ジュースに関する特許のRを用いたテキストマイニングによる出願人解析

(株)三菱ケミカルリサーチ
北川　道成

はじめに

- ＩＰランドスケープと言われる特許情報の可視化が盛んにおこなわれるようになってきている。
- 可視化にはテキストマイニング(テキストアナリシス)の手法が用いられる。
- テキストマイニングはテキストを数値化したのちにデータマイニングの手法で解析を行う。
- データマイニングには様々な手法がある。
- 本検討では統計解析ソフトＲを用いて、ジュースの特許を解析し出願人ごとの傾向を可視化した。
- 特許情報の数値化の際の重みづけ、解析対象単語選択による可視化への影響を検討した。

解析の準備１－検索

- 検索
  - 対象公報：公開系公報
    - 公開 1993年～2018年02月01日
    - 公表 1993年～2018年02月01日
    - 再公表 1993年～2018年02月01日
  - 検索式
    - A23L2/02/ANY IPC
      AND (アサヒ+伊藤園+カゴメ+キリン+ コカ・コーラ+ +サッポロ+サントリー+デルモンテ+ポッカ+ヤクルト)/ 出願人・権利者
      AND　19930101:/公報発行日
      →465件

解析の準備２－データ加工

- 共願の扱い
  - それぞれの会社に1件として割り振った
    - 特開2013-188191：アサヒビール(株)、カゴメ(株)
    - 特開2007-089433：伊藤園・日本デルモンテ
- 権利者変更の扱い
  - 現在ジュースメーカー以外の出願人は元出願人として扱う
    - 特開2006-166880：現在は三菱重工だが、元はポッカと東洋製作所共願→サッポロ・ポッカとして扱う
- 合計467公報(2公報は上の共願による重複)を解析

解析の準備３－名寄せ

- 出願人表記、企業統合が有ったため以下の通り名寄せを行った。
  - アサヒ：　アサヒビール(株)、アサヒ飲料(株)、アサヒグループホールディングス(株)、アサヒグループ食品→66件
  - 伊藤園：(株) 伊藤園→83件
  - カゴメ：カゴメ(株) →84件
  - キリン：キリン(株)、キリン・トロピカーナ(株)、キリンホールディングス(株)、キリンフードテック、キリン協和フーズ(株)→39件
  - コカ・コーラ：ザコカ・コーラカンパニー、日本コカ・コーラ(株)→39件
  - サッポロ：ポッカサッポロフード＆ビバレッジ(株)、サッポロビール(株)→36件
  - サントリー：サントリーホールディング(株)、サントリー食品インターナショナル(株)、サントリー酒類(株)→63件
  - デルモンテ：日本デルモンテ(株)→33件
  - ヤクルト：(株)ヤクルト本社、ヤクルトヘルスフーズ(株)→24件

解析の準備４－txtファイルに書き出し

- Excelのマクロを用い、要約を1公報毎1ファイルに書き出した。
  - ファイル名：出願人略称番号.txt
    - (例)アサ01.txt
      →アサヒビールの1番目の公報の要約
- 作成したtxtファイルをRMeCabで形態素解析を行い単語－文書行列を作成した。

解析

- 作成したtxtファイルをRMeCabで形態素解析
- 解析対象品詞は以下の通り
  - POS1=名詞,形容詞,動詞
  - POS2=一般,固有名詞,自立
- 以下単語は解析から除外した（stopワードの設定）
  - １つ|一つ|ある|する|できる|ない|なる|よい|よる|易い|該|含む|好ましい|手段|課題|目的|与える|工程|有す|優れる|得る|特徴|用いる|感じる|方法|範囲|上記|量|良い|選ぶ|係る|条件下|値|所定|示す|行う|ｍｇ|ｋｇ|ｐｐｂ|Ｌ
- 形態素解析の単語出現頻度(スコア)を出願人毎に合計し、合計頻度（スコア）の上位1-100位/5-100位の単語を解析対象とした。
- 解析は以下の解析を行った
  - ワードクラウド
  - 多次元尺度法
    - 距離のパラメータ：
      "euclidean","maximum","manhattan","canberra","binary","minkowski"
  - 主成分分析
  - クラスタリング(k-means法)

検討事項

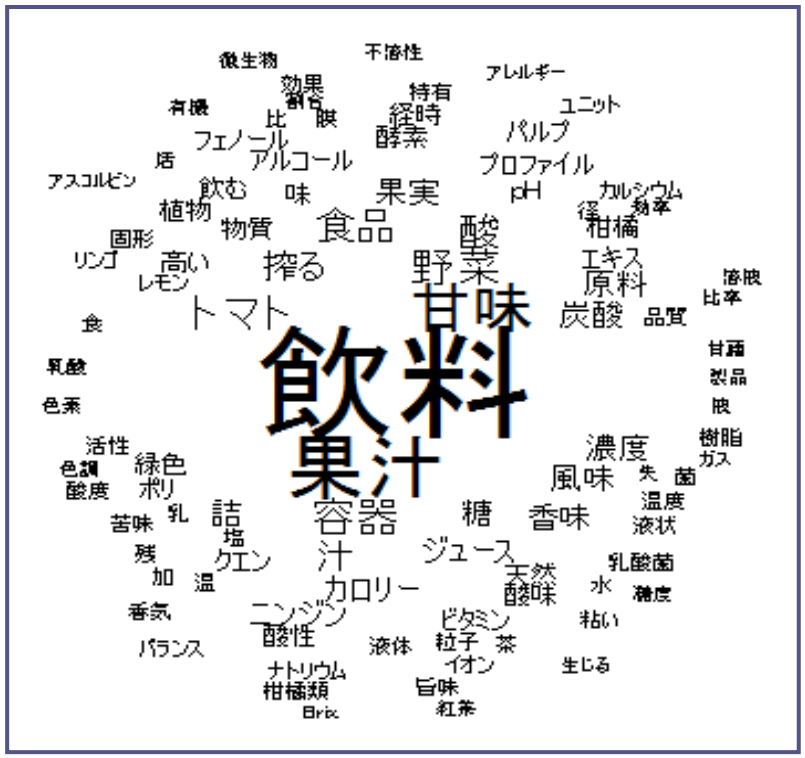
以下の事項を検討した。

- 単語の重み付けの影響
  - 重み付けなし
  - tf*idf
- 解析単語設定方法
  - 上位1-100位
  - 上位5-100位
- 多次元尺度法プロット時の距離のパラメータ
  - euclidean
  - maximum
  - manhattan
  - canberra
  - binary
  - minkowski

