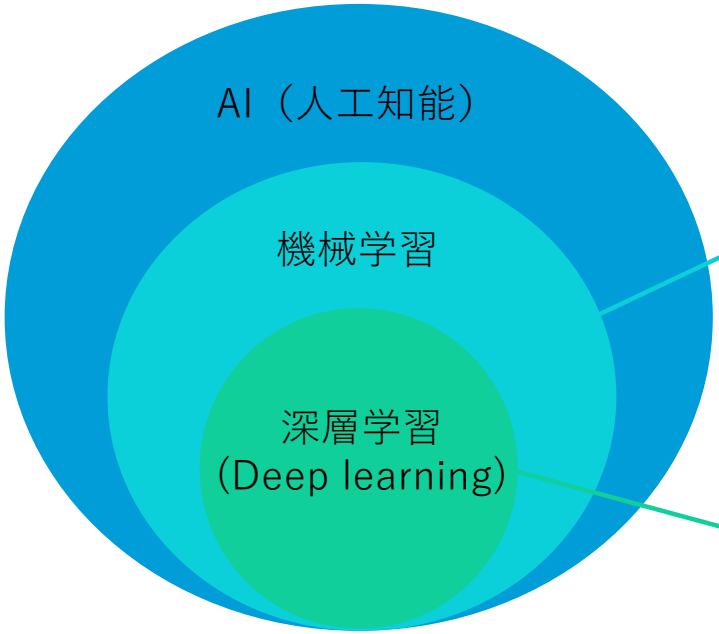
AIで読む『小説家になろう』：Transformerによる物語ジャンルの可視化と分類

