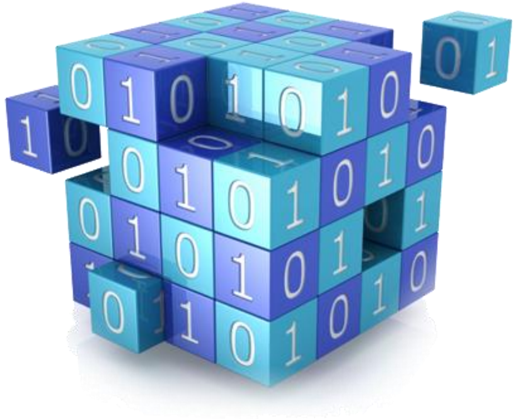
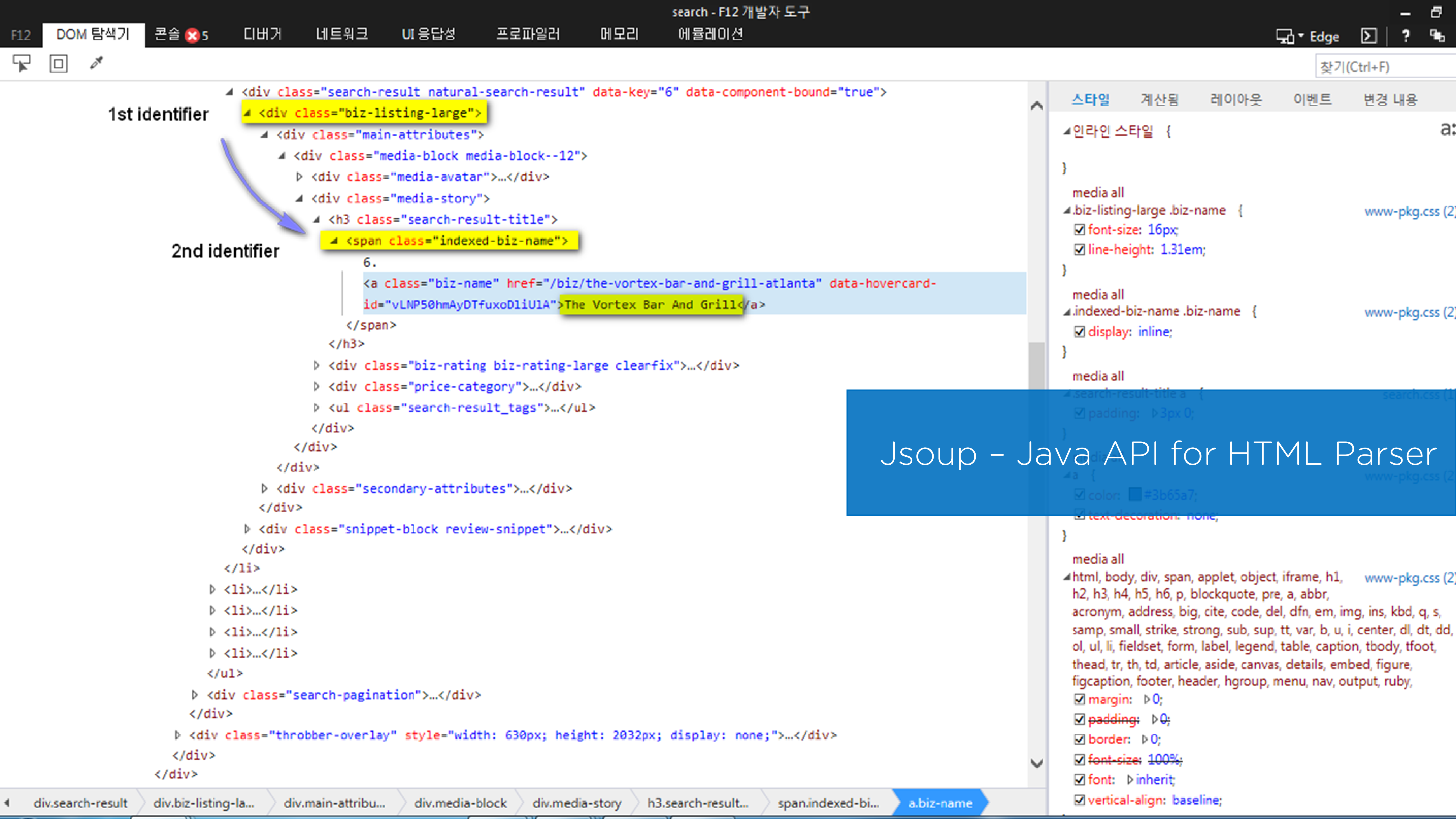


