BERTによる文の類似度分析

BERTによる文の類似度分析プログラム

目的：複数の文章を入力すると、**どの文とどの文が意味的に似ているか**を自動的に分析・可視化する

できること

- 文の意味的な**近さ**を**数値化**する
- **似ている文同士を自動的にグループ分け**する
- 結果を3種類の図で視覚的に確認できる

使用例

アンケートの自由記述回答の分類、文章の主題の分析、似た意味の文を検索

プログラムが使う4つの技術

BERT（ベルト）

- 文章の「意味」を数値（ベクトル）に変換する
- 例：「手を洗う」→ [0.23, -0.45, 0.67, ...]（768個の数値）

コサイン類似度

- 2つの文がどれくらい似ているかを0〜1で表す
- 1に近い = 意味が似ている、0に近い = 意味が異なる

K-meansクラスタリング

- 似たものを自動的にグループ分けする手法

シルエット係数

- グループ分けの品質を0〜1で評価する指標
- 1に近い = 良いグループ分け、0に近い = 不明瞭なグループ分け

```
=== 入力文章（正規化後） ===
文1: 手を洗って清潔にする
文2: 手にけがを負った
文3: 手が冷たくなってきた
文4: 手のひらに汗をかく
文5: 手袋をはめる
文6: 手首を痛めた
文7: 問題解決の手を考える
文8: 良い手が見つからない
文9: 次の手を打つ
文10: 有効な手段を講じる
文11: 打開策の手がない
文12: 対処する手を探す
文13: 新しい手法を試す
文14: 別の手を使う
文15: 右手を動かす
文16: 左手を上げる
文17: 手を振って挨拶する
文18: 手の指を動かす
文19: 手を伸ばして取る
文20: 手をたたく
```

サンプルテキスト

```
=== 入力方法の選択 ===
0: サンプルテキストを使用
1: テキストファイルをアップロード
3: 終了
選択してください (0, 1, 2, or 3): 0
```

メニュー

=== 文間のコサイン類似度 ===

	文1	文2	文3	文4	文5	文6	文7	文8	文9	文10	文11	文12	文13	文14	文15	文16	文17	文18	文19	文20
文1:	1.000	0.897	0.921	0.949	0.939	0.910	0.915	0.911	0.880	0.909	0.898	0.901	0.910	0.895	0.889	0.899	0.938	0.838	0.933	0.936
文2:	0.897	1.000	0.916	0.920	0.915	0.968	0.889	0.902	0.867	0.884	0.900	0.892	0.898	0.888	0.859	0.897	0.914	0.811	0.906	0.926
文3:	0.921	0.916	1.000	0.931	0.927	0.927	0.927	0.943	0.872	0.900	0.938	0.907	0.936	0.893	0.862	0.891	0.914	0.815	0.918	0.943
文4:	0.949	0.920	0.931	1.000	0.939	0.921	0.912	0.908	0.899	0.911	0.904	0.901	0.916	0.921	0.913	0.927	0.946	0.869	0.947	0.963
文5:	0.939	0.915	0.927	0.939	1.000	0.934	0.933	0.936	0.902	0.930	0.920	0.916	0.929	0.925	0.884	0.921	0.938	0.832	0.948	0.960
文6:	0.910	0.968	0.927	0.921	0.934	1.000	0.905	0.926	0.866	0.893	0.912	0.893	0.913	0.895	0.866	0.903	0.908	0.805	0.911	0.930
文7:	0.915	0.889	0.927	0.912	0.933	0.905	1.000	0.949	0.903	0.928	0.939	0.950	0.956	0.917	0.846	0.876	0.915	0.805	0.919	0.934
文8:	0.911	0.902	0.943	0.908	0.936	0.926	0.949	1.000	0.893	0.935	0.974	0.938	0.949	0.916	0.827	0.871	0.909	0.768	0.916	0.930
文9:	0.880	0.867	0.872	0.899	0.902	0.866	0.903	0.893	1.000	0.914	0.910	0.913	0.897	0.946	0.873	0.895	0.904	0.848	0.933	0.918
文10:	0.909	0.884	0.900	0.911	0.930	0.893	0.928	0.935	0.914	1.000	0.944	0.948	0.940	0.925	0.841	0.870	0.913	0.785	0.920	0.926
文11:	0.898	0.900	0.938	0.904	0.920	0.912	0.939	0.974	0.910	0.944	1.000	0.939	0.938	0.925	0.839	0.877	0.907	0.784	0.912	0.926
文12:	0.901	0.892	0.907	0.901	0.916	0.893	0.950	0.938	0.913	0.948	0.939	1.000	0.935	0.917	0.839	0.862	0.907	0.805	0.912	0.912
文13:	0.910	0.898	0.936	0.916	0.929	0.913	0.956	0.949	0.897	0.940	0.938	0.935	1.000	0.915	0.831	0.871	0.915	0.777	0.915	0.942
文14:	0.895	0.888	0.893	0.921	0.925	0.895	0.917	0.916	0.946	0.925	0.925	0.917	0.915	1.000	0.892	0.902	0.908	0.857	0.940	0.934
文15:	0.889	0.859	0.862	0.913	0.884	0.866	0.846	0.827	0.873	0.841	0.839	0.839	0.831	0.892	1.000	0.952	0.904	0.969	0.923	0.916
文16:	0.899	0.897	0.891	0.927	0.921	0.903	0.876	0.871	0.895	0.870	0.877	0.862	0.871	0.902	0.952	1.000	0.936	0.908	0.942	0.941
文17:	0.938	0.914	0.914	0.946	0.938	0.908	0.915	0.909	0.904	0.913	0.907	0.907	0.915	0.908	0.904	0.936	1.000	0.864	0.953	0.953
文18:	0.838	0.811	0.815	0.869	0.832	0.805	0.805	0.768	0.848	0.785	0.784	0.805	0.777	0.857	0.969	0.908	0.864	1.000	0.896	0.871
文19:	0.933	0.906	0.918	0.947	0.948	0.911	0.919	0.916	0.933	0.920	0.912	0.912	0.915	0.940	0.923	0.942	0.953	0.896	1.000	0.968
文20:	0.936	0.926	0.943	0.963	0.960	0.930	0.934	0.930	0.918	0.926	0.926	0.912	0.942	0.934	0.916	0.941	0.953	0.871	0.968	1.000