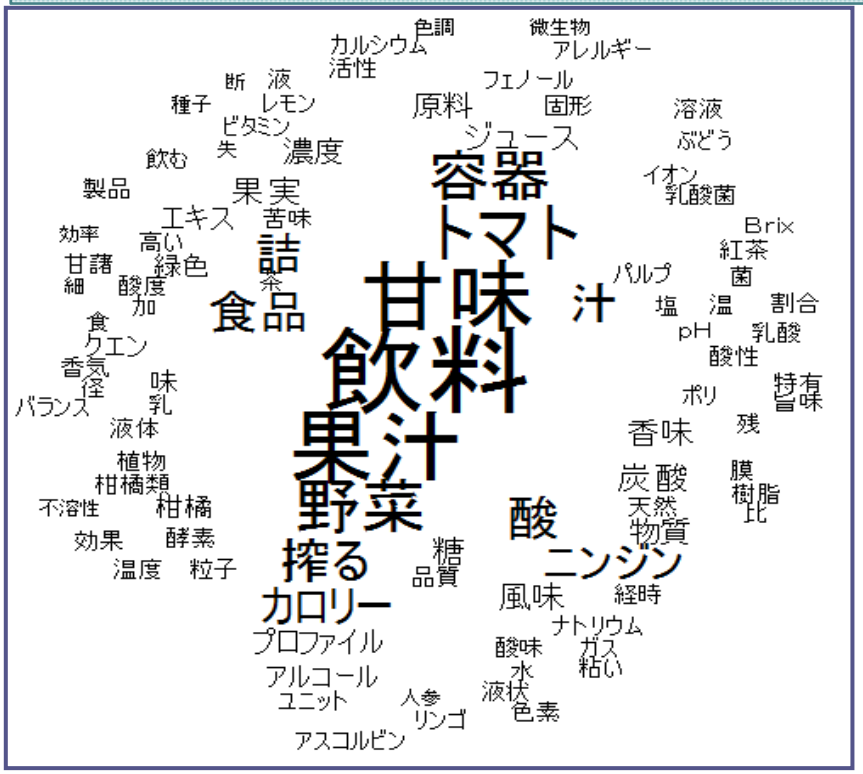
ワードクラウド
全出願人の合計

ワードクラウド
全出願人合計−1-100位

重み付けなし



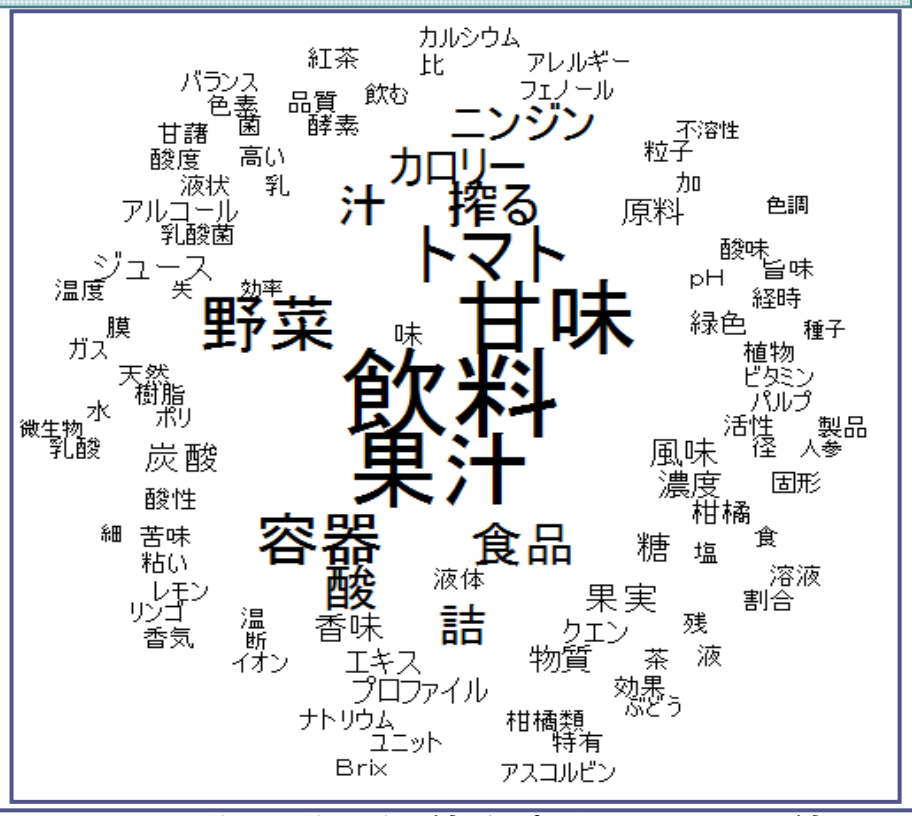
tf*idf



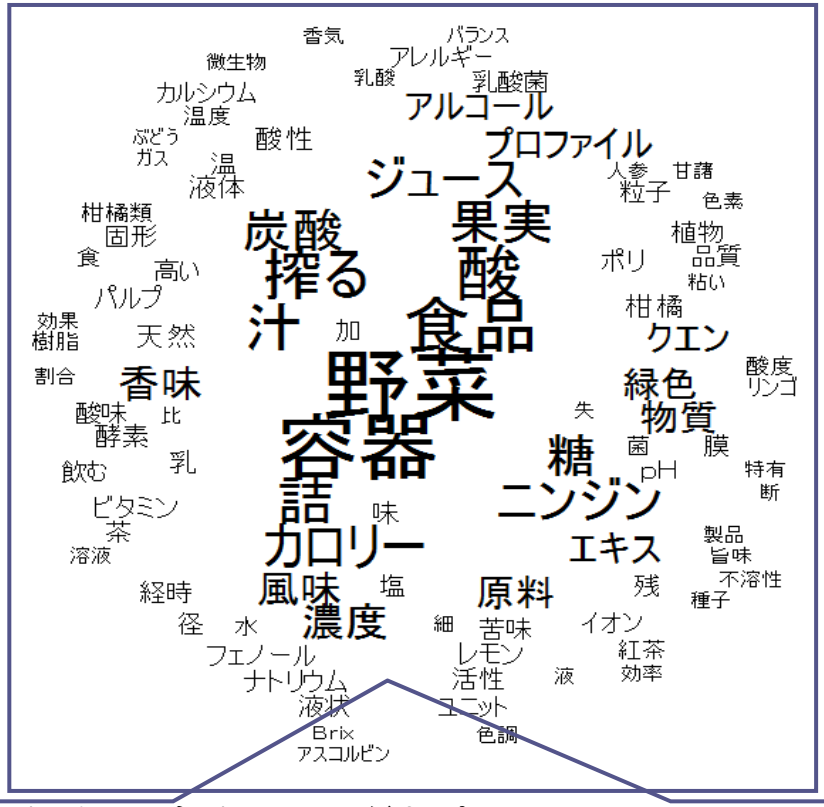
tf*idfの重み付けでは
多くの公報で共通的に使われる「飲料」の大きさが相対的に小さくなり、
特定の公報に出現する「野菜」、「甘味」、「ニンジン」、「搾る」、「カロリー」、「酸」等が強調され
