マイクロブログにおける リツイート行動の要因分析

～世界陸上と甲子園を題材に～

2014年 1月12日 卒論発表会
情報システム解析学科 尾崎研究室
小林竜也

研究動機

Twitter　身近かつ話題が多いソーシャルメディア

顔文字　連絡手段で日頃使う

スポーツに着目　陸上競技に携わっていた経験がある

関連研究

Bad News Travel Fast:
A Content-based Analysis of Interestingness on Twitter

WebSci '11: Proceedings of the 3rd International Conference on Web Science, (2011)

導入

Bad News Travel Fast

Tweet内に出現する属性(要素)
属性のReTweetに対する影響
影響の大きい属性の調査
回帰分析

コンテンツに注目(英文)

例:顔文字があるTweetはReTweet
されやすい

⇓ ⇓ ⇓

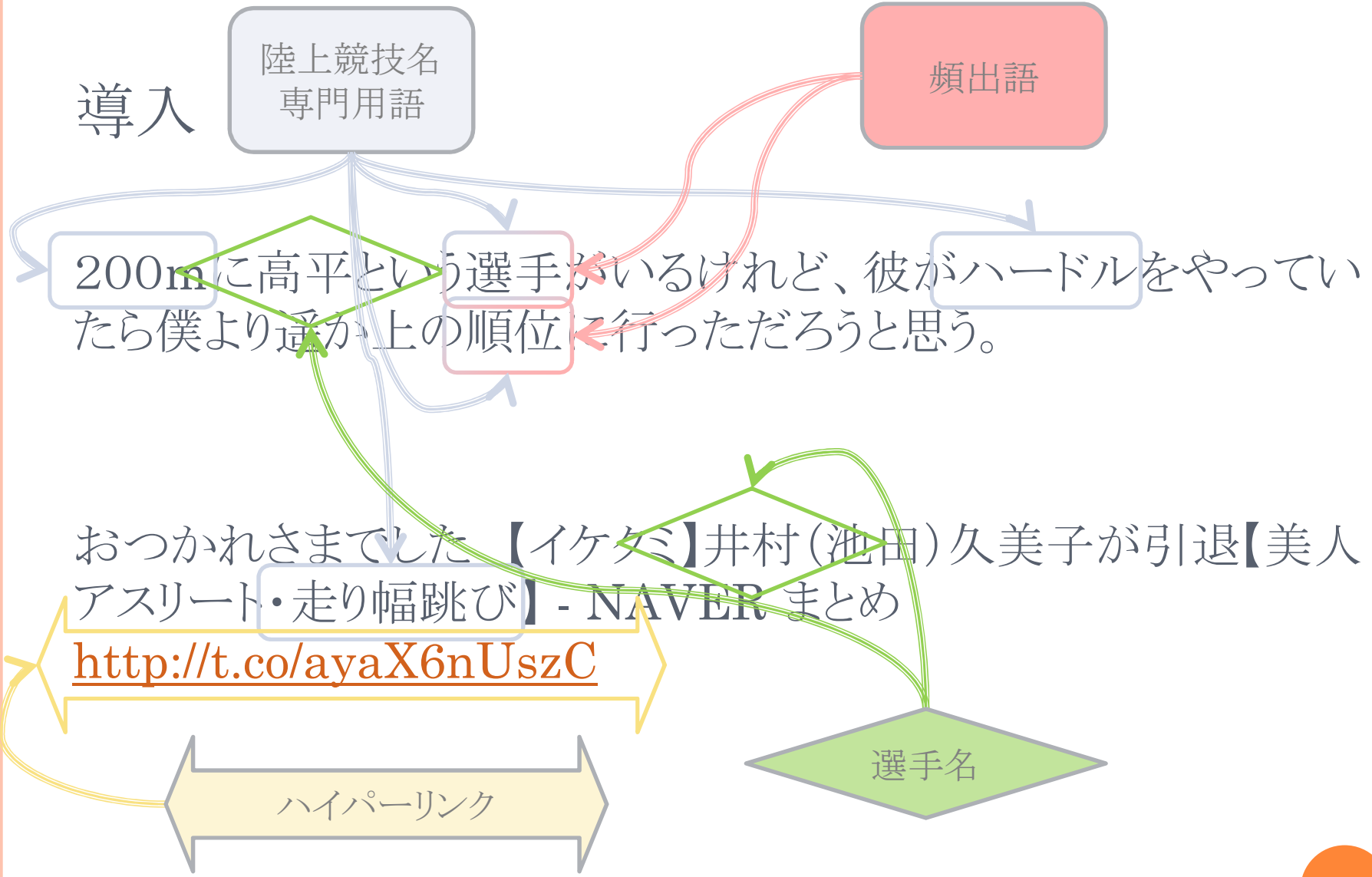
コンテンツのみに注目
日本語Tweet
世界陸上2013, 夏の甲子園
回帰分析, 決定木, 傾向スコア

