三国志で学ぶデータ分析

2019 年 12 月 7 日

概要

この記事は 2019/12/7 に開催された Japan.R の発表原稿である.

当初は 5 分間の LT の予定だったので記述統計の見方とかを話すつもりだったが, 20 分枠に変更したことに合わせてボリュームを増そうとしたらバランスが狂った感じになった.

注意: 今回の内容は日本で普及しているフォントで表示できない文字が含まれるなどの写植上の制約から, フォントを埋め込んだ pdf 版を公開している.

■キーワード　三國志, スクレイピング, 名寄せ処理, 自然言語処理 (?), 画像認識, ディープラーニング, 計量距離学習, 多変量解析

1　イントロダクション

1.1　三国志の背景

そこで今回取り上げる「三国志」について, 簡単に解説する.

魏から西晋の時代の歴史家である陳寿によって著された, 魏書・呉書・蜀書のいわゆる三国時代の歴史書を総称して三国志, 通称『正史三国志』と呼ばれる. これは正史, つまり当時の王朝によって正統な歴史書と認定された書物であるから, 必ずしも「真実」が描かれているとは限らない. 現在に残る正史三国志は, 南朝時代の裴松之の註解が付されているが, 王朝が変わった後世であることもあってより批判的である.

一方で, 『三国志演義』とは正史三国志や, それにまつわる無数の民間伝承や演劇「三国志平話」を羅貫中が編纂したものである. 本場である中国ではそれ以降も多くのバージョンが作られ, 主要な底本も複数存在する. 20 世紀になってからも『反三国志』(周, 1919) といったメタフィクション作品が作られている. 三国志演義の成立史だけでも膨大な研究が存在するはずだが, ここではそれに触れない.

渡辺 (2011) によれば, 三国志演義の「演義」とは, 義を演繹する, 義を敷衍するという意味であり, 当時の中国における倫理とされていた儒教に規定される道徳心を民衆に教えるという意図がある. よって, 当時の社会, 政権の意図が大きく反映されており, 道徳に悖る行動をした人物はみじめに破滅し, 道徳に則った行動を取るものは讃えられるという勧善懲悪の筋書きになっている[*1]. これは陳寿による史書, いわゆる「正史三国志」とはかなり異なる記述である.

[*1] 是とする道徳心すらも, 長い中国の歴史の中で変遷しており, 時代によって人物描写も変化している. しかし今回はそこに深く入ることはしない. 詳しい話は 渡辺 (2011) を参照.

渡辺 (2011) によれば,『日本書紀』の編纂の時点で三国志の影響が見られると言うから, 日本にはかなり早くに伝わっていた. しかし近年の日本では吉川英治の『三国志』(吉川, 1939) が有名ではないだろうか. これは三国志演義をもとに吉川が脚色したものであり[*2], 中国本国の三国志演義や正史三国志に忠実な翻訳作品ではない. 横山光輝の漫画『三国志』も, 概ね吉川英治の内容に準拠している.

また, 漫画作品では横山光輝作品の他, 李學仁・王欣太の『蒼天航路』も有名である. これまで悪役とされることが多かった**曹操**(ソウソウ)を主役としている[*3]など, 従来の三国志人物像に対するメタな作風が特徴である. その他にも日本の大衆文化における三国志をモチーフにした創作には枚挙に暇がない[*4].

一方で, 歴史書としての三国志, つまり『正史三国志』が日本で紹介されたのは比較的最近であり, 少なくとも民間向けでは 1977 年に筑摩書房によって魏書の一部の翻訳[*5]が出版され, 82, 89 年に続いて魏書の残りと蜀書[*6], 呉書[*7]がそれぞれ刊行されている[*8]. また, 三国志演義よりも正史に取材して書かれた作品としては, 陳舜臣の『秘本三国志』(陳, 1974)[*9]北方謙三の『三国志』北方 (1996), 宮城谷昌光の『三国志』(宮城谷, 2004) がある[*10].

このように, 史書でも創作でも, 書かれた時代や地域によって三国志の人物の扱われ方が異なる.

1.2 コーエーテクモのゲーム『三國志』シリーズ

コーエーテクモ (旧, 光栄) 社はこの三国志をモチーフにしたゲーム『三國志』シリーズを発売している. 1 作目は 1985 年で, 最新のものは 2016 年の『三國志 13』である. コーエーテクモは「歴史シミュレーションゲーム」と銘打っているが, 作品によっては, 中国大陸に割拠する勢力の 1 つを操作し天下統一を目標とするターン制戦略ゲームであったり, 登場人物の一人となって立身出世を目指すロールプレイング・ゲーム的要素の強いゲームだったりもする.

『三國志英傑伝』『三國志孔明伝』『三國志曹操伝』といったナンバーのないタイトルもある.

また, 8 以降の作品では, おまけ要素として三国志外の時代の人物, 例えば**管夷吾** (管仲) や**楽毅**, **藺相如**といった春秋戦国時代の英雄や, 時系列では後になる南北朝時代の**高長恭** (蘭陵武王), モンゴルの**チンギス＝ハン** (成吉思汗), 南宋の**岳飛**などが登録されている[*11]. 一方で最新作の三国志 13 (2016 年発売) では戦国時代末期の人物が増えており, これは原泰久の漫画『キングダム』の人気を反映していると思われる. さらに 2020 年発売予定の最新作 14 では, 田中芳樹原作『銀河英雄伝説』のキャラクタを登場させるようだ[*12].

1.3 問題提起

正史と演義での人物の評価両方を取り入れようとすると, どうしても矛盾が生じる. 例えば, 演義では曹操は徹底して「奸雄」つまり小狡い悪党として描かれ, 一方で劉備は利益より義を優先する道徳の手本のような人

[*2] この脚色もまた, 中国人の儒教精神を反映して改版してきたのと同様に, 日本人の価値観を反映したものだろう.
[*3] とはいえ, 曹操を悪役とする作劇は, 日本に老いては吉川版三国志の時点でかなり緩和されている気がする.
[*4] 『天地を喰らう』はもはやおっさんしか知るまい.
[*5] 今鷹真・井波律子訳 (1977)『三国志魏書』世界古典文学全集 24A, ISBN: 978-4-480-20324-3.
[*6] 今鷹真・小南一郎・井波律子訳 (1982)『三国志魏書・蜀書』世界古典文学全集 24B, ISBN: 978-4-480-20324-3.
[*7] 小南一郎訳 (1989)『三国志魏書・蜀書』世界古典文学全集 24C, ISBN: 978-4-480-20354-0
[*8] 現在は全 8 冊の文庫版『正史三国志』として流通している.
[*9] 全く関係ないが『インド三国志』も面白い
[*10] 宮城谷の三国志はほぼ史書をもとに記述を時系列順に編集し, 著者の人物評などを交えるという形式をとっている.
[*11] このへんの人選は田中芳樹の影響を受けている気がする.
[*12] 三國志 14:『銀河英雄伝説』コラボ情報

物として描かれる. しかし歴史はそう単純ではなく, 正史での記述は大きく食い違う. もちろんそれは, 魏とその後継王朝である西晋にとって都合の良いように描かれたという側面もある. しかしいま関心があるのは, なにが史実か, なにが真実かではなく「**人々の認識がどう変わったか**」である.

矛盾する複数の物語を公平に取り入れようとするならば, 人物の評価はいいとこどりにするか, 悪いところどりにするしかないだろう. よって, 正史三国志が日本人に膾炙されるようになれば, それまで三国志演義で悪役として描かれ評価の低かった人物たちの評価があがり, 結果として『三國志』シリーズでステータスの差別化ができなくなっていくと予想する. 今回は, この仮説を検証するまでの過程を「実践的なデータ分析のチュートリアル」として記録する.

1.4 先行・関連研究

たぶんこんなバカなこと考えるやつは過去にも例がないだろう. よって本研究の新規性・独自性は疑いようがない[*13].

2 前処理

三國志シリーズの登場人物のステータス情報は, インターネット上のいくつかの個人サイトから取得した.

- 三國志 1-7[*14], および 12: 瀬戸大将-三國志舞踏仙郷-
- 三國志 8: 武将リスト (web archive)
- 三國志 9: 三國志 9 武将一覧
- 三國志 10: 三国志 10 武将データ
- 三國志 11: 史実武将データ - 三國志 11 攻略 wiki
- 三國志 13: 武将一覧 - 三國志 13 攻略 WIKI

コーエーテクモ公式の資料集も存在するが, 全て紙媒体であり, 購入および転記のコストを考えて利用しなかった.

