Sparking Joy with NLP

Identifying the Tidiers and the Hoarders

Will Karnasiewicz – April 2019

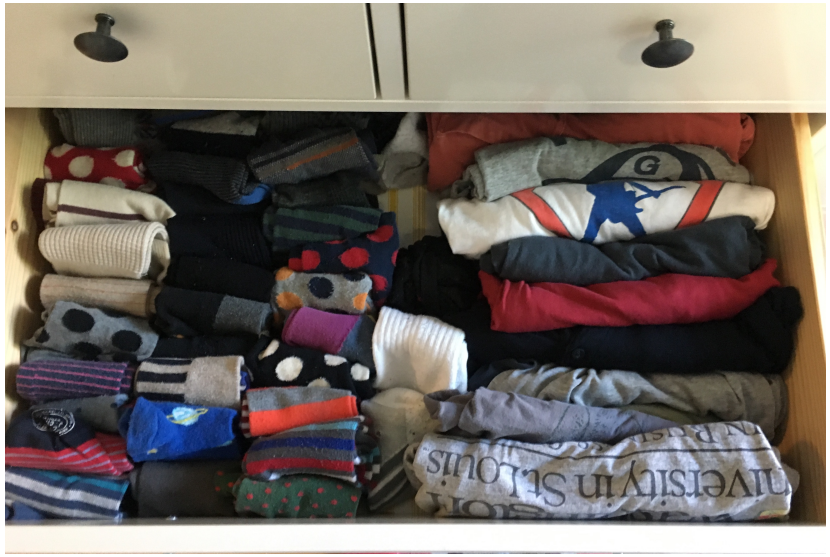
The Life-Changing Magic of Tidying Up

- Marie "KonMari" Kondo, professional organizing consultant
- "The Life-Changing Magic of Tidying Up" book first published in 2014
- Name one of Top 100 Most Influential People by TIME (2015)
- Published in 42 regions/countries, over 10 million copies sold
- Now worldwide phenomenon!
- Netflix series released January 1, 2019

The Konmari Method

1. Commit yourself to tidying up
2. Imagine your ideal lifestyle
3. Finish discarding first
4. Tidy by category, not by location
5. Follow the right order
6. Ask yourself if it **Sparks Joy**

Konmari Examples



My socks and t-shirts
(Brooklyn 2017)

Konmari Examples

Soil & water conservation organization in Ohio are Konmari fans too!

Reddit NLP: Problem Statements

Problem Statements

- Can a model identify posts from r/konmari vs. posts from r/hoarding?
- Can the model still work if I also include r/declutter and r/organization?
- What are the most important words?
- What posts does the model have trouble with?

Data Gathering & Preprocessing

Data Gathering

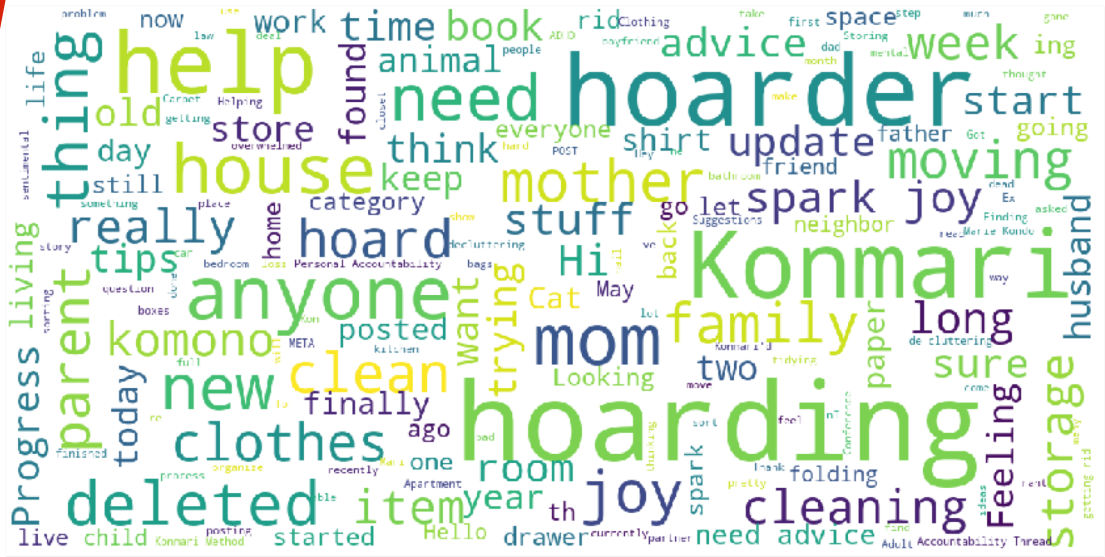
- Pulled data from Reddit API using PRAW
  - Limited in number of submissions pulled in each query
- Also pulled using Pushshift API to expand data set

