テキストマイニング

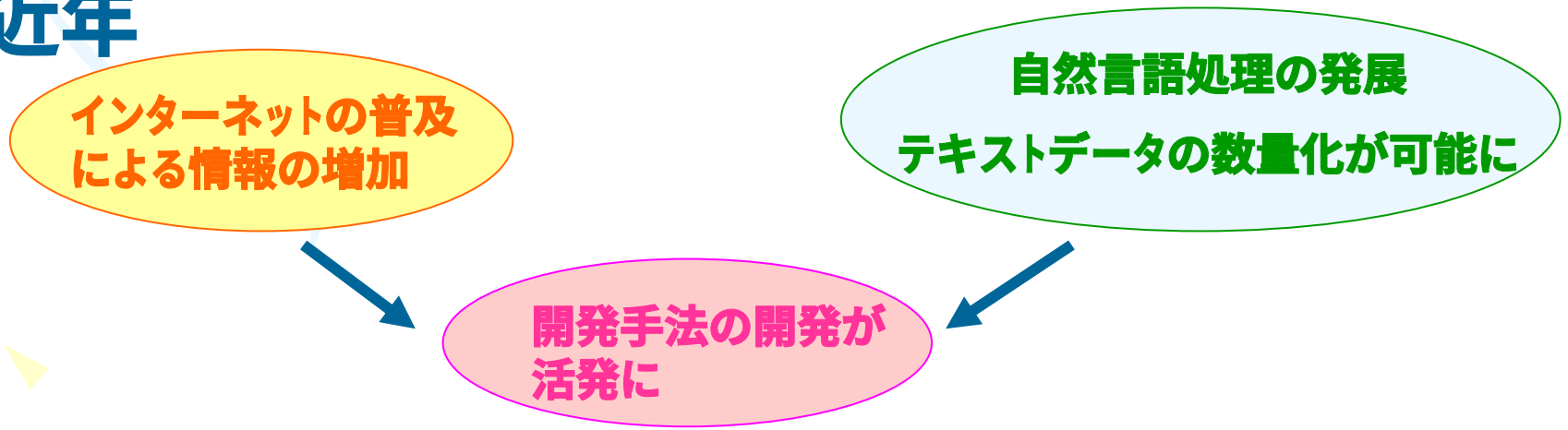
心理データ解析演習

2008/1/16

M1 山添愛

テキストマイニングとは？

- テキストデータを分析し、分析者にとって有益な知識や情報を取り出す技術
  - 主に企業による消費者調査の分析等に用いられる
    - Web上に書き込まれた情報
    - 自由記述式アンケート
    - コールセンターに集まる消費者の意見
- 近年



テキストマイニングとは？

1. **情報の抽出**
   - なるべく少ないノイズで必要な情報だけを集める
2. **抽出した情報の解析**
   - 正しく考察・理解するための手法
3. **解析結果の可視化**

- **ソフトの例**
  - Document Broker(**日立製作所**) 700**万円～**
  - SymfoWARE Text Mining Sever(**富士通**) 250**万円～**

安くて簡単な方法

- ChaSen
  - 形態素解析のフリーソフト
  - 奈良先端科学技術大学院大学情報学研究科自然言語処理学講座　松本研究室
  - ダウンロード
    - ChaSen本体とWinCha(Windowsから利用するためのアプリケーション)

    http://www.ohmsha.co.jp/data/link/4-274-06493-X/index.htm
    - cha21244.exeクリック→実行→自己解凍プログラム
  - やってみよう：「ワインの良し悪し.xls」
    - ワインの良し悪しをどのようなところから判断されますか？