ﾀﾞｲﾚｸﾄﾒｯｾｰｼﾞ	0, 1
ユーザー名	0, 1
ハッシュタグ	0, 1
URL	0, 1
！/？	0, 1
ポジ/ネガ単語	0, 1
ポジ/ネガ顔文字	0, 1
価数(正負の感情)	-5, +5
覚醒(気分の感情)	-5, +5
支配(強弱の感情)	-5, +5
単語	0, 1
文章	0, 1

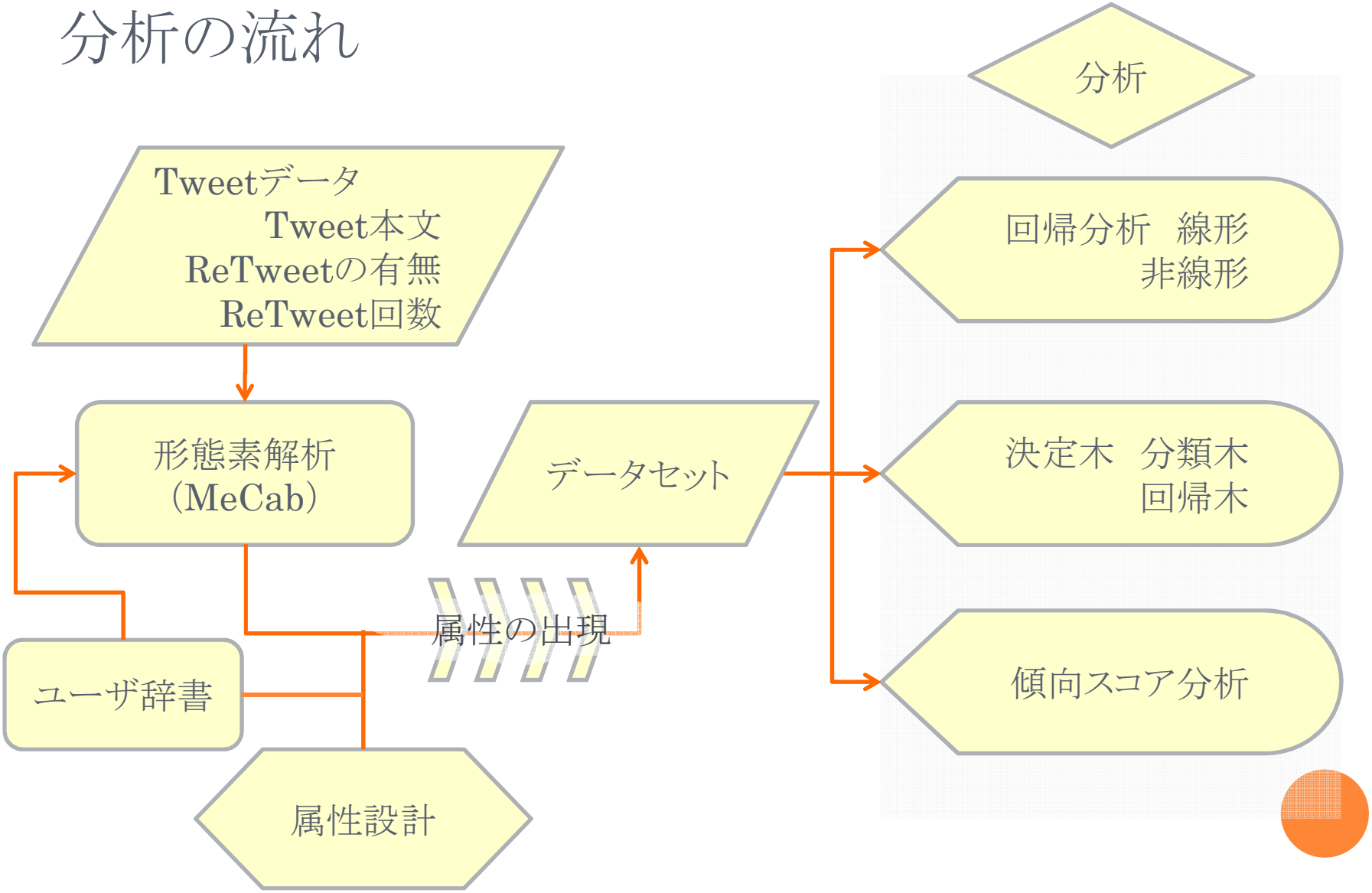
導入

Tweet本文の中から
ReTweetに関係の強い要素を分析する

導入


陸上競技名
専門用語
頻出語
200mに高平という選手がいるけれど、彼がハードルをやっていたら僕より遥か上の順位に行っただろうと思う。
おつかれさまでした【イケケミ】井村(池田)久美子が引退【美人アスリート・走り幅跳び】- NAVERまとめ
http://t.co/ayaX6nUszC
ハイパーリンク
選手名

分析の流れ


分析
Tweetデータ
Tweet本文
ReTweetの有無
ReTweet回数
回帰分析　線形
非線形
形態素解析
（MeCab）
データセット
決定木　分類木
回帰木
属性の出現
ユーザ辞書
傾向スコア分析
属性設計

