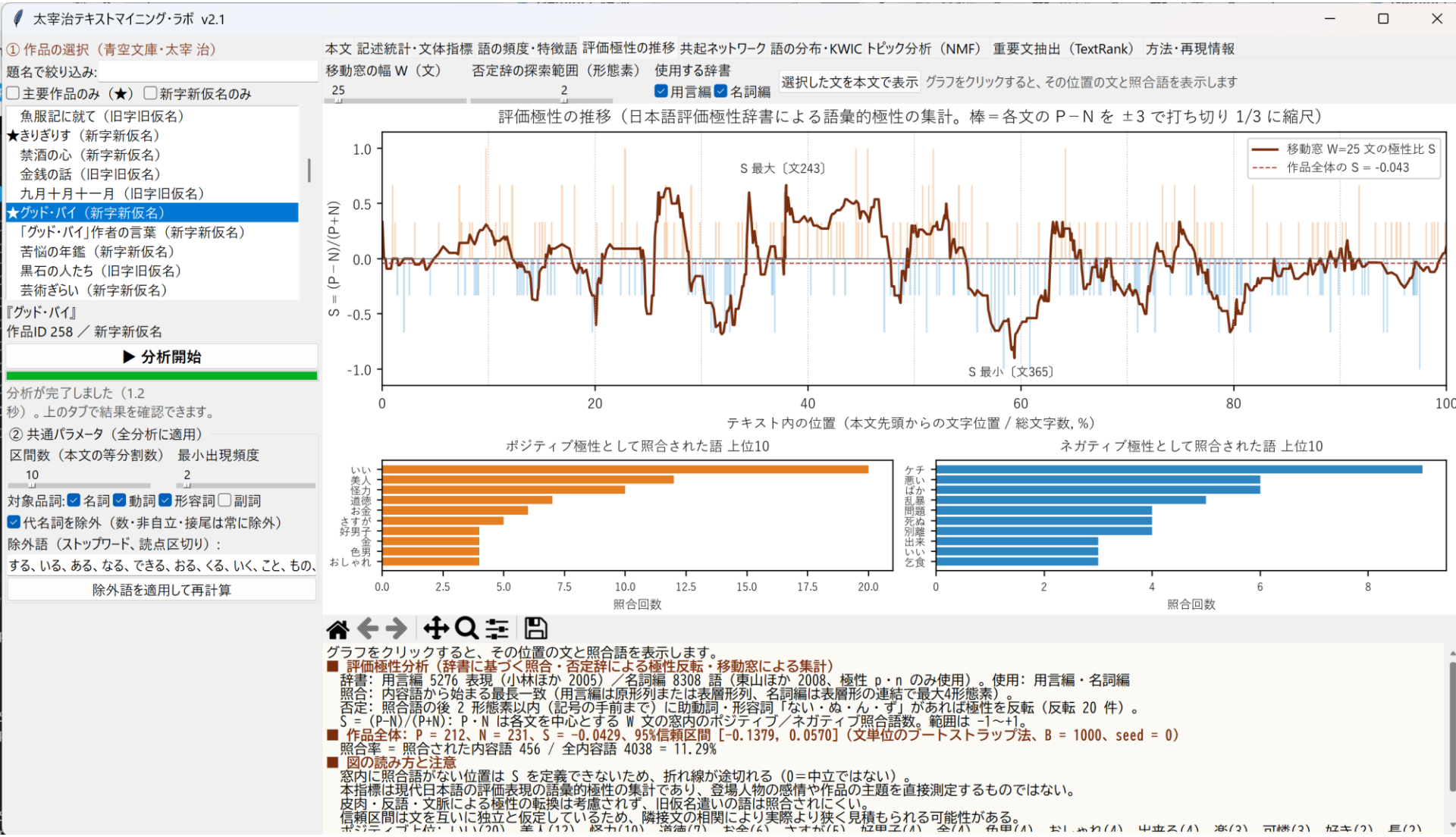