手順

Chasenによる形態素解析

↓

品詞情報をもとにキーワードを取り出す

↓

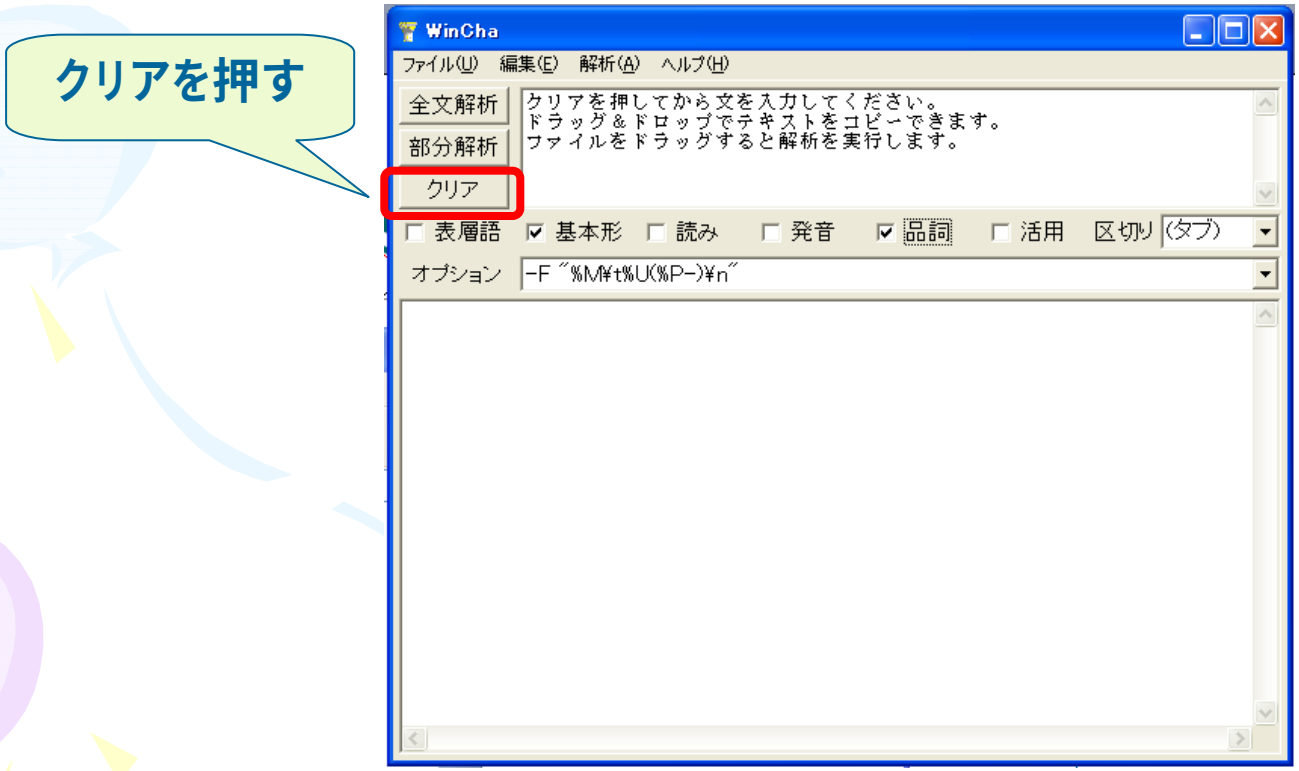
キーワードの出現頻度をヒストグラム化

↓

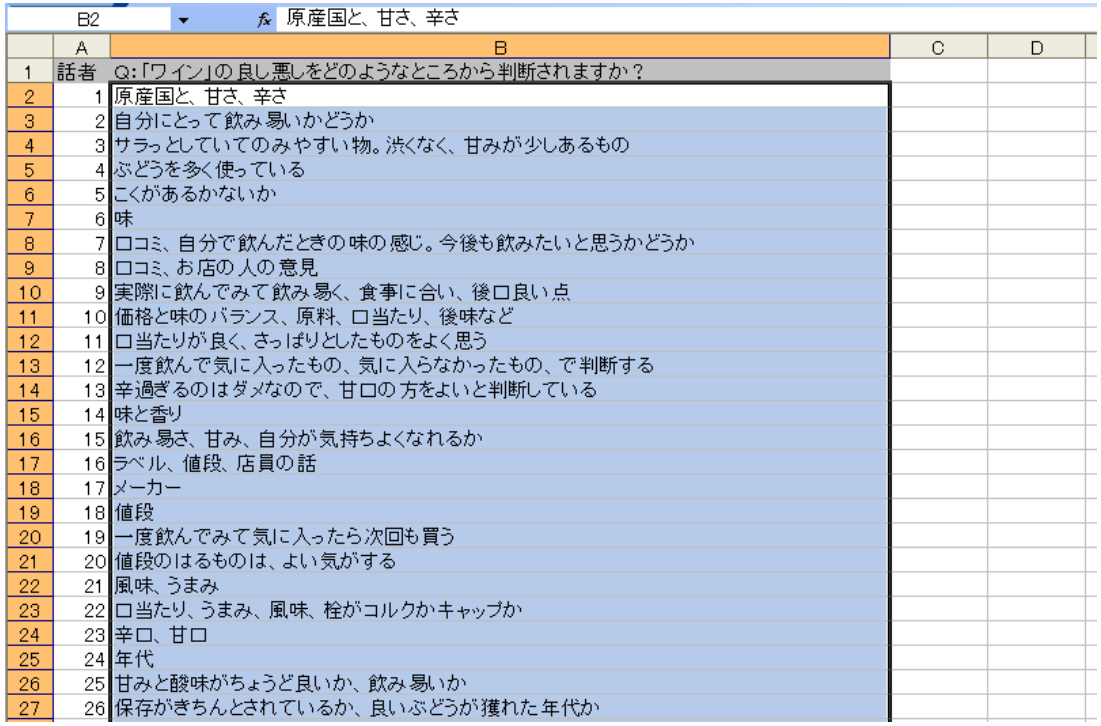
分析

ChaSenによる形態素分析

- ChaSen起動
- 「基本形」と「品詞」のみにチェック，区切りを「タブ」に



- 「ワインの良し悪し.xls」の解析したい部分「列B」を選択してコピー



B2 fx 原産国と、甘さ、辛さ

	A	B	C	D
1	話者	Q:「ワイン」の良し悪しをどのようなところから判断されますか？		
2	1	原産国と、甘さ、辛さ		
3	2	自分にとって飲み易いかどうか		
4	3	サラっとしていてのみやすい物。渋くなく、甘みが少しあるもの		
5	4	ぶどうを多く使っている		
6	5	こくがあるかないか		
7	6	味		
8	7	口コミ、自分で飲んだときの味の感じ。今後も飲みたいと思うかどうか		
9	8	口コミ、お店の人の意見		
10	9	実際に飲んでみて飲み易く、食事に合い、後口良い点		
11	10	価格と味のバランス、原料、口当たり、後味など		
12	11	口当たりが良く、さっぱりとしたものをよく思う		
13	12	一度飲んで気に入ったもの、気に入らなかったもの、で判断する		
14	13	辛過ぎるのはダメなので、甘口の方をよいと判断している		
15	14	味と香り		
16	15	飲み易さ、甘み、自分が気持ちよくなれるか		
17	16	ラベル、値段、店員の話		
18	17	メーカー		
19	18	値段		
20	19	一度飲んでみて気に入ったら次回も買う		
21	20	値段のはるものは、よい気がする		
22	21	風味、うまみ		
23	22	口当たり、うまみ、風味、栓がコルクかキャップか		
24	23	辛口、甘口		
25	24	年代		
26	25	甘みと酸味がちょうど良いか、飲み易いか		
27	26	保存がきちんとされているか、良いぶどうが獲れた年代か		