属性の設計

属性名	属性値	説明	
頻出語	0,1	頻出語上位 100件	
ハイパーリンク	0,1	URLを表す文字列	
ユーザーネーム	0,1	ユーザーネームを表す文字列	
顔文字	0,1	日本で使われる顔文字	・Webより
感情語 喜	0,1	喜 を表す感情語	
怒	0,1	怒 を表す感情語	・感情表現辞典より
哀	0,1	哀 を表す感情語	
恐	0,1	恐 を表す感情語	
陸上用語	0,1	陸上競技に関する用語	・JAAFより
競技名	0,1	陸上競技の競技名	・経験を基に自作
TOP8選手名	0,1	今世界陸上各種目TOP8の選手名	・TBS公式より
日本人選手名	0,1	今大会の出場した日本人選手名	
野球用語	0,1	野球関連の用語	・自作, 共有

TWEETデータ　属性出現数

属性	世界陸上	甲子園
ReTweet / 総数	27268 / 67839	21984 / 41139
ReTweet最大数	4502	2653
ハイパーリンク	9193	11861
ユーザーネーム	27504	21863
顔文字	439	81
喜	747	482
怒	13	3
哀	78	62
恐	846	80
陸上用語	49968	-----
競技名	23660	-----
TOP8選手名	31387	-----
日本人選手名	19165	-----
野球用語	-----	14901

回帰分析

- 回帰式を用い，目的変数が説明変数によってどれだけ説明できるかを分析すること，その値を求めること

- 線形回帰　目的変数　⇒　ReTweetの回数

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

目的変数 Y　⇒　ReTweetの回数

説明変数 X　⇒　各属性の値

- 非線形回帰（ロジスティック回帰）

目的変数　⇒　ReTweetの有無（有の確率）

回帰分析結果　線形回帰分析：RT数

weka

retweetCnt =

20.7058 * t0 +
-13.4222 * t1 +
22.719 * t2 +
3.9621 * t3 +
-15.6674 * t4 +
52.1776 * t5 +
2.2807 * t6 +
15.7838 * t7 +
20.4603 * t8 +
-5.704 * t10 +
105.7826 * t11 +
-7.0532 * t12 +
-5.2608 * t13 +
-8.6726 * t14 +
52.0654 * t15 +
11.7158 * t16 +
23.1146 * t17 +
5.1599 * t18 +
54.7705 * t19 +
-6.8915 * t20 +
42.8766 * t22 +
24.7807 * t23 +
124.6392 * t24 +
25.1639 * t26 +
12.331 * t27 +
-17.2761 * t29 +
25.0986 * t30 +
-56.3881 * t31 +
33.3831 * t32 +
56.228 * t33 +
11.0857 * t34 +
-8.6447 * t35 +
9.8008 * t36 +
-4.7705 * t37 +
9.6509 * t38 +
19.1076 * t39 +
17.5127 * t40 +
30.5441 * t41 +
78.2019 * t42 +
-7.9234 * t44 +
-20.3449 * t45 +
-13.8902 * t46 +
60.7493 * t47 +
23.8211 * t48 +
10.7017 * t49 +
31.5712 * t50 +
148.4615 * t51 +
-13.0119 * t52 +
-54.766 * t53 +
20.4052 * t54 +
-21.5052 * t55 +
-13.4582 * t56 +
-9.9675 * t57 +
46.262 * t58 +
76.212 * t59 +
52.9423 * t60 +
12.1573 * t61 +
29.0598 * t62 +
-21.9166 * t63 +
74.8852 * t64 +
18.412 * t65 +
-8.3461 * t66 +
101.3152 * t67 +

回帰分析結果　線形回帰分析：RT数

weka

52	アメリカ	-54	裕二
105	daijapan	76	競歩
52	400m	52	途中
54	速報	74	最終
124	niigata	101	良子
-56	銅	60	高瀬
56	金メダル	51	今季
78	今日	120	ウクライナ
60	(-	526	心配
148	4×100m		

為末 大
公式アカウント

久保倉 里美
所属 新潟A・RC

世界陸上お馴染みの
キャスター
織　田　裕　二

回帰分析結果　非線形回帰分析：RT有無

weka

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...

Variable	Class 1
男子	0.9313
決勝	-0.0007
日本	0.3778
選手	0.3847
女子	0.6161
アメリカ	0.3575
予選	0.1484
ジャマイカ	0.4296
福士	0.3548
マラソン	-0.491
モスクワ	-0.053
daijapan	4.0388
記録	0.2269
大会	0.1436
世界	0.6997
400m	0.5648
進出	0.3966
入賞	0.5681
200m	0.2567
速報	0.4664
金	0.2868
桐生	0.0486
織田	0.6772
優勝	0.536
niigata	0.978
通過	1.2697
時間	0.6698
木崎	-0.4053
イギリス	0.5687
川内	0.4702
応援	0.079
銅	-0.257
アリソン	0.1841
金メダル	0.4032
種目	0.6038
野口	-0.1797
銅メダル	0.2636
飯塚	-0.1072
新谷	-0.3811
山縣	0.4831
棄権	0.6803
獲得	0.4014
今日	1.4515
室伏	-1.0024
失格	0.6134
スタート	0.9092
五輪	1.01
(-	0.9077
最高	0.4622
km	1.2845
+	0.5528

