Dataset Format Analysis and Column Classification for CompleteSearch

Colloquium

Freiburg, 17th October 2018

Olivier Puraye

Objective

board_games.csv

```
rank;names;designer;category;bgg_url;min_players;max_players;avg_time;year;av
1;Gloomhaven;Isaac Childres;Adventure#Exploration#Fantasy#Fighting#Miniatures
2;Pandemic Legacy: Season 1;Rob Daviau#Matt Leacock;Environmental#Medical;htt
3;Through the Ages: A New Story of Civilization;Vlaada Chvátil;Card Game#Civi
4;Twilight Struggle;Ananda Gupta#Jason Matthews;Modern Warfare#Political#Warg
5;Terraforming Mars;Jacob Fryxelius;Economic#Environmental#Industry / Manufac
6;Terra Mystica;Jens Drögemüller#Helge Ostertag;Civilization#Economic#Fantasy
7;Scythe;Jamey Stegmaier;Civilization#Economic#Fighting#Miniatures#Science Fi
8;7 Wonders Duel;Antoine Bauza#Bruno Cathala;Ancient#Card Game#City Building#
9;Great Western Trail;Alexander Pfister;American West;https://boardgamegeek.c
```

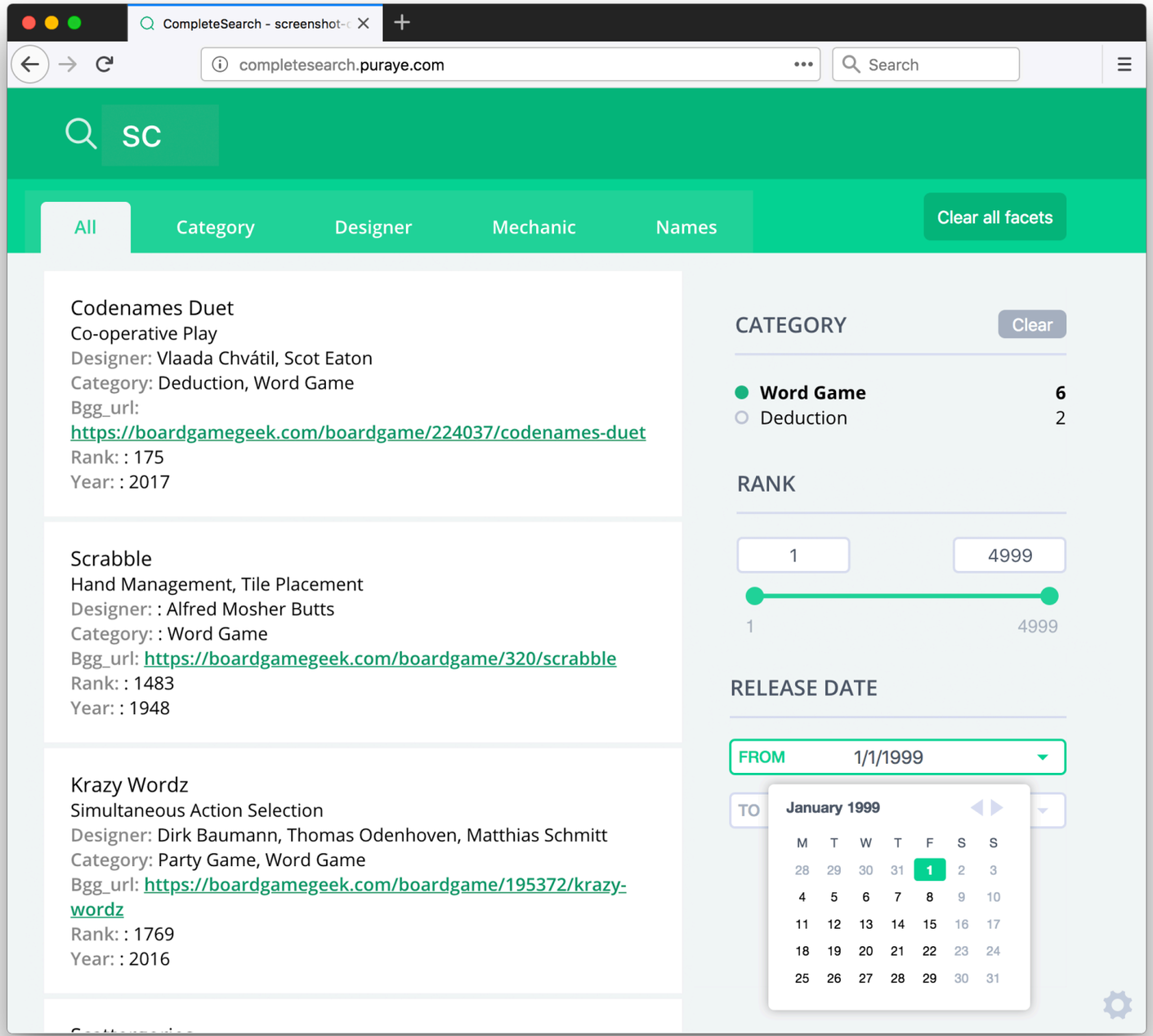
Tabular Dataset (CSV, TSV, …)

Objective

board_games.csv

```
rank;names;designer;category;bgg_url;min_players;max_players;avg_time;year;av
1;Gloomhaven;Isaac Childres;Adventure#Exploration#Fantasy#Fighting#Miniatures
Selection#Storytelling#Variable Player Powers
2;Pandemic Legacy: Season 1;Rob Daviau#Matt Leacock;Environmental#Medical;htt
3;Through the Ages: A New Story of Civilization;Vlaada Chvátil;Card Game#Civi
4;Twilight Struggle;Ananda Gupta#Jason Matthews;Modern Warfare#Political#Warg
5;Terraforming Mars;Jacob Fryxelius;Economic#Environmental#Industry / Manufac
6;Drögemüller#Helge Ostertag;Civilization#Economic#Fantasy#Territory Building
7;Scythe;Jamey Stegmaier;Civilization#Economic#Fighting#Miniatures#Science Fi
8;7 Wonders Duel;Antoine Bauza#Bruno Cathala;Ancient#Card Game#City Building#
9;Great Western Trail;Alexander Pfister;American West;https://boardgamegeek.co
```

Tabular Dataset (CSV, TSV, …)