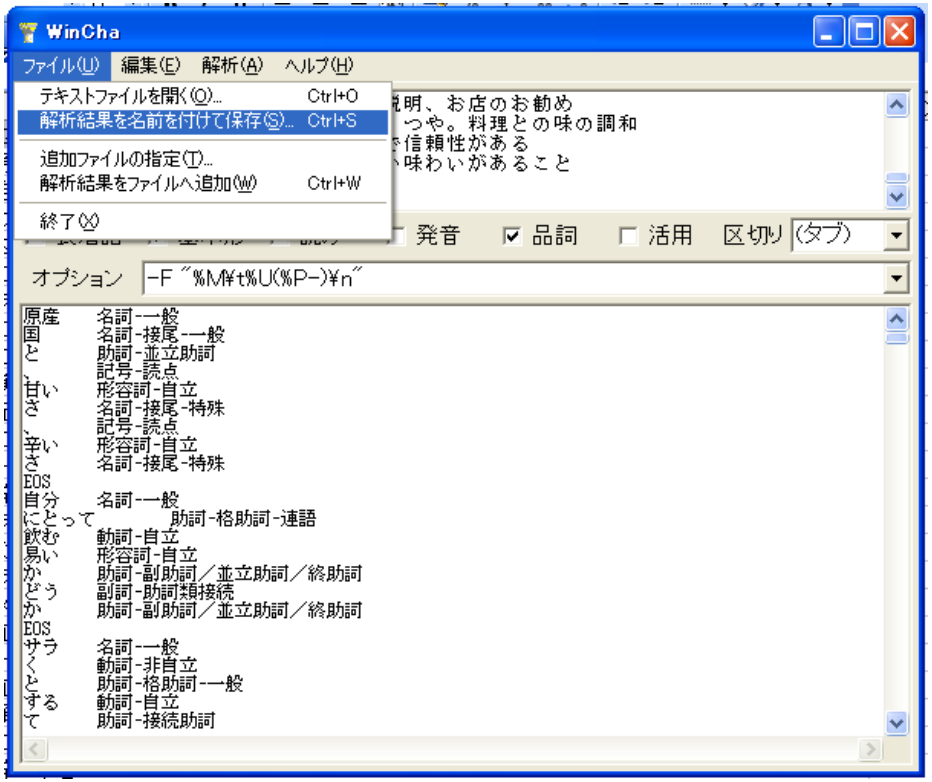
- WinChaのメニューから「編集」→「貼り付け」
- 「全文解析」ボタン

- **形態素に分割し，品詞情報が付加されたものが出力される**

一度，保存

Textファイル
で保存される

Excelで開いたものが
「ChaSen結果.xls」



品詞情報をもとにキーワードを取り出す

- 品詞情報の整理
  「データ」→「並べ替え」→
  「列B」を最優先
- 必要な品詞の抽出
  - 形容詞ー自立
  - 形容詞ー接尾
  - 形容詞ー非自立
  - 動詞ー自立
  - 副詞ー助詞類接続
  - 名詞ーサ変接続
  - 名詞ー一般
  - 名詞ー形容動詞語幹
  - 名詞ー固有名詞ー組織
  - 上記以外は削除