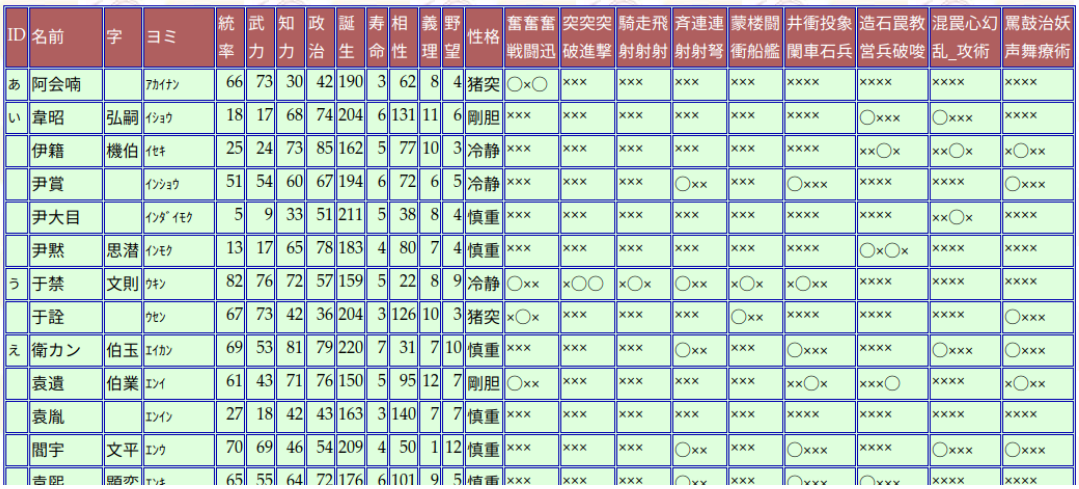
2.1 スクレイピング

まずは `rvest` パッケージで各ページを取得した. `rvest` はパイプ演算子でスクレイピングした html (xml) ノードデータを取得できるため, 使い勝手が良いパッケージである. 取得したページを `rvest` や `tidyverse` を使い整然データとする.

しかし, これらのサイトは全て管理者が異なり, 非公式のものであるからフォーマットも違うため, 作品ごとに異なる工程が必要である. 多くは `<table>` タグを使って掲載されているため, `rvest::html_table()` 関数を使えば概ねうまくいくが, 特に手間がかかったのは, 表が整然化されていない三國志 9 と, 表の背景色でデータを表現していた三國志 12 のページである. 前者は一つのセルに複数の項目が文字列として入っていた (図 1) ため, `stringr::str_split_fixed()` など文字列を処理するパッケージを駆使して分解する必要があった. 後者は, 1 名の人物あたり 2 行で掲載し, なおかつ一部の項目を文字ではなく背景色の塗りつぶしで表現していた

[*13] もちろんこれはジョークである. 研究の新規性・独自性とは, 研究の開拓に対する貢献を伴ったものでなければならない. 単に突飛なだけ, 誰もやらなかったものを初めてやった, だけでは研究の価値を主張したことにならない.

[*14] 『三國志』シリーズの正式名称は, 10 作目まではローマ数字だが, ここでは便宜上全てアラビア数字で表記する.



ID	名前	字	ヨミ	統率	武力	知力	政治	誕生	寿命	相性	義理	野望	性格	奮奮奮戦闘迅	突突突破進撃	騎走飛射射射	斉連連射射弩	蒙楼闘衝船艦	井衝投象闇車石兵	造石罠教営兵破唆	混罠心幻乱_攻術	罵鼓治妖声舞療術
あ	阿会喃		ｱｶｲﾅﾝ	66	73	30	42	190	3	62	8	4	猪突	○×○	×××	×××	×××	×××	××××	××××	××××	××××
い	韋昭	弘嗣	ｲｼｮｳ	18	17	68	74	204	6	131	11	6	剛胆	×××	×××	×××	×××	×××	××××	○×××	○×××	××××
	伊籍	機伯	ｲｾｷ	25	24	73	85	162	5	77	10	3	冷静	×××	×××	×××	×××	×××	××××	××○×	××○×	×○××
	尹賞		ｲﾝｼｮｳ	51	54	60	67	194	6	72	6	5	冷静	×××	×××	×××	○××	×××	○×××	××××	××××	○×××
	尹大目		ｲﾝﾀﾞｲﾓｸ	5	9	33	51	211	5	38	8	4	慎重	×××	×××	×××	×××	×××	××××	××××	××○×	××××
	尹黙	思潜	ｲﾝﾓｸ	13	17	65	78	183	4	80	7	4	慎重	×××	×××	×××	×××	×××	××××	○×○×	××××	××××
う	于禁	文則	ｳｷﾝ	82	76	72	57	159	5	22	8	9	冷静	○××	×○○	×○×	○××	×○×	×○××	××××	××××	××××
	于詮		ｳｾﾝ	67	73	42	36	204	3	126	10	3	猪突	×○×	×××	×××	×××	○××	××××	××××	××××	○×××
え	衛カン	伯玉	ｴｲｶﾝ	69	53	81	79	220	7	31	7	10	慎重	×××	×××	×××	○××	×××	○×××	××××	○×××	○×××
	袁遺	伯業	ｴﾝｲ	61	43	71	76	150	5	95	12	7	剛胆	○××	×××	×××	×××	×××	××○×	×××○	××××	×○××
	袁胤		ｴﾝｲﾝ	27	18	42	43	163	3	140	7	7	慎重	×××	×××	×××	×××	×××	××××	××××	××××	××××
	閻宇	文平	ｴﾝｳ	70	69	46	54	209	4	50	1	12	慎重	×××	×××	×××	○××	×××	○×××	××××	○×××	○×××
	袁煕	顕奕	ｴﾝｷ	65	55	64	72	176	6	101	9	5	慎重	×××	×××	×××	○××	×××	○×××	○×××	××××	××××

図1　三國志 9 の人物一覧ページ

(図 2). そのため, テキスト情報とタグの属性をそれぞれ別に取得し, 結合する必要があった.

さらに, ここで一部名寄せ処理を行っている. というのも, 三国志には数組の同姓同名の人物がいるからである.

- **張温**: 東漢 (後漢) の高級官僚と, 孫呉に仕えた人物
- **張闓**: 陶謙に仕えた武将と, 袁術に仕えた武将
- **張南**: 袁紹に仕えたのち曹操に降伏した武将と, 蜀の武将
- **馬忠**: 呉の孫権に仕えた武将と, 蜀の武将
- **李豊**: 袁術に仕えた武将と, 蜀漢の武将李厳の子, そして魏の人物

これらは識別できなければならないため, 名前の末尾に「孫呉」「東漢」などと所属勢力を括弧書きで追加した. ただし元のページでは必ずしもどの人物か同定されていないため, 生没年や字の有無, ステータスの数値等から判断した.

この作業のため, 各作品内で名前に重複がないか確認したところ, これ以外にも名前を誤って重複しているものが見られた. 能力値や字, 生没年などから推理して修正した. この時点で, **7,115 件**, **1,120 名**の人物データが入手できた.

ここで説明した処理は `scraping.R` と `tidying.R` でなされている.

2.2　さらなる名寄せ処理

今回の情報源は, 複数の個人サイトによるもので, フォーマットも全く異なる. さらに, 表記にもかなりゆらぎがある. 単なる誤変換であるもの, 原典である『正史三国志』と『三国志演義』の間でもすでに食い違っているものなど, 原因は様々である. 使用するデータの品質向上のため, 当初は手動でいくつかの方法を試した.