CompleteSearch

Dataset

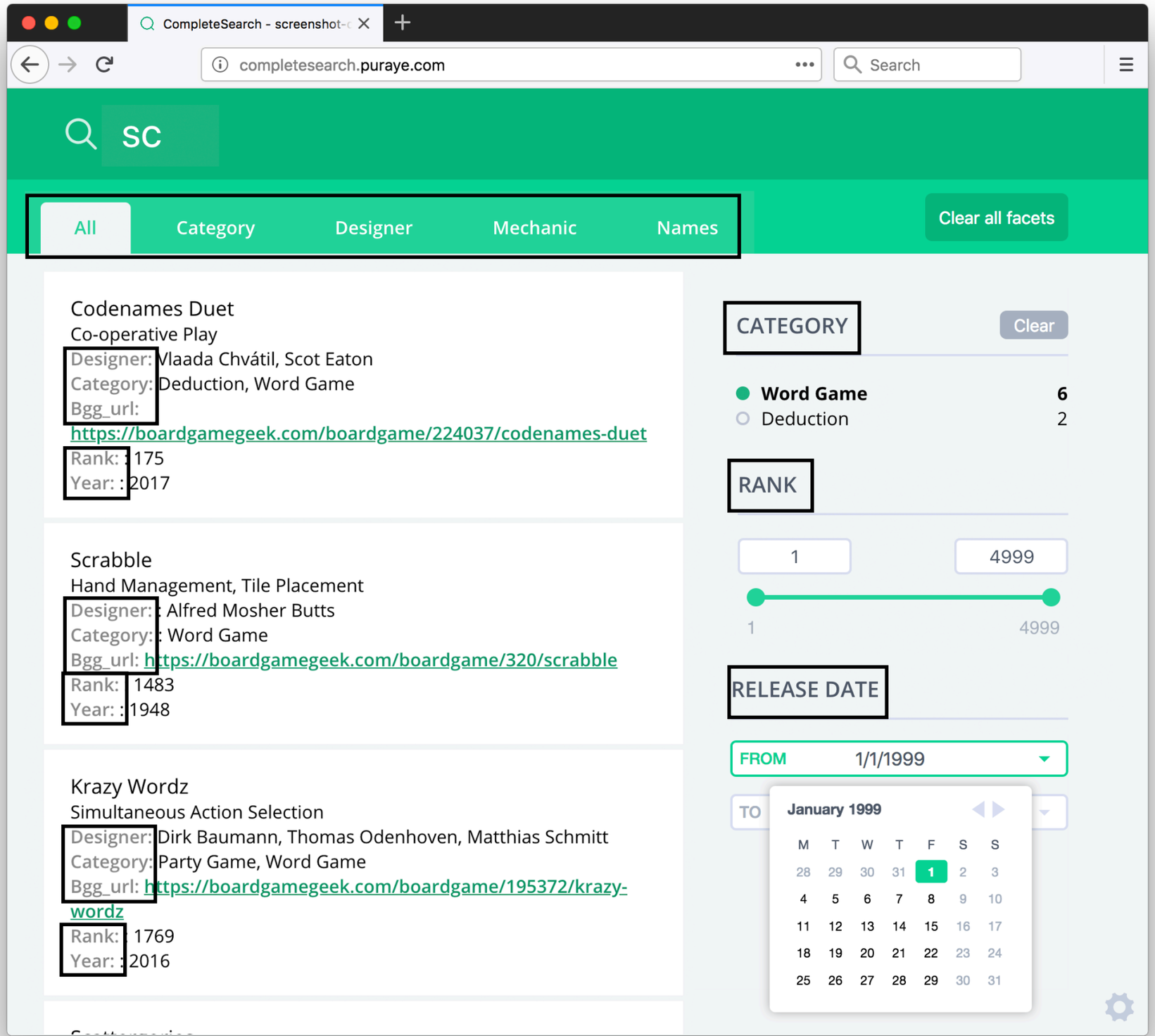
board_games.csv

Header Row with Column Labels

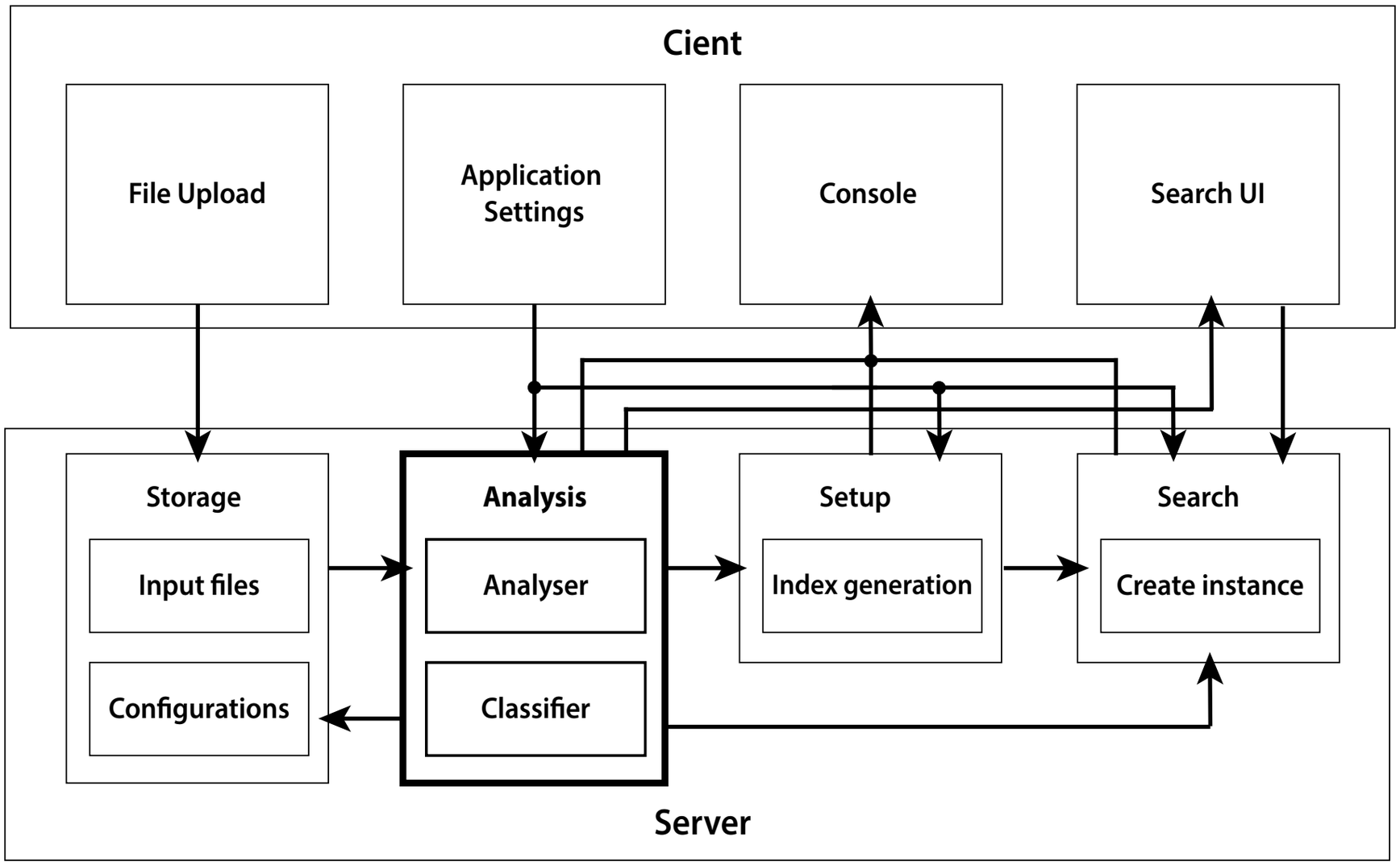
```
rank;names;designer;category;bgg_url;min_players;max_players;avg_time;year;av
1;Gloomhaven;Isaac Childres;Adventure#Exploration#Fantasy#Fighting#Miniatures
2;Pandemic Legacy: Season 1;Rob Daviau#Matt Leacock;Environmental#Medical;htt
3;Through the Ages: A New Story of Civilization;Vlaada Chvátil;Card Game#Civi
4;Twilight Struggle;Ananda Gupta#Jason Matthews;Modern Warfare#Political#Warg
5;Terraforming Mars;Jacob Fryxelius;Economic#Environmental#Industry / Manufac
6;Terra Mystica;Jens Drögemüller#Helge Ostertag;Civilization#Economic#Fantasy
7;Scythe;Jamey Stegmaier;Civilization#Economic#Fighting#Miniatures#Science Fi
8;7 Wonders Duel;Antoine Bauza#Bruno Cathala;Ancient#Card Game#City Building#
9;Great Western Trail;Alexander Pfister;American West;https://boardgamegeek.c
```