以後はtf*idfの重み付けで解析を行った。

ワードクラウド
全出願人合計–tf*idf

1-100位



5-100位

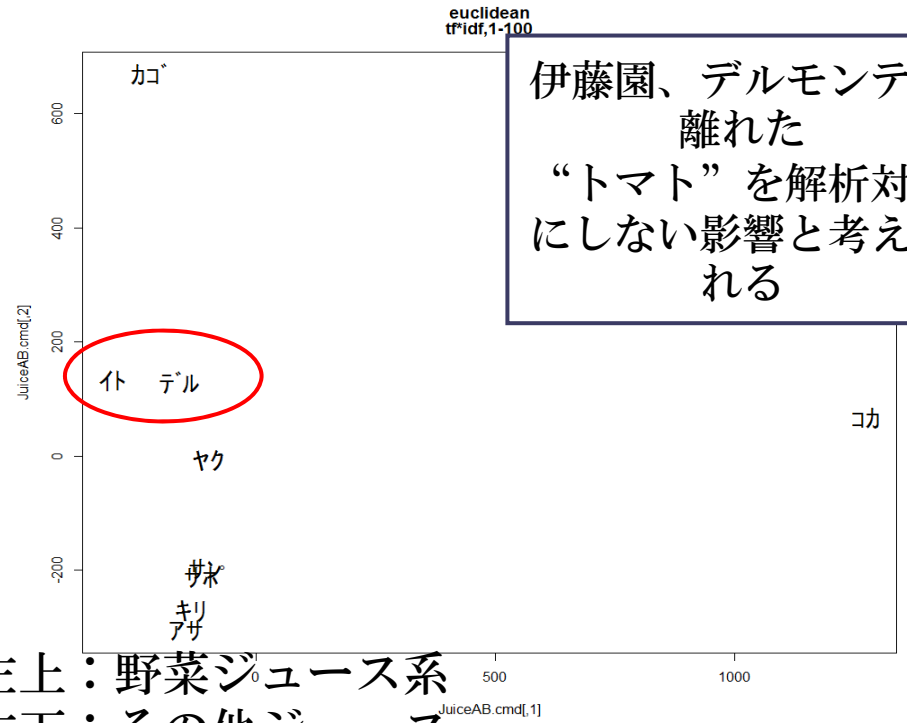


tf*idfではトマトが4位となり、5-100位のワードクラウドには現れない
飲料 果汁 甘味 トマト 野菜 容器 食品 酸 搾る汁
1860 1502 1486 1199 1067 923 784 762 724 710
「トマト」は今回の解析には重要なキーワードであり、
1-100位と5-100位では結果に影響すると考え、多次元尺度法プロットで比較を行う。

