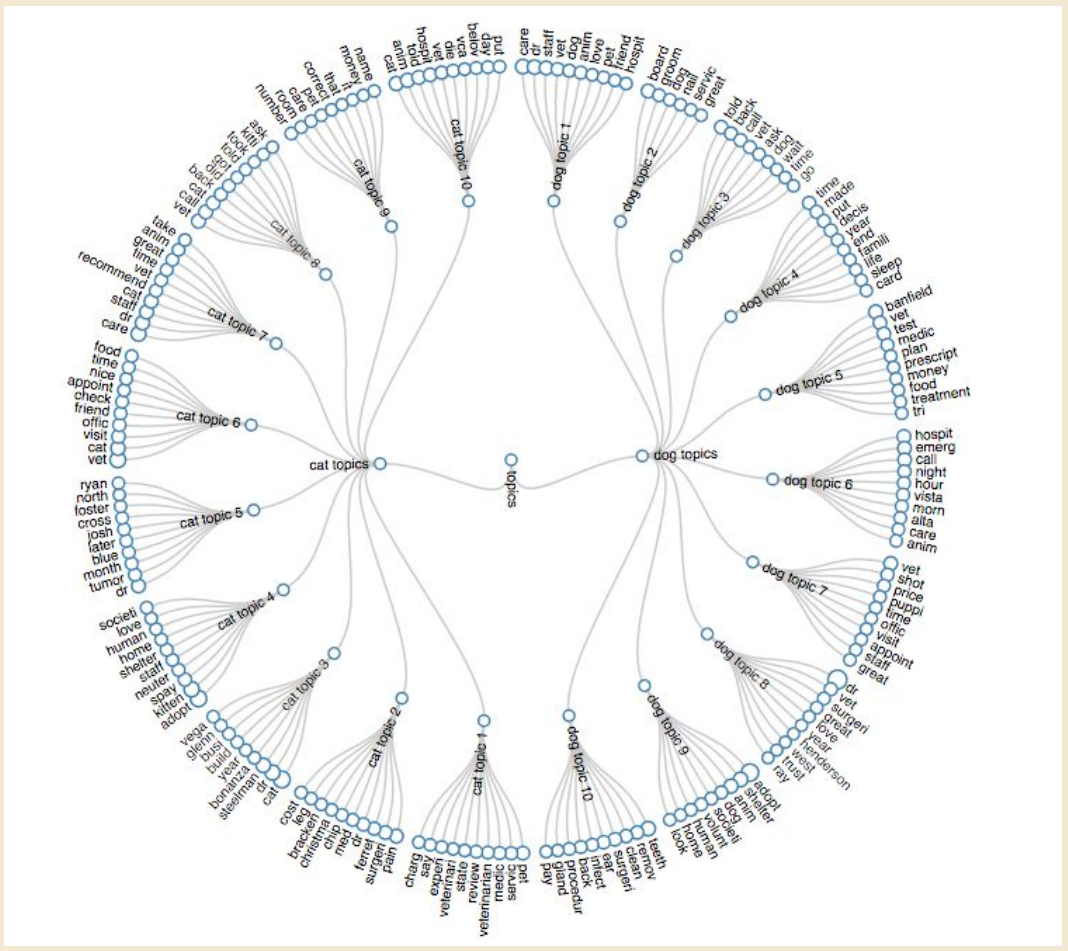
T1: Exploration of Data Set

Dog people vs Cat people



Igor Milla

13.09.2015

GETTING DATASETS

Original Yelp dataset was imported into *mongodb*, and as original data was in the JSON format it was straightforward process. After I wrote a *node.js* script to generate next files:

- reviews.txt – contains reviews of all Veterinarian businesses.
- cat-reviews.txt – contains all reviews of veterinarian businesses which have mention of cat (or kitten, kittens, cats, etc)
- dog-reviews.txt – contains all reviews of veterinarian businesses which have mention of dog (or puppy, dogs, etc)

APPLYING TOPIC MODEL

I used LDA topic model from *MeTA* toolkit, with next configuration:

> *[[analyzers]]*
> *method = "ngram-word"*
> *ngram = 1*
> *filter = "default-unigram-chain"* # as workaround for sentence segmentation tags
>
> *[lda]*
> *inference = "gibbs"*
> *max-iters = 1000*
> *alpha = 0.1*
> *beta = 0.1*
> *topics = 10*
> *model-prefix = "lda-model"*

PREPARING DATA

With a little help of regex, and one more node.js script I converted default outputs from MeTA lda into JSON format suitable for visualisation with d3.js

SOURCE

https://github.com/igormilla/dataminingcapstone

RESULTS

Task 1.1

I used tree-mapping for visualizing what people were talking in reviews of veterinarian businesses. Colors represent different topics, and sizes of the cells inside of the colored blocks reflects the weight of that particular node in that topic.

From this visualisation we can see that most people sharing how long they were waiting, how much they paid, how good and caring was a doctor, or what was the procedure their pet had in the veterinarian they reviewed.

Task 1.2

For this diagram I also used tree-mapping with the same configurations as for the previous task. The upper part shows topics from reviews that were mentioning dogs, and bottom part from reviews that were mentioning cats.

Diagram from the cover page, represents exactly the same thing, but using radial tree layout, with node sizes reflecting node weight in the topic.

Interesting to point that mentioning of money(cost, price, etc) occurring more in the reviews made by “cat people”, which makes me think that cat owners are less used to spend much money on their pet. And as I always considered myself a “cat person” i’m now asking myself: cat, as a pet, it’s not a decision of preference for some people, but raiser a choice made basing on money people willing to spend on their pet in the long run?