Tabular Dataset (CSV, TSV, …)

Search UI


CompleteSearch - screenshot-
completesearch.puraye.com
Search
sc
All
Category
Designer
Mechanic
Names
Clear all facets
Codenames Duet
Co-operative Play
Designer: Vlaada Chvátil, Scot Eaton
Category: Deduction, Word Game
Bgg_url: https://boardgamegeek.com/boardgame/224037/codenames-duet
Rank: 1175
Year: 2017
Scrabble
Hand Management, Tile Placement
Designer: Alfred Mosher Butts
Category: Word Game
Bgg_url: https://boardgamegeek.com/boardgame/320/scrabble
Rank: 1483
Year: 1948
Krazy Wordz
Simultaneous Action Selection
Designer: Dirk Baumann, Thomas Odenhoven, Matthias Schmitt
Category: Party Game, Word Game
Bgg_url: https://boardgamegeek.com/boardgame/195372/krazy-wordz
Rank: 1769
Year: 2016
CATEGORY
Clear
Word Game 6
Deduction 2
RANK
1
4999
1
4999
RELEASE DATE
FROM 1/1/1999
TO
January 1999
M T W T F S S
28 29 30 31 1 2 3
4 5 6 7 8 9 10
11 12 13 14 15 16 17
18 19 20 21 22 23 24
25 26 27 28 29 30 31

Application architecture



Configuration parameters (1)

- **column-separator:** separator which delimits columns in the dataset

```
rank;names;designer;category;bgg_url;min_players;max_players;avg_time;year;av
1;Gloomhaven;Isaac Childres;Adventure#Exploration#Fantasy#Fighting#Miniatures
2;Pandemic Legacy: Season 1;Rob Daviau#Matt Leacock;Environmental#Medical;htt
3;Through the Ages: A New Story of Civilization;Vlaada Chvátil;Card Game#Civi
4;Twilight Struggle;Ananda Gupta#Jason Matthews;Modern Warfare#Political#Warg
5;Terraforming Mars;Jacob Fryxelius;Economic#Environmental#Industry / Manufac
```

rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575
3	Through the Ages: A New Story of Civilization	Vlaada Chvátil	Card Game#Civilization#Economic	https://boardgamegeek.com/boardgame/182028/through-ages-new-story-civilization	2	4	240	2015	8.65702
4	Twilight Struggle	Ananda Gupta#Jason Matthews	Modern Warfare#Political#Wargame	https://boardgamegeek.com/boardgame/12333/twilight-struggle	2	2	180	2005	8.35188
5	Terraforming Mars	Jacob Fryxelius	Economic#Environ	https://	1	5	120	2016	8.38331

Configuration parameters (2)

- **subitem-separator**
  - separator which delimits lists of subitems within a column item
  - same subitem separator for all columns in the dataset

rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575
3	Through the Ages: A New Story of Civilization	Vlaada Chvátil	Card Game#Civilization#Economic	https://boardgamegeek.com/boardgame/182028/through-ages-new-story-civilization	2	4	240	2015	8.65702
4	Twilight Struggle	Ananda Gupta#Jason Matthews	Modern Warfare#Political#Wargame	https://boardgamegeek.com/boardgame/12333/twilight-struggle	2	2	180	2005	8.35188
5	Terraforming Mars	Jacob Fryxelius	Economic#Environ	https://	1	5	120	2016	8.38331

Configuration parameters (3)

- **allow-subitems**
  - array of columns that allow lists of subitems

- **Example:** allow-subitems = [designer, category]

rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575
3	Through the Ages: A New Story of Civilization	Vlaada Chvátil	Card Game#Civilization#Economic	https://boardgamegeek.com/boardgame/182028/through-ages-new-story-civilization	2	4	240	2015	8.65702
4	Twilight Struggle	Ananda Gupta#Jason Matthews	Modern Warfare#Political#Wargame	https://boardgamegeek.com/boardgame/12333/twilight-struggle	2	2	180	2005	8.35188
5	Terraforming Mars	Jacob Fryxelius	Economic#Environ	https://	1	5	120	2016	8.38331

Configuration parameters (4)