Preprocessing:

- Removing duplicate posts
- Combining title text and self text (post) into 'fulltext'
- Binarizing target 'subreddit' column
  - (1 if konmari, 0 otherwise)
- Stemming / Lemmatization / Removing certain words?

Word cloud of full dataset

(All words incl.)



Notice that name of subreddits are frequent words

Modeling Summary

(all words incl.)

Model 1:

- Tfidf vectorizer, Naïve Bayes
  - 0.99 training score, 0.95 testing score (0.95 cv)

Model 2:

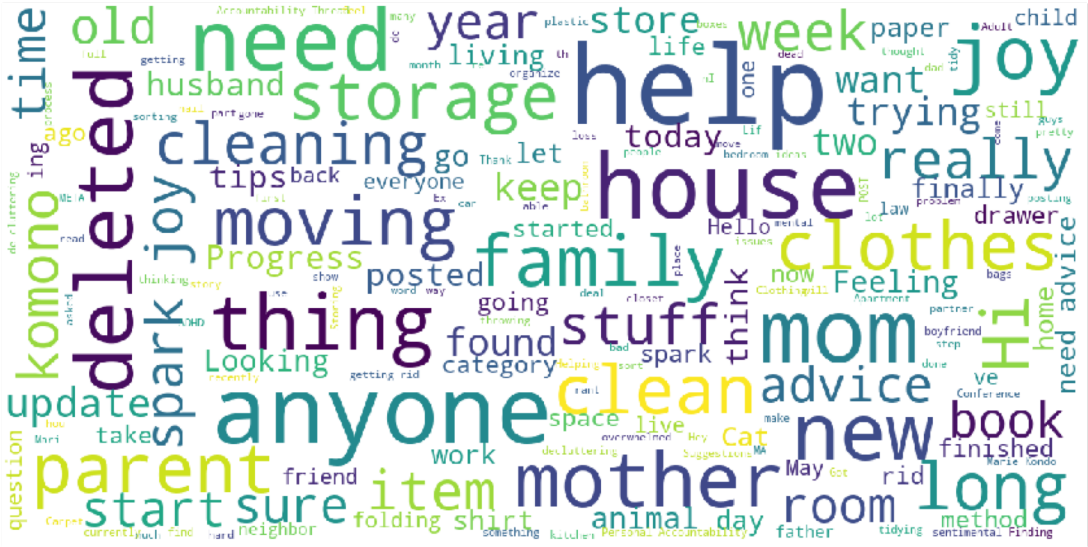
- Tfidf vectorizer, Logistic Regression
  - 0.9997 training score, 0.96 testing score (0.96 cv)

Model 3: (*if you hate deciding …)*

- Tfidf vectorizer, Voting Classifier
  - 0.99 training score, 0.95 testing score
- All models use (1,2) n-gram range and english stop words

Word cloud of full dataset

(No subreddit name words)



Sorry to all the mothers out there…

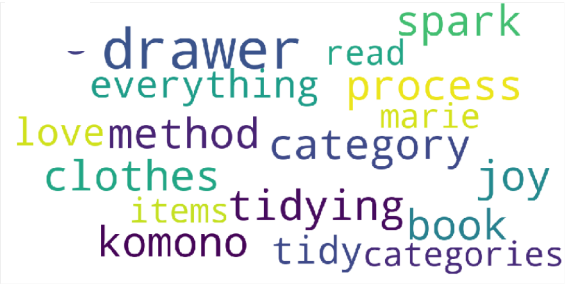
Modeling Summary

(Removed subreddit names)

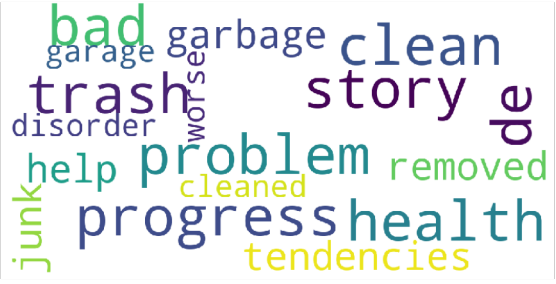
Dropped 'konmari' and 'hoard' variations from corpus.

