Analysis of the use of text fields in the Consumer Reports database

prmr 19 January 2014

I analyzed all product records for the use of the following text fields:

- Description
- Highs
- bottomLine
- Review
- Summary

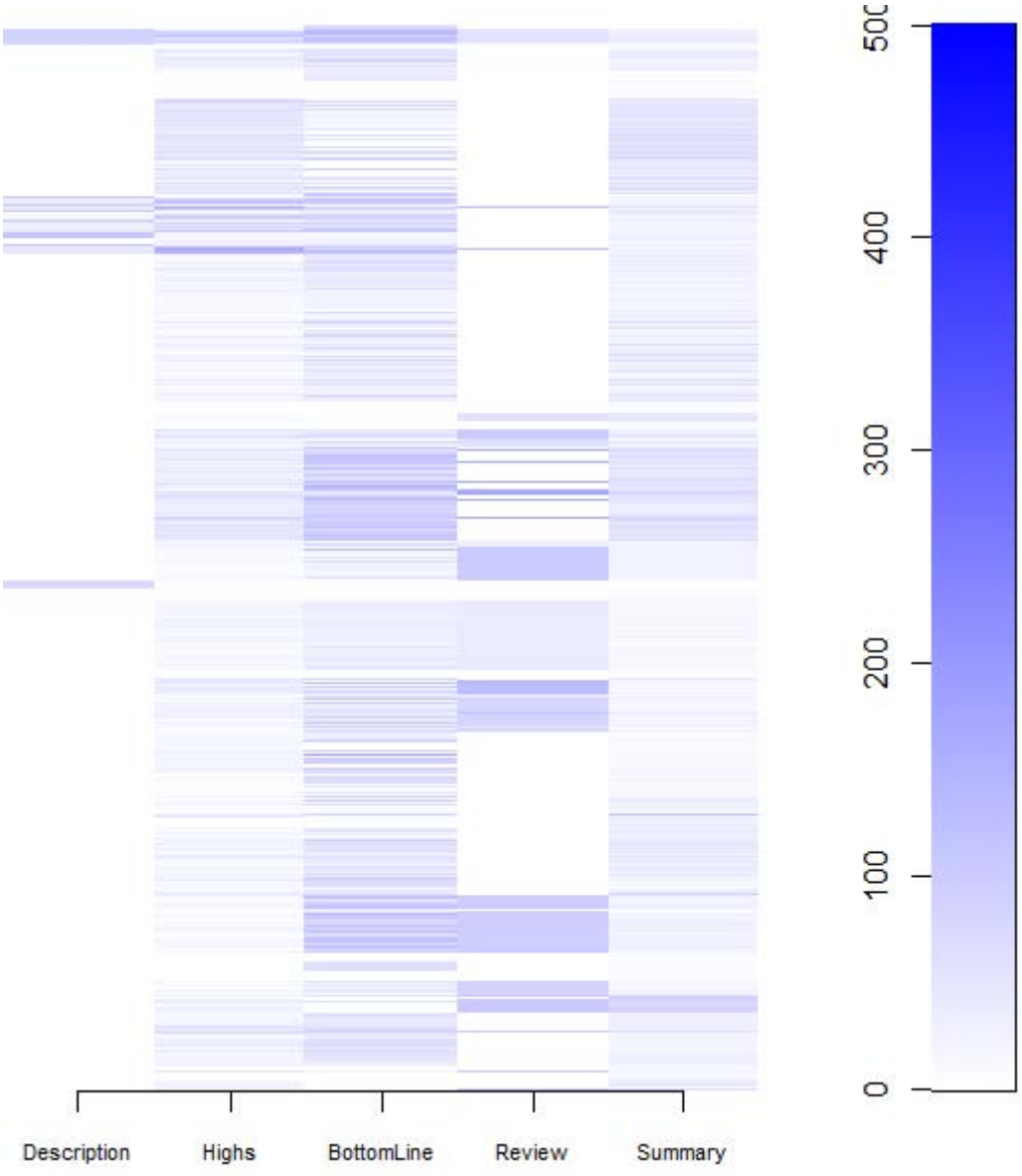
Visualization

For each product, I computed the number of words in each text field. The number of words was computed using the java String.split("\\s") method, i.e., any blank separated two tokens. I report the results per product category using a head-map type of matrix, where the level of blue saturation represents the number of word for a particular field. Each row represents a product. Next to each title, I report the number of product records that had any text over the total number of products analyzed, together with the percentage. The matrix only shows records for where there is some text. In other words, all record with no text are not accounted for in the visualizations.