多次元尺度法でのプロット
解析単語選択・距離算出法の影響

多次元尺度法–解析単語選択の影響
tf*idf euclidean

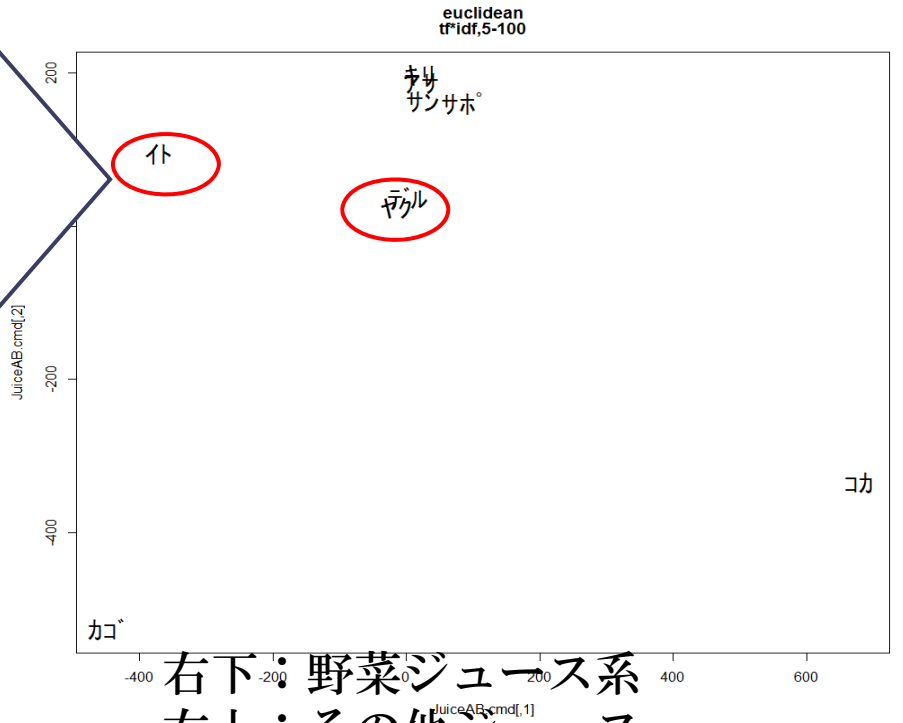
1-100位



伊藤園、デルモンテが離れた
“トマト”を解析対象にしない影響と考えられる

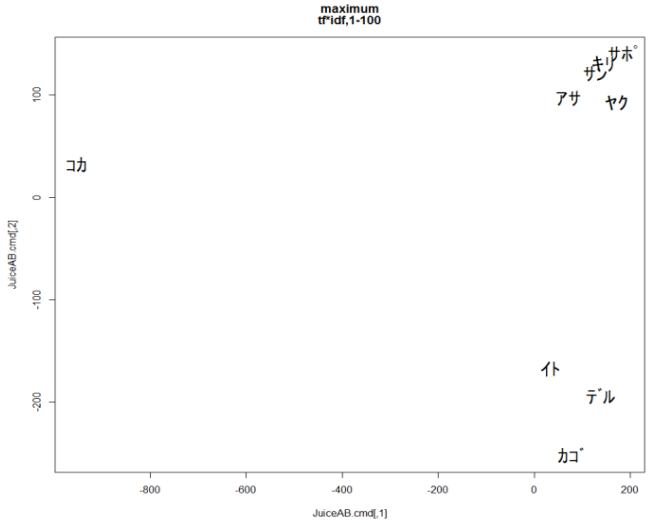
左上：野菜ジュース系
左下：その他ジュース
右：コカコーラ

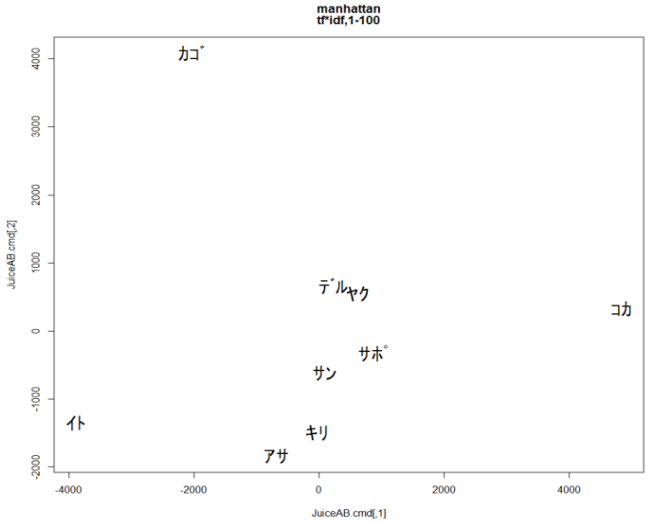
5-100位

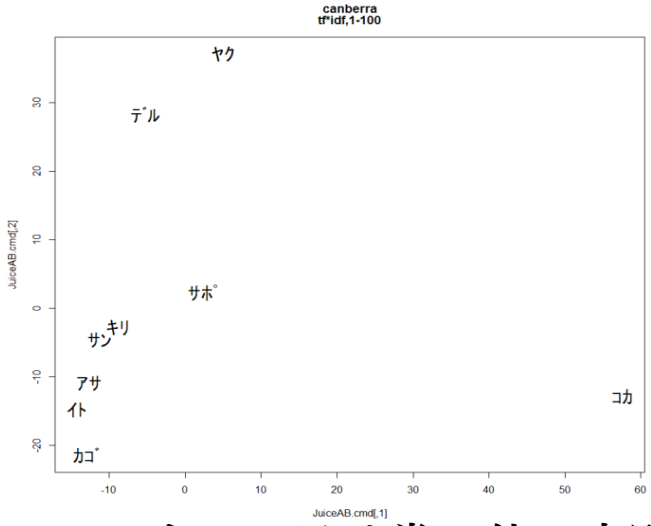


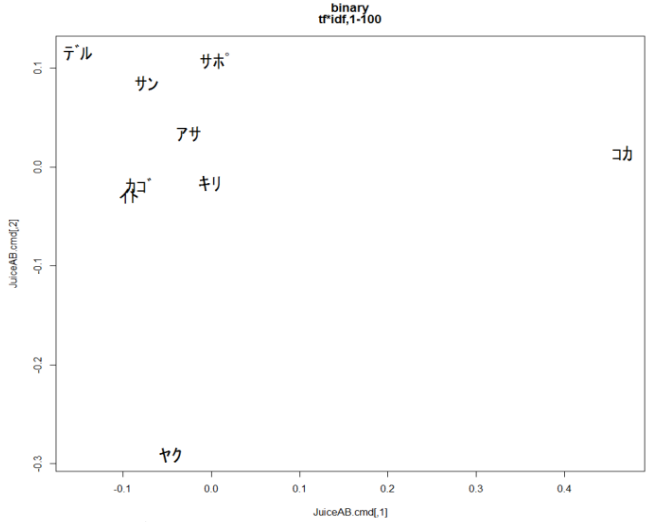
右下：野菜ジュース系
右上：その他ジュース
右：コカコーラ

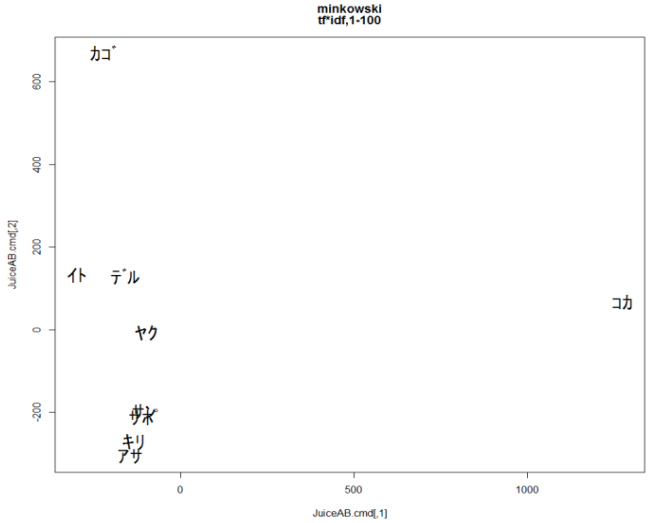
多次元尺度法–距離算出法依存性
tf*idf 1-100位





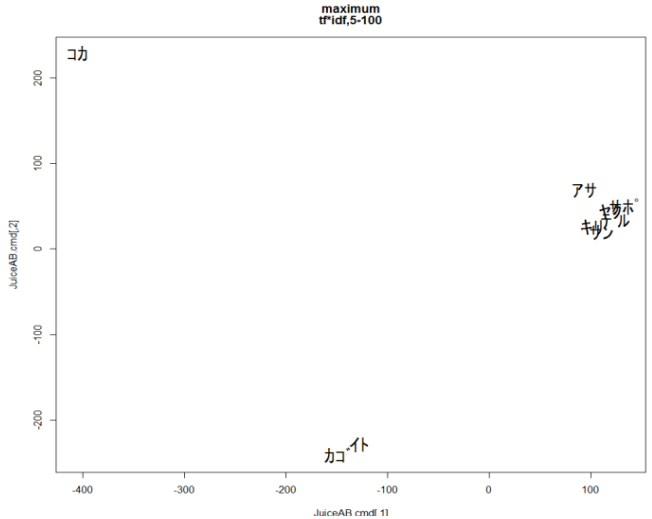