Yelp Scraper: Frequent Menu Finder

Data Mining Application on Yelp

Gu Young Chung, Keun S. Lee, Jongyeon Kim

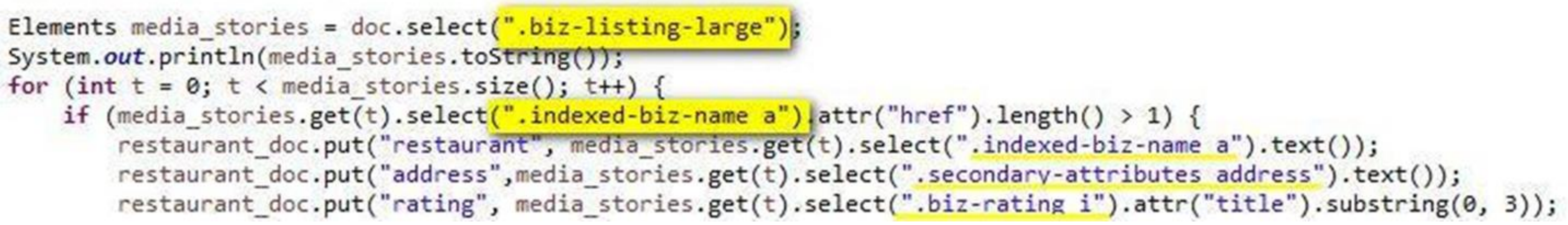


Jsoup – Java API for HTML Parser

```
◢ <div class="search-result natural-search-result" data-key="6" data-component-bound="true">
  ◢ <div class="biz-listing-large">
    ◢ <div class="main-attributes">
      ◢ <div class="media-block media-block--12">
        ▷ <div class="media-avatar">…</div>
        ◢ <div class="media-story">
          ◢ <h3 class="search-result-title">
            ◢ <span class="indexed-biz-name">
                6.
                <a class="biz-name" href="/biz/the-vortex-bar-and-grill-atlanta" data-hovercard-
                id="vLNP50hmAyDTfuxoD1iU1A">The Vortex Bar And Grill</a>
              </span>
            </h3>
          ▷ <div class="biz-rating biz-rating-large clearfix">…</div>
          ▷ <div class="price-category">…</div>
          ▷ <ul class="search-result_tags">…</ul>
          </div>
        </div>
      </div>
    ▷ <div class="secondary-attributes">…</div>
    </div>
  ▷ <div class="snippet-block review-snippet">…</div>
  </div>
</li>
▷ <li>…</li>
▷ <li>…</li>
▷ <li>…</li>
▷ <li>…</li>
</ul>
▷ <div class="search-pagination">…</div>
</div>
▷ <div class="throbber-overlay" style="width: 630px; height: 2032px; display: none;">…</div>
</div>
</div>
```

```
◢인라인 스타일  {
}
media all
◢.biz-listing-large .biz-name  {                      www-pkg.css (2)
  ☑ font-size: 16px;
  ☑ line-height: 1.31em;
}
media all
◢.indexed-biz-name .biz-name  {                       www-pkg.css (2)
  ☑ display: inline;
}
media all
◢.search-result-title a  {                            search.css (1)
  ☑ padding: ▷3px 0;
}
◢a  {                                                 www-pkg.css (2)
  ☑ color: #3b65a7;
  ☑ text-decoration: none;
}
media all
◢html, body, div, span, applet, object, iframe, h1,   www-pkg.css (2)
h2, h3, h4, h5, h6, p, blockquote, pre, a, abbr,
acronym, address, big, cite, code, del, dfn, em, img, ins, kbd, q, s,
samp, small, strike, strong, sub, sup, tt, var, b, u, i, center, dl, dt, dd,
ol, ul, li, fieldset, form, label, legend, table, caption, tbody, tfoot,
thead, tr, th, td, article, aside, canvas, details, embed, figure,
figcaption, footer, header, hgroup, menu, nav, output, ruby,
  ☑ margin: ▷0;
  ☑ ~~padding:~~ ▷~~0;~~
  ☑ border: ▷0;
  ☑ ~~font-size: 100%;~~
  ☑ font: ▷inherit;
  ☑ vertical-align: baseline;
```

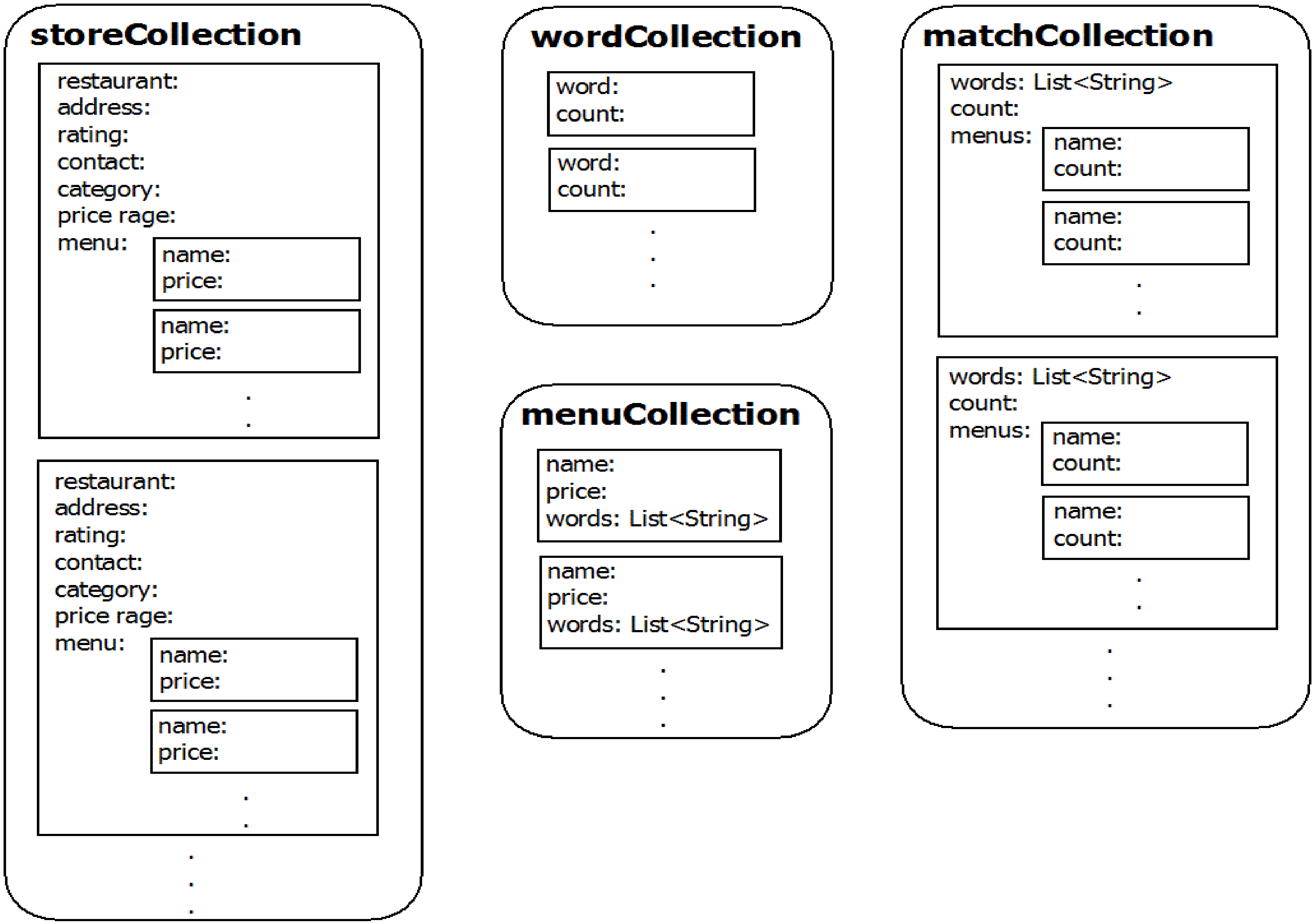
Jsoup example: “Restaurant”, “Address”, “Rating”



```
Elements media_stories = doc.select(".biz-listing-large");
System.out.println(media_stories.toString());
for (int t = 0; t < media_stories.size(); t++) {
    if (media_stories.get(t).select(".indexed-biz-name a").attr("href").length() > 1) {
        restaurant_doc.put("restaurant", media_stories.get(t).select(".indexed-biz-name a").text());
        restaurant_doc.put("address",media_stories.get(t).select(".secondary-attributes address").text());
        restaurant_doc.put("rating", media_stories.get(t).select(".biz-rating i").attr("title").substring(0, 3));
```

Restaurant: The Vortex Bar And Grill
Address: 878 Peachtree St NE Atlanta, GA 30309
Rating: 4 Stars

MongoDB



storeCollection

restaurant:
address:
rating:
contact:
category:
price rage:
menu:
name:
price:
name:
price:
⋮

restaurant:
address:
rating:
contact:
category:
price rage:
menu:
name:
price:
name:
price:
⋮

⋮

wordCollection

word:
count:

word:
count:

⋮

menuCollection

name:
price:
words: List<String>

name:
price:
words: List<String>

⋮

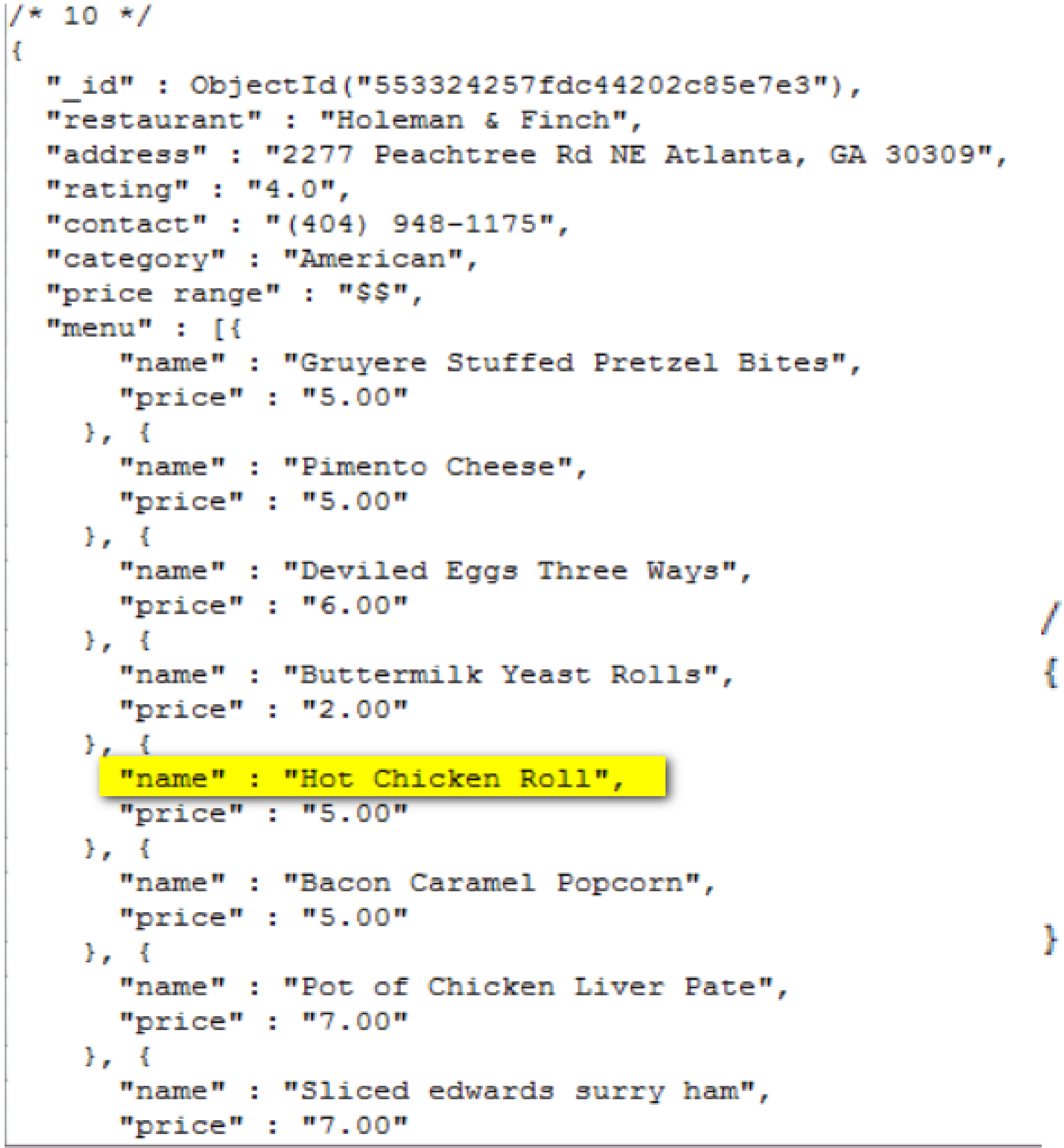
matchCollection

words: List<String>
count:
menus:
name:
count:
name:
count:
⋮

words: List<String>
count:
menus:
name:
count:
name:
count:
⋮

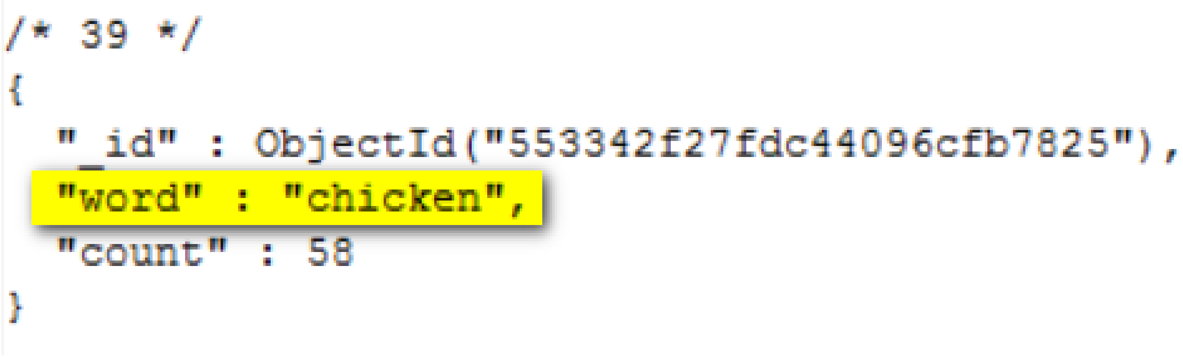
⋮

storeCollection



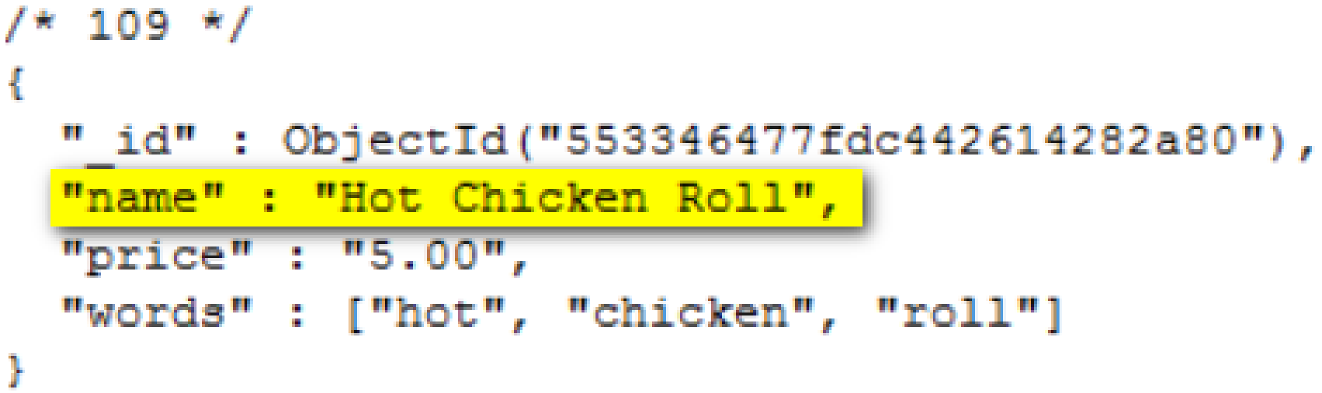
```
/* 10 */
{
  "_id" : ObjectId("553324257fdc44202c85e7e3"),
  "restaurant" : "Holeman & Finch",
  "address" : "2277 Peachtree Rd NE Atlanta, GA 30309",
  "rating" : "4.0",
  "contact" : "(404) 948-1175",
  "category" : "American",
  "price range" : "$$",
  "menu" : [{
      "name" : "Gruyere Stuffed Pretzel Bites",
      "price" : "5.00"
    }, {
      "name" : "Pimento Cheese",
      "price" : "5.00"
    }, {
      "name" : "Deviled Eggs Three Ways",
      "price" : "6.00"
    }, {
      "name" : "Buttermilk Yeast Rolls",
      "price" : "2.00"
    }, {
      "name" : "Hot Chicken Roll",
      "price" : "5.00"
    }, {
      "name" : "Bacon Caramel Popcorn",
      "price" : "5.00"
    }, {
      "name" : "Pot of Chicken Liver Pate",
      "price" : "7.00"
    }, {
      "name" : "Sliced edwards surry ham",
      "price" : "7.00"
```