The main observation is that text fields are not used in a consistent pattern across categories, and in many cases also within categories. For the money field, I didn't find any text.

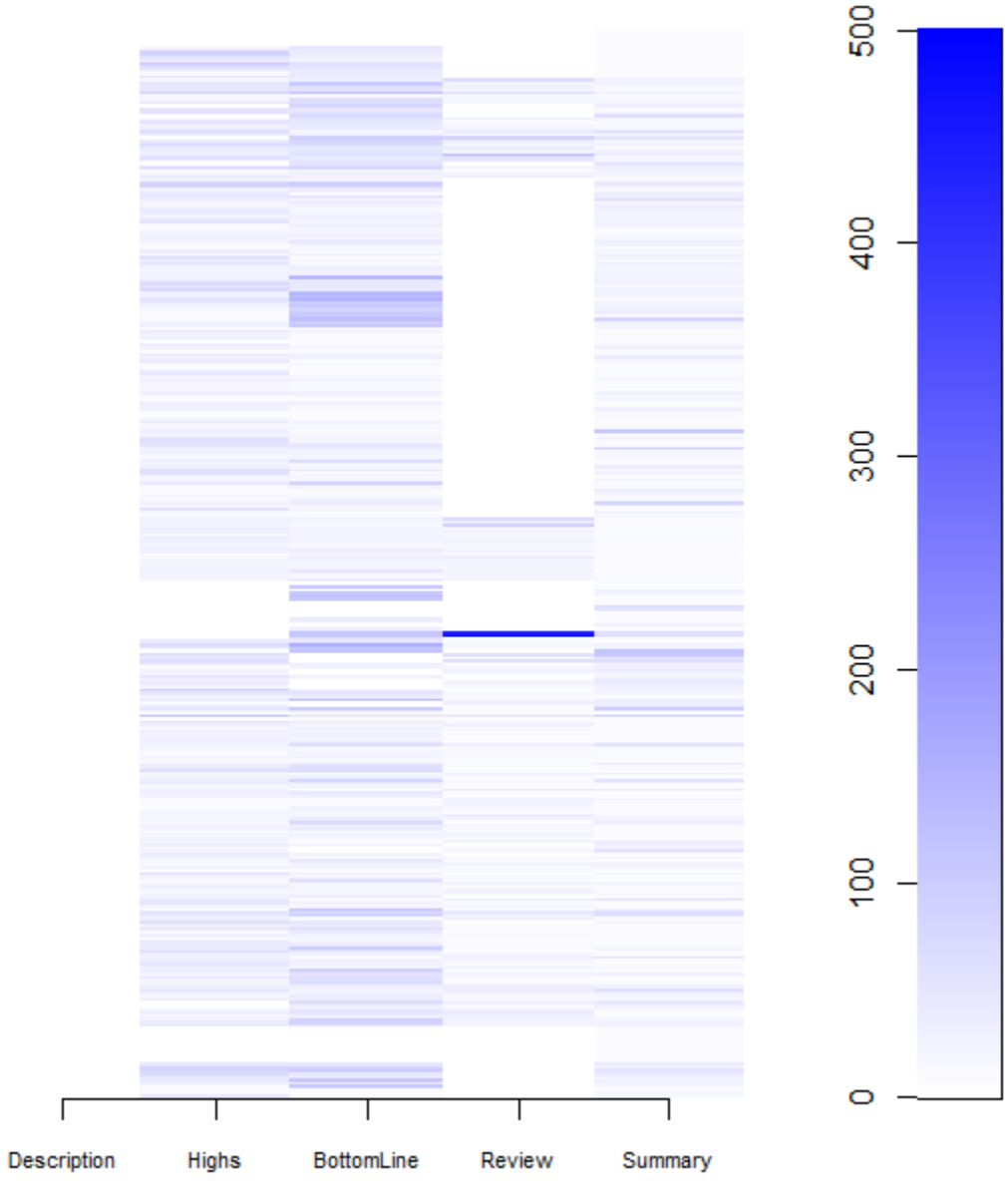
Appliances (2022/15116 = 13.4%)

A large number of appliances do not have a description field. Reviews are only present about half the time.