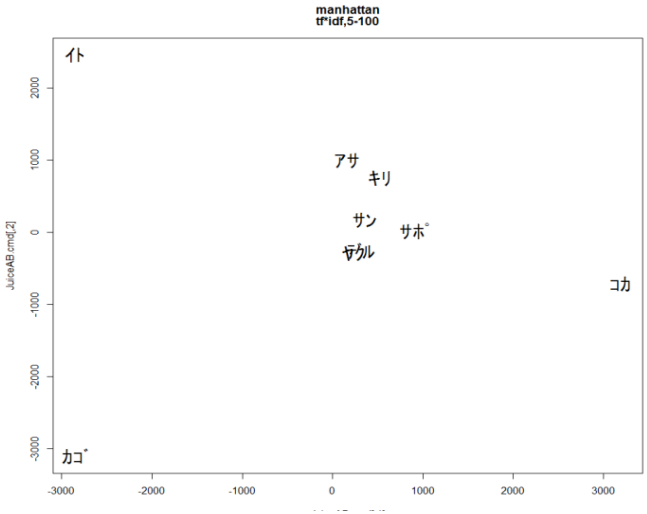


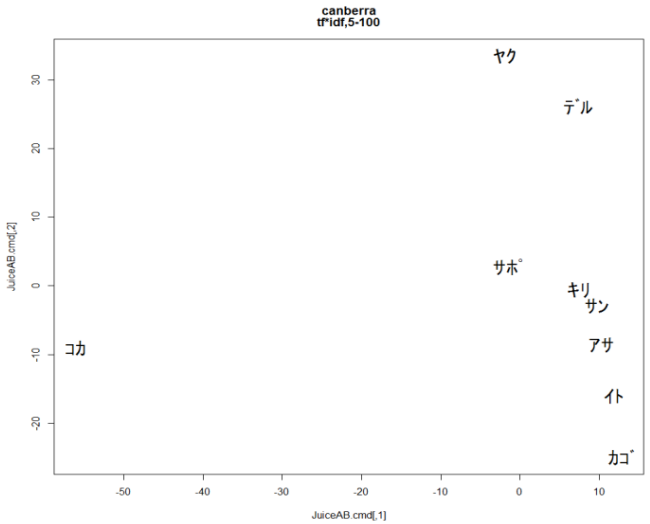


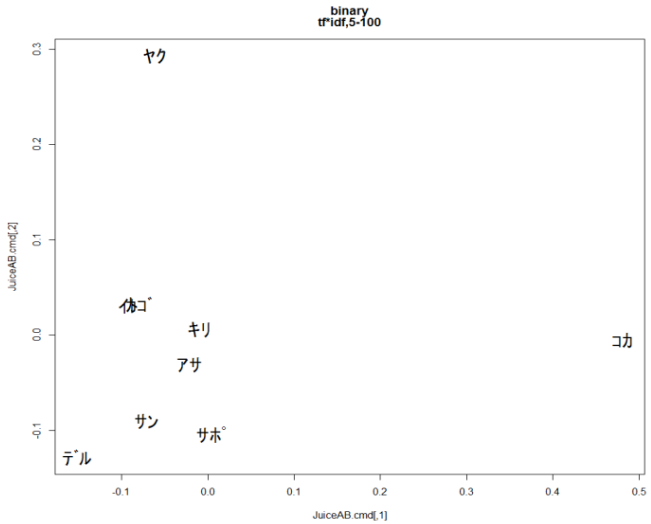
コカコーラは常に他の出願人と離れて位置している。

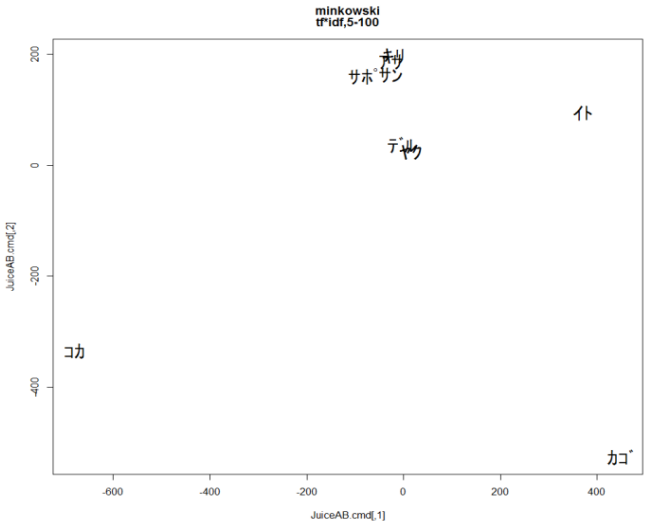
多次元尺度法–距離算出法依存性
tf*idf 5-100位







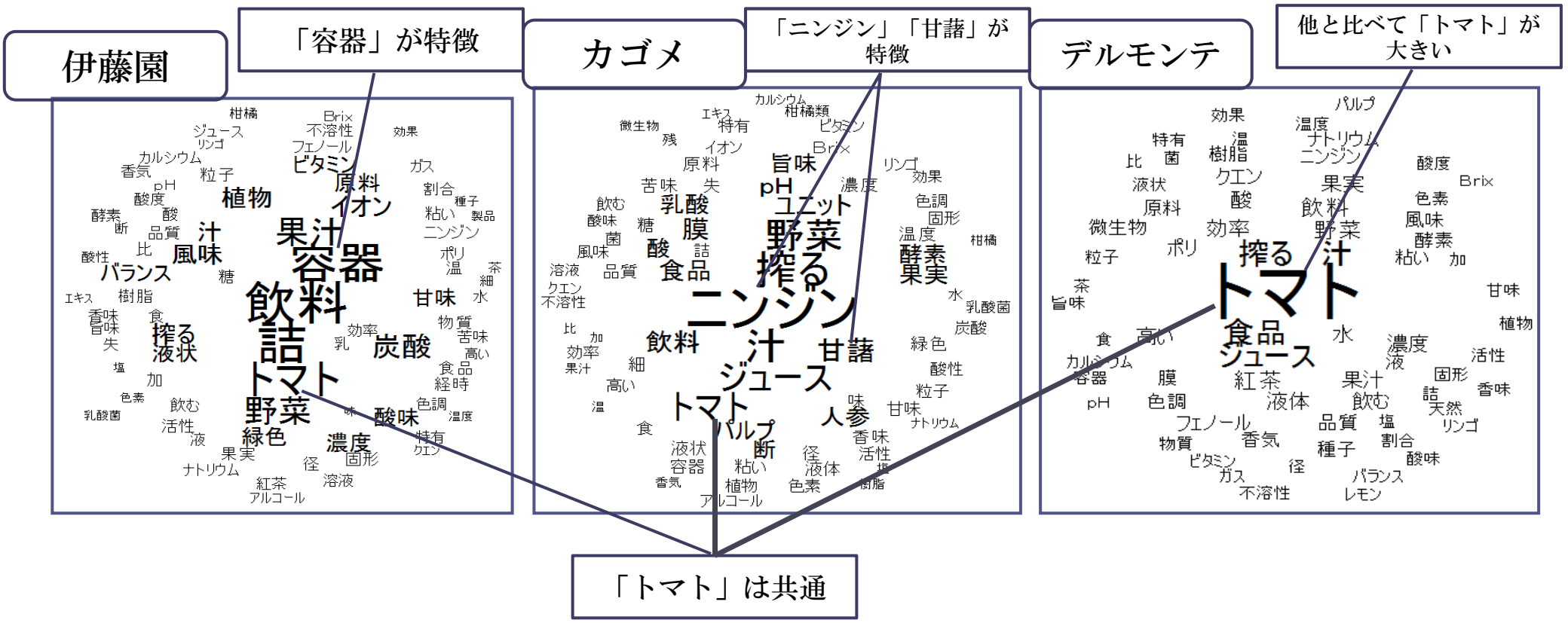




コカコーラは常に他の出願人と離れて位置している。

ワードクラウド
出願人毎の傾向

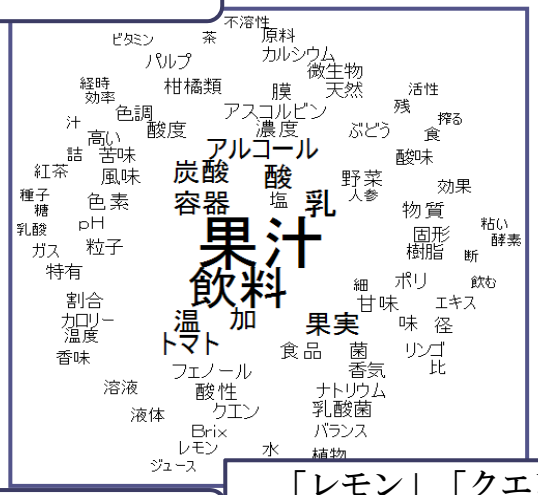
ワードクラウド–トマトジュース系 1-100位–tf*idf



伊藤園の容器は「容器詰飲料」が形態素解析で切れた単語。

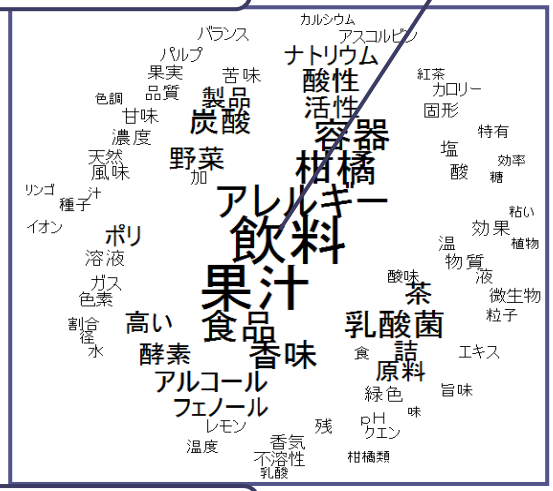