大阪大学人文学研究科

言語文化専攻助教

黄 晨雯

Transformerの位置付け



線形回帰
ロジックスティック回帰
サポートベクトルマシン
階層クラスタリング
トピックモデル等

ニューラルネットワーク

RNN、LSTM、CNN
Transformer
（2017年に論文『Attention is All You Need』
にて発表されたアーキテクチャ）

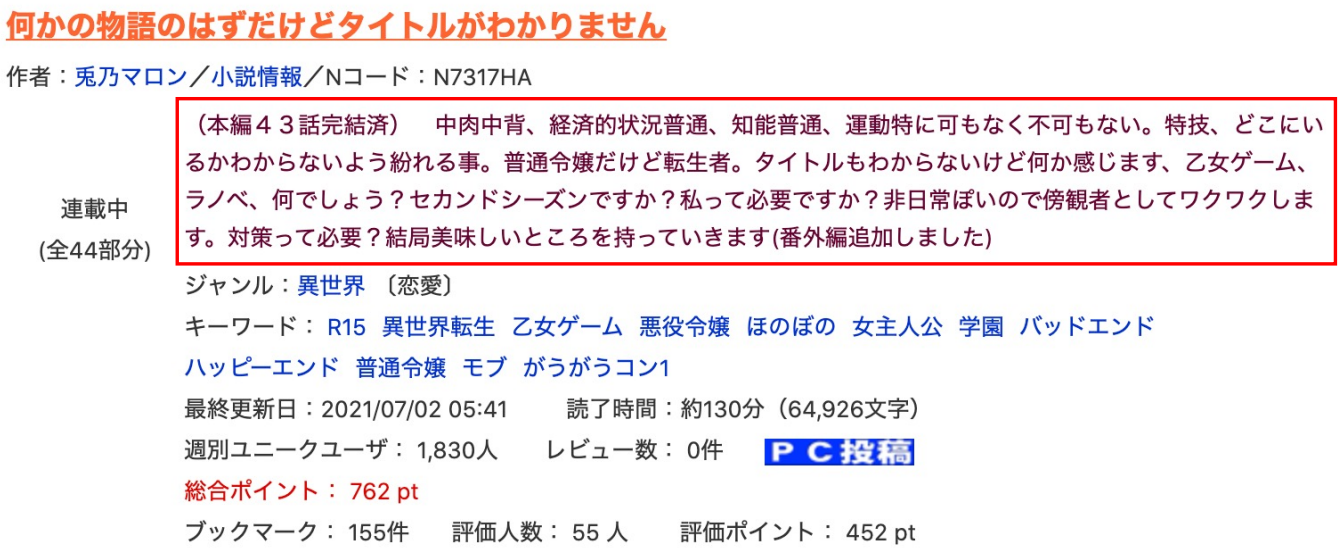
Transformerができること

- テキスト生成
- テキスト分類
- 固有表現抽出
- 文章要約
- 質問応答
- 音声処理
- その他

オンライン小説のあらすじを入力文とし、モデルを微調整する。
→ 文体や内容の傾向に基づいたジャンル分類の可能性を探る。

分析対象（メタ情報）

- 『小説家になろう』：https://yomou.syosetu.com/search.php



何かの物語のはずだけどタイトルがわかりません

作者：兎乃マロン／小説情報／Nコード：N7317HA

連載中
(全44部分)

（本編４３話完結済）　中肉中背、経済的状況普通、知能普通、運動特に可もなく不可もない。特技、どこにいるかわからないよう紛れる事。普通令嬢だけど転生者。タイトルもわからないけど何か感じます、乙女ゲーム、ラノベ、何でしょう？セカンドシーズンですか？私って必要ですか？非日常ぽいので傍観者としてワクワクします。対策って必要？結局美味しいところを持っていきます(番外編追加しました)

ジャンル：異世界 〔恋愛〕

キーワード： R15 異世界転生 乙女ゲーム 悪役令嬢 ほのぼの 女主人公 学園 バッドエンド ハッピーエンド 普通令嬢 モブ がうがうコン1

最終更新日：2021/07/02 05:41　　読了時間：約130分（64,926文字）

週別ユニークユーザ： 1,830人　　レビュー数： 0件　　ＰＣ投稿

総合ポイント： 762 pt

ブックマーク： 155件　　評価人数： 55 人　　評価ポイント： 452 pt

分析対象（メタ情報）

- なろうAPI（デベロッパー向け）

https://dev.syosetu.com/man/api/