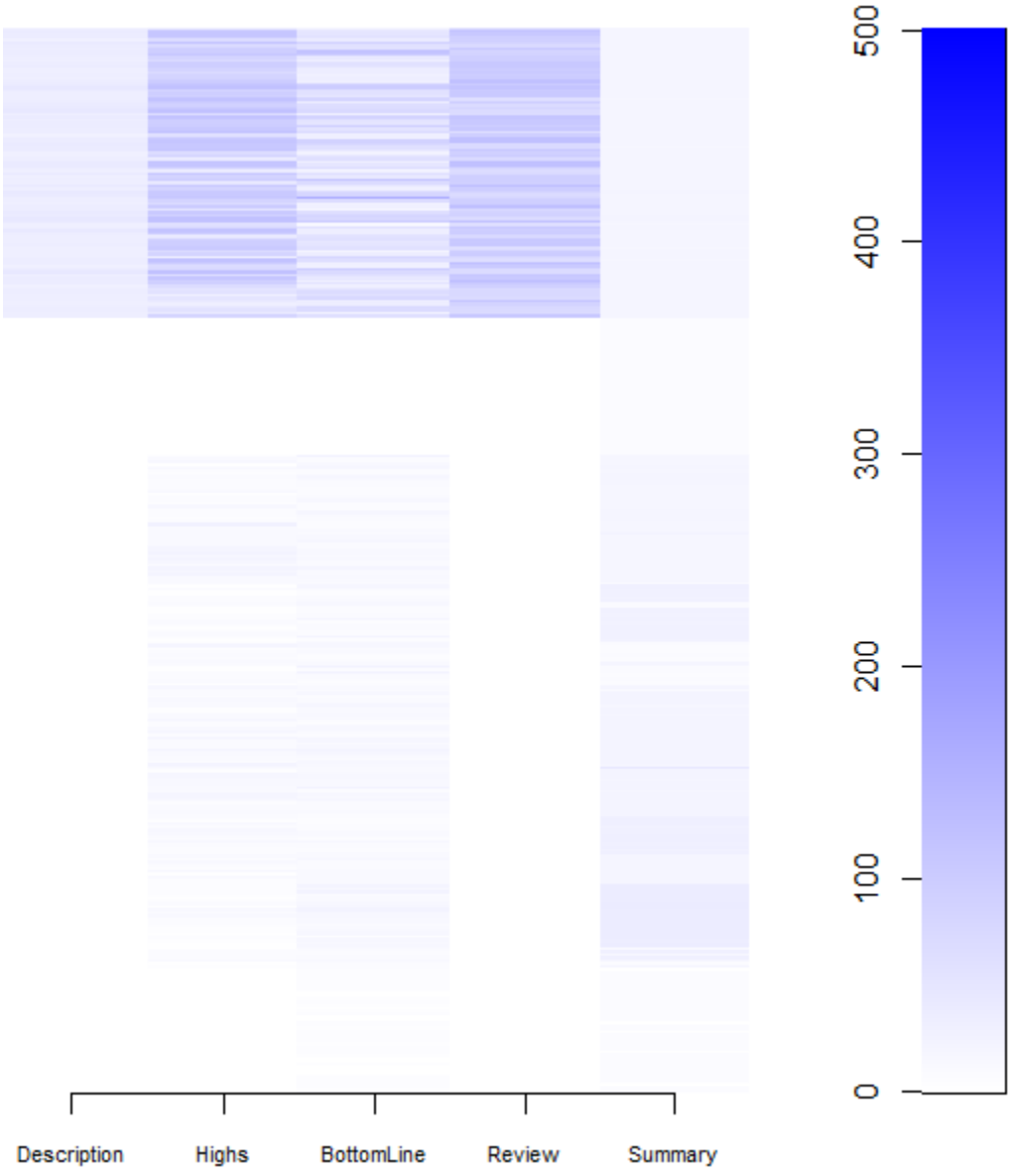
Babies and Kids (332/334 = 96.5%)

The description field is not used.