ワードクラウド–非野菜ジュース系 1-100位–tf*idf

アサヒ



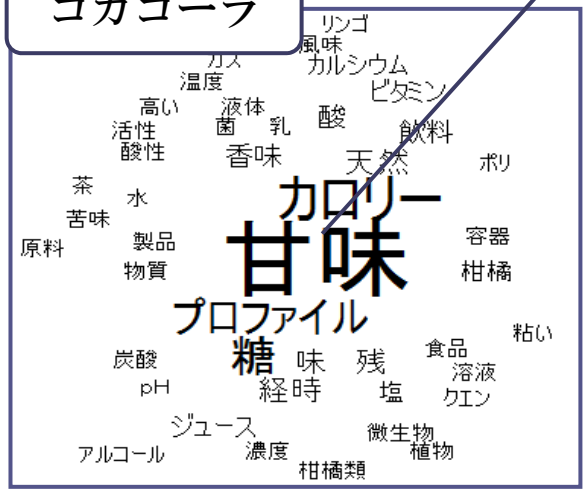
キリン

「アレルギー」が特徴的



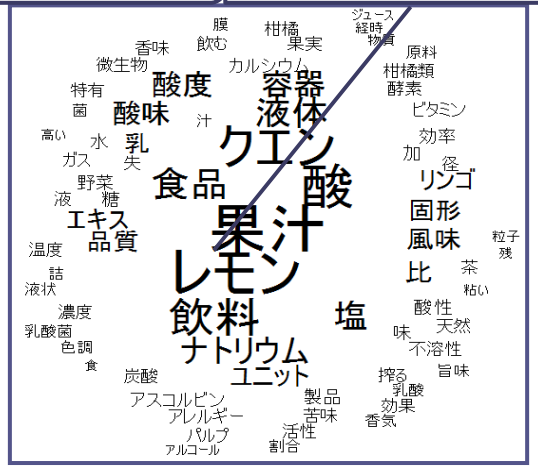
コカコーラ

「甘味」「カロリー」「糖」が特徴的



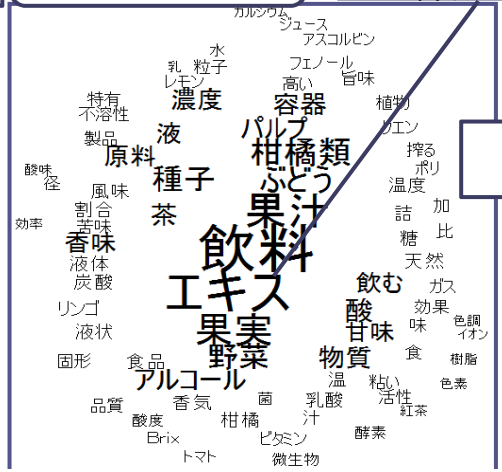
サッポロ

「レモン」「クエン（酸）」が特徴的。ポッカを買収。



サントリー

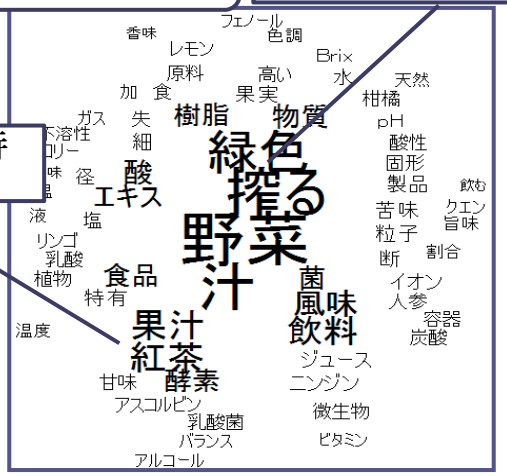
「エキス」が特徴的。



「紅茶」が特徴的。

ヤクルト

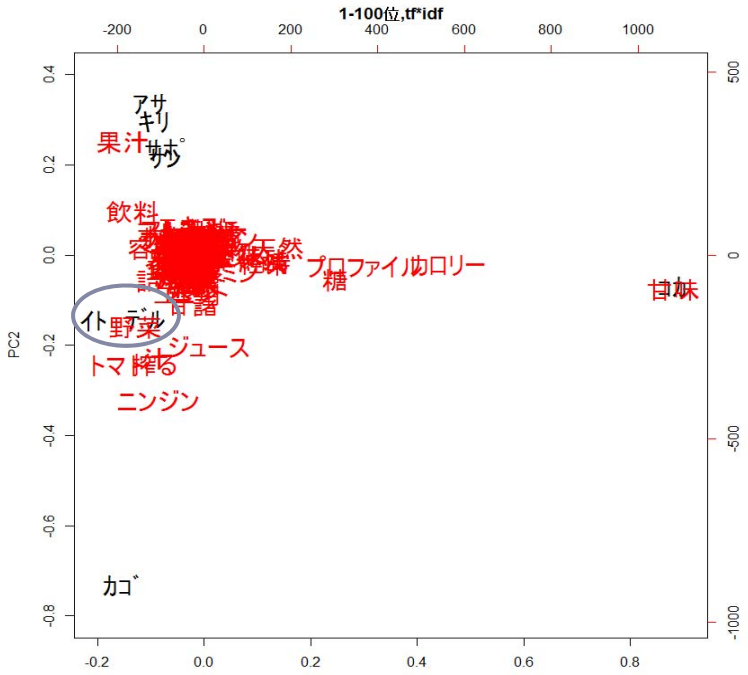
「トマト」がないが、野菜ジュースを販売



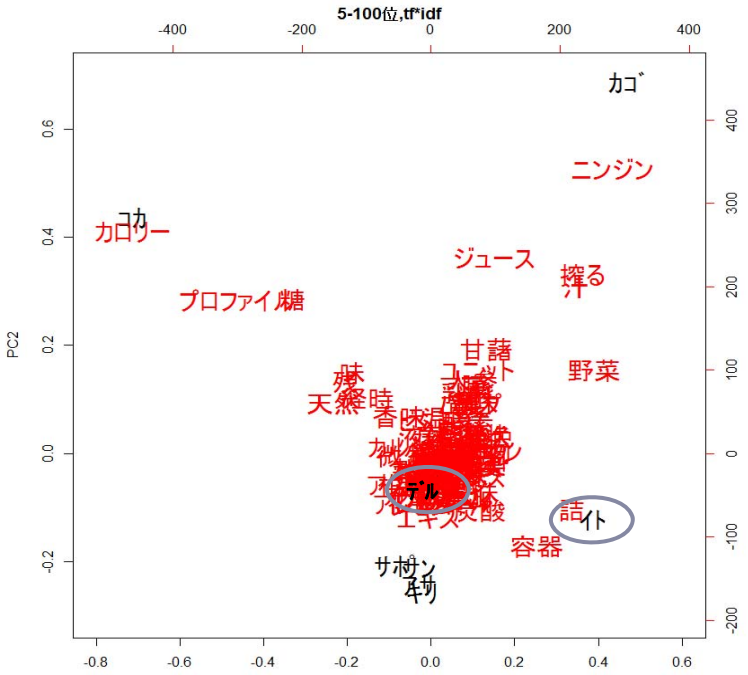
主成分分析
解析単語選択の影響

主成分分析–tf*idf

1-100位



5-100位