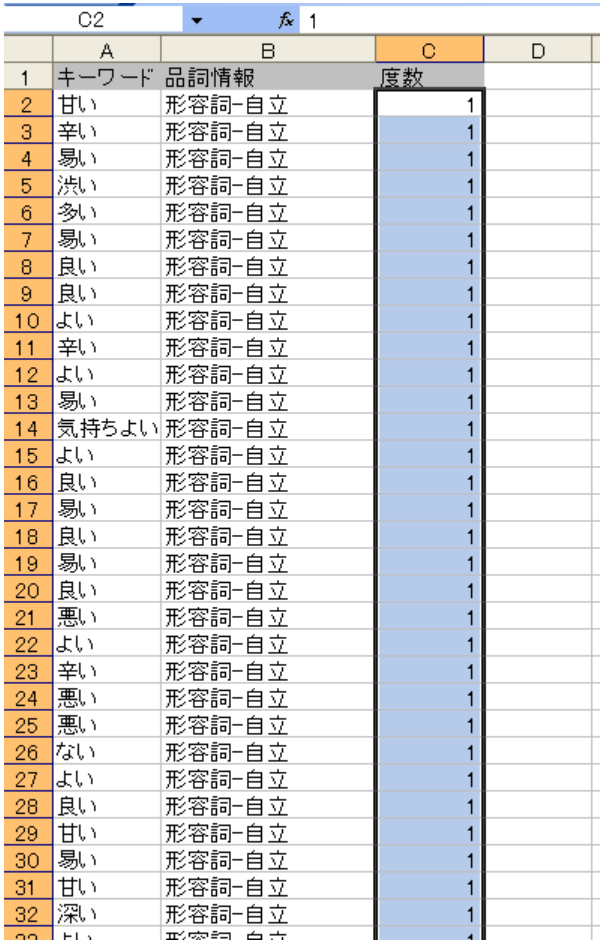
「整理後」
シート

A1 甘い

	A	B	C	D	E
1	甘い	形容詞-自	立		
2	辛い	形容詞-自	立		
3	易い	形容詞-自	立		
4	渋い	形容詞-自	立		
5	多い	形容詞-自	立		
6	易い	形容詞-自	立		
7	良い	形容詞-自	立		
8	良い	形容詞-自	立		
9	よい	形容詞-自	立		
10	辛い	形容詞-自	立		
11	よい	形容詞-自	立		
12	易い	形容詞-自	立		
13	気持ちよい	形容詞-自	立		
14	よい	形容詞-自	立		
15	良い	形容詞-自	立		
16	易い	形容詞-自	立		
17	良い	形容詞-自	立		
18	易い	形容詞-自	立		
19	良い	形容詞-自	立		
20	悪い	形容詞-自	立		
21	よい	形容詞-自	立		
22	辛い	形容詞-自	立		
23	悪い	形容詞-自	立		
24	悪い	形容詞-自	立		
25	ない	形容詞-自	立		
26	よい	形容詞-自	立		
27	良い	形容詞-自	立		
28	甘い	形容詞-自	立		
29	易い	形容詞-自	立		
30	甘い	形容詞-自	立		
31	深い	形容詞-自	立		

キーワードの出現頻度をヒストグラム化

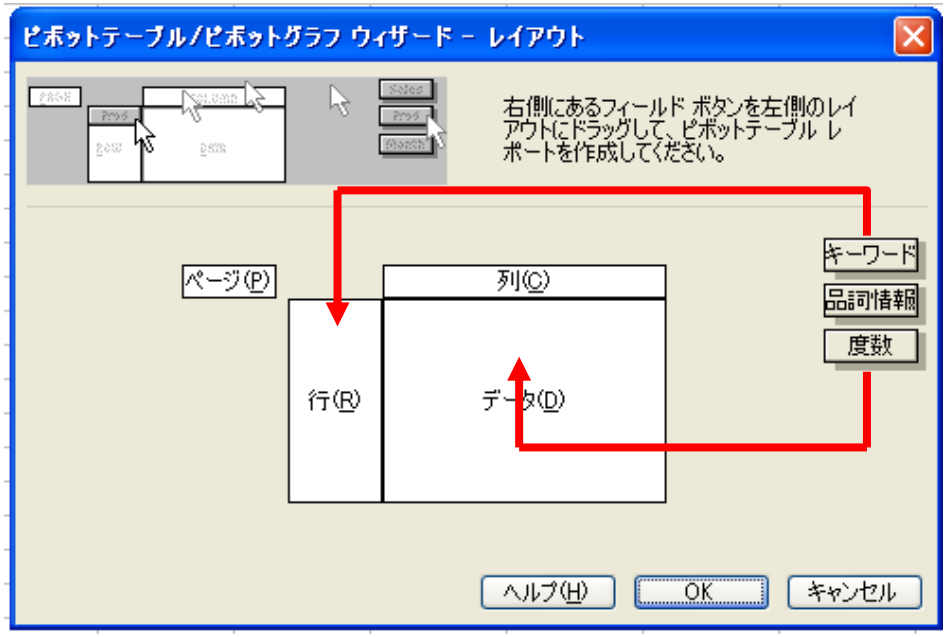
- 各列にラベルをつける
- [度数]の列を「1」で埋める



C2 fx 1

	A	B	C	D
1	キーワード	品詞情報	度数	
2	甘い	形容詞-自立	1	
3	辛い	形容詞-自立	1	
4	易い	形容詞-自立	1	
5	渋い	形容詞-自立	1	
6	多い	形容詞-自立	1	
7	易い	形容詞-自立	1	
8	良い	形容詞-自立	1	
9	良い	形容詞-自立	1	
10	よい	形容詞-自立	1	
11	辛い	形容詞-自立	1	
12	よい	形容詞-自立	1	
13	易い	形容詞-自立	1	
14	気持ちよい	形容詞-自立	1	
15	よい	形容詞-自立	1	
16	良い	形容詞-自立	1	
17	易い	形容詞-自立	1	
18	良い	形容詞-自立	1	
19	易い	形容詞-自立	1	
20	良い	形容詞-自立	1	
21	悪い	形容詞-自立	1	
22	よい	形容詞-自立	1	
23	辛い	形容詞-自立	1	
24	悪い	形容詞-自立	1	
25	悪い	形容詞-自立	1	
26	ない	形容詞-自立	1	
27	よい	形容詞-自立	1	
28	良い	形容詞-自立	1	
29	甘い	形容詞-自立	1	
30	易い	形容詞-自立	1	
31	甘い	形容詞-自立	1	
32	深い	形容詞-自立	1	