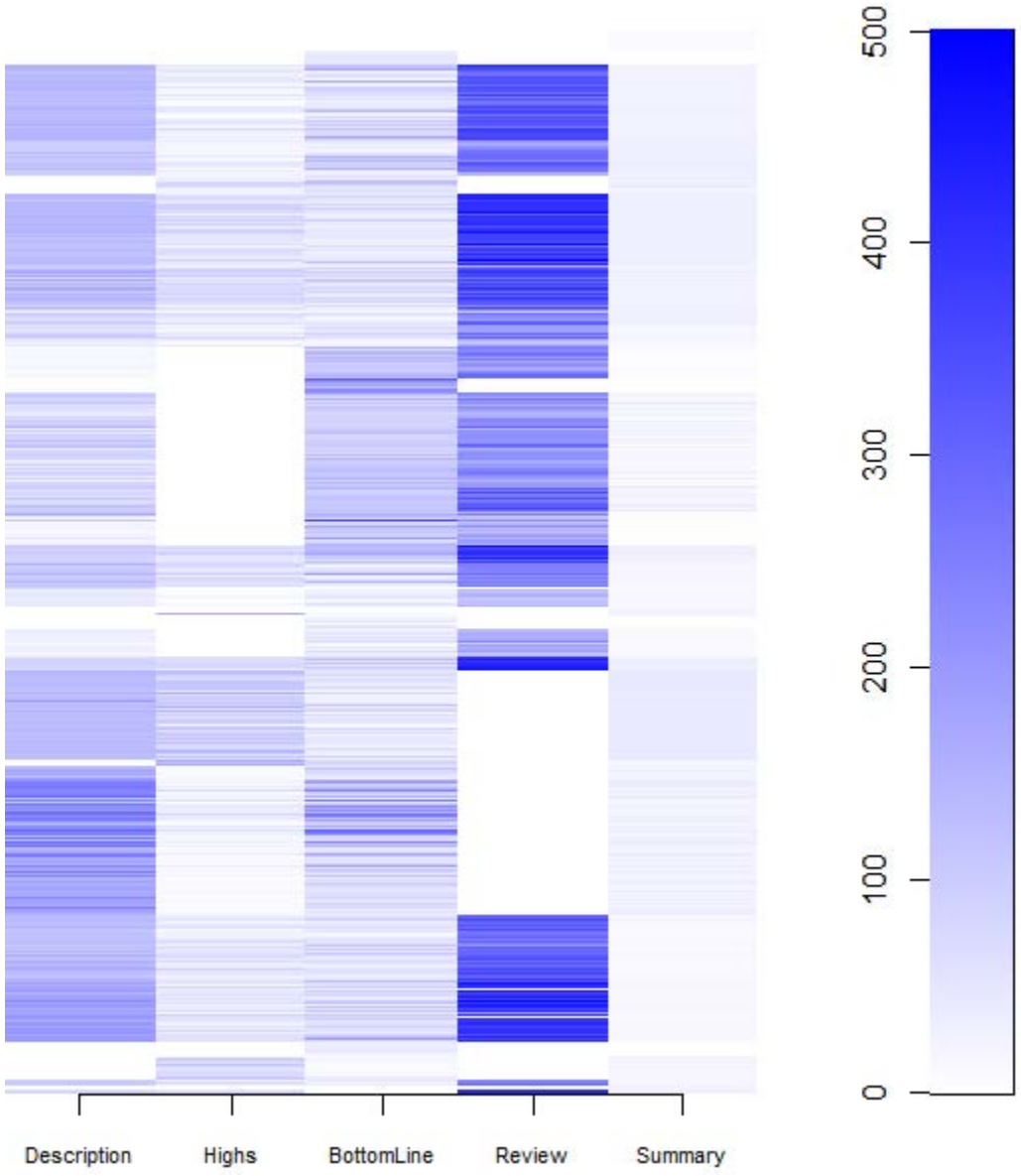
Cars (359/416 = 86.3%)

About a fifth of the products have detailed text across all fields, and the rest only a few words an no description.