文同士のコサイン類似度

2つの文がどれくらい似ているかを0〜1で表す

文ペアの類似度（例：文1と文2 = 0.856）

- 0.9以上：非常に似ている
- 0.7〜0.9：やや似ている
- 0.7未満：あまり似ていない

文番号順ヒートマップ

- **縦軸と横軸に文番号**、交点の色が類似度
- **赤い＝似ている**（類似度が高い）
- **青い＝似ていない**（類似度が低い）
- 対角線は常に赤（自分自身との比較）

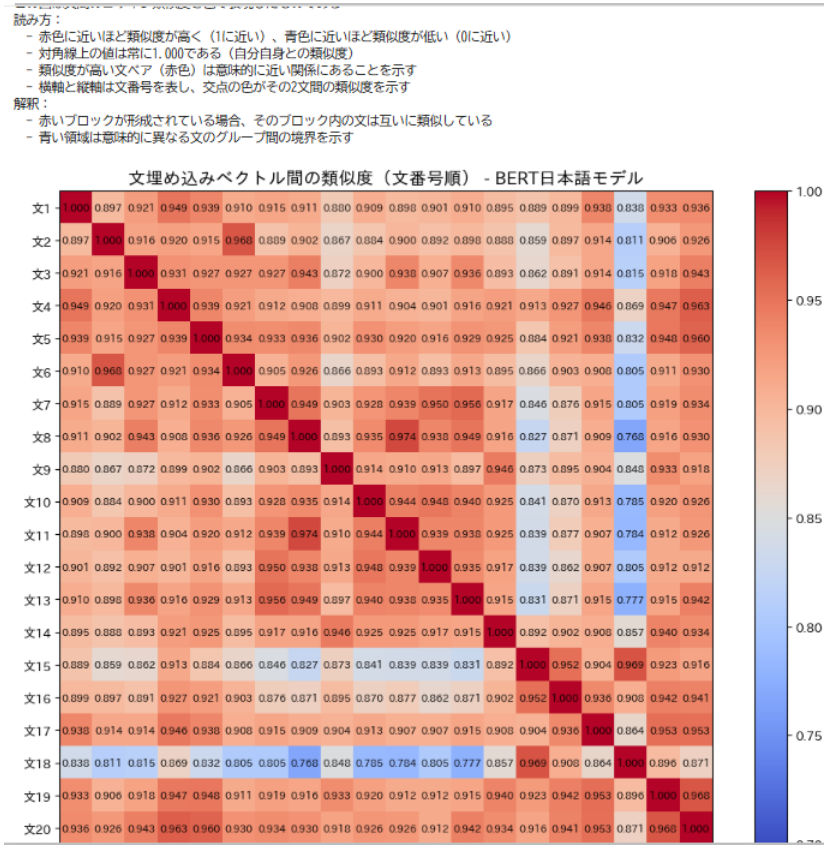


図1：文番号順ヒートマップ

2次元配置図

- **文**を**平面上の点**として配置
- **距離が近い＝意味が似ている**
- 色＝自動的に分類されたグループ
- **同じ色の点は同じグループに属する**