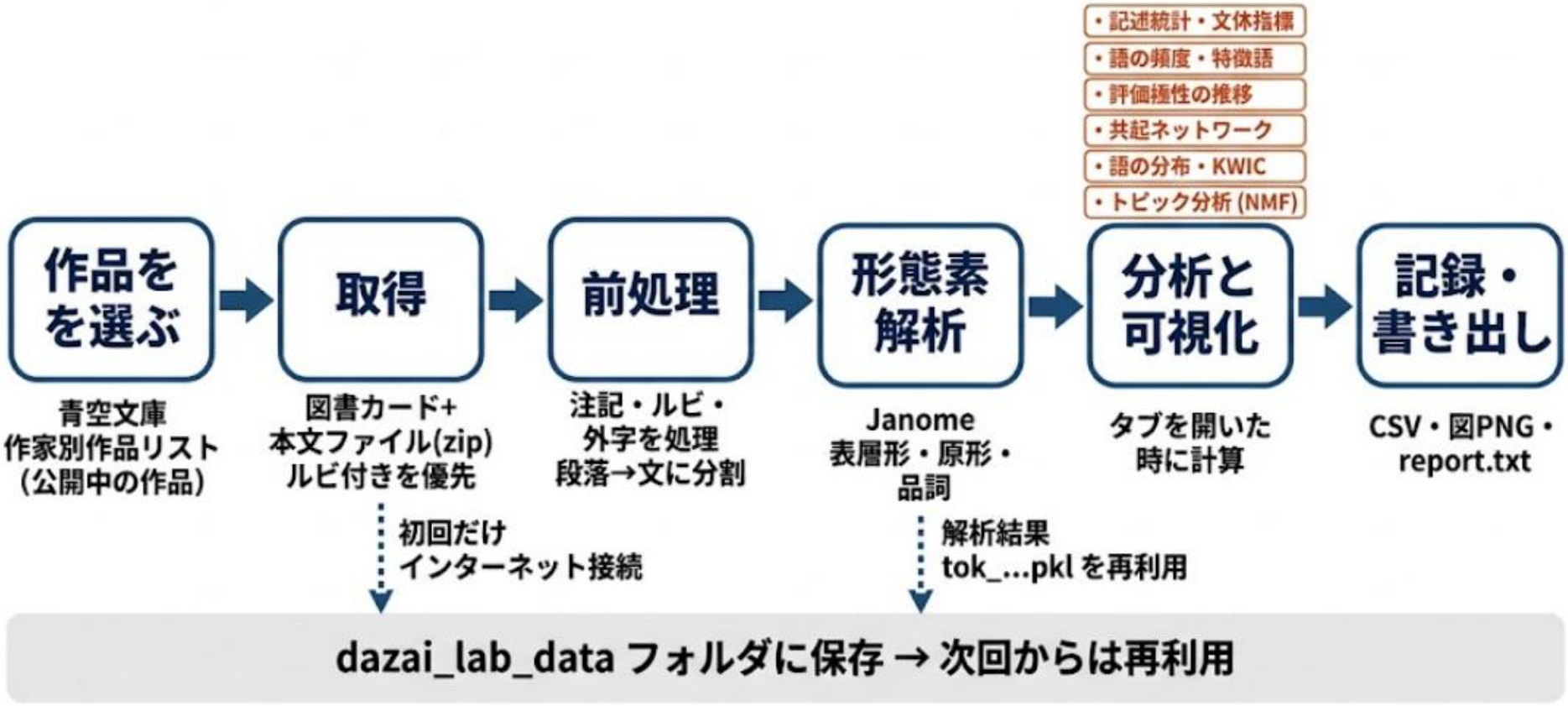
太宰治テキストマイニング



