# Neuro Linguistic Programming

Mikkel Werling

Victor Møller Poulsen

December 2020

# 1 Abstract

To run: modeling_tree.py

5-50, (end with 51, not 50)

Whole grid: 1, 0.1, 0.01 for both.

Mikkel mode: eta = 1, all other combinations.

# 2 Introduction

## 2.1 From qualitative survey to quantitative data

### 2.1.1 Smarthome - what's the buzz?

### 2.1.2 Building on Marco et. al (2020)

The project is building on the work by Marco et. al (2020), which investigated . . .

Previous work was already done, they had some ideas

### 2.1.3 Additions to the framework

More data, more reproducible

### 2.1.4 Introduction to the hypothesis

On the basis of the paper by Marco et. al (2020), we investigated two research questions...

### 2.1.5 Why Topic modelling

Presentation of topic modelling and motivation for topic modelling

# 3 Data

Datadump for all of reddit, using two subreddits

Include plot for how different documents can be made on the basis of parent ids and posts

Using these different metrics, we branch out into H1 and H2

# 4 Methods

Here we also need to branch out into H1 and H2

## 4.1 Topic Modelling

Grid search for alpha, eta and k.

Using LDA, underlying assumptions and hyperparameters

### 4.1.1 Different Metrics to Evaluate Quality of Topics

Coherence as well as Cao Juan and Arun to get different measures of quality. We also looked at interpretability as...

Plots of the different evaluation metrics

NOTE: We could potentially run the whole grid over New Years, so that we can plot everything

## 4.2 Thoughts on choosing "atomic units"

Tree was higher in coherence, while thread was higher in the measure from Arun. So we had to take a more qualitative dive. Here we found that tree had several categories that seemed random (3 random words categories). Thread had overlap across categories, but seemed to capture more specific conversations. These were mainly due to different companies that people used. Moreover, the trees were made by pasting the submission text, which gives the submission text a lot of weight in the topic models. Here we focus on the common denominators between the two. NOTE: Here we could consider two plots - both showing the similarity and the disparity between thread and tree.

## 4.3 Topics and their distributions

We followed the division of the topics proposed in the qualitative study. Overall, we find that comfort and lighting as well as control and connectivity are the central themes of posts on the subreddits.

# 5 Results

## 5.1 H1

Do we find the same topics as the qualitative study?

- We find topics that resemble those found in the qualitative study to a large degree

- Light, connectivity and comfort dominates the topics (plot in topic_prevalence.py)

- First indications that most topics are concerned with "how to" and not a discussion board

## 5.2 H2

# 6 Discussion

## 6.1 H2

Are people even remotely concerned about privacy, security and trust? No.

- Trust and privacy are very rarely used on the forums. From investigating samples of trust and privacy, they do seem to talk about privacy concerns regarding their data. But as this proportion of posts are so small, we will not look further into it. It is unclear whether this is evidence against people being concerned about data privacy concerns or whether it is evidence for them just not discussing privacy concerns on Reddit.

- Security is the only word of interest which is used regularly. However, security is not used to relay concerns about the security of one's data, but rather the security of one's house. The topics where the word security appears in the top 10 words,

are regarding either security cameras or security systems. Potentially, security cameras could be a cause for concern among users regarding their privacy. We therefore sampled submissions, which had the security camera topic as the most dominant topic, as well as using the word "security". The overall picture was very clear; users are not worried about security cameras watching them, they are worried about which one to buy. This is a case of the tendency of the forum, to be very practically oriented and not a forum for discussion.

## 6.2   Mallet as a possible method instead of Gensim

## 6.3   Reddit as a "how-to" site

From the topics extracted from the corpus, it seems like the subreddits are primarily used to ask questions concerning specific products and how to implement them. Importantly for H2, it is used to much less degree as a discussion forum (for instance regarding security).

## 6.4   Use of seeded topic models?

We initially thought that seeded topic models would help us disentangle different senses of security. But it does not seem to be warranted - we do not find indications that security is used in the more abstract sense. However, security does come up in several different categories. Notably, it comes up in topics related to cameras, which could indicate worries about cameras surveying them or just interest in security cameras. To investigate this, we sample some of the most representative documents of the topics and investigate the tendency. NOTE: Right now this is placed in the bottom of "doc_from_topic.py", but could be in its own document. But the findings are very clear - people are worried about buying the right camera, not whether it is filming them.