■三国志以外の登場人物を除外する　既に述べたように, 春秋戦国時代や, 魏晉時より後代の人物が隠し要素として存在する. 三国志演義が史書とは異なる創作であり, 真実がなんであるかを問題としない以上, 2 世紀末の中国で**チンギス＝ハン**が覇を唱えようが**織田信長**が乱入しようが, **皇帝(カイザー)ラインハルト**率いる宇宙艦隊が遠征し

武将名	字	特技								義理	勇猛	相性	誕生	登場	没年	寿命	口調	格付け
		統率	武力	知力	政治	合計	兵科	戦法										
あかいなん	-	商才	耕作	名士	兵心	練兵	収集	人脈	監視	補修								
阿会喃	-	65	74	26	33	198	騎兵	攻撃強化		2	2	62	190	217	225	36	威厳男	★
いせき	きはく	商才	耕作	名士			収集			弁舌								
伊籍	機伯	29	24	80	86	219	弓兵	破壊力弱化		3	0	77	162	189	226	65	策士男	★★
いんしょう	-	商才							監視									
尹賞	-	51	44	62	66	223	弓兵	弓攻撃強化		2	0	72	194	213	260	67	丁寧男	-
いんもく	しせん	商才					収集										兵器	
尹黙	思潜	26	15	66	77	184	槍兵	知力上昇		2	0	80	183	212	239	57	能吏男	★
うきん	ぶんそく					練兵			監視					水練		攻城		
于禁	文則	83	78	74	57	292	弓兵	弓軍強射		1	1	22	159	184	221	63	勇将男	★★
えいかん	はくぎょく			名士						弁舌					遠射		兵器	
衛瓘	伯玉	69	46	79	78	272	弓兵	弓攻撃強化		2	0	31	220	239	291	72	策士男	★
えんいん	-		耕作															
袁胤	-	32	14	39	41	126	槍兵	防御強化		2	0	140	163	184	199	37	丁寧男	-
えんき	けんえき						収集											
袁熙	顕奕	66	51	63	65	245	弓兵	射程弱化		2	0	101	176	190	207	32	策士男	-

図2　三國志 12 の人物一覧ページ

てこようが, 原則を言えばあらゆる創作を「三国志」として認めなければならない. しかし今回はあくまで, 三国志の人物の評価の変遷を知るのが目的である. こういった企画で採用される人物はその時代を代表する英雄であるため, しばしば非常に高いステータス値が設定されている. そういう人物が後期の作品では数十人ほど登録されており, 要約統計などに対する影響はかなり大きい. よって, 今回は『三国志演義』『正史三国志』『反三国志』および『花関索伝』で言及される人物[*15]だけを対象とすることにした. この処理によって **179 名**が除外された.

■漢字が使われていない名前を検査する　まず, 正規表現で漢字以外の使われている人名を探した. 機種依存文字をカタカナ等で置き換えていたものを見つけた手動で修正した. 有名な例では, UTF-8 が普及する以前は**張郃**の「郃」の字に対応した文字コードがなかったため, インターネット上でしばしば「合 β」と表記されていた. この方法では, **龐徳**(ホウトク), **賈詡**(カク), **郝昭**(カクショウ), など同様の原因で, 補正すべき表記のゆらぎが発生している人名を **122 件**発見した.

■3 文字以上の名前を検査する　三国志の時代では, 多くの人名は姓名が 1 字づつであることが多く, 3 字以上の名前は珍しい. **夏侯, 諸葛, 司馬, 公孫**など 2 字の姓は限られている. 名が 2 字以上になるのも**郭攸之**(カクユウシ), **戯志才**(ギシサイ)などどかなり限られる. それ以外で 3 字以上の名前の多くは, **於夫羅, 卑弥呼, 都市牛利**など, 非漢民族の発音を当てたものと思われる. そこで, 名前が 3 字以上のものも手作業で確認してもさほど手間にならないと判断し確認した. その結果, 以下のような表記のゆらぎを **19 件**見つけた. 事例の一部を抜粋する.

- **許劭**(キョショウ)**/許子将**. **子将**は字である[*16].
- **金環三結**(キンカンサンケツ)**/金環結**: 後者は三国志 3 でのみ見られた. 人名に 3 字までの制約があったのだと思われる.
- **祝融**(シュクユウ)**/祝融夫人**: これは誤りではないが, 同一人物の表記が異なるとその後の処理に支障を来す. 「**夫人**」を除外した.

[*15] 実際には『反三国志』に由来する人物は, 馬雲騄(バウンリョク)だけである.

[*16] この表記は 95 年発売の『三国志 5』にのみ見られた. 陳 (1974) では字で表記ししているので, これの影響か.

- **秦宜禄(シンギロク)/秦誼**: そもそも史書で表記のゆらぎがある.
- **邢道栄(ケイドウエイ)/刑道栄**:「邢」をカタカナで置き換えるケースは既に見たが,「刑」で置き換えているケースも発見.
- **劉豹(リュウヒョウ)/左賢王**: 左賢王は南匈奴の称号. 作中のテキストから, 史書で左賢王の地位にあった**劉豹**と同定される. **劉豹**は**於扶羅(オフラ)**の子.

ただし, **許劭/許子将**や, **秦宜禄/秦誼**の組み合わせは, 三国志の知識がなければただちには分からない.

さらに, 本来の意図ではないが, 3 字以上の人名に誤記を見つけた. その抜粋は以下.

- **毌丘倹(カンキュウケン)/母丘倹**: 子弟である**秀**, **甸**にも同様の誤りが見られた.
- **諸葛瑾(ショカツキン)/諸葛謹**: オウ偏の**瑾**の字があまり使われないための誤記と思われる.
- **太史亨(タイシキョウ)/太史享**: 字の違いは微妙である.

■出現頻度の少ない人名を検査する　字数の多い名前での表記のゆらぎはすでに確認できた. しかし, 3 字以上の名前だけを見ても, これだけ表記にゆらぎがあるならば, 2 字の名前でも同様にゆらぎがあると予想できる. そこで, シリーズ全作品のデータを結合した上で, 出現回数が 2 回以下のものを確認した. これで, 誤字をいくらか発見できると考えた. しかし, 実際には知名度の低い人物が多くピックアップされただけであり, ここから表記のゆらぎを見つけるのは難しい. 誤記・誤変換ならばソートしても対になる人名が近くにくるとも限らない.

■そして機械学習へ　そこでさらなる名寄せ処理として, どうやって互いに類似する人名を取り出すか, ということを考える.

多くの自然言語処理の研究では, 文章を対象としている. しかし, すでに述べたように人名のほとんどが 2 字, 多くとも 4 字である. 形状の似ている文字を見つけるということから, 画像認識の技術を応用できないか考えてみる. 画像認識の一種としての手書き文字の認識は昔から研究されている. しかし, これは癖のある字をどう認識するかという教師あり学習の問題として扱われることが多いため, 今回の問題と合致しない.

今回の問題設定に合致するような先行研究がなかなか見つけられないため, 自分なりのアイディアとして, 人名の文字を画像データと見なし, 画像間の類似度を計算することで似たような字を見つける, と言う方法を採用した. これは表記ゆれを確実かつ漏れなく発見できるわけではないが, 総当りよりも効率よく見つけられると考えられる.

画像として表示するにはフォントが必要である. 入力者がどのフォントを使っていたかは特定できない. また, 一部の人名は標準的な日本語フォントに対応していないものもある. 具体的には, 呉の景帝の太子の一人である「孫(ソンワン)竷」である.「竷」の字は Unicode では CJK 統合漢字拡張 B のカテゴリに含まれており[*17], これに対応する日本語フォントは花園明朝 B である. よって, 文字画像には花園明朝 A および B を使うことにした.

まず, 2 つの人名文字列のビットマップ情報[*18]に変換する. それから, ビットマップ情報から特徴量を取り出す.

特徴量の取り出し方は, 今回 2 通りの方法を試した.

1. ビットマップ単位の情報をそのまま使う.

[*17] Unicode のグリフや対応フォントの情報は, FileFormat.Info や グリフウィキ で確認できる.

[*18] 今回は既に 3 字以上の人名の名寄せを手動で行ったが, より汎用的な性能を確認したいため, ここでは 3 字以上の人名も含めて実施してみる. そのため, 画像のサイズは 4 文字分で固定し, 字数の少ない人名は横に引き伸ばしてレンダリングしたものを使う.

name1	name2	Manhattan	Euclidean
干麋	于麋	4.94	4.79
車**胄**	車冑	4.87	4.93
王凌	王**淩**	4.81	5.03
夏侯威	夏侯咸	4.73	4.79
呉**鋼**	呉綱	4.65	4.79
薛**翊**	薛翍	4.59	4.58
邢道栄	**刑**道栄	4.58	4.00
全禕	金禕	4.55	4.47
王匡	士匡	4.52	4.45
劉璜	劉**潰**	4.46	4.47

表1　M 類似度上位 10 件, 誤字を強調

2. 鴨下他 (1998) の方法に即して特徴量を作成する.

(1) の方法では, 特徴量は $32 \times 128 = 4096$ 次元の数値となる[*19]. (2) の方法は, ピクセルの並びの全ての行・列それぞれに対して, 背景色・文字色の変化の回数 (これを「微分」と呼ぶ), 文字色の割合 (これを「積分」と呼ぶ) を計算する方法である. これによって, $(32 + 128) \times 2 = 320$ 次元の特徴量が得られる (実際に使用したのは 317 次元).