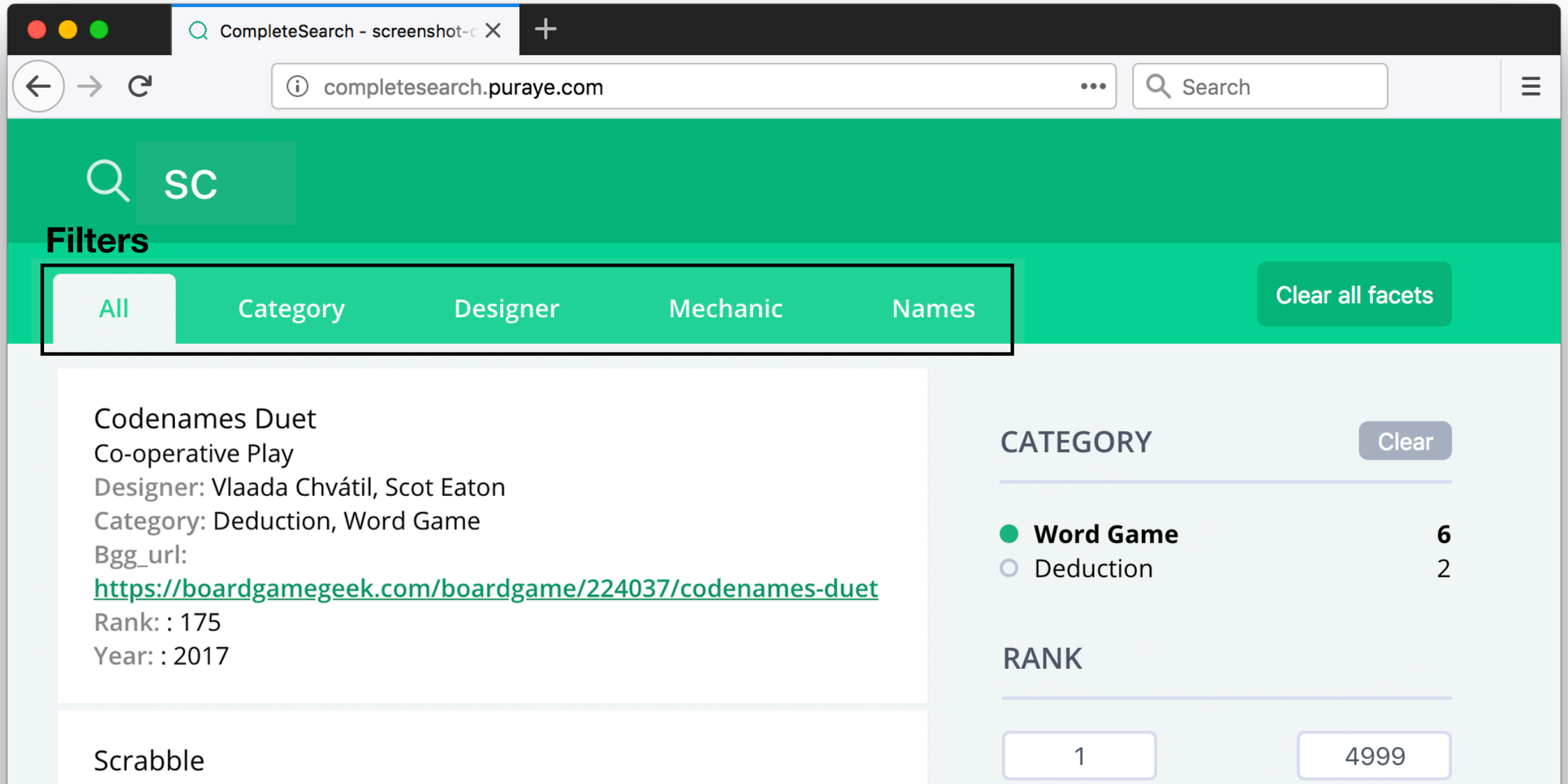
- **full-text**
  - array of columns that should be searched by simple query

- **Example:** full-text = [names, designer, category, year]

rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575
3	Through the Ages: A New Story of Civilization	Vlaada Chvátil	Card Game#Civilization#Economic	https://boardgamegeek.com/boardgame/182028/through-ages-new-story-civilization	2	4	240	2015	8.65702
4	Twilight Struggle	Ananda Gupta#Jason Matthews	Modern Warfare#Political#Wargame	https://boardgamegeek.com/boardgame/12333/twilight-struggle	2	2	180	2005	8.35188
5	Terraforming Mars	Jacob Fryxelius	Economic#Environ	https://	1	5	120	2016	8.38331

Configuration parameters (5)

- **filter**
  - array of columns that should support filtering
  - Filtering by a column restricts the search query to that specific column
  - Filters represented by tabs in search user interface

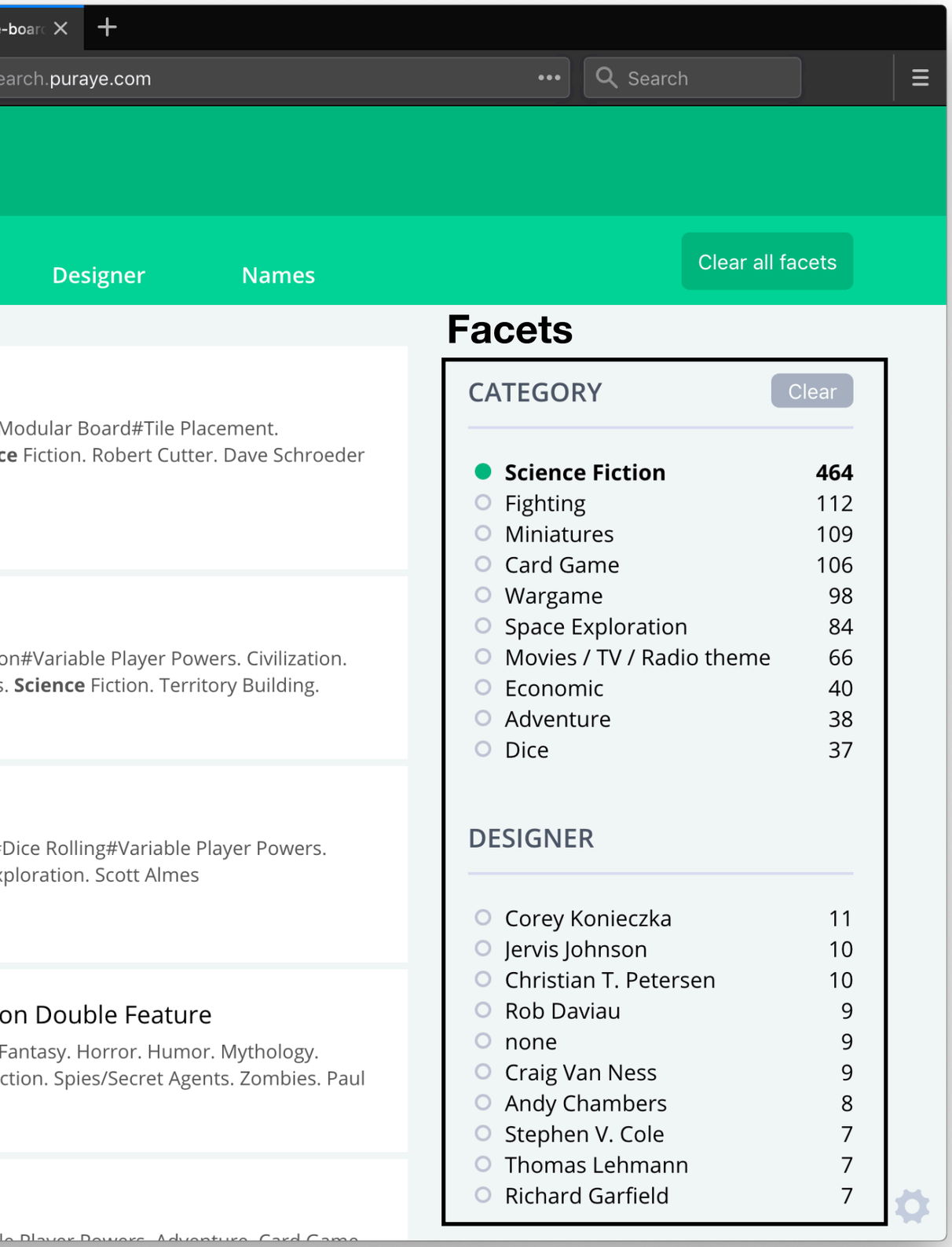


Configuration parameters (5)