wordCollection



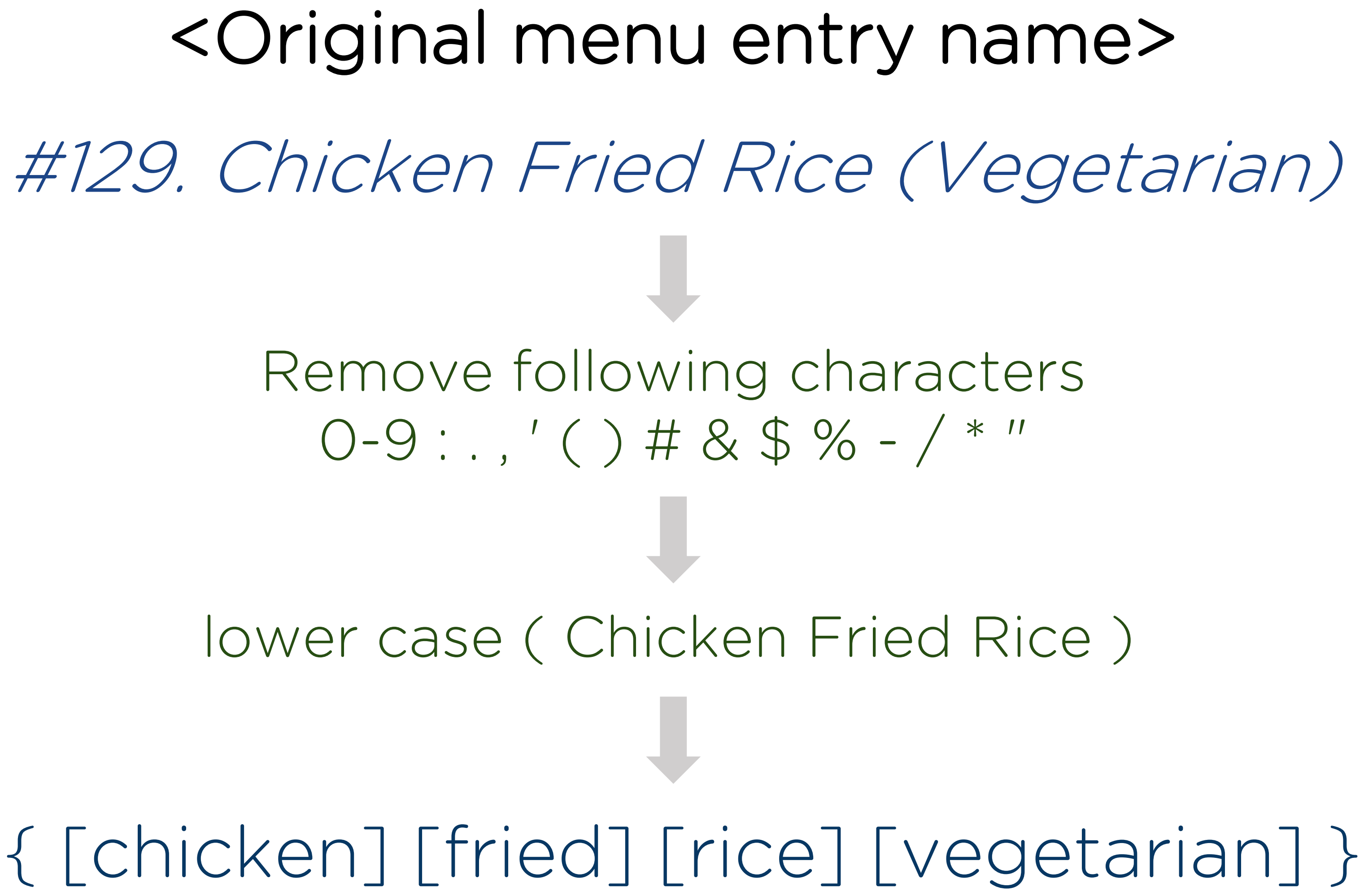
```
/* 39 */
{
  "_id" : ObjectId("553342f27fdc44096cfb7825"),
  "word" : "chicken",
  "count" : 58
}
```

menuCollection



```
/* 109 */
{
  "_id" : ObjectId("553346477fdc442614282a80"),
  "name" : "Hot Chicken Roll",
  "price" : "5.00",
  "words" : ["hot", "chicken", "roll"]
}
```