Pythonスクリプトでスクレイピング

```
title: デビルエンジェルAYA
ncode: N7449HE
userid: 738795
writer: 八上 みつか
story: |
  関東某県の山中にある新興都市。
  そこは企業が『ＪＯＹ使い』という能力者を秘密裏に開発するための施設だった。
  昼間は普通の学園都市であるが、夜は制限を解かれた能力者たちが争い合う危険な街へと変貌する。
  何も知らずに高校からこの街で一人暮らしを始めた主人公、星野空人。
  彼は同じ新入生でありながら最強クラスのＪＯＹ能力を持ちヒーローに憧れる少女、赤坂綺と出会い一目惚れをする。
  それぞれ特徴のある三つの学園、夜間に勢力争いをする少年能力者グループ、そして街を支配する企業の闇。
  さまざまな陰謀渦巻く街で、空人は時に争いに巻き込まれながらも、綺に並び立つ男になるため日夜奮闘していく。
biggenre: 2
genre: 202
gensaku:
keyword: >
  R15 残酷な描写あり 青春
  異能力バトル 学園 現代
  男主人公 群像劇 超能力
general_firstup: 2021-09-07 00:00:00
general_lastup: 2021-09-22 02:00:00
novel_type: 1
end: 1
general_all_no: 138
length: 795817
time: 1592
isstop: 0
isr15: 1
isbl: 0
isgl: 0
iszankoku: 1
istensei: 0
istenni: 0
pc_or_k: 2
global_point: 30
daily_point: 12
weekly_point: 4
monthly_point: 18
quarter_point: 18
yearly_point: 18
fav_novel_cnt: 5
impression_cnt: 1
review_cnt: 0
all_point: 20
all_hyoka_cnt: 2
sasie_cnt: 0
kaiwaritu: 29
novelupdated_at: 2021-09-22 02:00:01
updated_at: 2021-09-22 02:02:04
```