最期に, 2 つの文字画像の特徴量ベクトル $\boldsymbol{x}, \boldsymbol{y}$ について, 距離 $d(\boldsymbol{x}, \boldsymbol{y})$ を計算する. 今回はこれを min-max 正規化したものを, 類似度 s として, 値の大きい順にならべる.

$$s(x, y) := \frac{d(x, y) - \min d}{\max d - \min d}$$

なお, このような類似度の求め方は**テンプレートマッチング**と呼ばれる (糟谷・山名, 2006). $d(x, y)$ には, ユークリッド距離

$$d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y})},$$

マンハッタン距離

$$d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|_1 = \sum_k |x_k - y_k|,$$

で計算した. 両者は次元の大きさが全く異なるが, 提示された結果はかなり似ている. 上位 30 件を確認して発見した表記のゆらぎを表1に抜粋する.

特に紛らわしいのは表2である. これは文字を拡大しないと気づきづらい.

新たに多くの表記ゆれを発見できたが, 一方で**誤検知**もある. 表1 では, 夏侯威(カコウイ)と夏侯咸(カコウカン), 全禕(ゼンイ)と金禕(キンイ), 王匡(オウキョウ)と士匡(シキョウ)の組み合わせは別人物である.

1 字同じだけでもかなり一致度が高くなってしまう. 結果として勘でやったほうが修正の必要な箇所を多く見つけられたので, より精度が必要である. 一方で, 鴨下他 (1998) はかなり古い研究で, 文字のビット数が小さ

[*19] どの文字画像でも変化のないピクセルは情報を持たないため除外し, 実際には 4025 次元を使った.

正	誤	解説
車冑(シャチュウ)	車冑	「冑」の下
関彝(カンイ)	関彝	「米糸」と「米分」
鍾会(ショウカイ)	鐘会	鐘(カネ)ではない

表2　発見できた紛らわしい表記のゆらぎ例

く, さらに特徴量を大きく削減するなど計算量を削減しているが, 上記の結果とあまり変わらない結果が得られた.

そもそもなぜ表記ゆらぎが起きるかと言えば, 登録時点でのミス, 原作時点でのミスである. 前者は音や形状の似た字への誤変換, 普及している日本語フォントではカバーしていない, あるいは IME が対応していない字 (いわゆる機種依存文字) の代用, 後者は同一文献や, 創作物ごとのゆらぎがある[*20].

例: 原作からしてゆらぎがある[*21]

- **李堪**(リカン)と**李湛** (三国志演義と吉川三国志)
- **楊脩**(ヨウシュウ)と**楊修**
- **雷銅**(ライドウ)と**雷同**
- **陳羣**(チングン)と**陳群**
- **田豫**(デンヨ)と**田予**

例: 機種依存文字の影響で間違えやすい字: 部首が違う

- **劉璟**(リュウカイ) (正) **劉潰** (誤): 「璟」は日本語ではほぼ使われない
- **王凌**(オウリョウ) (正) と**王淩** (誤): ニスイ偏が正しい.
- **鍾会**(ショウカイ) (正) と**鐘会** (誤): カネではない[*22].
- **歩騭**(ホシツ) (正) と **歩隲** (誤): コザト偏の位置

似ているが別人の例として, 既に紹介したもの以外にも以下のようなものがある.

- **鄧艾**(トウガイ)と**鄧芝**(トウシ)
- **桓楷**(カンカイ)と**桓階**(カンカイ)

以上の傾向から, 字形の平均的な一致度ではなく, 部首単位での類似を考慮して類似度を計算することができれば効率的であると思われる. また, 教師データも ground-truth なモデルも用意できないため, 「なるべく少ない労力で, たまたまでもうまく表記ゆらぎを見つけられるような類似度の求め方」が得られれば良い.

以上の処理は, `image_recognition.R` で実行している.

[*20] まれなケースとして, 字 (あざな) が使われている場合もあるが, 当時の名の多くは 1 字である一方, 字の多くは 2 文字であり, 文字数が多いため手作業でもすぐに見つけることができた.

[*21] 初期の作品はハードの制約から, より簡単な表記を選んだとも考えられる.

[*22] 現代の簡体字では統一して同じ字として扱われる.

2.3 ディープラーニングでなんとかできないか?

このセクションは昨日思いついて試してみたけど時間がたりなかったので書きかけです. ディープラーニングじたいほとんどやってないので話半分で読んで欲しい.

画像認識と言えば最近はニューラルネットワークを使った話が流行っているので, 何か応用できるものがないか探してみた. 機械学習の問題としてみれば教師なし学習で, かつ 2 点間の類似度を出せるものがよい. ここまで試したのは 2 つの文字画像のピクセル $\boldsymbol{x}, \boldsymbol{y}$ 間の距離である. 例えばユークリッド距離で,

$$d(\boldsymbol{x}, \boldsymbol{y}) := \sqrt{\|\boldsymbol{x} - \boldsymbol{y}\|_2}$$

を 2 つの画像の類似度としてきた. しかしこれでは限界があることがわかったので, なんらかの適切な特徴量変換器 $\boldsymbol{f}$ を挟んで,

$$s(\boldsymbol{x}, \boldsymbol{y}) := \sqrt{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\|_2}$$

のような類似度計算ができるようになればいい. 機械学習の研究では, これを**計量距離学習** (metric learning) という[*23].

ここでいくつか関連しそうな研究を紹介しておく.

Wang et al. (2014), Hoffer and Ailon (2015), Sanakoyeu et al. (2018), Turpault et al. (2019) などを参考にすると最近は計量距離学習では triplet network と呼ばれるモデルが流行しているらしい.

Zhang and Komachi (2019) では, CHASE プロジェクトのデータベースから, 文字の部首情報を取り出して教師なしニューラル機械翻訳 (UNMT) をしている[*24]. しかしこれは画像認識ではない

Liu et al. (2017) は音素も考慮しているが, 今回は日本語での入力の問題なので少し違う. あと教師あり学習.

”In words, this encodes the pair of distances between each of x+ and x− against the reference x.“

Wang et al. (2014), Hoffer and Ailon (2015) 前者は多クラス分類だが, 後者はランキング問題

なお私は計量距離学習というトピックをこれまで全く知らなかった. 基本的な考え方を理解するために今回初めて Bellet (2013), Bellet et al. (2014) などを参照した程度である (よって見落としているだけということもありうる). このサーベイ・チュートリアル資料で紹介されているアイディアの多くは教師ありないし半教師あり学習だが, 今回は教師ラベルを作るのが面倒な場合はどうするかというのが問題である. ここでは主に Turpault et al. (2019) の提案する半教師あり学習[*25]をもとに試してみる. まず, 従来的な 2 点の比較は双生児 (siamese) ネットワークと呼ばれる:

$$s_{siamese}(\boldsymbol{x}, \boldsymbol{y}) := \|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2 .$$

一方で, 基準点 (anchor あるいは query と呼ばれる) $\boldsymbol{x}^a$ に対して正例 $\boldsymbol{x}^p$, 負例 $\boldsymbol{x}^n$ の 3 対 (triplet) $(\boldsymbol{x}^a, \boldsymbol{x}^p, \boldsymbol{x}^n)$ を考慮したのが triplet network である.

$$s_{triplet}(\boldsymbol{x}, \boldsymbol{x}^p, \boldsymbol{x}^n) := \begin{bmatrix} \|f(\boldsymbol{x}^a) - f(\boldsymbol{x}^p)\|_2 \\ \|f(\boldsymbol{x}^a) - f(\boldsymbol{x}^n)\|_2 \end{bmatrix}$$

[*23] 要はクラスタリングのことだと思うのだが, この単語を見かけるようになったのは最近になってからな気がする.