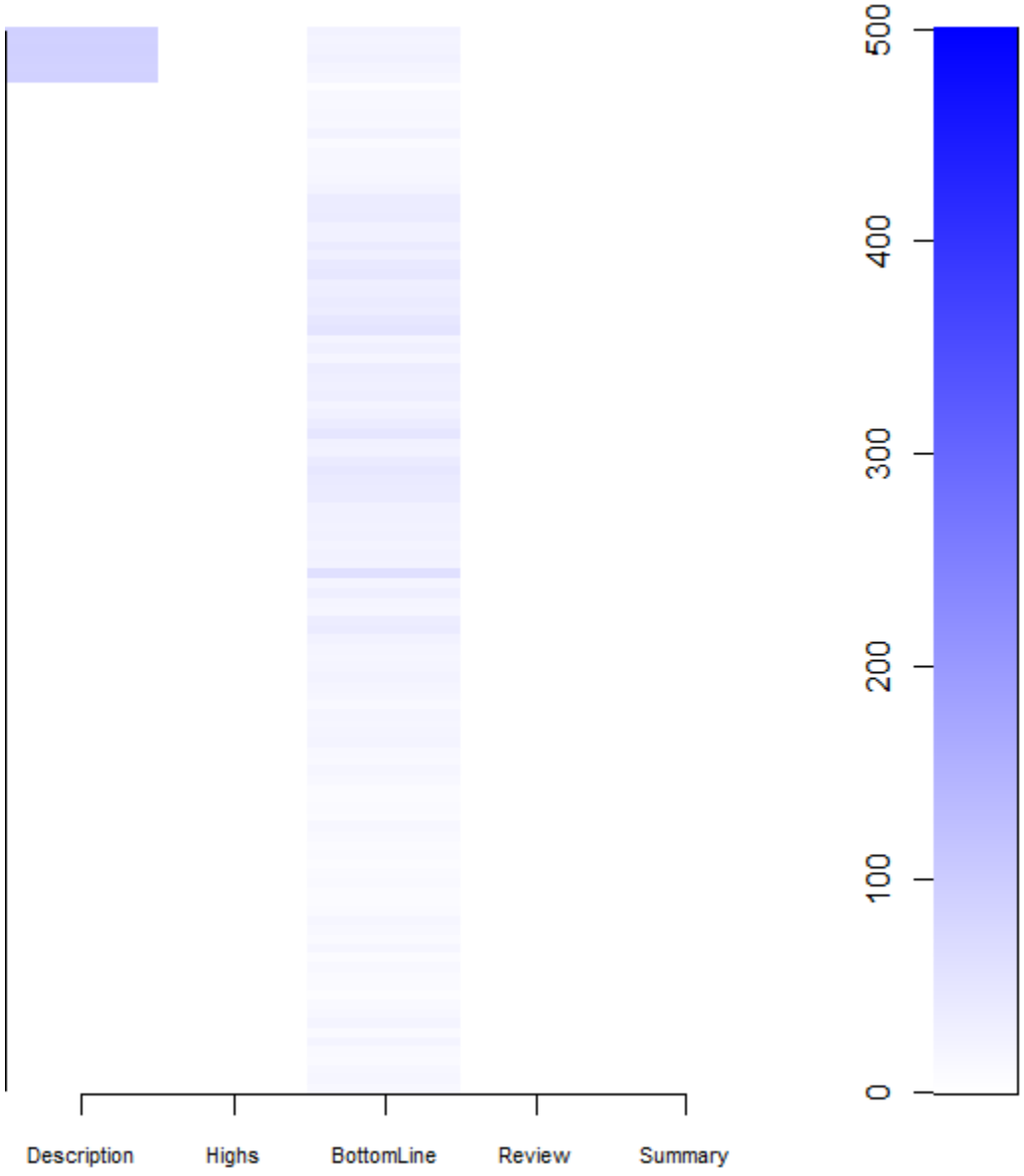
Electronics and Computers (1359/4254 = 32.0%)

Extensive textual information. Some highs and reviews missing, fortunately without overlap.