Normalization on Data Format



<Original menu entry name>

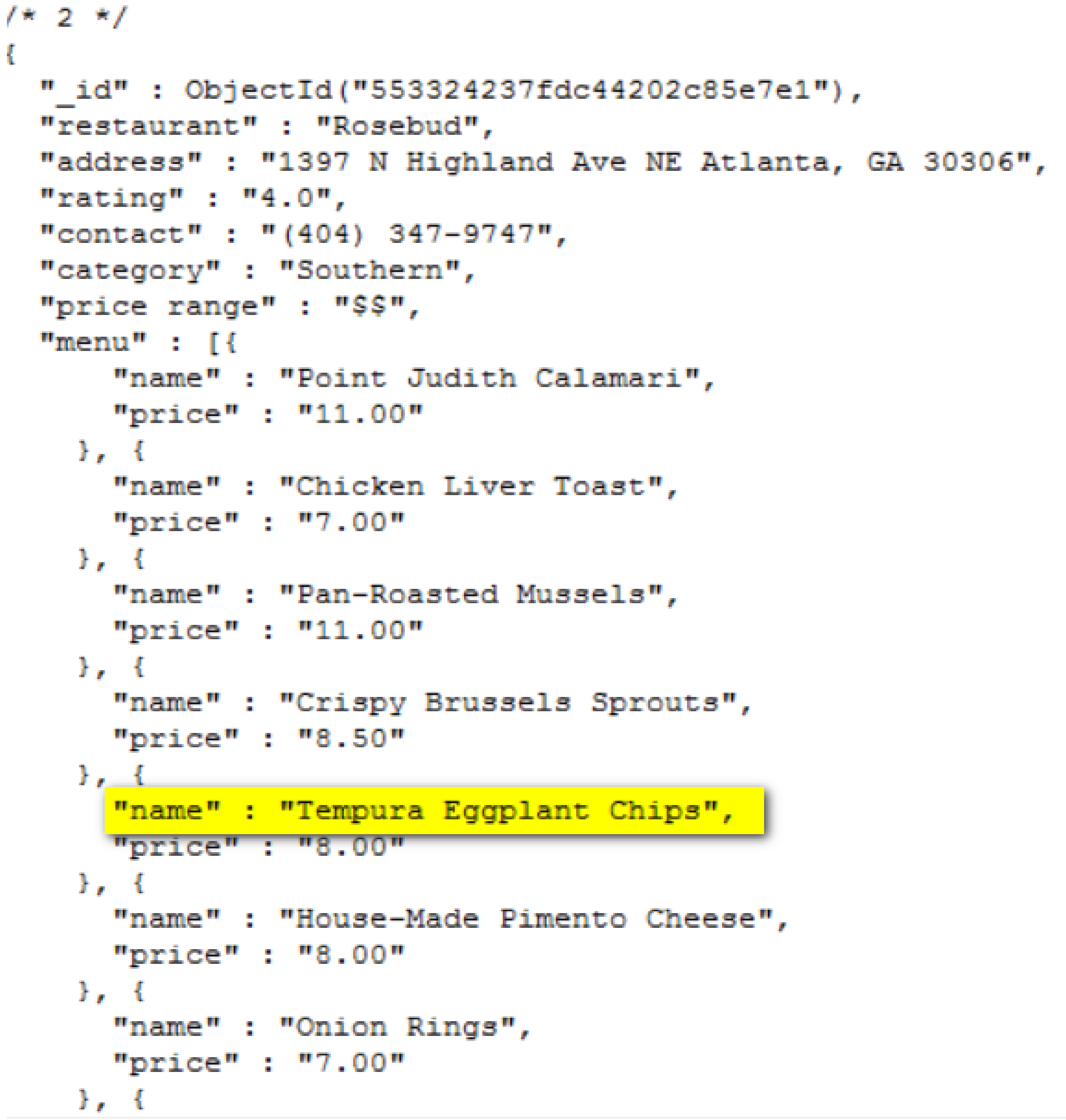
#129. Chicken Fried Rice (Vegetarian)

Remove following characters
0-9 : . , ' ( ) # & $ % - / * "

lower case (Chicken Fried Rice)

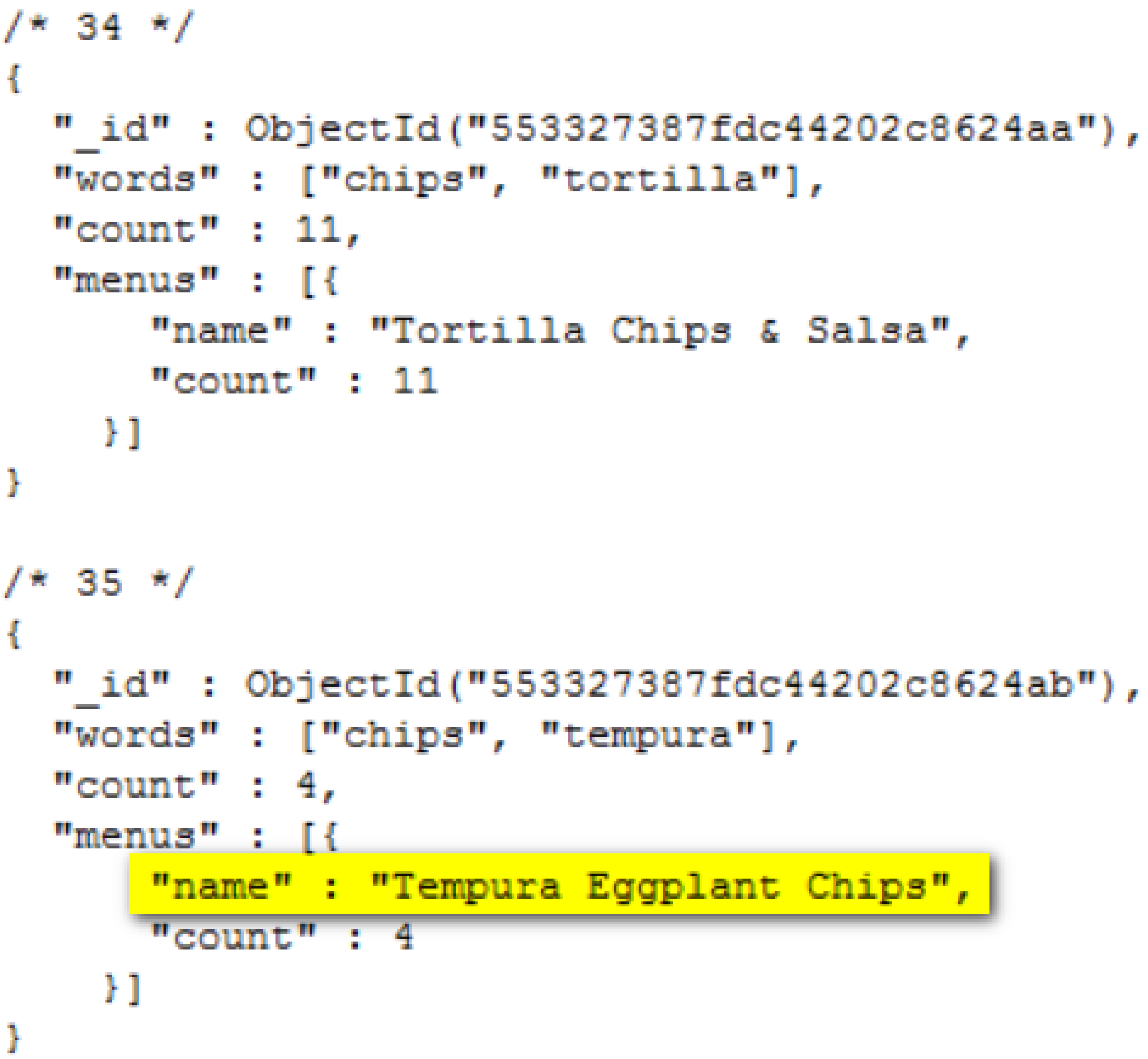
{ [chicken] [fried] [rice] [vegetarian] }

storeCollection



```
/* 2 */
{
  "_id" : ObjectId("553324237fdc44202c85e7e1"),
  "restaurant" : "Rosebud",
  "address" : "1397 N Highland Ave NE Atlanta, GA 30306",
  "rating" : "4.0",
  "contact" : "(404) 347-9747",
  "category" : "Southern",
  "price range" : "$$",
  "menu" : [{
      "name" : "Point Judith Calamari",
      "price" : "11.00"
    }, {
      "name" : "Chicken Liver Toast",
      "price" : "7.00"
    }, {
      "name" : "Pan-Roasted Mussels",
      "price" : "11.00"
    }, {
      "name" : "Crispy Brussels Sprouts",
      "price" : "8.50"
    }, {
      "name" : "Tempura Eggplant Chips",
      "price" : "8.00"
    }, {
      "name" : "House-Made Pimento Cheese",
      "price" : "8.00"
    }, {
      "name" : "Onion Rings",
      "price" : "7.00"
    }, {
```