回帰分析結果　非線形回帰分析：RT有無

weka

4.03	daijapan	-1.35	西塔
1.26	通過	1.29	ウクライナ
1.45	今日	1.39	心配
-1.00	室伏	1.19	事
1.01	五輪	1.26	本日
1.28	km	1.70	拓己
-1.68	4×100m	1.05	解説
1.01	仁美	-1.08	顔文字
1.47	良子	1.19	URL
1.48	mr		

圧倒的

西塔拓己
名字と名前
で正負逆

4×100m
正から負に

mrはリレー
競技を指す

決定木

- 分岐する過程を階層化，樹形図で表したグラフ
- 根に近いものがより影響力を持つ

- 分類木

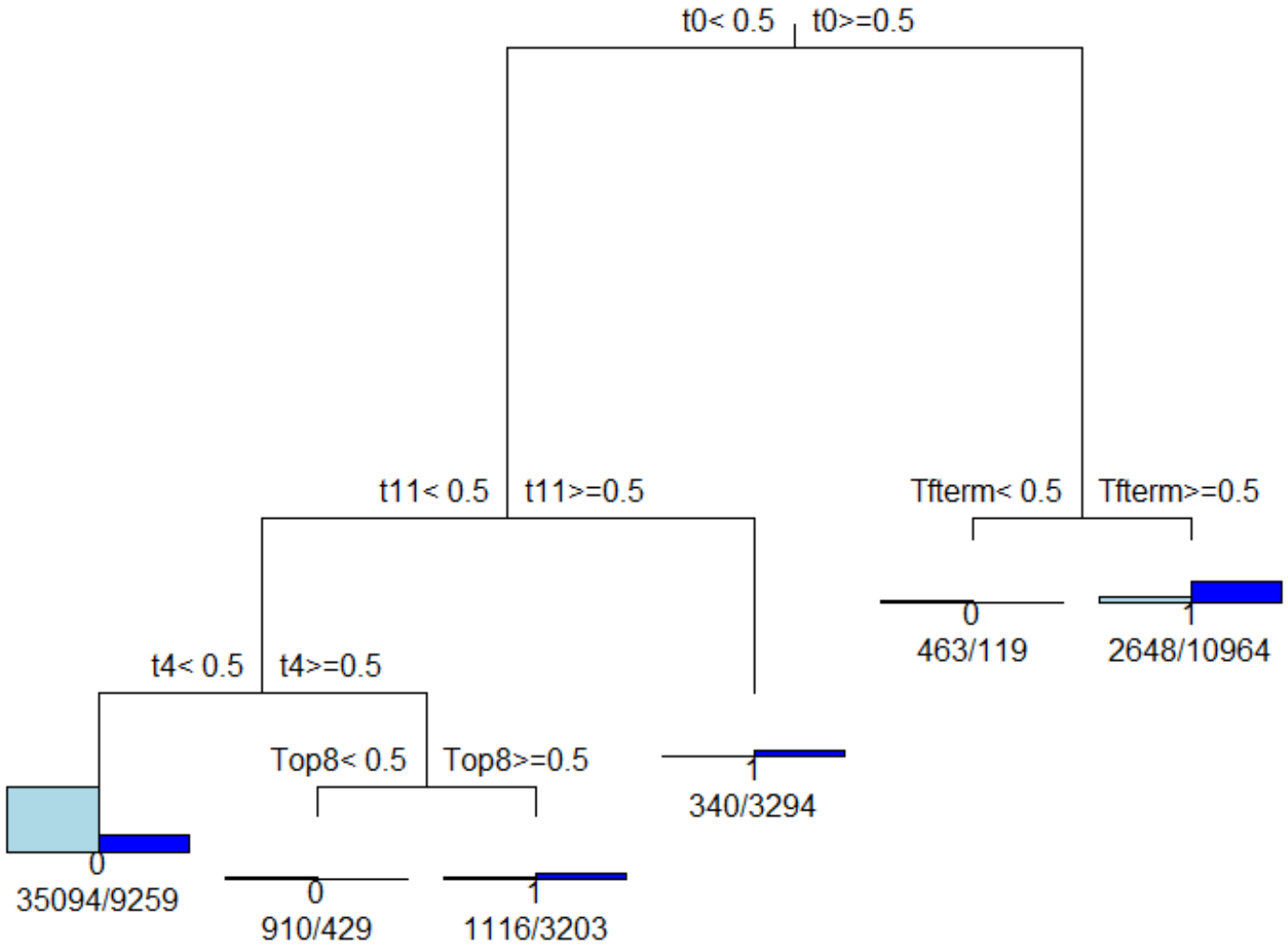
  目的属性　⇒　カテゴリー型（リツイートの有無）

- 回帰木

  目的属性　⇒　数値型（リツイートの回数）

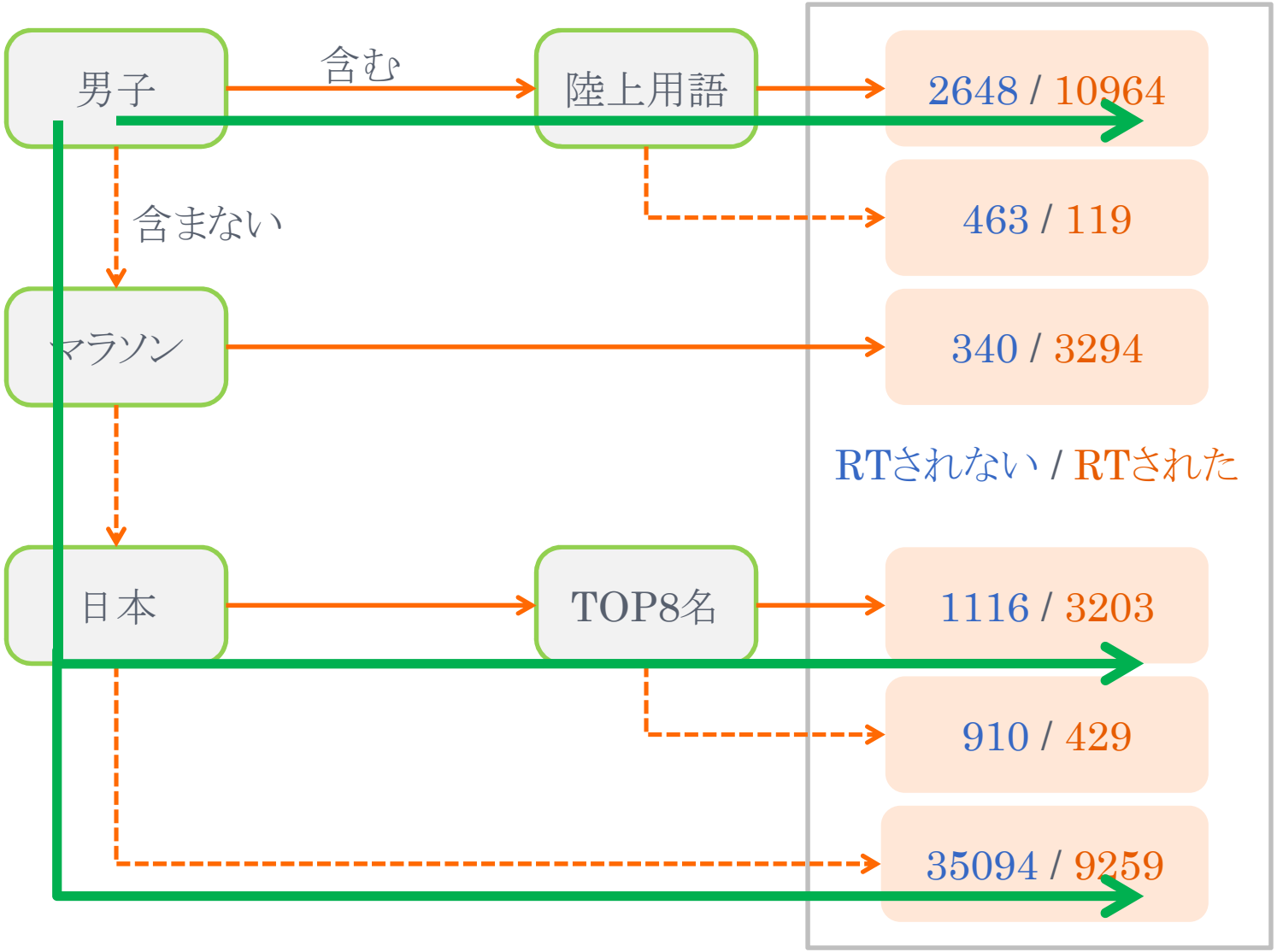
R言語

決定木結果　分類木：RT有無，されたTWEET数



