジー構造で、各データが相互にリンクされている多言語対応のKnowledge Database
- 特定のクラスに属する、Wikipediaの編集履歴を獲得

Written Workより下のSubclassに属する2000年以降に出版されたWikidata (書籍)が対象（全43,058 Entity）
各Entityがどれだけの言語のWikipediaにページが作成されているかを調査





アルゴリズム

時系列順に各文章を適用

① Wikipediaの編集履歴の前処理

i : 編集履歴をSentenceごとに分割

2011年6月24日追加

「8番目の物語（＝事実上の第8巻）」と銘打たれている。文はト書きと台詞で構成されており、第1巻 - 第7巻のような小説のスタイルはとられていない。またジョン・ティファニー、ジャック・ソーンとローリングの共著名義である。」

分割

2011/06/24	「8番目の物語（＝事実上の第8巻）」と銘打たれている
2011/06/24	文はト書きと台詞で構成されており、第1巻 - 第7巻のような小説のスタイルはとられていない。
2011/06/24	またジョン・ティファニー、ジャック・ソーンとローリングの共著名義である。

ii :言語に関わらず時系列順に並び替え

Date	Lang	Sentence
2011/06/24	ja	またジョン・ティファニー、ジャック・ソーンとローリングの共著名義である。
2011/06/24	es	Además, EA produjo en 2003 un juego de simulación de Quidditch: Harry Potter: Quidditch World Cup.
2011/06/24	en	Some of the translators hired to work on the books were well-known authors before their work on Harry Potter, such as Viktor Golyshev, who oversaw the Russian translation of the series' fifth book.
⋮	⋮	⋮
2019/02/23	fr	Le trio retourne donc au château, très vite attaqué par Voldemort.
2019/02/24	pt	Arrecadando em bilheterias aproximadamente 7,7 bilhões de dólares;
2019/02/24	zh	學者與記者的解讀更多元，有的還包括政治。大致上，對於小說系列主體的解讀有

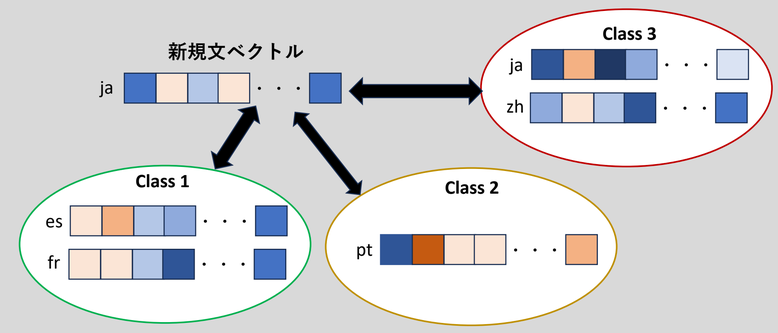
② 文章の埋め込み表現獲得

またジョン・ティファニー、ジャック・ソーンとローリングの共著名義である。

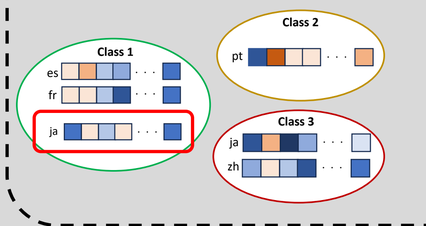
→ **Multilingual Sentence Transformers** →