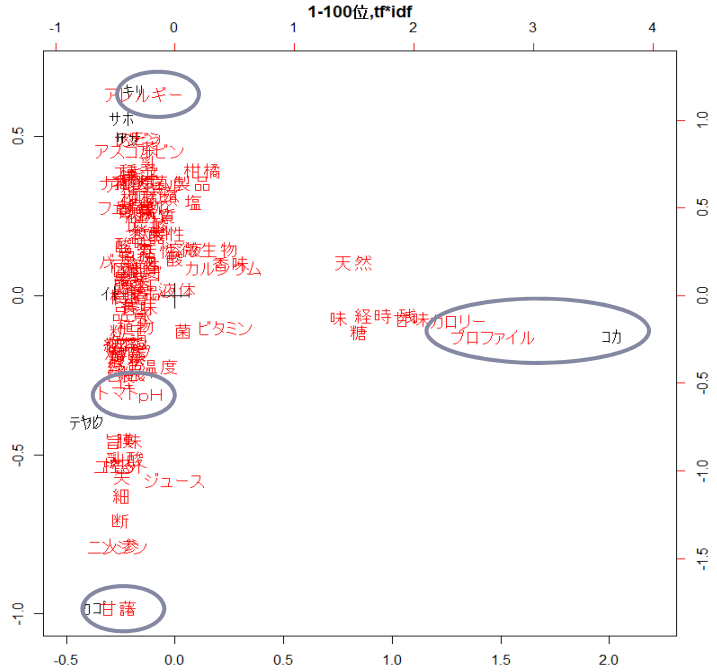


甘味はコカコーラ。野菜、トマト、ニンジンは伊藤園、デルモンテ、カゴメ。
果汁はアサヒ、キリン、サッポロ、サントリー。

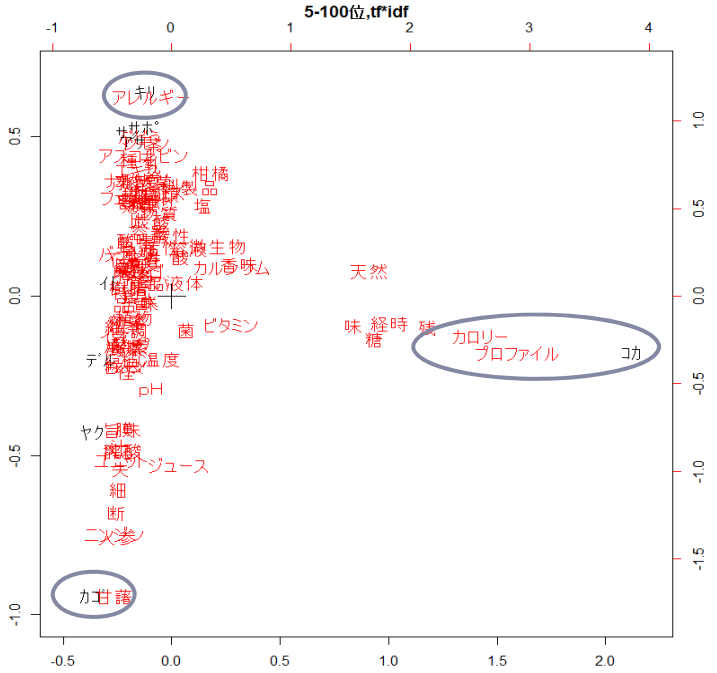
対応分析
解析単語選択の影響

対応分析−tf*idf

1-100位



5-100位



キリン：アレルギー、カゴメ：甘藷、コカコーラ：カロリー
1-100位、5-100位ともに傾向は同じ。

K-means法
1-100位

- 中心数=３でクラスタリング
  1. 伊藤園　カゴメ
  2. コカコーラ
  3. アサヒ　キリン　サッポロ サントリー　デルモンテ ヤクルト

- 中心数=２でクラスタリング
  1. コカコーラ
  2. アサヒ　伊藤園　カゴメ キリン　サッポロ サントリー　デルモンテ ヤクルト
- 中心数=５でクラスタリング
  1. アサヒ キリン サントリー
  2. サッポロ デルモンテ ヤクルト
  3. コカコーラ
  4. カゴメ
  5. 伊藤園

K-means法
5-100位

- 中心数=３でクラスタリング
  1. アサヒ　キリン　コカコーラ　サッポロ サントリー　デルモンテ ヤクルト
  2. 伊藤園
  3. カゴメ

- 中心数=２でクラスタリング
  1. アサヒ　キリン　コカコーラ　サッポロ サントリー　デルモンテ ヤクルト
  2. 伊藤園　カゴメ
- 中心数=５でクラスタリング
  1. サッポロ デルモンテ ヤクルト
  2. コカコーラ
  3. アサヒ キリン サントリー
  4. カゴメ
  5. 伊藤園

まとめ

- 今回はジュースの公開特許公報の要約を用いて形態素解析後の文書単語行列作成の際の重み付け、単語選択が出願人の解析に及ぼす影響に付いて検討した。
- 頻度分布で上位に現れる語は共通に現れる語のため特徴を表さないと考え除くと、単語によっては解析に大きな影響を及ぼす事が解った。
- 多次元尺度構成法、k-means法等解釈が難しい結果もあるが、今後は解釈を追加していきたい。
- 重み付けに関しては他にもいろいろな手法が挙げられている、これらについて今後さらなる検討をしていきたい。