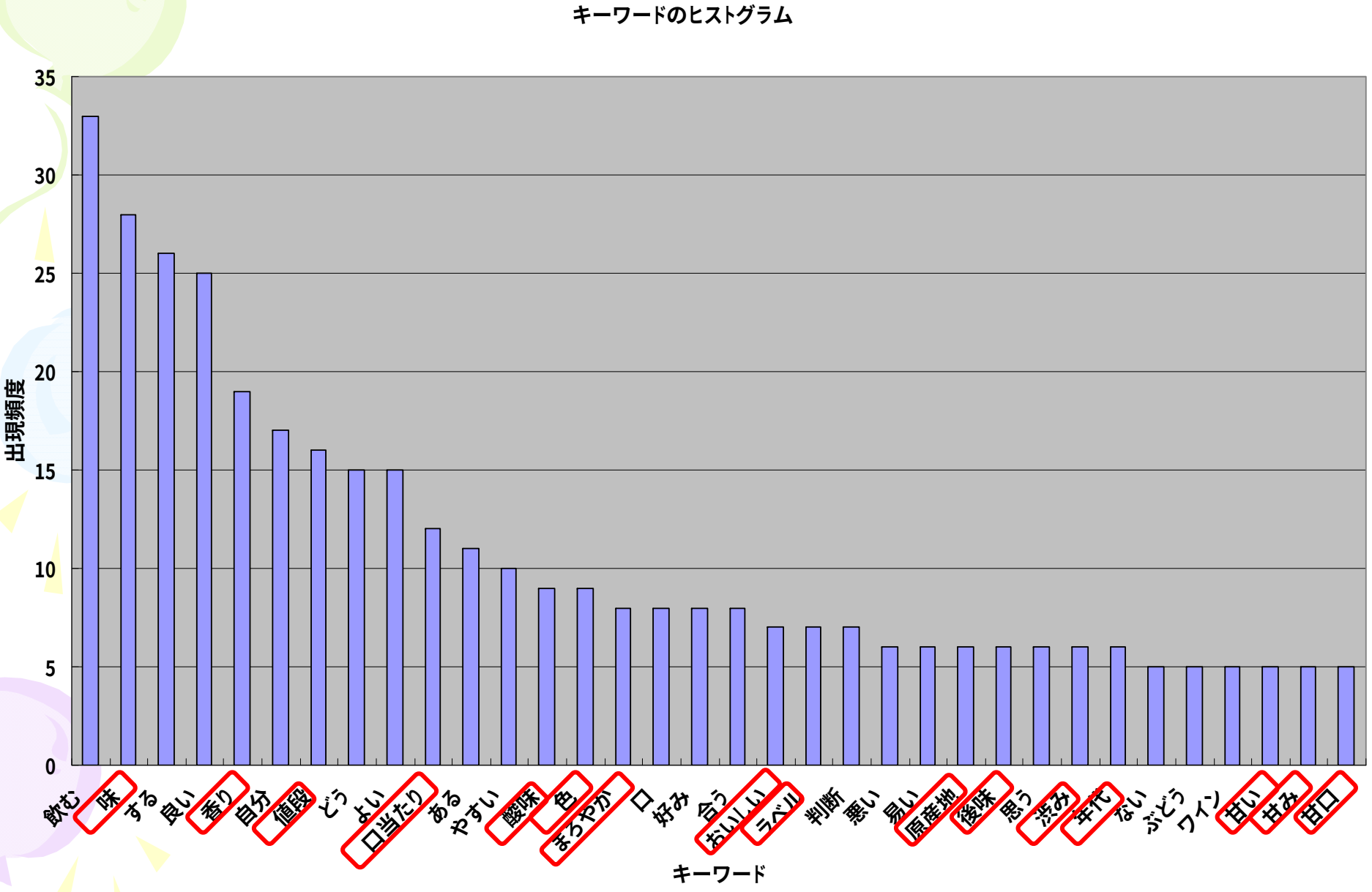
- データの範囲を指定
  →「データ」
  →「ピポットテーブル」
- ウィザード 1/3
  - Excelのリスト/データベース
  - ピポットテーブル
- ウィザード 2/3
  - そのまま「次へ」
- ウィザード 3/3
  - 「レイアウト」
  - 右図参照