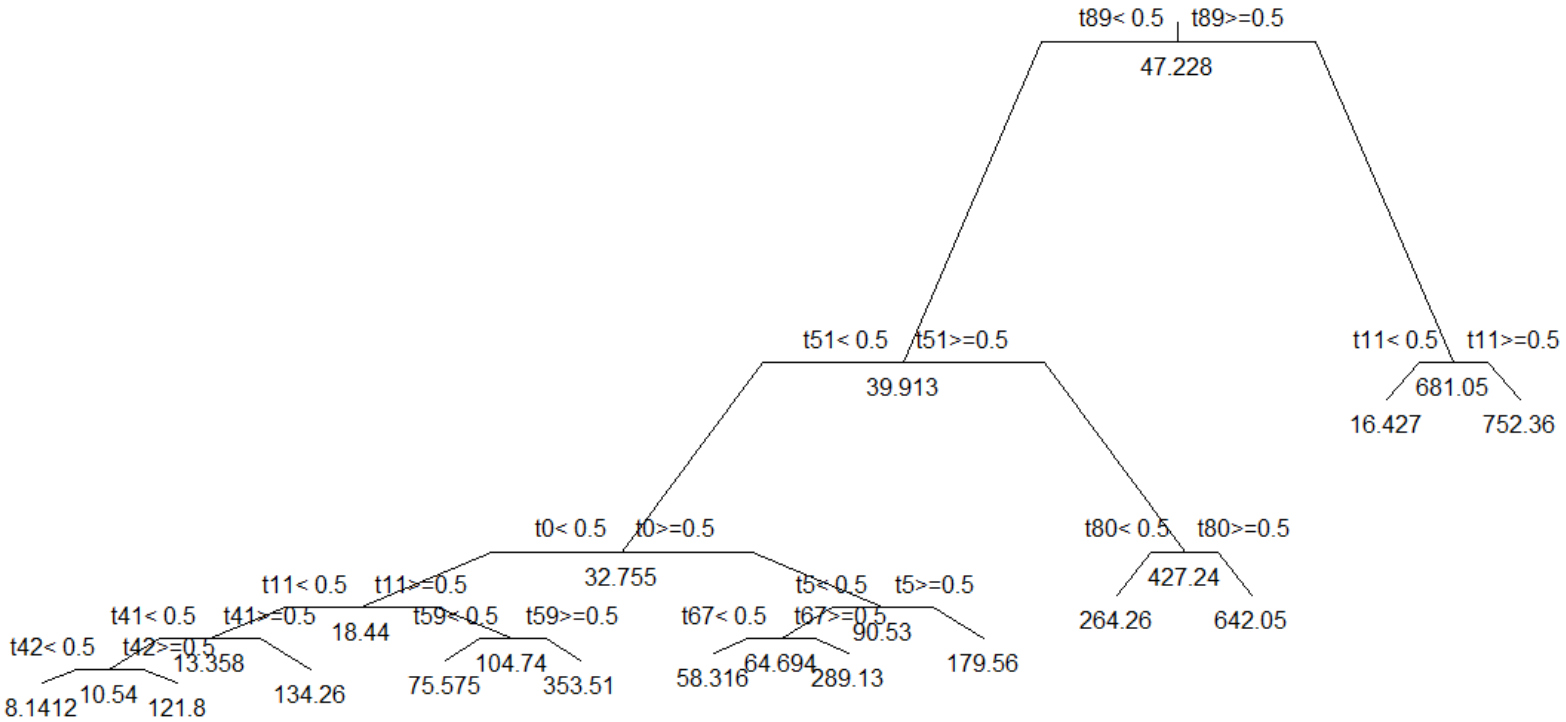
R言語

決定木結果　分類木：RT有無，されたTWEET数



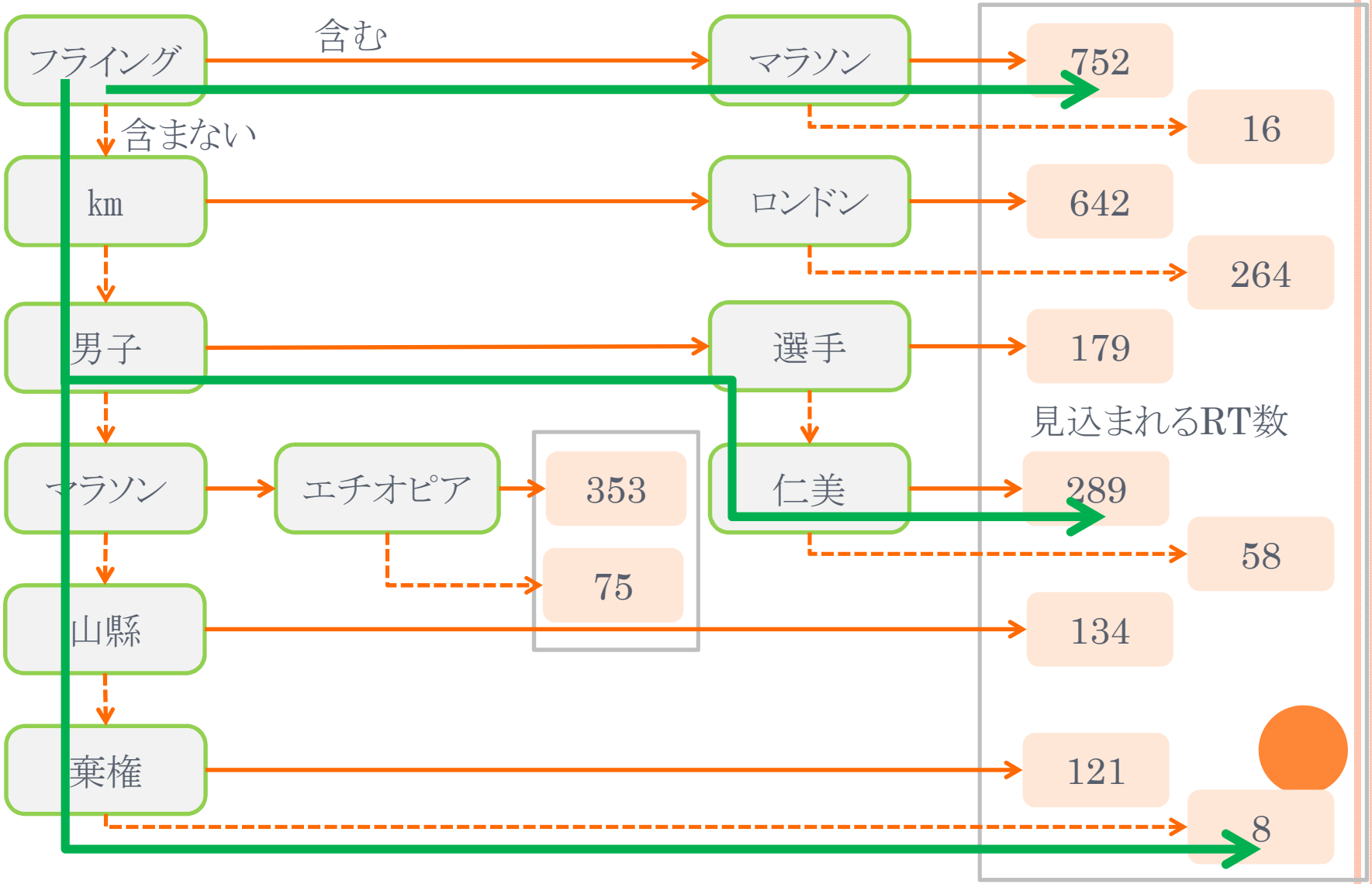
R言語

決定木結果　回帰木:RT数

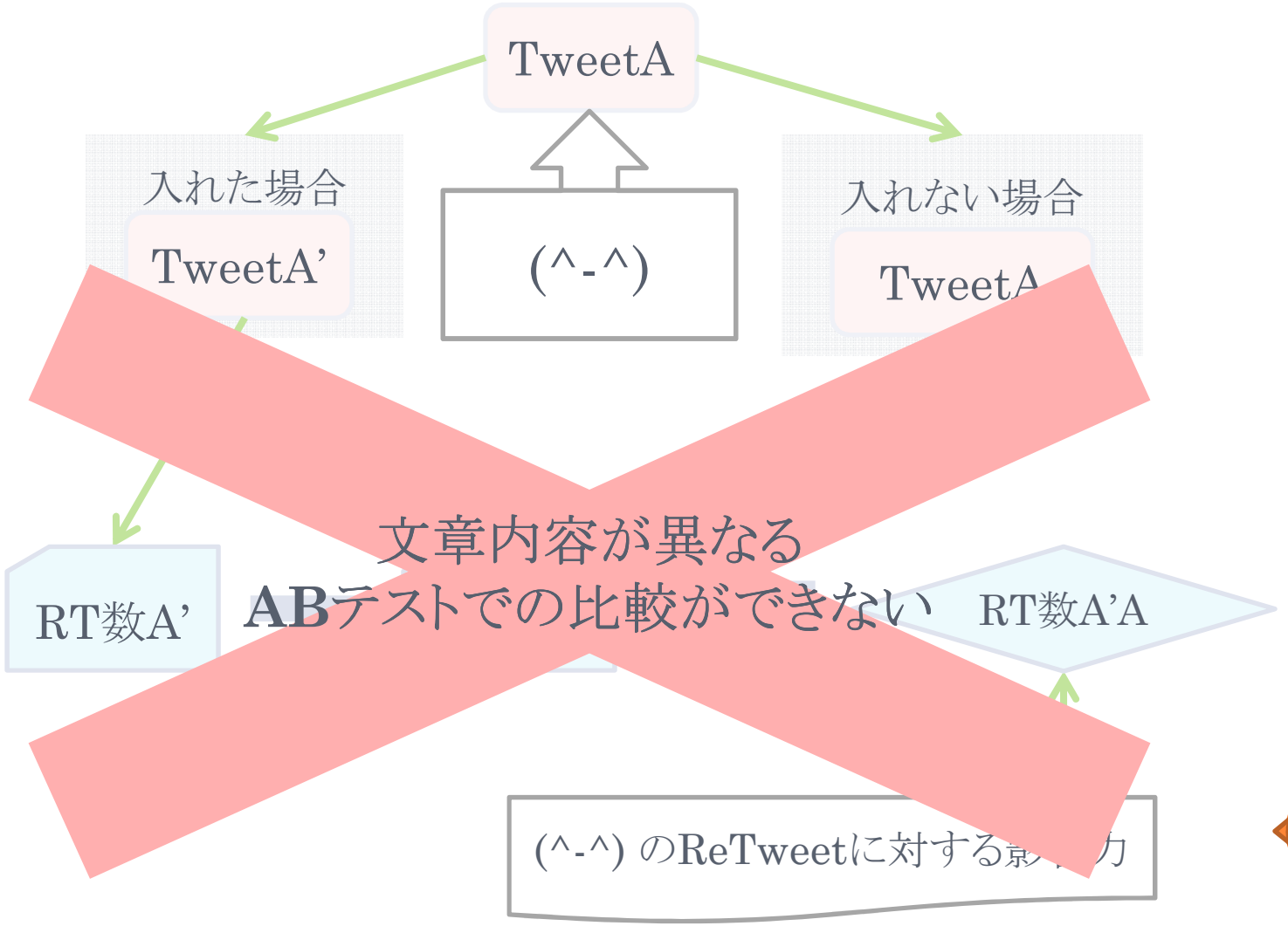


R言語

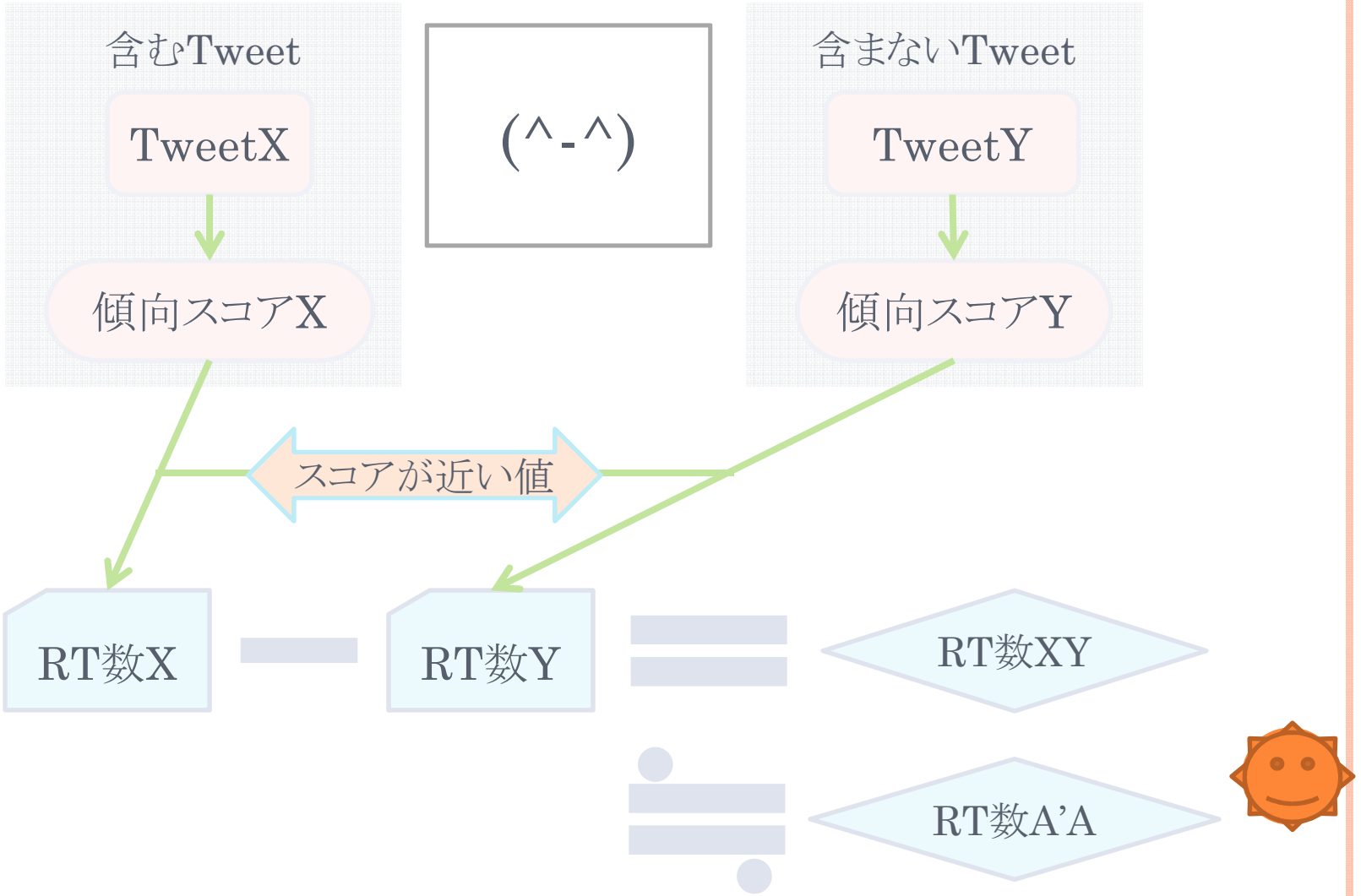
決定木結果　回帰木:RT数



傾向スコア　具体図：ABテスト



傾向スコア　具体図：観測データからの因果関係の導出



傾向スコア分析結果　RT数

221	マラソン	-6	新
272	入賞	417	ロンドン
205	金	-8	末
306	川内	681	フライング
222	野口	-3	笑
426	km	-6	顔文字
-2	自己	-0	怒
-4	最終	-1	哀
201	仁美	622	恐
384	mr		

次回開催地

ルール改正
前大会ボルト
の失格

手に汗握る
鳥肌，がくぶる
震え，どきはら
心配，冷や汗

考察

- 顔文字，感情語が少ない
  —出現数，影響，辞書内の数

- ReTweetの有無と回数に同じ働きをしない
- 専門用語よりも頻出語
- 同義語も表記の差で変わる

- 人名が有効とは限らない
- 日本人が活躍する競技は正負の要素に入る

まとめと今後の課題

- 日本人が関係するものは頻出としても多い
- この単語ならReTweetされ，かつ伸びるとは言えない
- 分析手法によって正負が逆転する場合も存在する

- 辞書，判定の見直し
- データセット，分析の見直し
- 分析結果の正確度を調査
- 他条件を考慮してみる
- 題材を他の分野