分析対象（メタ情報）

- メタデータコーパス構築

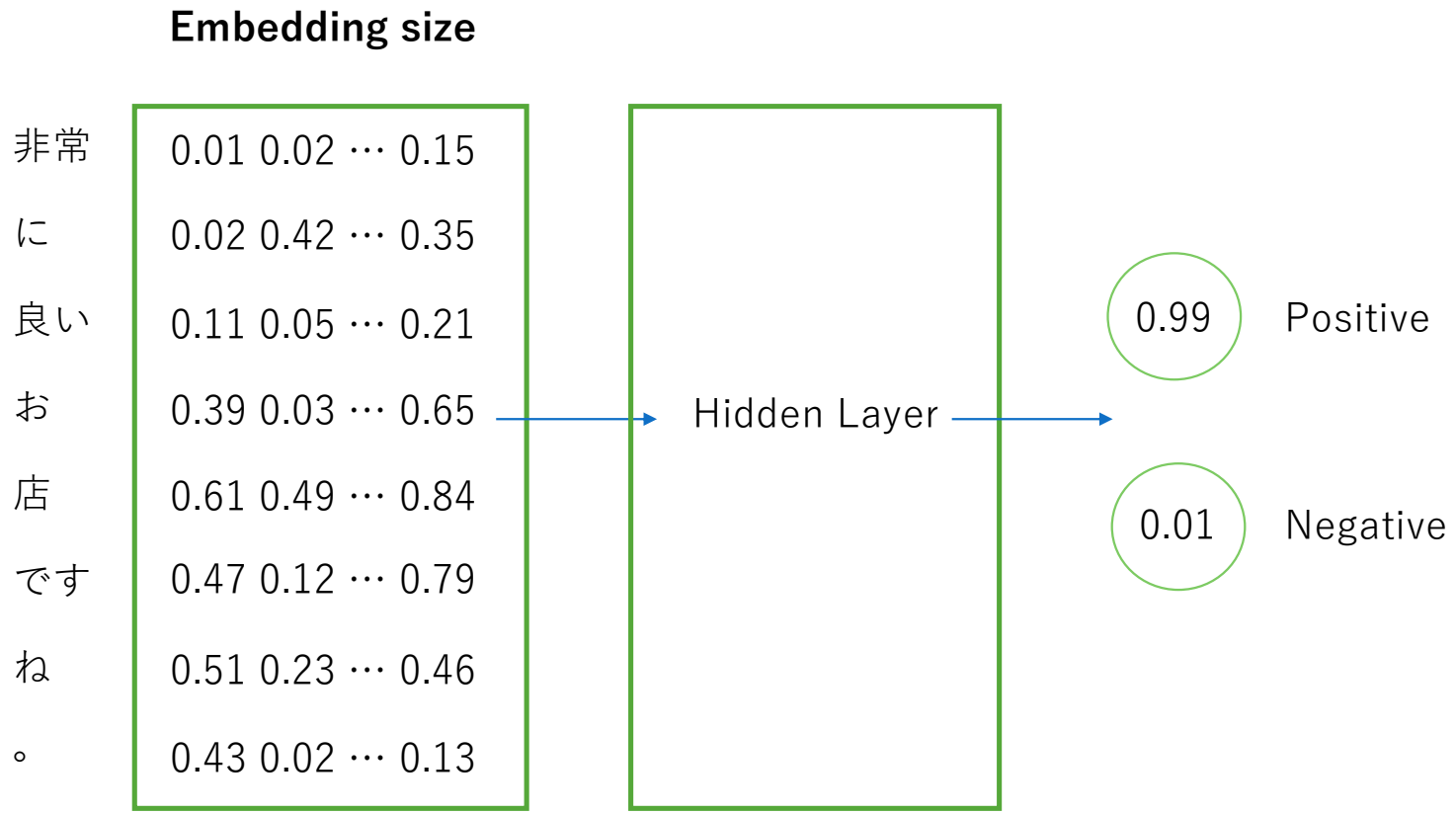
	title	writer	story	Group	Genre	Year	global_point
0	【WEB版】うちの弟子がいつのまにか人類最強になっていて、なんの才能もない師匠の俺が、それを...	アキライズン	「貴方様が剣聖アリス様のお師匠様、大剣聖タクミ様で御座いますね」「え、いや、人違いじゃないか...	文芸	コメディー〔文芸〕	2019	29464
1	水属性の魔法使い	久宝　忠	【好きラノ2021年上期　新作部門第2位！】書籍版第二巻「水属性の魔法使い　第一部　中央諸国...	ファンタジー	ハイファンタジー〔ファンタジー〕	2020	256230
2	【三章開始！】創成魔法の再現者　～『魔法が使えない』と実家を追放された天才少年、魔女の弟子と...	みわもひ	「貴様は出来損ないだ、二度と我が家の敷居を跨ぐなぁ！」魔法が全ての国、とりわけ貴族だけが生ま...	ファンタジー	ハイファンタジー〔ファンタジー〕	2021	162034
3	【アイテム無消費】だけが売りの俺。ダンジョン最下層へと追放される～だがそこで手に入れたアイテ...	まんじ	「くそっ！くそっ！くそくそくそ！」地面に跪き、怨嗟の声を上げる若者の名は――セドリ。【アイテ...	ファンタジー	ハイファンタジー〔ファンタジー〕	2021	39846
4	【9月書籍・コミカライズ発売！】植物魔法チートでのんびり領主生活始めます～前世の知識を駆使し...	りょうとかえ	9月書籍第3巻、コミカライズ第2巻発売！ページ下部のバナーから公式サイトへと移動できます。貴...	ファンタジー	ハイファンタジー〔ファンタジー〕	2019	169186