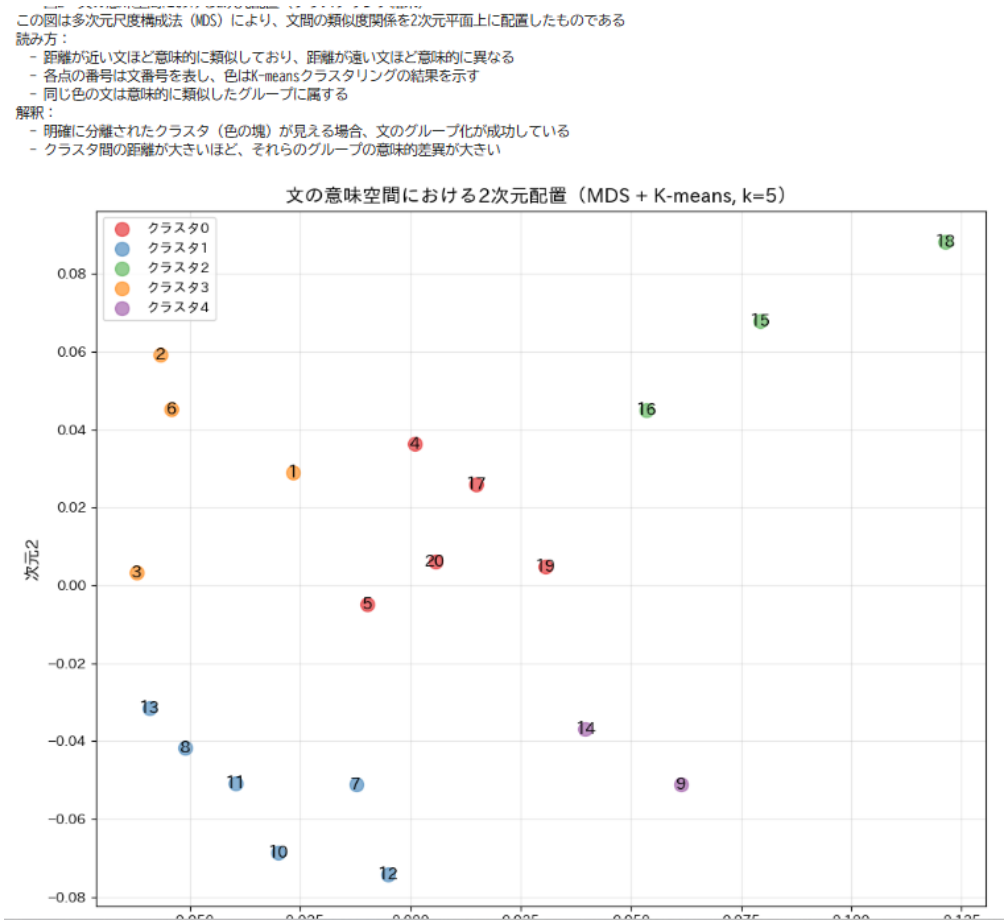
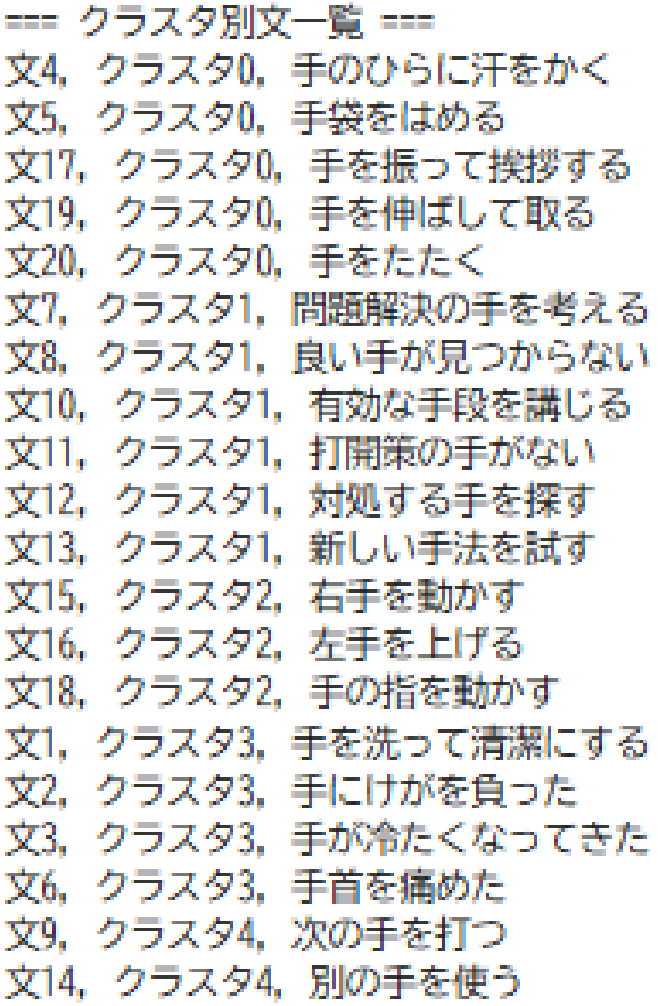


図2：2次元配置図

クラスタ別文一覧

- **どの文**が**どのグループに分類されたか**の一覧
- クラスタ番号順に表示される



クラスタ別文一覧

表示される情報

- 平均類似度：**全体的な文の類似性**
- 標準偏差：**類似度のばらつき**
- 最小/最大類似度：**最も似ていない/似ているペア**
- 最適クラスタ数：**自動決定されたグループ数（2〜10）**
- シルエット係数：**グループ分けの品質（0.5以上が良好）**