[*24] 実装は Python の https://github.com/vincentzlt/textprep に依存

[*25] Turpault et al. (2019) は画像認識ではなく音声認識のテーマである

プログラム 1 SPARQL の例

```
PREFIX␣dbpedia:␣<http://ja.dbpedia.org/resource/>
PREFIX␣dbp-owl:␣<http://dbpedia.org/ontology/>
PREFIX␣rdf:␣<http://www.w3.org/2000/01/rdf-schema#>
PREFIX␣category-ja:␣<http://ja.dbpedia.org/resource/Category:>
SELECT␣DISTINCT␣?article,␣?text
WHERE␣{
␣␣␣␣␣␣␣␣?article␣dbp-owl:wikiPageWikiLink␣category-ja:三国志の登場人物␣.
␣␣␣␣␣␣␣␣?article␣rdf:comment␣?text␣.
}
```

これら 3 点の相対的な距離をもとに学習するというのが triplet network のアイディアになる. さらに, Wang et al. (2014) に従って triplet 損失を

$$L_{triplet}(\boldsymbol{x}^a, \boldsymbol{x}^p, \boldsymbol{x}^n; \delta) := \lfloor \|f(\boldsymbol{x}^a) - f(\boldsymbol{x}^p)\|_2 - \|f(\boldsymbol{x}^a) - f(\boldsymbol{x}^n)\|_2 + \delta \rfloor$$

で定義する.

しかし今回は教師ラベルがないため, $\boldsymbol{x}^p$, $\boldsymbol{x}^n$ をどう選べばいいかが分からない. そこで, Turpault et al. (2019) の提案するように, 特徴量 $\boldsymbol{x}$ の距離で正例負例を与える.

2.4 補足: DBpedia を利用した二重チェック

教師なし学習による探索だけでは心もとないので, Wikipedia の記事を使った二重チェックを行った. DBpediaとは, Wikipedia を構造化したデータベースで, SPARQL によってデータを取得できる. 例えば プログラム 1 のようなクエリになる.

R では, SPARQLパッケージが用意されている.

3 本題: 三国志の翻訳文献の充実は「三國志」シリーズの人材の無個性化につながっているか?

以上で一旦名寄せ処理を切り上げる.

要約統計量を計算するのに役立つのが skimr パッケージである. 要約統計量を表示する関数は, 組み込みの summary() を始めいくつもあるが, skimr は

- summary() よりも見やすい
- group_by() したデータを与えるとグループ別集計してくれる

といった便利さからおすすめする. ただし日本語の情報が少ない. 私の知る限り『niszet の日記』が唯一言及しているブログで, しかも現在はさらに仕様が変わっている.

仕様が変わったのは表示する統計量を決める部分である. 以前は skim_with() でグローバルに変更していた

プログラム 2 skimr に歪度と尖度を追加する

```
library(moments)
my_skim␣<-␣skim_with(numeric␣=␣sfl(skew␣=␣skewness,␣kurto␣=␣kurtosis,␣hist␣=␣NULL),␣append␣=␣T)
DATA_FRAME␣%>%␣group_by(title)␣%>%␣my_skim()
```

skim_variable	title	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.skew	numeric.kurto
武力	1	0	1	-8e-17	1	-1.7	-0.87	0.0054	0.87	1.7	-0.0038	1.8
武力	2	0	1	2.5e-17	1	-2.2	-0.83	0.1	0.71	1.9	-0.093	2.1
武力	3	0	1	-6.2e-17	1	-2.7	-0.55	0.15	0.56	2.3	-0.46	3
武力	4	0	1	1.2e-16	1	-2.4	-0.6	0.23	0.67	1.9	-0.56	2.6
武力	5	0	1	-1.3e-16	1	-2.3	-0.72	0.3	0.69	1.7	-0.64	2.5
武力	6	0	1	-1.6e-16	1	-2.1	-0.78	0.21	0.73	2.1	-0.32	2.3
武力	7	0	1	-4.3e-17	1	-2.3	-0.66	0.21	0.75	1.9	-0.5	2.3
武力	8	0	1	-3.6e-17	1	-2.3	-0.61	0.3	0.64	2	-0.7	2.6
武力	9	0	1	-5.5e-17	1	-2.3	-0.81	0.38	0.67	1.8	-0.66	2.3
武力	10	0	1	-1.5e-16	1	-2.4	-0.75	0.32	0.71	1.9	-0.67	2.5
武力	11	0	1	1.3e-16	1	-2.2	-0.87	0.36	0.72	1.8	-0.57	2.2
武力	12	0	1	4.2e-19	1	-2.2	-0.8	0.35	0.71	1.7	-0.66	2.3
武力	13	0	1	-1.1e-16	1	-2.3	-0.86	0.37	0.71	1.9	-0.58	2.2
知力	1	0	1	9.9e-17	1	-1.7	-0.87	0.0046	0.87	1.8	-0.0044	1.8
知力	2	0	1	1.5e-16	1	-2	-0.82	0.05	0.87	1.9	-0.091	1.9
知力	3	0	1	-2e-16	1	-2.5	-0.64	0.051	0.62	2.5	0.0014	2.8
知力	4	0	1	1.4e-16	1	-2.3	-0.69	0.12	0.63	2.1	-0.24	2.4
知力	5	0	1	-4e-17	1	-2.4	-0.8	0.073	0.74	1.9	-0.13	2.3
知力	6	0	1	-3.1e-17	1	-2.2	-0.75	0.15	0.7	2.1	-0.34	2.4
知力	7	0	1	-1.2e-16	1	-2.2	-0.79	0.18	0.79	2	-0.19	2.2
知力	8	0	1	-1.2e-16	1	-2.4	-0.79	-0.012	0.72	2.3	0.0015	2.2
知力	9	0	1	4.6e-17	1	-2.8	-0.75	0.28	0.67	2	-0.56	2.8

表3　skimr による表 (一部)

が, 現在は関数ジェネレーターのような仕様になっている (プログラム 2). 出力例が表3である.

しかし, 今回は見るべき項目が多いので, グラフで見やすくする必要がある. まずは, 各作品で, 新しく登録された人物と除外された人物が何人かを表してみる. 図 3 では, 前作から追加された人物が in, 逆に除外された人物を out, 続投している人物を keep で表した. つまり, in + keep が各作品に登場する人数である.

図 3 からは, 4, 12 で前作より減っているものの, 基本的に登場人物が増えていることがわかる. よって, 少しづつ正史三国志に記述のある人物が増えていることが分かる[*26].

図 3 を含め, 以降の画像は全て ggplot2 だけで作成した. ただし, カラーパレットは ggtheme のものを使うと良い. ここでは colorblind シリーズを使用している.

ここからは, いくつかの切り口からデータを見ていく. まずは, 4 名の人物について, シリーズを通してステータスがどう変化しているかを見る. 主要人物は記述が多く, 演義と正史での評価の差異を細かく説明するのが大変である. そこで, 主要人物ではないが, 差異の分かりやすい人物を挙げる.

- 華雄 (カユウ)
  - 正史: 「胡軫の配下として孫堅軍に討たれた」としか書かれていない (呉書孫堅伝)
  - 演義: 董卓配下の猛将で, 孫堅を敗走させる. しかし関羽に即敗北する (三国志演義)

[*26] 今回のデータは, 各人物が三国志演義と正史三国志いずれに登場しているかをはっきり示す情報を含んでいない. 万全を期すならば詳細に調査すべきだが, 名寄せ処理と同様の理由で今回は割愛した.

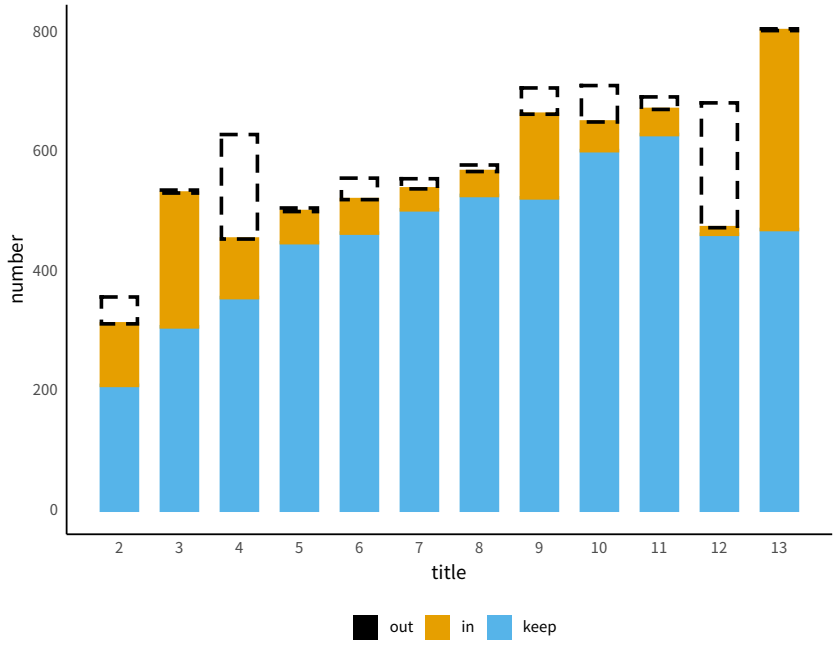


図3　登録・除外フロー

- 関興（カンコウ）
  - 正史: 「父関羽の死後, 将来を嘱望されるも数年後病死」のみ (蜀記)
  - 演義: 父の仇討ちに成功し, 数度の北伐で活躍
- 曹真（ソウシン）
  - 正史: 諸葛亮の北伐に対する防衛を指揮し, 二度退ける
  - 演義: 北伐では終始諸葛亮に翻弄され, 最期は罵倒され憤死した
- 李通（リ ツウ）
  - 正史: 曹操の本拠地, 豫州南部を守り抜く (李通伝)
  - 演義: 馬超と一騎打ちし即敗北する

図 4 では, 「演義で活躍の盛られている」代表である華雄, 関興はシリーズを通してあまり変化していない. すくなくとも低下しているようには見えない. 一方で, 李通, 曹真は徐々に上昇しているように見える.