モデルの訓練（教師あり）

- **教師あり学習**

入力文：

非常に良いお店ですね。



モデルの訓練（教師あり）

- 正解ラベル：Genre列のジャンル

- 訓練データ：あらすじ（80%）
- テストデータ：あらすじ（20%）

- 事前学習済モデル：

  `tohoku-nlp/bert-base-japanese-v3`

ジャンル	ラベル
コメディー〔文芸〕	0
ハイファンタジー〔ファンタジー〕	1
ローファンタジー〔ファンタジー〕	2
現実世界〔恋愛〕	3
アクション〔文芸〕	4
異世界〔恋愛〕	5
VRゲーム〔SF〕	6
空想科学〔SF〕	7
ヒューマンドラマ〔文芸〕	8
歴史〔文芸〕	9
宇宙〔SF〕	10
パニック〔SF〕	11
純文学〔文芸〕	12
ノンジャンル〔ノンジャンル〕	13
推理〔文芸〕	14
ホラー〔文芸〕	15

モデルの評価

- テストデータを用いたモデルの評価結果

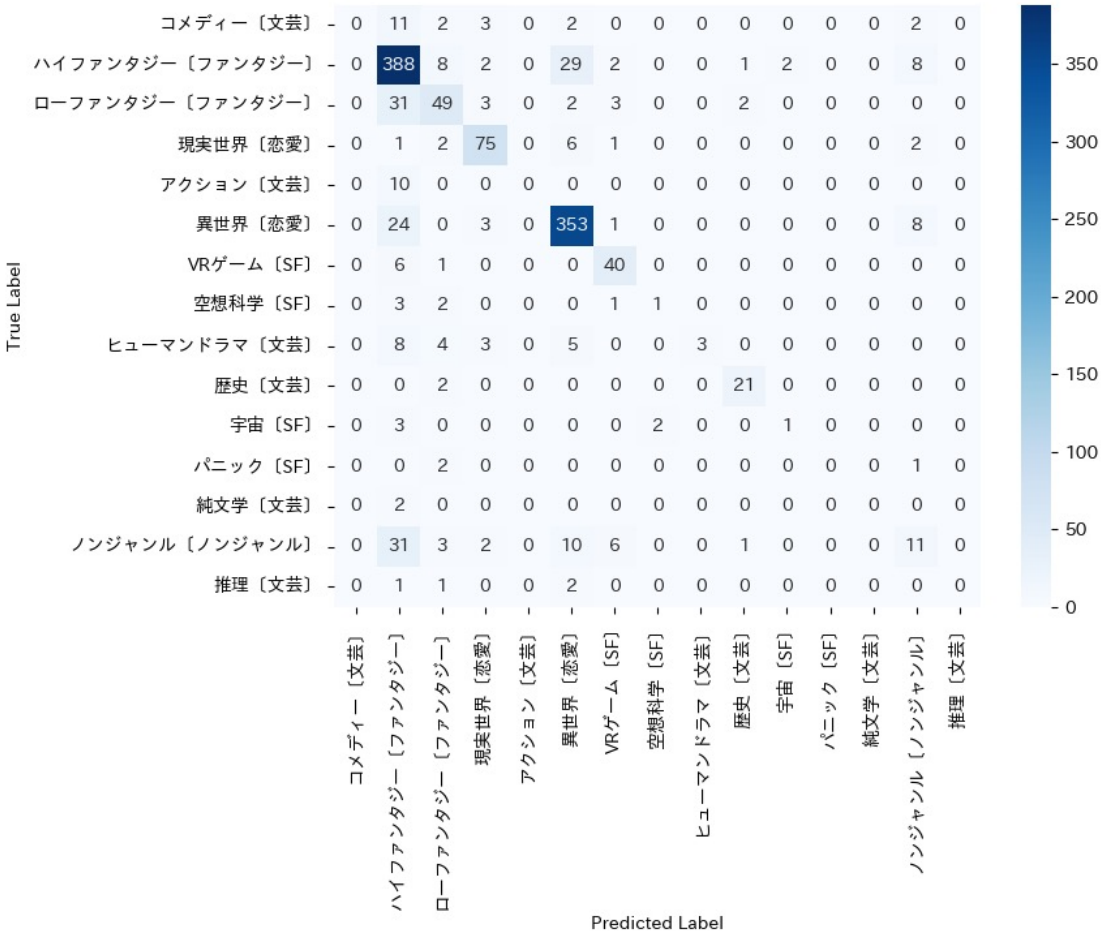
	指標	スコア
評価損失（誤差）	eval_loss	0.87029701
評価用モデル準備時間	eval_model_preparation_time	0.008
正解率	eval_accuracy	0.77530864
精度と再現率の調和平均	eval_f1	0.74625348
評価にかかった時間	eval_runtime	41.9998
評価サンプル数/秒	eval_samples_per_second	28.929
評価ステップ数/秒	eval_steps_per_second	1.81