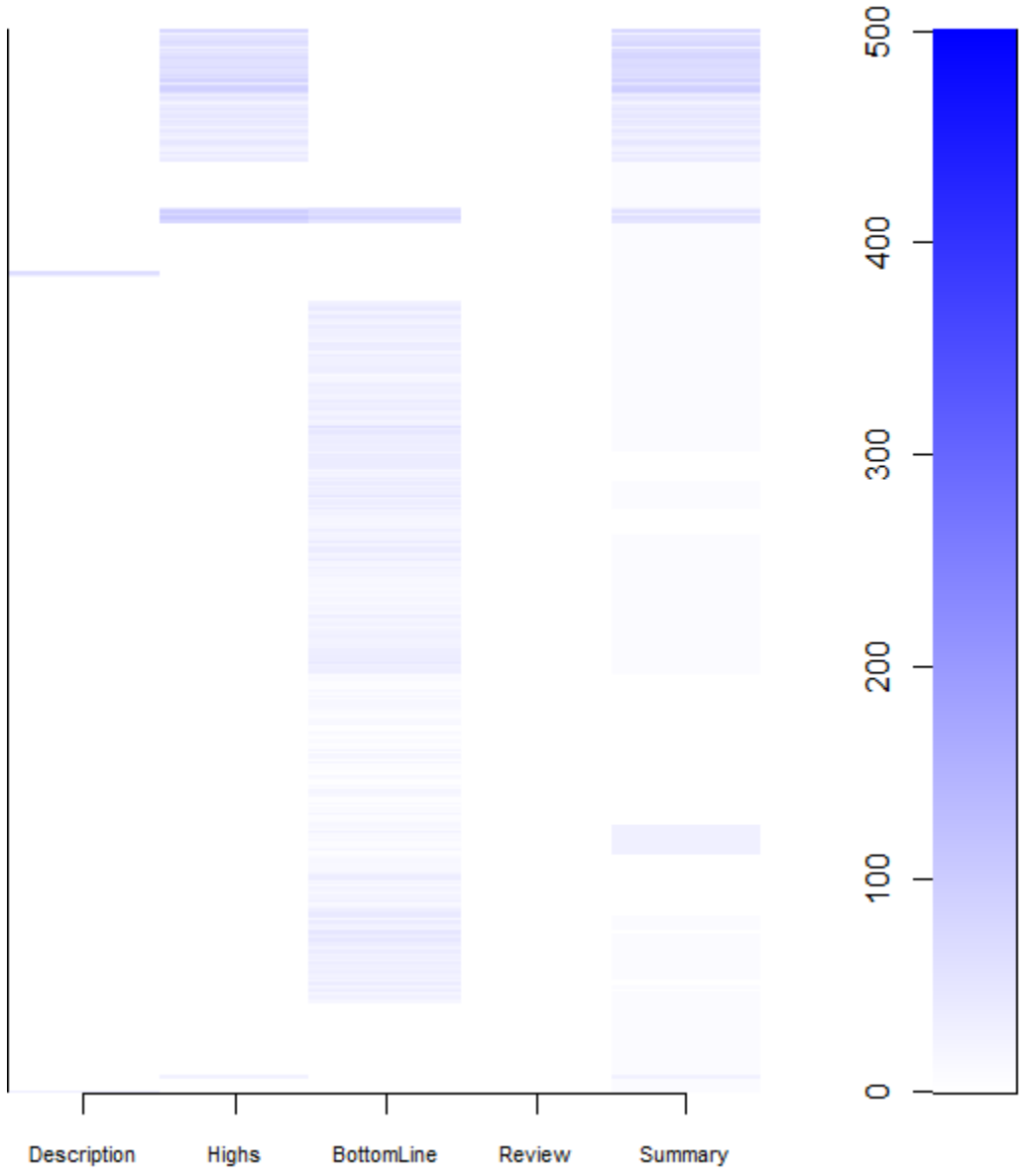
Food (114/254 = 44.9%)

Only the bottom line field is used here, except for a few items with a description.



Health (415/570 = 72.8%)

Here mostly bottom line and summary.



Home & Garden (1169/3058 = 38.2%)

Here mostly high, bottom line, and summary. A few items have very long reviews.