ということは, もしこれが全体の傾向にも当てはまるのなら, 分布にも現れるはずである. そこで, シリーズの作品ほとんどで存在するステータス項目である, 「武力」「知力」「魅力」「政治」を確認してみる. 分布確認にはヒストグラムもいいが, ここでは `geom_violin()` を使ってバイオリン図を作図した (図5-図 8).

このセクションは `analysis.R` で実施している.

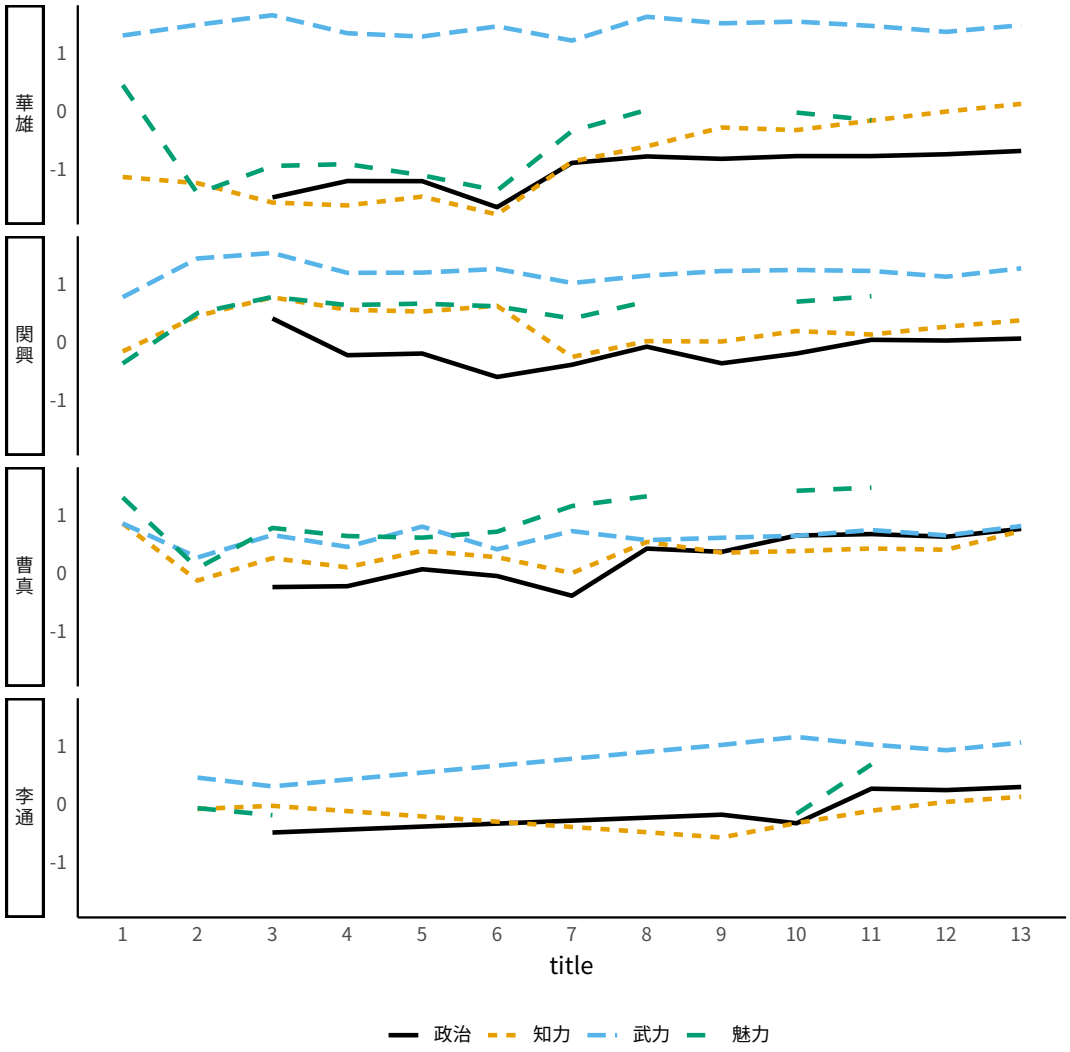


図4　4 名のシリーズを通した変化

3.1　補足: データビジュアライゼーションの教科書について

グラフの書き方にも流儀がある. 3D 円グラフはやめよう[*27], ユーレイ棒グラフはやめよう[*28], という話は昔から喚起されているので知っている人も多いかもしれない. そして体系的なグラフ作成のルールというのがあるのだが, それ含めてをここで説明するのは大変だ. そのうち挑戦してみたくはある.

グラフの書き方に関する本は Tufte (2001) が古典的? であるが, 最近のものとして Healy (2018) は Tufte (2001) の思想を受け継ぎつつ全ての図に対して作図した `ggplot2` によるコードを公開しているため, タイトル通り "practical" である. 一方で, これらはいずれも日本語訳がない[*29]. Tufte 流の理論に則ったという本では,

[*27] 3D 円グラフを使うのはやめよう | Okumura's Blog Wonder Graph Generator, 森藤・あんちべ (2014)
[*28] ユーレイ棒グラフ? | Okumura's Blog - 奥村研究室
[*29] Tufte の名前は日本語文献でもちらほらみられるようになったが, そもそも著作は未だに翻訳されていない. だれかやりましょう?

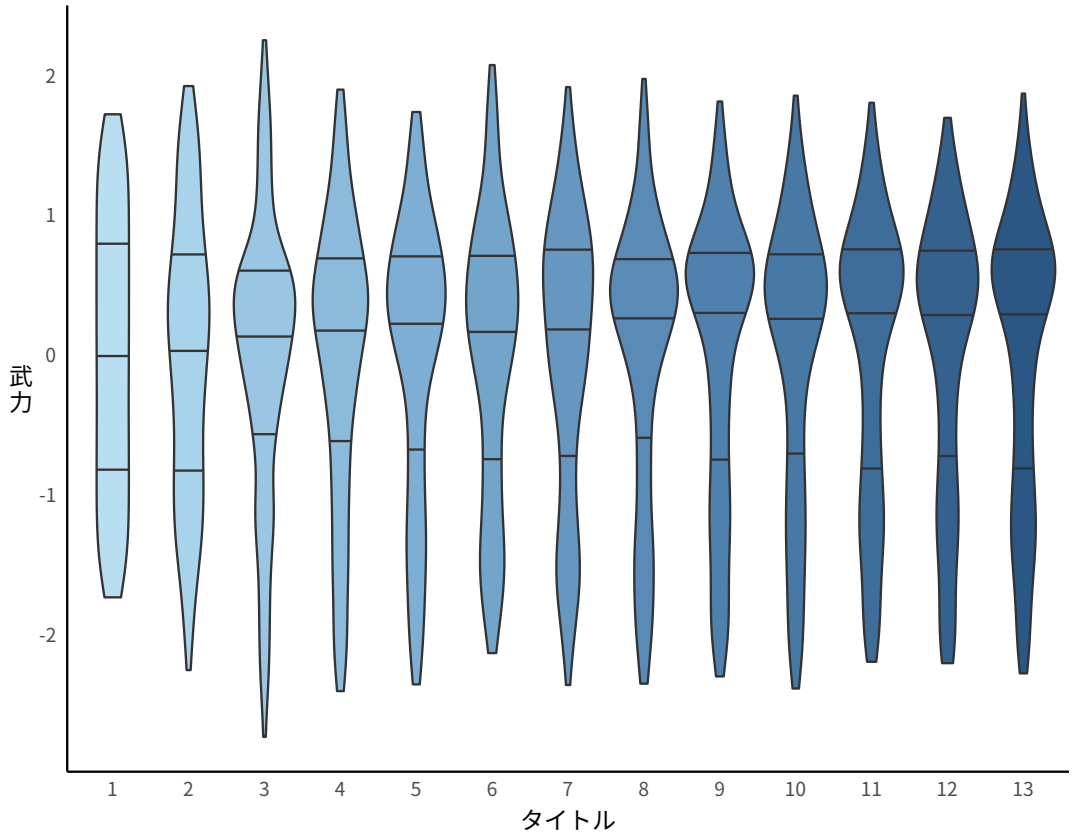


図5　武力の分布

藤・渡部 (2019) が比較的近い. ただしグラフの例は紹介されているものの実際にどのようなソフトウェアで作成するかといったことは書かれていない.