青空文庫の太宰治作品を1編選ぶと、取得から分析・書き出しまでを1つの画面で行う



どの分析にも、図の下に計算の根拠
（式・数値・注意点）が表示される

青空文庫の書式を取り除いて本文だけにし、形態素解析の後、条件に合う語だけを数える

前処理

富士《ふじ》	➡ 富士	ルビを除去 (語と読みは一覧に残す)
[#…]	➡ (削除)	入力者注を除去
※[#…U+番号…]	➡ 元の文字、	外字を復元 復元できない場合 → ≡
いろ／＼	➡ いろ〳〵	くの字点を置換

改行で段落に分割 ➡ 。!? の位置で文に分割 (直後の閉じ括弧を含む)

形態素解析

私は富士を見ていた。

私	は	富士	を	見	て	い	た	。

原形：見る　原形：いる

語は原形で数える
(「いた」は「いる」として数える)

語のふるい

対象品詞か？
(初期値：名詞・動詞・形容詞)
- 私 ×代名詞
- いる ×除外語

↓

代名詞・除外語でないか？
(数・非自立・接尾は常に除外)
- 私 ×代名詞
- いる ×除外語

↓

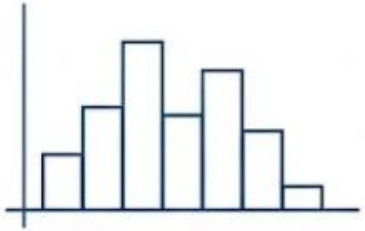
最小出現頻度以上か？
(初期値 2 回)
- 富士 → ○
- 見る → ○

品詞・除外語の設定を使う分析：頻度・特徴語／共起ネットワーク／トピック分析／重要文抽出

使わない分析：語彙指標 (TTR・MATTR・Yule の K) ／評価極性の照合／KWIC 検索

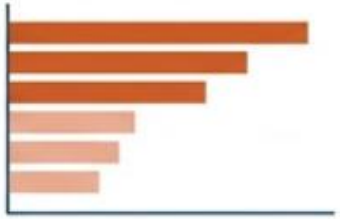
１つの作品を７つの角度から数値化する（横軸に本文の位置をとる図には、区間の境界を点線で示す）

記述統計・文体指標



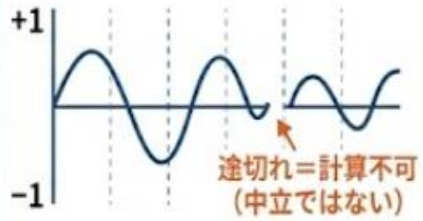
文の長さ・字種・品詞・文末・一人称
語彙の多様性：MATTR、
反復の多さ：Yule の K

語の頻度・特徴語



出現回数の多い語
ある区間に特に多い語を G^2 で順位づけ(濃い色＝補正後に有意)

評価極性の推移



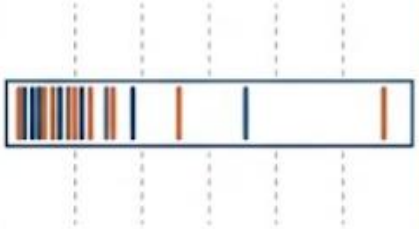
S＝(P−N)/(P+N)
P＝肯定的な語の数、N＝否定的な語の数（移動窓ごと）

共起ネットワーク



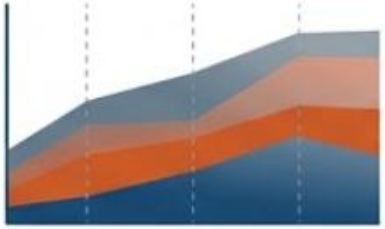
円の大きさ＝その語を含む文の数
線の太さ＝ジャカード係数、
色＝語のまとまり

語の分布・KWIC



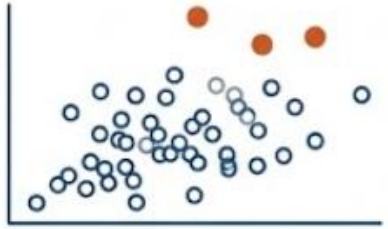
どこに現れるか
DP：0＝均等、1＝一部に集中

トピック分析 (NMF)



本文をＤ個の部分に分け、k個の話題の構成比の推移を示す

重要文抽出 (TextRank)