- **filter**
  - array of columns that should support filtering
  - Filtering by a column restricts the search query to that specific column
- **Example:** filter = [names, designer, category]

rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575
3	Through the Ages: A New Story of Civilization	Vlaada Chvátil	Card Game#Civilization#Economic	https://boardgamegeek.com/boardgame/182028/through-ages-new-story-civilization	2	4	240	2015	8.65702
4	Twilight Struggle	Ananda Gupta#Jason Matthews	Modern Warfare#Political#Wargame	https://boardgamegeek.com/boardgame/12333/twilight-struggle	2	2	180	2005	8.35188
5	Terraforming Mars	Jacob Fryxelius	Economic#Environ	https://	1	5	120	2016	8.38331

Configuration parameters (6)



- **facet**
  - array of columns that can be used to further refine the search results by specifying explicit values for these columns

Configuration parameters (6)

- **facet**
  - array of columns that can be used to further refine the search results by specifying explicit values for these columns
- **Example:** facets = [designer, category, min_players, max_players, avg_time, year, avg_rating]

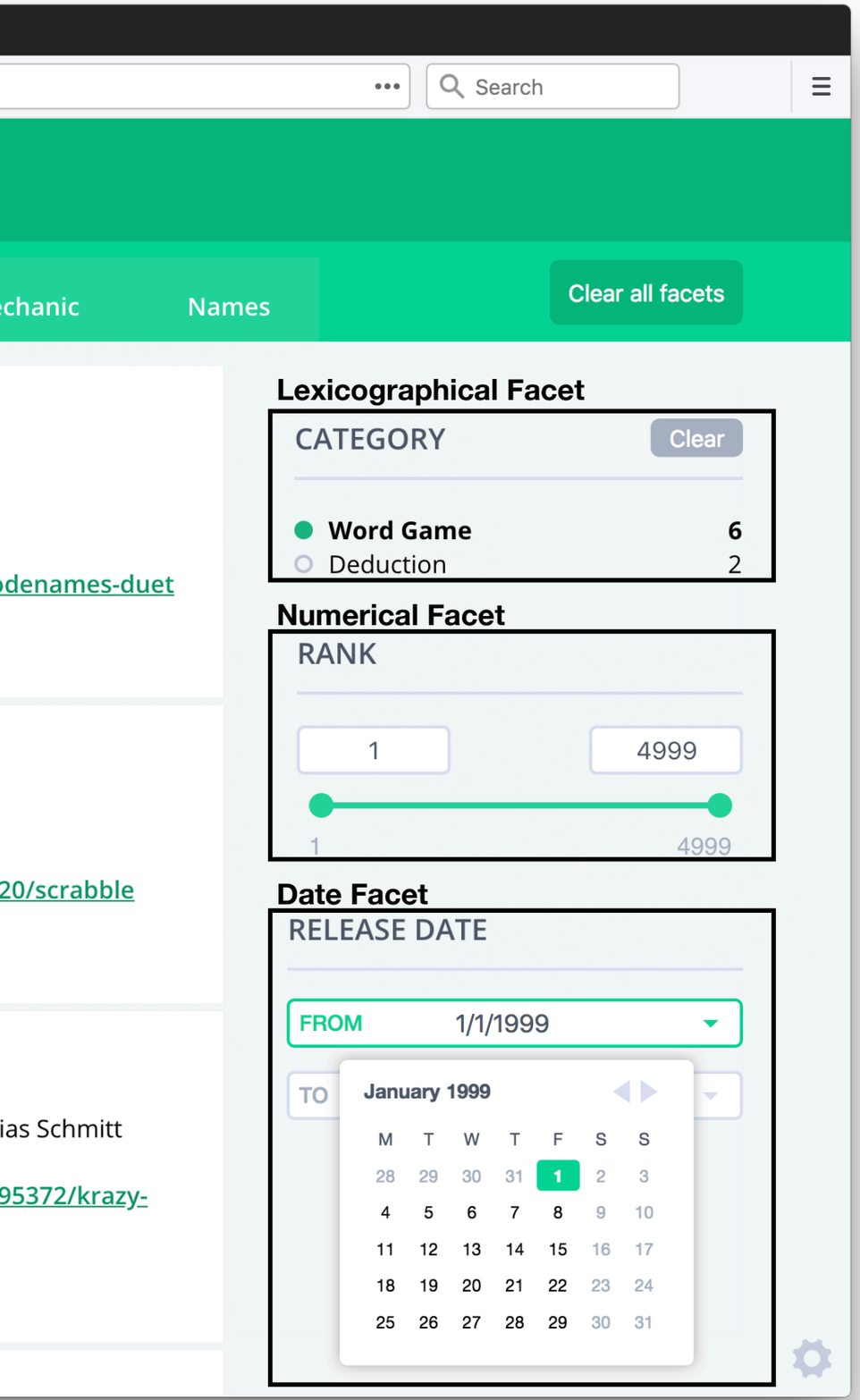
rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575
3	Through the Ages: A New Story of Civilization	Vlaada Chvátil	Card Game#Civilization#Economic	https://boardgamegeek.com/boardgame/182028/through-ages-new-story-civilization	2	4	240	2015	8.65702
4	Twilight Struggle	Ananda Gupta#Jason Matthews	Modern Warfare#Political#Wargame	https://boardgamegeek.com/boardgame/12333/twilight-struggle	2	2	180	2005	8.35188
5	Terraforming Mars	Jacob Fryxelius	Economic#Environ	https://	1	5	120	2016	8.38331

Configuration parameters (7)