4　まとめ

フォントレンダリングのことなんもわからん

CJK 統合漢字なんもわからん

原典の自然言語処理とか面白そうだと思ったけど日本語ソースの人名ですら名寄せこんなめんどくさいのでやったら絶対死んでた[*30]

参考文献

Bellet, Aurélien (2013) “Tutorial on Metric Learning,” October, retrieved from *here*.

Bellet, Aurélien, Amaury Habrard, and Marc Sebban (2014) “A Survey on Metric Learning for Feature Vectors and Structured Data,” *arXiv:1306.6709 [cs, stat]*, February, arXiv: 1306.6709.

[*30] 三国志演義は様々な説話を集めて編纂されたので, 話によって文体や人名の呼び方が違ったりする. 例えば, 関羽, 関雲長, 関公, 関将軍, など. ましてや正書法の整備されていない中近世.

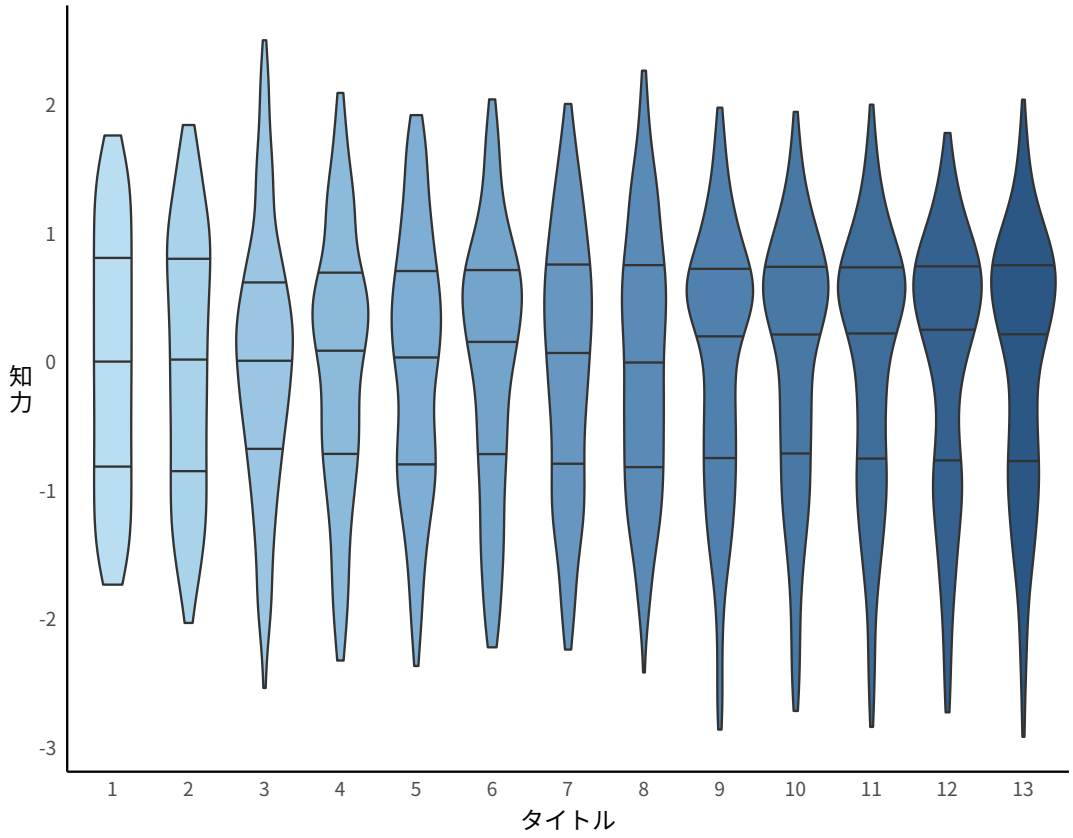


図6　知力の分布

Healy, Kieran (2018) *Data Visualization: A Practical Introduction*, Princeton, NJ: Princeton University Press, retrieved from *here*.

Hoffer, Elad and Nir Ailon (2015) "Deep Metric Learning Using Triplet Network," in Feragen, Aasa, Marcello Pelillo, and Marco Loog eds. *International Workshop on Similarity-Based Pattern Recognition*, Vol. 9370, pp. 84–92, Cham: Springer International Publishing, DOI: 10.1007/978-3-319-24261-3_7.

Liu, Ming, Vasile Rus, Qiang Liao, and Li Liu (2017) "Encoding and Ranking Similar Chinese Characters," *Journal of Information Science and Engineering*, Vol. 33, pp. 1195–1211, retrieved from *here*.

Sanakoyeu, Artsiom, Miguel A. Bautista, and Björn Ommer (2018) "Deep Unsupervised Learning of Visual Similarities," *Pattern Recognition*, Vol. 78, pp. 331–343, June, DOI: 10.1016/j.patcog.2018.01.036.

Tufte, Edward R. (2001) *The Visual Display of Quantitative Information*, Cheshire, Conn: Graphics Press, 2nd edition.

Turpault, Nicolas, Romain Serizel, and Emmanuel Vincent (2019) "Semi-Supervised Triplet Loss Based Learning of Ambient Audio Embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 760–764, Brighton, United Kingdom: IEEE, May, DOI: 10.1109/ICASSP.2019.8683774.

Wang, Jiang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu (2014) "Learning Fine-Grained Image Similarity with Deep Ranking," in *2014 IEEE*

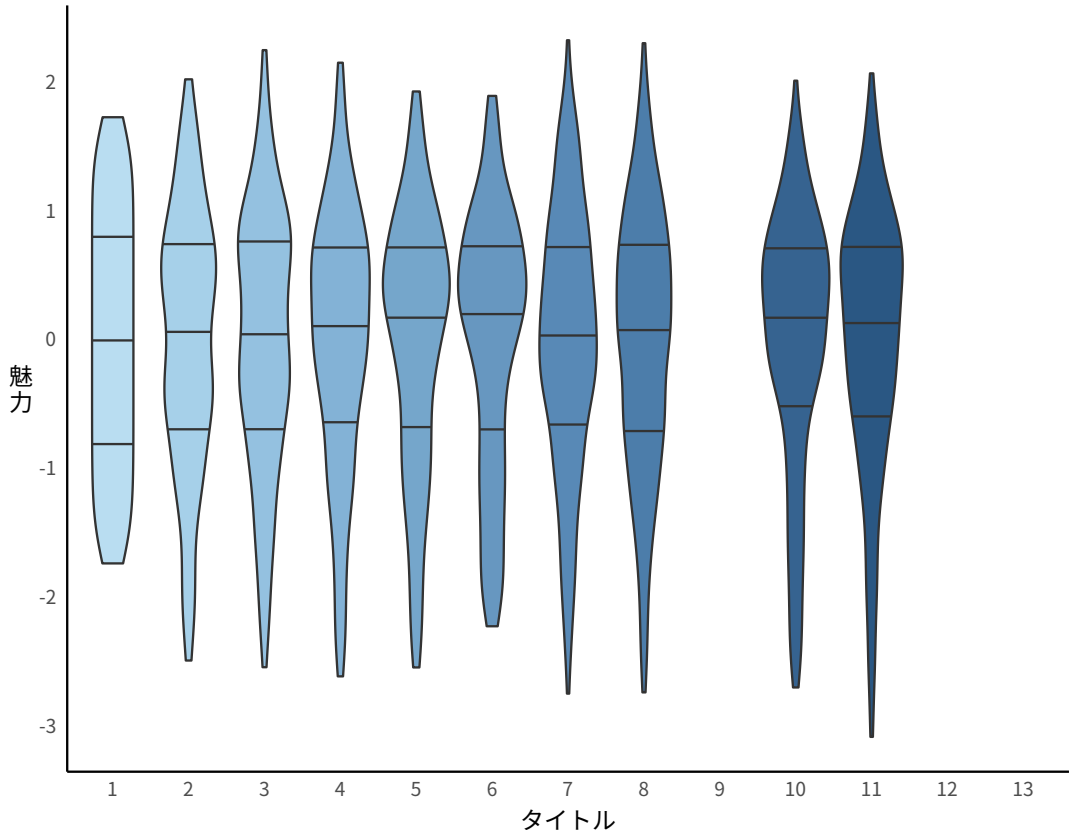


図7　魅力の分布

Conference on Computer Vision and Pattern Recognition, pp. 1386–1393, Columbus, OH, USA: IEEE, June, DOI: 10.1109/CVPR.2014.180.