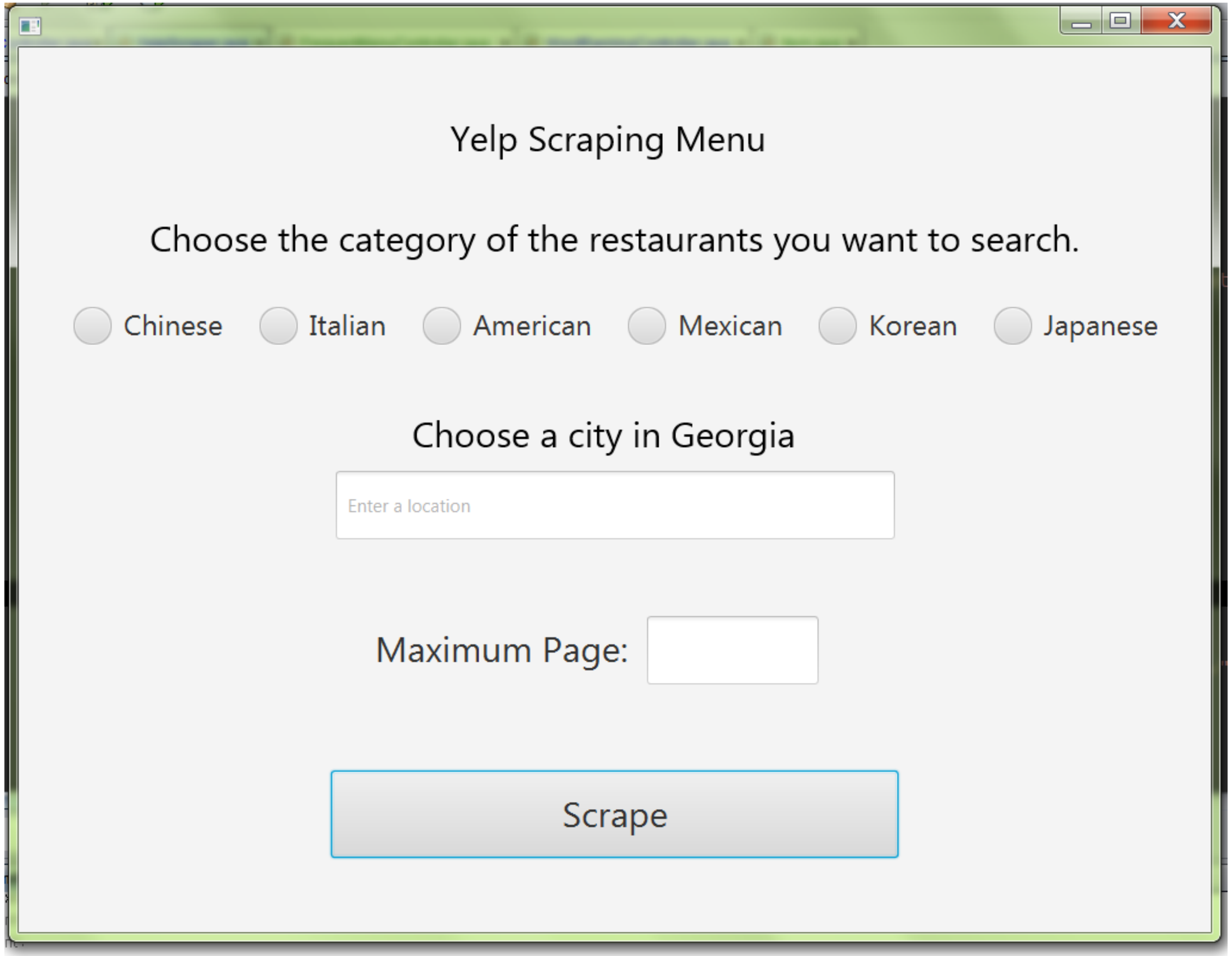
matchCollection



```
/* 34 */
{
  "_id" : ObjectId("553327387fdc44202c8624aa"),
  "words" : ["chips", "tortilla"],
  "count" : 11,
  "menus" : [{
      "name" : "Tortilla Chips & Salsa",
      "count" : 11
    }]
}

/* 35 */
{
  "_id" : ObjectId("553327387fdc44202c8624ab"),
  "words" : ["chips", "tempura"],
  "count" : 4,
  "menus" : [{
      "name" : "Tempura Eggplant Chips",
      "count" : 4
    }]
}
```

Main Menu



Yelp Scraping Menu

Choose the category of the restaurants you want to search.

○ Chinese ○ Italian ○ American ○ Mexican ○ Korean ○ Japanese

Choose a city in Georgia

Enter a location

Maximum Page:

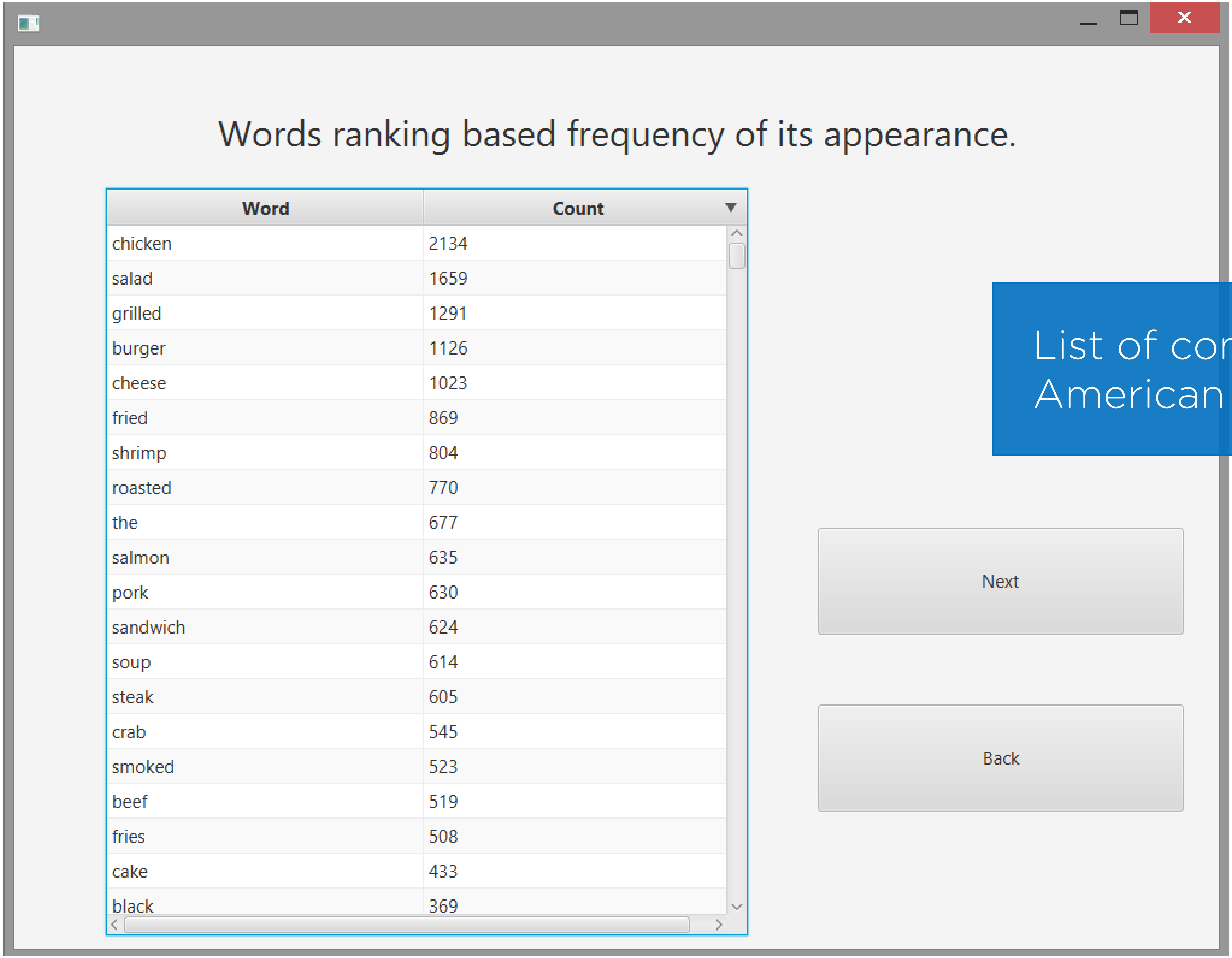
User Options:

Choose between 6 categories of restaurants

Type which city in GA to search nearby for restaurants

Limit the maximum number of pages the application will crawl

.

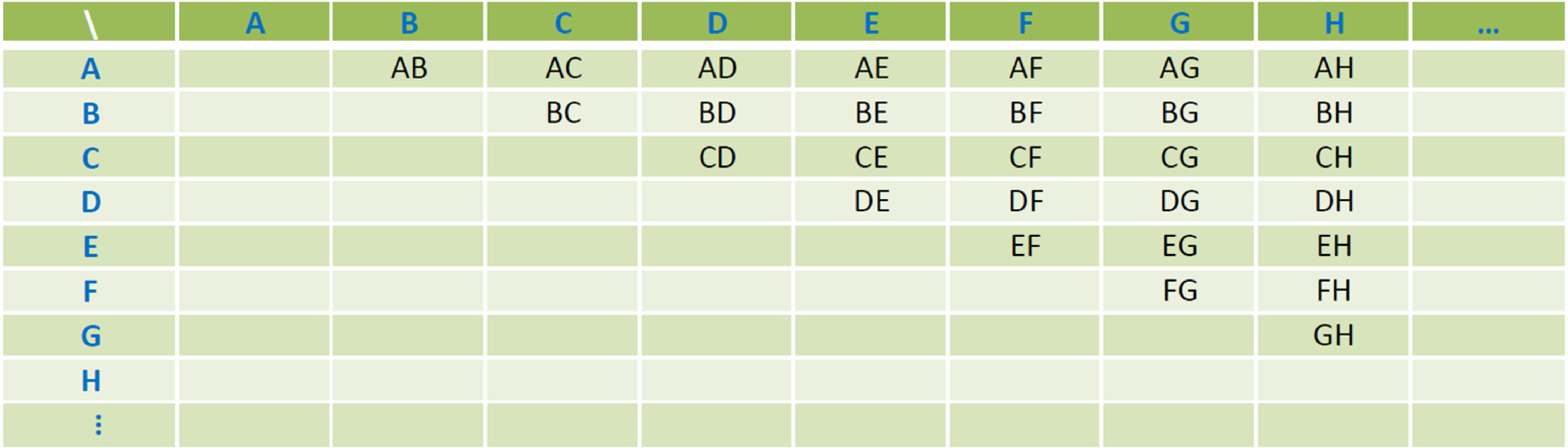


Words ranking based frequency of its appearance.

Word	Count
chicken	2134
salad	1659
grilled	1291
burger	1126
cheese	1023
fried	869
shrimp	804
roasted	770
the	677
salmon	635
pork	630
sandwich	624
soup	614
steak	605
crab	545
smoked	523
beef	519
fries	508
cake	433
black	369

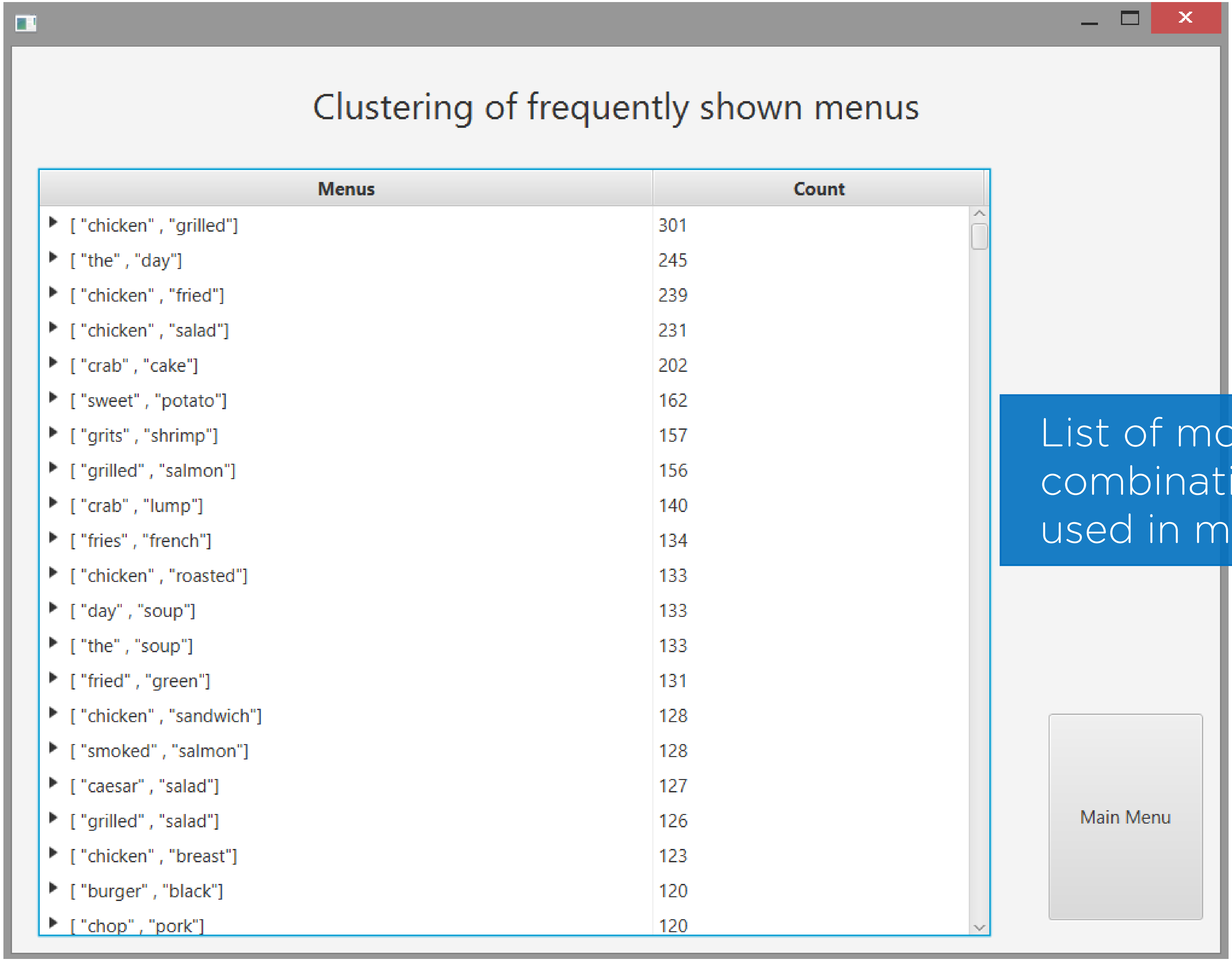
List of common words used in American restaurants’ menu