- **ordering**
  - array which describes how the different columns will be ordered.
  - Supported ordering: lexicographical, numerical, by date
  - By default column entries are ordered lexicographically
- **Example:** ordering = [rank:1.0, min_players:1.0, max_players:1.0, avg_time:3.0, year:4.0, avg_rating:1.5]

rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575
3	Through the Ages: A New Story of Civilization	Vlaada Chvátil	Card Game#Civilization#Economic	https://boardgamegeek.com/boardgame/182028/through-ages-new-story-civilization	2	4	240	2015	8.65702
4	Twilight Struggle	Ananda Gupta#Jason Matthews	Modern Warfare#Political#Wargame	https://boardgamegeek.com/boardgame/12333/twilight-	2	2	180	2005	8.35188

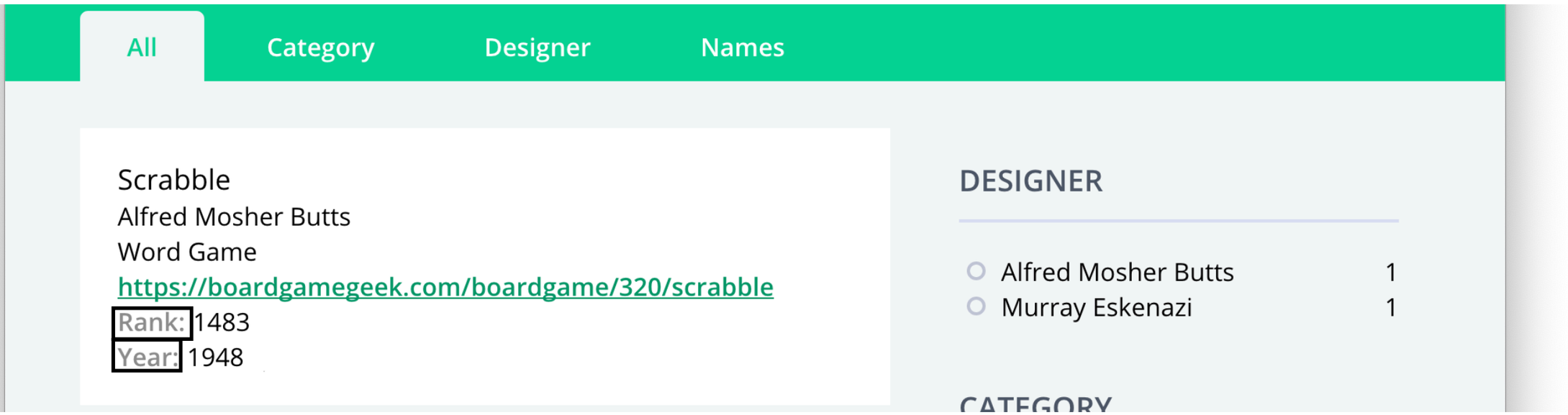
Configuration parameters (7)



- **ordering**
- different facet interface components for different orderings

Configuration parameters (8)

- **label**
  - arrays of columns which entries should be prefixed by the column name in search results to improve their informative value

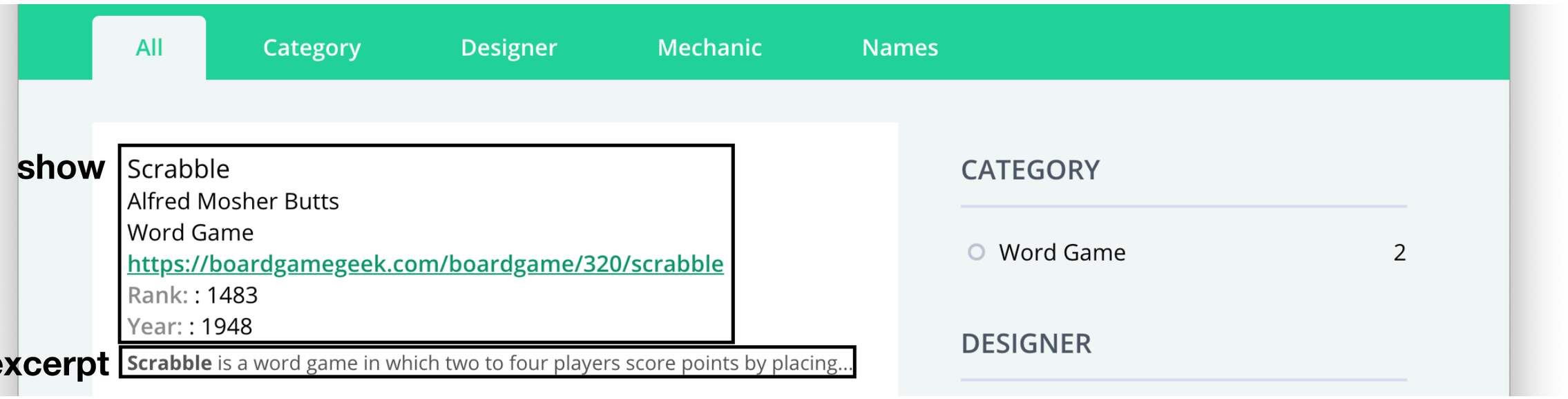


- **Example:** label = [rank, min_players, max_players, avg_time, avg_rating]

rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575

Configuration parameters (9)

- **show & excerpt**
  - arrays of columns which entries should be prefixed by the column name in search results to improve their informative value



- **Example:** show = [rank, names, designer, category, bgg_url, year], excerpt = []

rank	names	designer	category	bgg_url	min_players	max_players	avg_time	year	avg_rating
1	Gloomhaven	Isaac Childres	Adventure#Exploration#Fantasy#Fighting#Miniatures	https://boardgamegeek.com/boardgame/174430/gloomhaven	1	4	150	2017	9.0131
2	Pandemic Legacy: Season 1	Rob Daviau#Matt Leacock	Environmental#Medical	https://boardgamegeek.com/boardgame/161936/pandemic-legacy-season-1	2	4	60	2015	8.66575

Analysis Steps

1. Input File Analyser

- extracts relevant features from the input dataset

2. Column Classifier

- uses a classification algorithm to derive the CompleteSearch configuration parameters from the features we collected by the Analyser

Analyser (1)

1. Column Separator detection

```
rank;names;designer;category;bgg_url;min_players;max_players;avg_time;year;av
1;Gloomhaven;Isaac Childres;Adventure#Exploration#Fantasy#Fighting#Miniatures
2;Pandemic Legacy: Season 1;Rob Daviau#Matt Leacock;Environmental#Medical;htt
3;Through the Ages: A New Story of Civilization;Vlaada Chvátil;Card Game#Civi
4;Twilight Struggle;Ananda Gupta#Jason Matthews;Modern Warfare#Political#Warg
5;Terraforming Mars;Jacob Fryxelius;Economic#Environmental#Industry / Manufac
```

- Supported separators: `{",", "\t", ";", ".", "|", ":", "#", "/"}`
- in CSV file the column separator count is the same in every row