Zhang, Longtu and Mamoru Komachi (2019) "Chinese-Japanese Unsupervised Neural Machine Translation Using Sub-Character Level Information," February, arXiv: 1903.00149.

鴨下隆志・奥村健一・高橋和仁・増村正男・矢野宏 (1998)「文字認識におけるマハラノビスの距離による判定の研究」,『品質工学会』, 第 6 巻, 第 4 号, 39–45 頁, retrieved from *here*.

北方謙三 (1996)『三国志』, 角川春樹事務所.

陳舜臣 (1974)『秘本三国志』, 文藝春秋.

糟谷勇児・山名早人 (2006)「二種類の SVM を用いたオンライン類似数式文字識別」,『電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解』, 第 105 巻, 第 614 号, 55–60 頁, 2 月, retrieved from *here*.

藤俊久仁・渡部良一 (2019)『データビジュアライゼーションの教科書』, 秀和システム, 東京, retrieved from *here*, OCLC: 1103469309.

森藤大地・あんちべ (2014)『エンジニアのためのデータ可視化「実践」入門: D3.js による Web の可視化』, retrieved from *here*, OCLC: 1022205495.

宮城谷昌光 (2004)『三国志』, 文藝春秋.

吉川英治 (1939)『三國志』, 大日本雄辯會講談社.

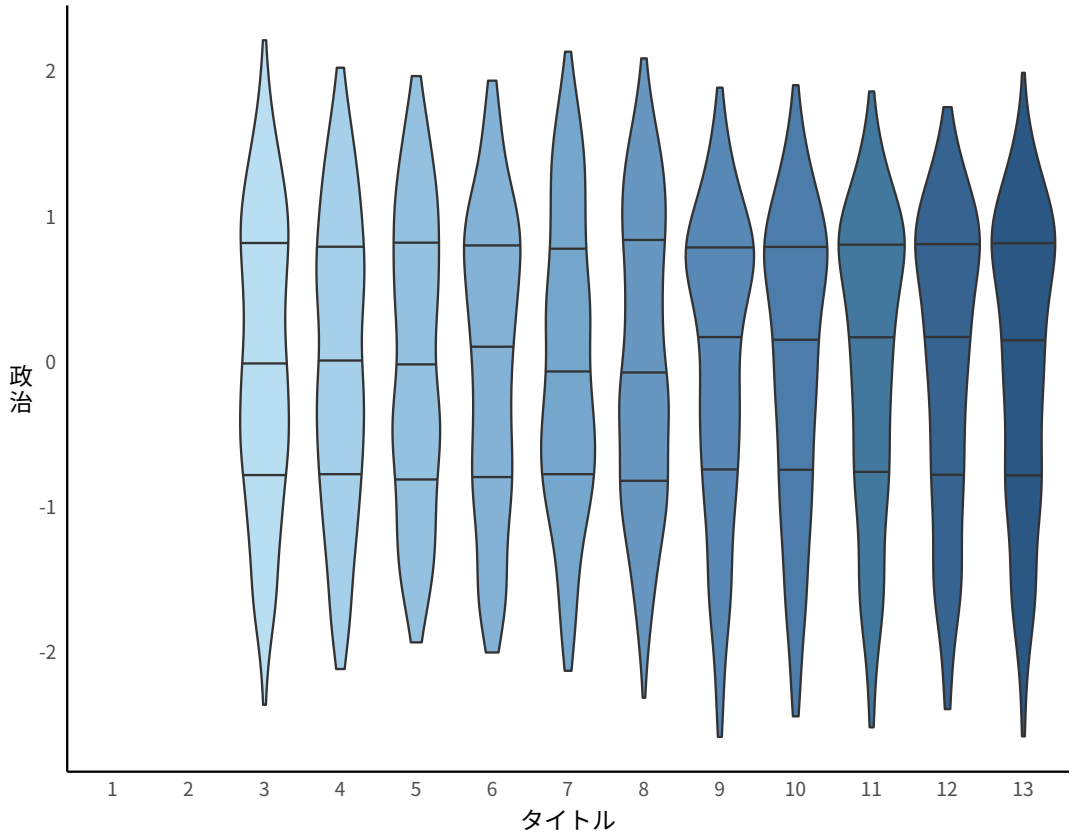


図8　政治の分布

周大荒 (1919)『反三國演義』，捷幼出版社，台北，（渡辺精一訳，『反三国志 (上, 下)』，講談社，1991 年），今戸栄一編訳『超・三国志』1991 年, 光栄.

渡辺義浩 (2011)『三国志: 演義から正史, そして史実へ』，中央公論新社，東京，OCLC: 752021927.

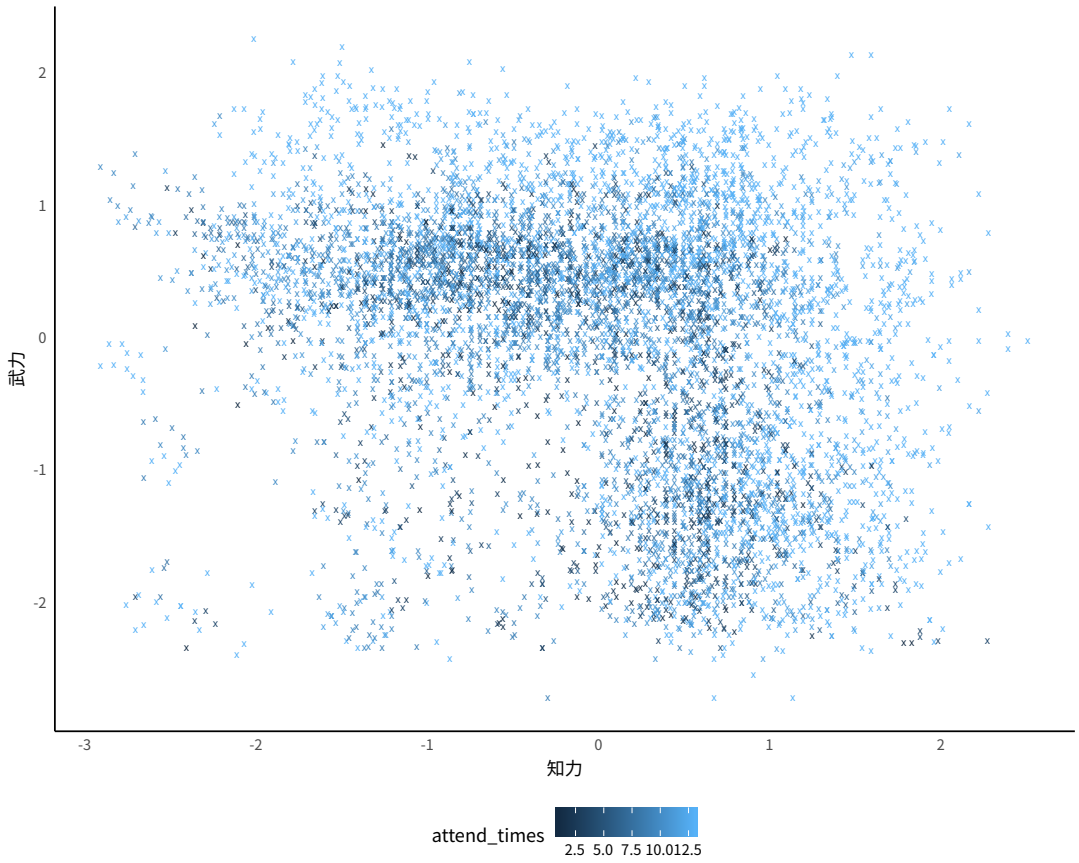


図9　シリーズ登場回数との３軸