Pairing of Common Word for Menu Clustering



\	A	B	C	D	E	F	G	H	…
A		AB	AC	AD	AE	AF	AG	AH	
B			BC	BD	BE	BF	BG	BH	
C				CD	CE	CF	CG	CH	
D					DE	DF	DG	DH	
E						EF	EG	EH	
F							FG	FH	
G								GH	
H									
⋮									

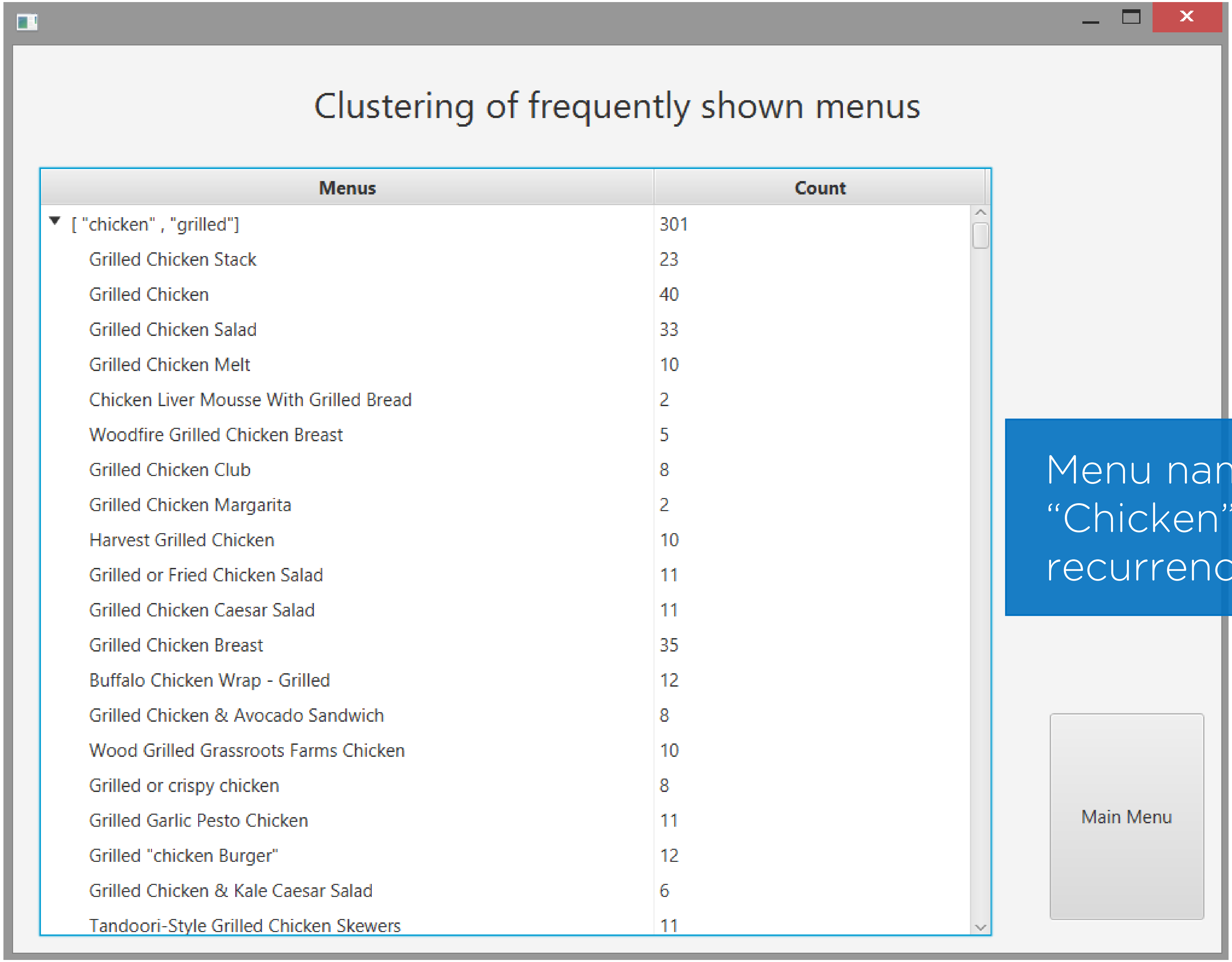
This yields 100 words => 100 x 101 / 2 – 100 =
= 4950 different combinations



Clustering of frequently shown menus

Menus	Count
▸ ["chicken" , "grilled"]	301
▸ ["the" , "day"]	245
▸ ["chicken" , "fried"]	239
▸ ["chicken" , "salad"]	231
▸ ["crab" , "cake"]	202
▸ ["sweet" , "potato"]	162
▸ ["grits" , "shrimp"]	157
▸ ["grilled" , "salmon"]	156
▸ ["crab" , "lump"]	140
▸ ["fries" , "french"]	134
▸ ["chicken" , "roasted"]	133
▸ ["day" , "soup"]	133
▸ ["the" , "soup"]	133
▸ ["fried" , "green"]	131
▸ ["chicken" , "sandwich"]	128
▸ ["smoked" , "salmon"]	128
▸ ["caesar" , "salad"]	127
▸ ["grilled" , "salad"]	126
▸ ["chicken" , "breast"]	123
▸ ["burger" , "black"]	120
▸ ["chop" , "pork"]	120

List of most common combinations of words that are used in menu



Clustering of frequently shown menus

Menus	Count
▼ ["chicken" , "grilled"]	301
Grilled Chicken Stack	23
Grilled Chicken	40
Grilled Chicken Salad	33
Grilled Chicken Melt	10
Chicken Liver Mousse With Grilled Bread	2
Woodfire Grilled Chicken Breast	5
Grilled Chicken Club	8
Grilled Chicken Margarita	2
Harvest Grilled Chicken	10
Grilled or Fried Chicken Salad	11
Grilled Chicken Caesar Salad	11
Grilled Chicken Breast	35
Buffalo Chicken Wrap - Grilled	12
Grilled Chicken & Avocado Sandwich	8
Wood Grilled Grassroots Farms Chicken	10
Grilled or crispy chicken	8
Grilled Garlic Pesto Chicken	11
Grilled "chicken Burger"	12
Grilled Chicken & Kale Caesar Salad	6
Tandoori-Style Grilled Chicken Skewers	11

Menu names that contain “Chicken” and “Grilled” and their recurrence counts