Analyser (2)

3. Column Parsing

3.1. Item Index Generation

- Avoids reprocessing reoccurring items

ColumnA
alpha
beta
alpha
alpha
beta
gamma

i	Item	Occurrence
0		1
1	alpha	3
2	beta	2
3	gamma	1

Analyser (3)

3. Column Parsing

3.2. Column-based feature determination

- **Fill rate**

$$fill\ rate = \frac{\sum_{i=0}^{n} occurrence(item_i) - occurrence(empty\ item)}{\sum_{i=0} occurrence(item_i)}$$

i	Item	Occurrence
0		1
1	alpha	3
2	beta	2
3	gamma	1

$$fill\ rate = \frac{1+3+2+1-1}{1+3+2+1} = \frac{6}{7} \approx 0.86$$

Analyser (4)

3. Column Parsing

3.2. Column-based feature determination

- **Uniqueness**

$$uniqueness = \frac{n}{\sum_{i=1}^{n} occurrence(item_i)}$$

i	Item	Occurrence
0		1
1	alpha	3
2	beta	2
3	gamma	1

$$uniqueness = \frac{3}{3+2+1} = \frac{3}{6} = 0.5$$

Analyser (5)

3. Column Parsing

3.3. Item-based feature determination

- **Check item against various common data type/formats** by a running it through a set of boolean pattern matchers:
  - **Numeric-value** matcher (“123”, “1,23”, “1.23”, “1.000,23”, “1,000.23”)
  - **Incremental-index** matcher
  - **Boolean** matcher (“0”, “1”, “true”, “false”, “Y”, “N”, “yes”, “no”)
  - **Value-with-unit** matcher (“$10”, “10m”, “10m^2”)
  - **Phone-number** matcher (“+352 123 456 - (12)”)
  - **Date** matcher (“d.m.yy”, “dd/mm/yy”, “mm-dd-yyyy”, “yyyy/mm/dd”)
  - **Timestamp** matcher (“20180101T235959Z”, “2018-01-01T23:59:59+00:00”)
  - **Email** matcher (“foo@email.com”, “foo.bar@email.co.uk”)
  - **URL** matcher (“http://www.foo.com”, “https://foo.bar.co.uk”, “sftp://foo.com”)
  - **JSON** matcher (‘{“foo”: true, “bar”: false}’)
  - **XML** Matcher (‘ bar ’, ‘’)

Analyser (6)

3. Column Parsing

3.3. Item-based feature determination

- **Column Score for each pattern matcher**

$$columnPropertyScore = \frac{\sum_{i=1}^{n} matcher(item_i) \cdot occurrence(item_i)}{\sum_{i=1}^{n} occurrence(item_i)}$$

i	Item	Occurrence	Numeric value matcher
0		1	0
1	1993.02.01	1	0
2	123,456	1	1
3	987	2	1

Example:

$$numericValueScore = \frac{0*1+1*1+1*2}{1+1+2} = \frac{3}{4} = 0.75$$

- **Additionally properties:**
  - Item Length
  - Word count
  - Character type occurrence (letter, digit, symbol)
  - Letter/Digit ratio

Analyser (7)

3. Column Parsing

3.4. Feature independence

- Column classification will rely on the “Naive Bayes” algorithm, which makes the assumptions that different features are independent from each other

➜ Problem: Mutually exclusive properties are not independent

➜ Only predominate property is retained

➜ Reduction to two features:

  - mutually exclusive property type
  - mutually exclusive property score

Analyser (8)

3. Column Parsing

3.4. Subitem separator detection

- Item split into subitems for each supported separator
  `{",", "\t", ";", ".", "|", ":", "#", "/"}`
- Separator can be invalidated by either of the following rules:
  - Separator is first or last character of an item
  - Two same separators occur directly next to each other
  - Separator following by space ➜ symbol likely part of of a sentence
- Example:

```
Economic#Environmental#Industry / Manufacturing#Territory Building
```

Subitem separator	Subitems
#	Economic, Environmental,Industry / Manufacturing, Territory Building
/	

Analyser (9)

3. Column Parsing

3.4. Subitem separator detection

- The results subitem indexes are evaluate in the same matter as the item index previously:
    - Column-based feature determination
    - (Sub)item-based feature determination (Characterisation, Column scores)
- Additionally: list occurrence and subitem count per item

Analyser (10)

4. File property summary

4.1. Noisy feature elimination

- At best only one subitem separators can be correct

  ➜ Find most likely separator

➜ Evaluation using the property scores from pattern matchers

- In best case, the subitems should have the same properties

➜ Property scores are optimal when we are either 0 or 1

Analyser (11)

4. File property summary

4.1. Noisy feature elimination

- Calculate subitem separator

  With m = count of matcher scores:

$$separator\ score = \prod_{i=1}^{m} \begin{cases} propertyScore_i, & \text{if } propertyScore_i > 0.5 \\ 1 - propertyScore_i, & \text{otherwise} \end{cases}$$

Example:

Subitem Separator	Numeric-Value Score	Date Score
,	0.8	0.1
.	0.4	0

$$separatorScore(`,`) = \ 0,8 * (1 - 0,1) = 0.72$$

$$separatorScore(`.`) = (1 - 0,4) * (1 - 0) = 0.6$$

Data collection & training set (1)

- Training data is made up of 50 different datasets, which were chosen with the goal to get a many different data formats and as many columns as possible
- Every column in the datasets was labelled manually with the different parameter classes
- Example: Board Games Dataset Labels

colName	full-text	filter	facet	subitem separator	allow-subitems	field-format	show	excerpt	ordering	url	email	label
rank	0	0	1	6	0	0	0	0	1	0	0	1
names	1	1	0	6	0	0	1	0	0	0	0	0
designer	1	1	0	6	1	0	1	0	0	0	0	1
category	1	1	1	6	1	0	1	0	0	0	0	0
bgg_url	0	0	0	6	0	0	0	0	0	1	0	0
min_players	0	0	1	6	0	0	0	0	1	0	0	1
max_players	0	0	1	6	0	0	0	0	1	0	0	1
avg_time	0	0	1	6	0	0	0	0	1	0	0	1
year	0	0	1	6	0	0	1	0	1	0	0	1
avg_rating	0	0	1	6	0	0	1	0	1	0	0	1

Data collection & training set (2)