混同行列の可視化

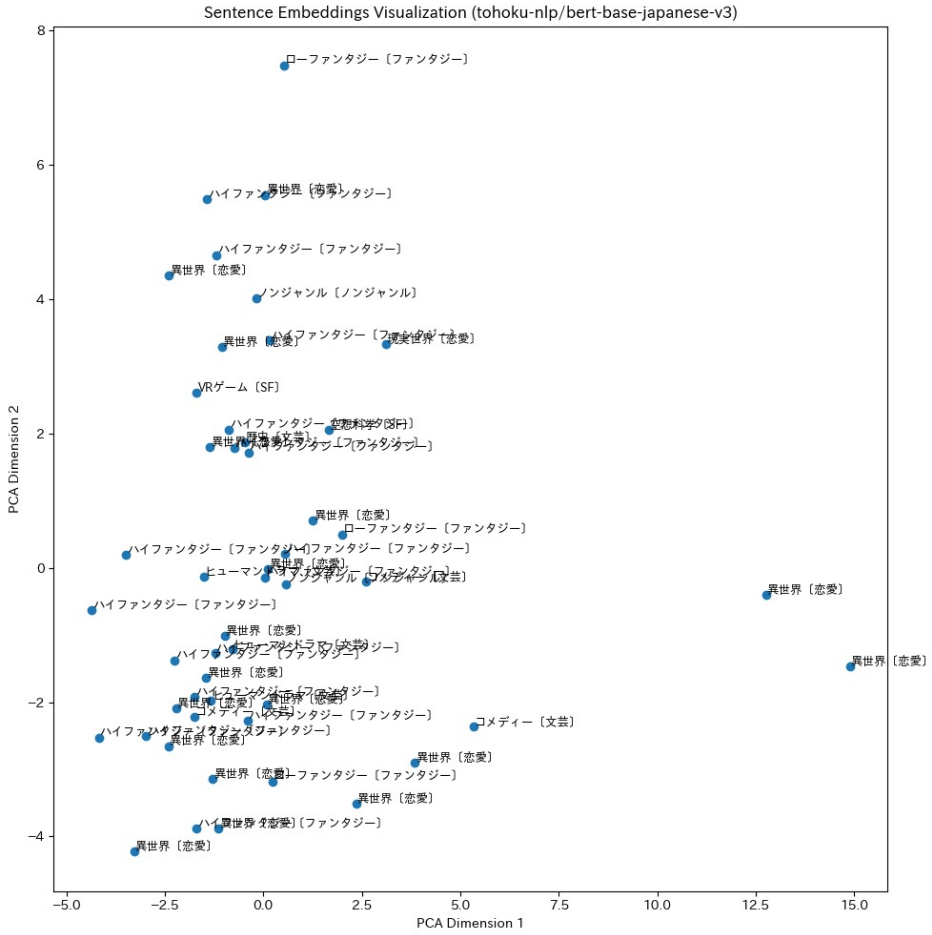
- 分類問題における予測結果の評価を行うために使用される行列

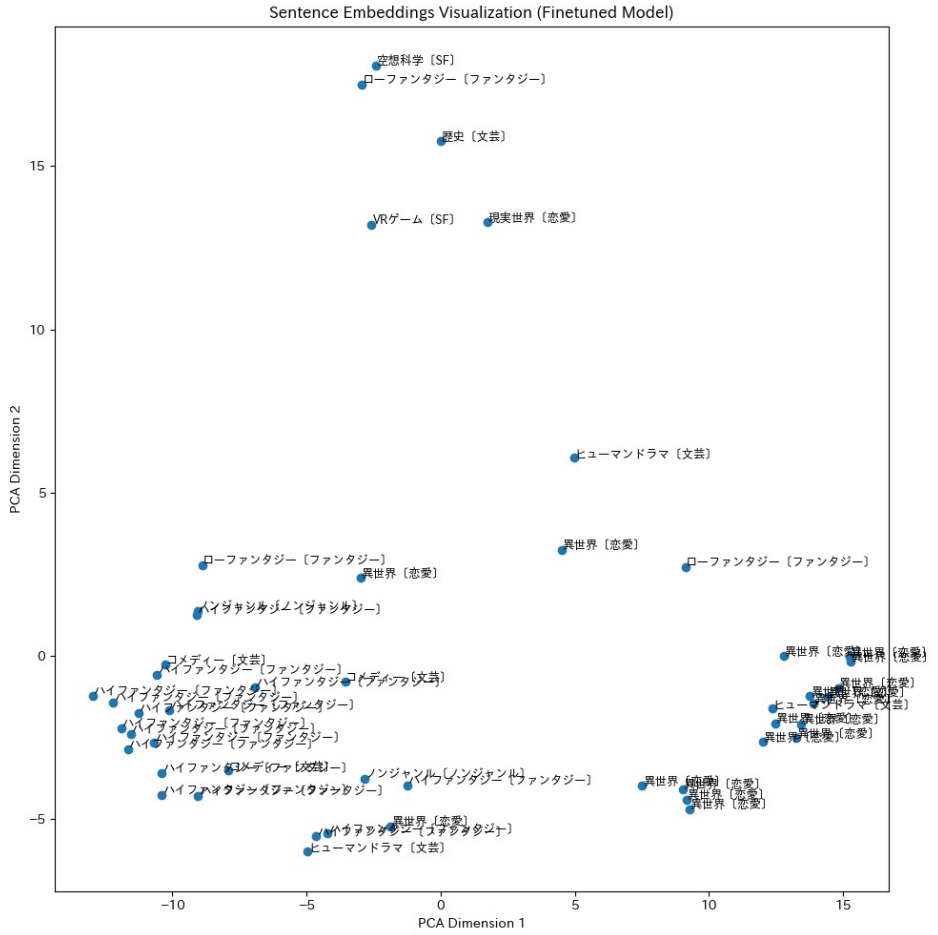