- 度数の大きい順に並べ替え
  - 集計結果のままではできない
  - 別の場所に貼り付け，並べ替え
- 度数が5以上のものについて棒グラフ化する

D1　合計 / 度数

	A	B	C	D	E	F
1				合計 / 度数		
2				キーワード	集計	
3	合計 / 度数			NG	1	
4	キーワード	集計		あう	2	
5	NG	1		あまり	3	
6	あう	2		ある	11	
7	あまり	3		あるく	1	
8	ある	11		いい	2	
9	あるく	1		いう	3	
10	いい	2		いつも	1	
11	いう	3		うまみ	2	
12	いつも	1		えぐい	1	
13	うまみ	2		おいしい	7	
14	えぐい	1		おしゃれ	1	
15	おいしい	7		お金	1	
16	おしゃれ	1		かかる	1	
17	お金	1		かかわる	1	
18	かかる	1		きめる	2	
19	かかわる	1		キャップ	1	
20	きめる	2		きれい	1	
21	キャップ	1		コク	2	
22	きれい	1		こく	2	
23	コク	2		コルク	1	
24	こく	2		さっぱり	2	
25	コルク	1		サラ	1	
26	さっぱり	2		スッキリ	1	
27	サラ	1		すっきり	1	
28	スッキリ	1		する	26	
29	すっきり	1		ダメ	1	
30	する	26		っぽい	1	
31	ダメ	1		つや	1	
32	っぽい	1		できる	2	


キーワードのヒストグラム
出現頻度
35
30
25
20
15
10
5
0
飲む
味
する
良い
香り
自分
値段
どう
よい
口当たり
ある
やすい
酸味
色
まろやか
口
好み
合う
おいしい
ラベル
判断
悪い
易い
原産地
後味
思う
渋み
辛口
ない
ぶどう
ワイン
甘い
甘み
甘口
キーワード

しかし，これだけでは不十分

- アンケートの分量の割に必要な情報が少ない
- 抽出されたキーワードのうち，どれについて満足しているのか，あるいは不満なのか不明

↓

- 定義形式の定型自由文アンケート
  - 「理想のワインとは（　）で，（　）で，（　）なものです」
  - 例：「(ボトルがきれい)で(香りが良いもの)です」
    - 無駄な情報が少なく，うまくいけばそのままキーワードに
    - 「理想の」をつけることでポジティブな内容に限定できる
    - 「普段飲んでいるワインは(　)で(　)で(　)なものです」の結果と比較することで，理想と現状の差が大きいものがわかる

もっと，いろいろ知りたい

- ネガティブな情報も知りたい
- どんな原因が，どんな結果を生むのか

↓

- 文章完成形式の定型自由文のアンケート
  - 「ワインは（　）ので，（　）から，（　）」
  - 例：「(アルコール度数が強い) ので，(酔いやすい) から，(困る)」
    「(様々な種類がある) ので，(色々選べる) から，(楽しい)」
    「(甘すぎず，苦すぎない) ので，(飲みやすい) から，(好き)」
    - クロス集計表に整理
    - 何が原因になりやすいか，その原因はどんな結果を生むのか
    - 何が結果になりやすいか，その毛かはどんな結果から得られるのか

最後に

- 注意点
  - 事前に知りたいことを明確にして，それに関する回答が得られるような質問の形にする
  - 「統計的に有意であるか」を示すものではなく，あくまで今後の方向性を決めるための手段
- 心理学への応用
  - 予備調査で参加者の内観をとる
  - 数値による結果を自由記述で裏付けて考察する

参考文献

- 林俊克　2002　Excelで学ぶテキストマイニング入門　オーム社