他の文と語を多く共有する文ほど高得点

読むときの注意

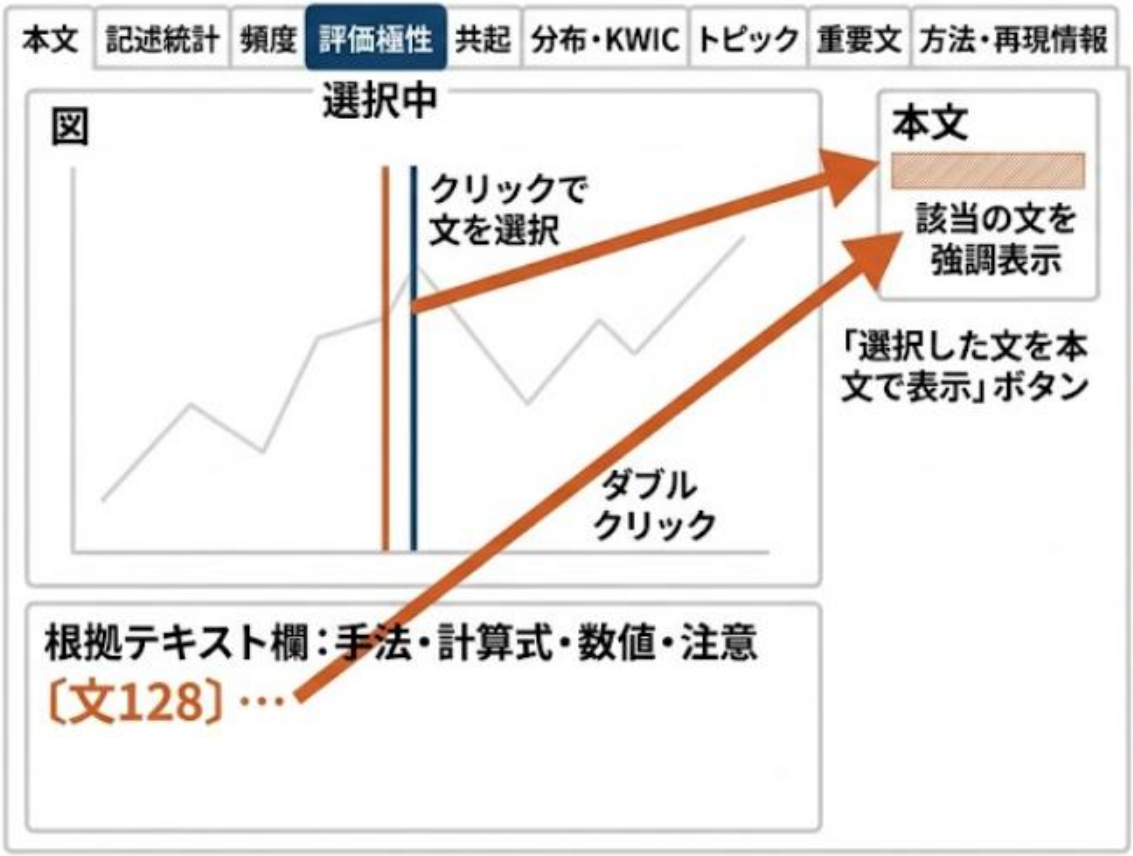
～！～

数値は手がかりであり、感情・主題・文学的な重要性を直接測るものではない
結果は本文の文脈で確かめる

数値の根拠を図の下に示し、気になった文はすぐ本文で確かめ、同じ条件で再現できる

①作品の選択

②共通パラメータ



乱数のシードを固定

共起の配置／NMFの初期値／
ブートストラップ法

記録（方法・再現情報タブ）

出典URL・取得日時／前処理の件数
ライブラリの版／全パラメータ
各分析の根拠／参考文献

書き出し

作品名_日付_時刻 フォルダ

CSV（Excelで文字化けしない）/
図PNG（150dpi）/ report.txt

同じ作品・同じパラメータ・
同じ版なら同じ結果