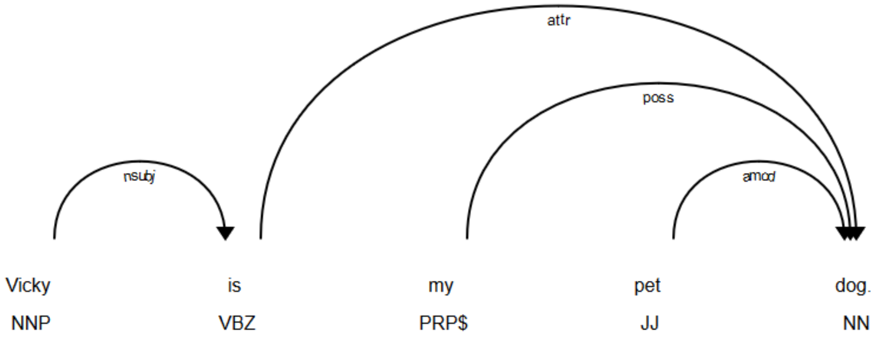
Notes

Approach

1. To understand, what is important for the customers, let's extract all of the words which they wrote (tokens) and count the amount of words occurrences: higher value will mean that the word is more important.
2. To understand, what customers like and dislike, let’s estimate the average review rating for each extracted token: higher value will mean that customers like the instance represented by the token.
3. Some of the tokens were not interpretable by themselves. Bigrams* were compiled for tokens annotation, also bigrams’ occurrences and ratings were estimated.
4. Some of the tokens were not informative, hence they were excluded from the analysis. Also, the tokens with occurrence below 30 were not considered for further research.
5. After these manipulations, 79 tokens left, and they were grouped/categorized.
6. The following categories were extracted: aroma; brand; effect (after product application); feature (product’s features); hydrating (this topic is very important for the customers, hence was separated from ‘effect’); package; usage (time and other peculiarities of the application).

*Bigrams – pairs of words (phrases) extracted using dependency tree.

Dependency tree example:



Deliverables

The archive includes several files, the main are:

- ‘3_tokens.xlsx’ – to extract what is important for customers and what they like/dislike;
- ‘4_reviews_categorized.xlsx’ –reviews with categories:
  - a review may belong to several categories;
  - ‘tokenTOP#’ column – the most popular token from the review.

Best Regards,
Artem Drofa
artem.drofa@gmail.com
+7 (915) 206-12-26
(including WhatsApp / Telegram)