True Label： 正解ラベル

Predicted Label：予測ラベル

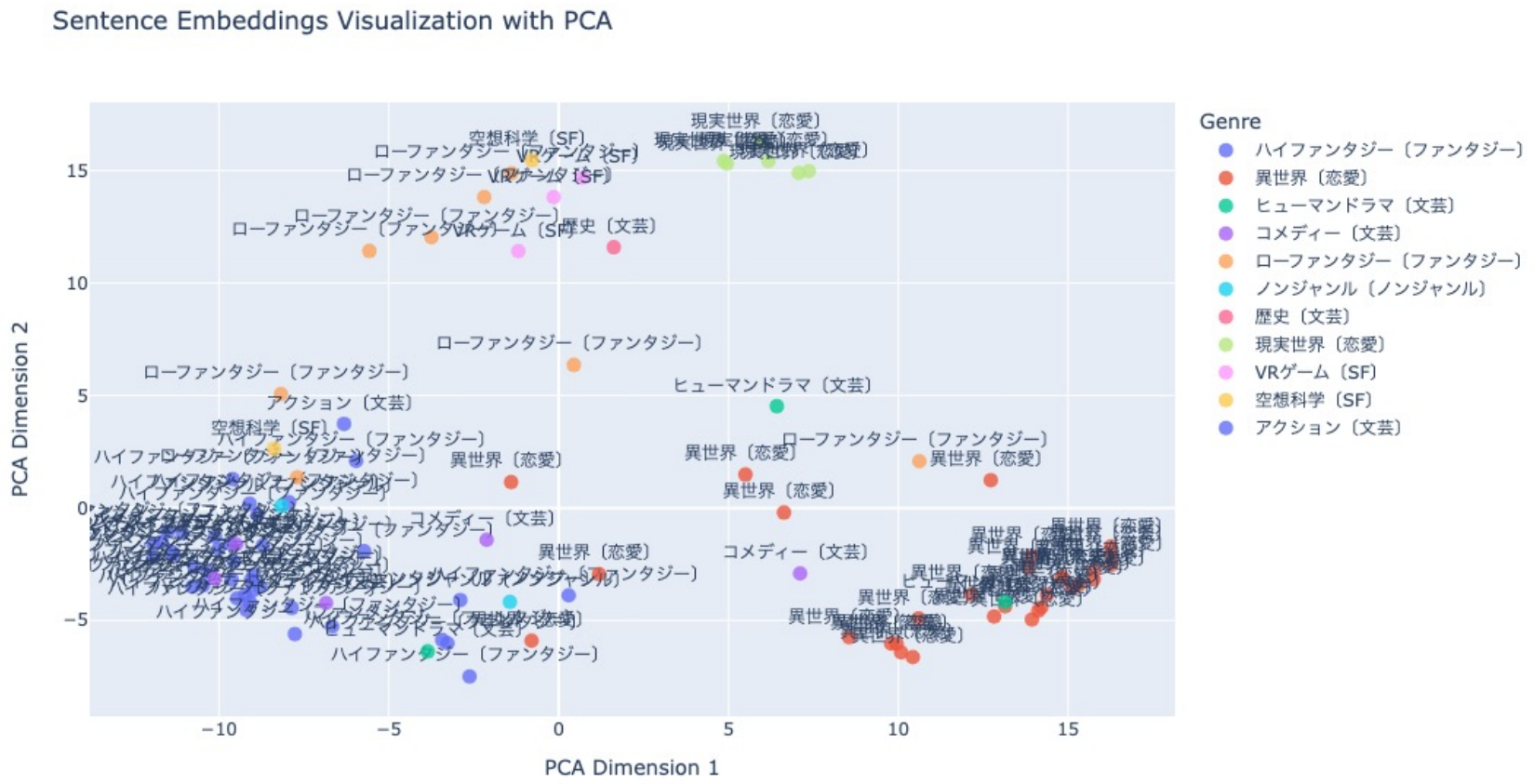