③ 文章のクラスタリング

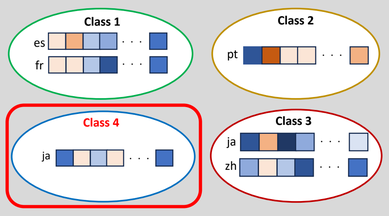
i 以前出現した文とのコサイン類似度



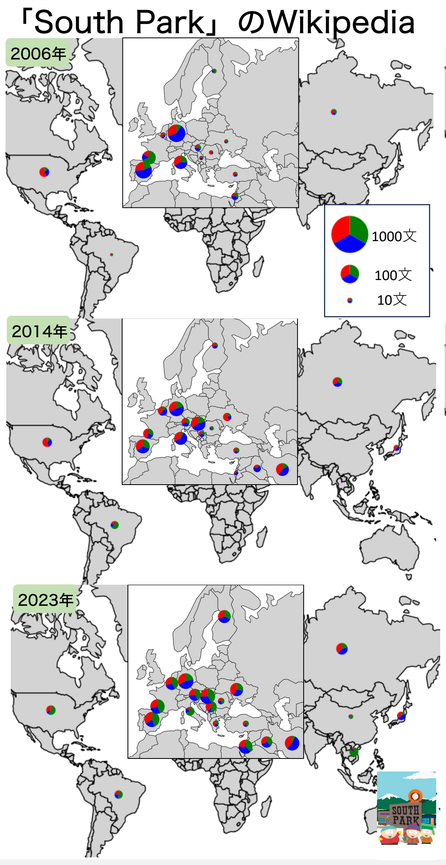
ii 既存クラスへの追加



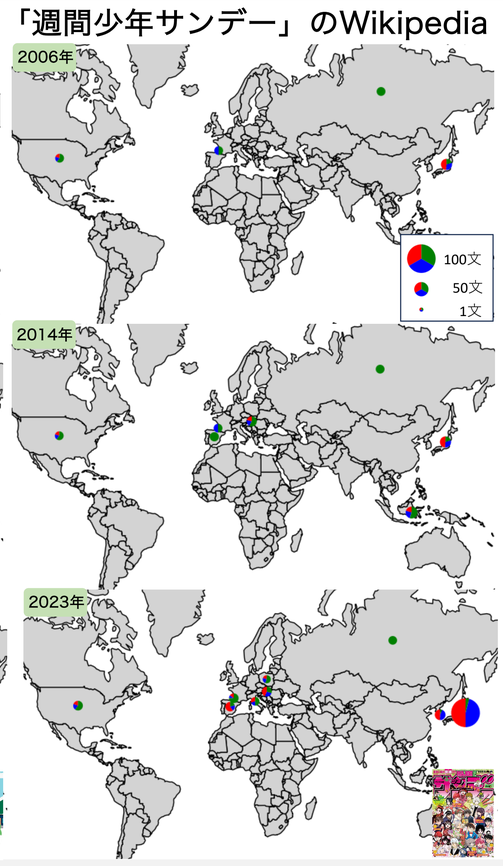
ii 新クラスの追加



「South Park」のWikipedia

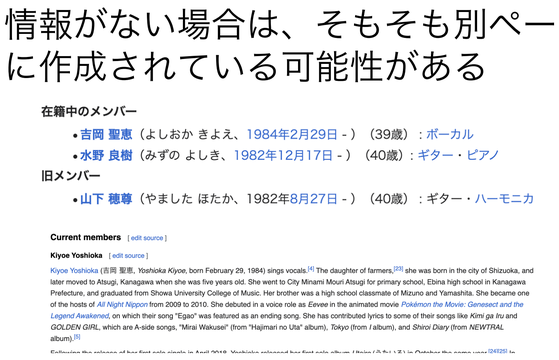


「週間少年サンデー」のWikipedia



課題/質問

- 情報がない場合は、そもそも別ページに作成されている可能性がある



在籍中のメンバー
- 吉岡 聖恵（よしおか きよえ、1984年2月29日 - ）（39歳）：ボーカル
- 水野 良樹（みずの よしき、1982年12月17日 - ）（40歳）：ギター・ピアノ

旧メンバー
- 山下 穂尊（やました ほたか、1982年8月27日 - ）（40歳）：ギター・ハーモニカ

Current members [edit source]

Kiyoe Yoshioka [edit source]

Kiyoe Yoshioka (吉岡 聖恵, *Yoshioka Kiyoe*, born February 29, 1984) sings vocals.[4] The daughter of farmers,[22] she was born in the city of Shizuoka, and later moved to Atsugi, Kanagawa when she was five years old. She went to City Minami Mouri Atsugi for primary school, Ebina high school in Kanagawa Prefecture, and graduated from Showa University College of Music. Her brother was a high school classmate of Mizuno and Yamashita. She became one of the hosts of *All Night Nippon* from 2009 to 2010. She debuted in a voice role as *Eevee* in the animated movie *Pokémon the Movie: Genesect and the Legend Awakened*, on which their song "Egao" was featured as an ending song. She has contributed lyrics to some of their songs like *Kimi ga Iru* and *GOLDEN GIRL*, which are A-side songs, "Mirai Wakuseil" (from "Hajimari no Uta" album), *Tokyo* (from *I* album), and *Shiroi Diary* (from *NEWTRAL* album).[5]

- 他に良い手法が無いか？
  - Entity Extractionは検討したが、表記揺れの問題で多言語のEntityを一致できない

  いきものがかり = Ikomonogakari
  = Ikimono-Gakari = 生き物係？
- **評価をどうするか？どのようにして論文として落とし込むか？**
  - Information Gapの解決という切り口はあるが…