<u>Modeled with Logistic Regression:</u>

- Tfidf vectorizer, Logistic Regression
  - Parameter changes: (1,1) n-gram range, No stop words
  - 0.9981 training score, 0.9350 testing score (0.9283 cv)



Konmari



Hoarding

Revisiting Problem Statements

Problem Statements Revisited

- Can a model identify posts from r/konmari vs. posts from r/hoarding? **YES**
- Can the model still work if I also include r/declutter and r/organization? Or if I remove certain words? **YES**
- What are the most important words? **Joy/method vs. clean/problem/trash** *(notice tone of Konmari)*
- What posts does the model have trouble with? **Crossposts? Advice gatherers?**

One More Thing: Testing the model with other text

With inspiration from Hussain, I wanted to see how the model performed with other text:

1. I found transcripts of each episode of Marie Kondo's Netflix show on https://www.springfieldspringfield.co.uk and web scraped
2. Fed the transcripts into model to see if it classifed as konmari

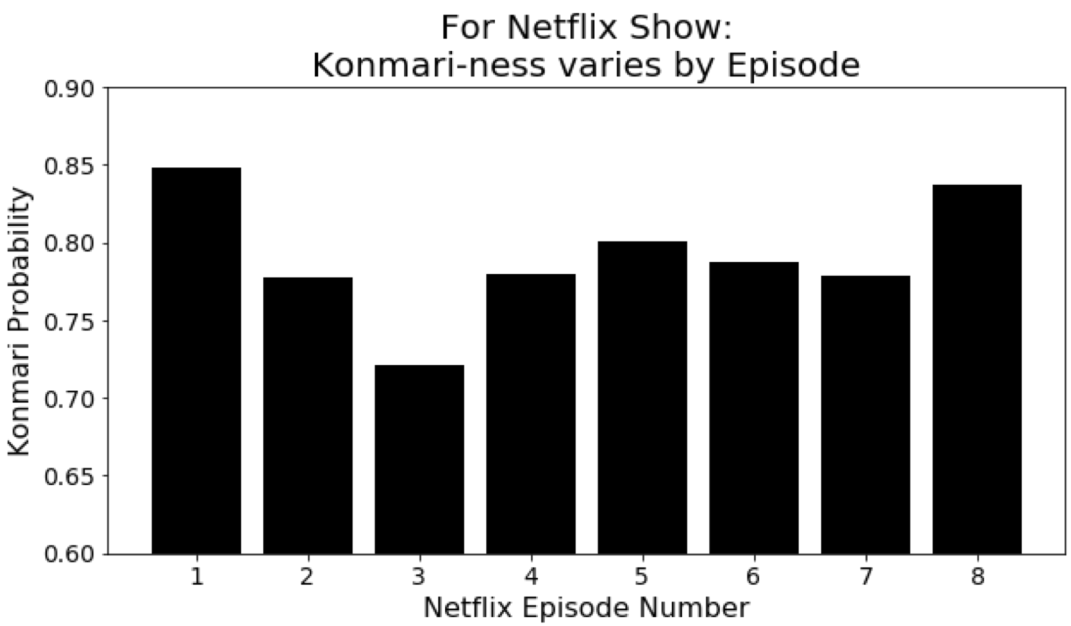
What were the results….?

Model works on Netflix episodes of Tidying Up with Marie Kondo

All episodes were correctly identified as Konmari!

(not terribly interesting)

Episodes have a varying level of Konmari-ness



(more interesting!)

Questions?

References

References

- Netflix "A Life Changing Magic of Tidying Up" (& GIPHY)
- http://www.mariekondobooks.com
- https://konmari.com/pages/about
- https://github.com/amueller/word_cloud
- https://www.springfieldspringfield.co.uk
- https://www.reddit.com