- Training set is formed by combining the Analyser output with the labels

  ➜ Training set contains a record for every column in the collected datasets

	Column Properties								Column Labels			
		Item Properties			Subitem Properties							
Column Name	Fill Rate	Uniqueness	Length	…	Subitem separator	Uniqueness	Length	…	full-text	filter	facet	…
rank	1	1	3,778		-1	0	0		0	0	1	
names	1	0,994199	11,300		-1	0	0		1	1	0	
designer	1	0,491698	19,715		6	0,445	13,445		1	1	0	
category	1	0,446689	29,210		6	0,010	10,041		1	1	1	
bgg_url	1	1	57,381		5	0,500	28,191		0	0	0	
min_players	1	0,002	1		-1	0	0		0	0	1	
max_players	1	0,006	1,041		-1	0	0		0	0	1	
avg_time	1	0,015	2,259		-1	0	0		0	0	1	
year	1	0,021	3,993		-1	0	0		0	0	1	
avg_rating	1	0,9881	6,867		3	0,487	2,933		0	0	1	

Naive Bayes

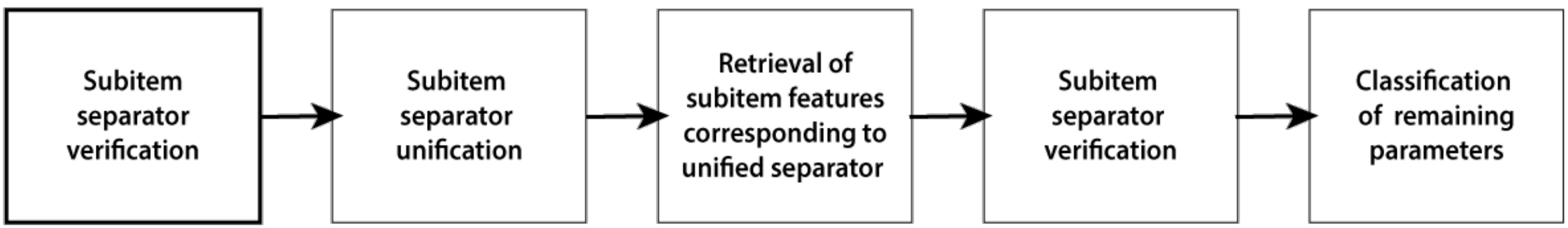
Naive Bayes assigns a problem instance $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, where $x_1, ..., x_n$ represent the values of the different features to a finite set of classes $\mathbb{C} = \{c_1, c_2, ..., c_k\}$

The probability of a problem instance being in class $c \in \mathbb{C}$ is expressed by the conditional probability $p(c \mid x_1, ..., x_n)$

$$p(c \mid \boldsymbol{x}) = \frac{p(c)}{p(\boldsymbol{x})} \prod_{i=1}^{n} p(x_i \mid c)$$

“Naive Bayes” makes the assumption that the different features are independent from each other

Classification (1)



- Binary classification using the *allow-subitems* labels to check which columns are likely to have a valid subitem separator

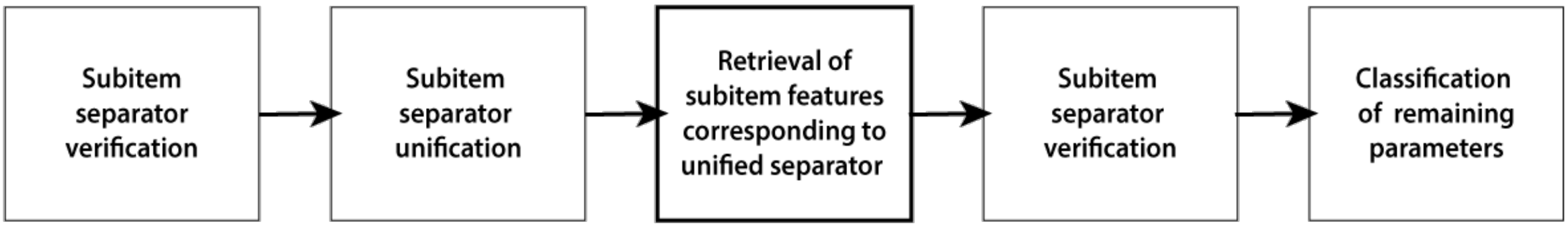
Classification (2)

- CompleteSearch only allows a single subitem separator for the entire file

  ➜ Find common separator from the separators that passed the verification step

  - valid separator that occurred most often
  - joint probability of valid separators from verification step as tie breaker

Classification (3)



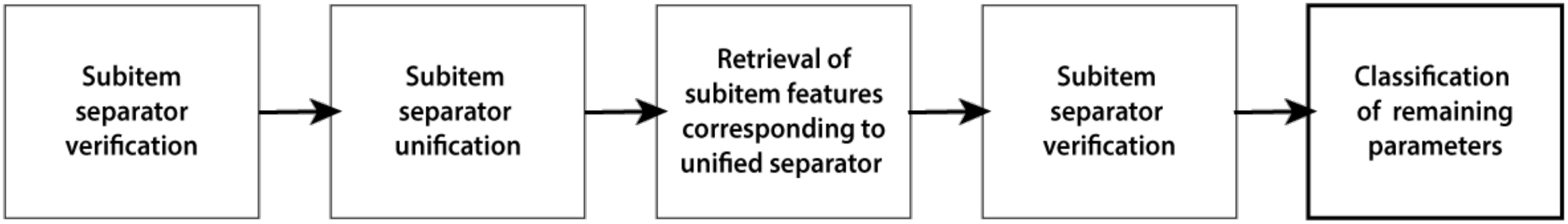
- Swap subitem feature data for columns where the subitem separator was not correctly chosen by Analyser

Classification (4)

- After updating subitem data, we determine to which columns the common subitem separator should be applied to

  ➜ CompleteSearch allow-subitems parameter

Classification (5)



Parameter	Classes
full-text	true, false
filter	true, false
facet	true, false
field-format	plain-text, JSON, XML
show	true, false
excerpt	true, false
ordering	lexicographical, numerical, date
url	true, false
email	true, false
label	true, false

- Classification of remaining parameters by using either the full item properties or the subitem properties depending on the results for the allow-subitems parameter

	Column Properties								
		Item Properties			Subitem Properties				
Column Name	Fill Rate	Uniqueness	Length	…	Subitem separator	Uniqueness	Length	…	allow-subitems
rank	1	1	3,778		-1	0	0		0
names	1	0,994199	11,300		-1	0	0		0
designer	1	0,491698	19,715		6	0,445	13,445		1
category	1	0,446689	29,210		6	0,010	10,041		1
bgg_url	1	1	57,381		5	0,500	28,191		0

Evaluation

Classification accuracy for test set made up by 25% of the labelled datasets

Parameter	**accuracy**
subitem-separator	0.833333
allow-subitems	0.993671
full-text	0.858650
filter	0.873418
facet	0.734177
field-format	1.000000
show	0.725738
excerpt	0.951477
ordering	0.970464
url	0.983122
email	0.989451
label	0.736287

Web App Demo

Questions?