あらすじ間の関係性可視化

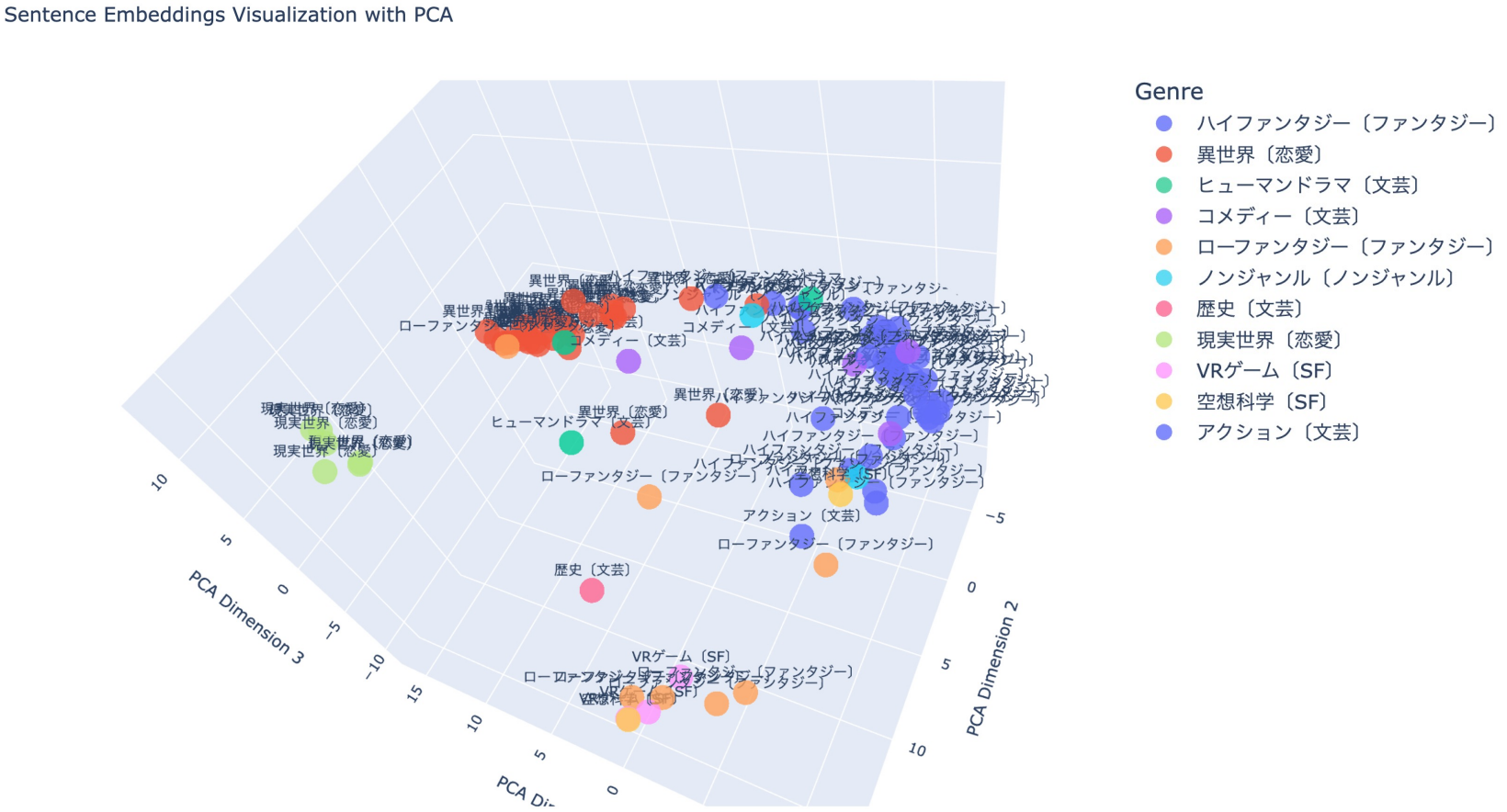




あらすじ間の関係性可視化



あらすじ間の関係性可視化



ご清聴ありがとうございました。