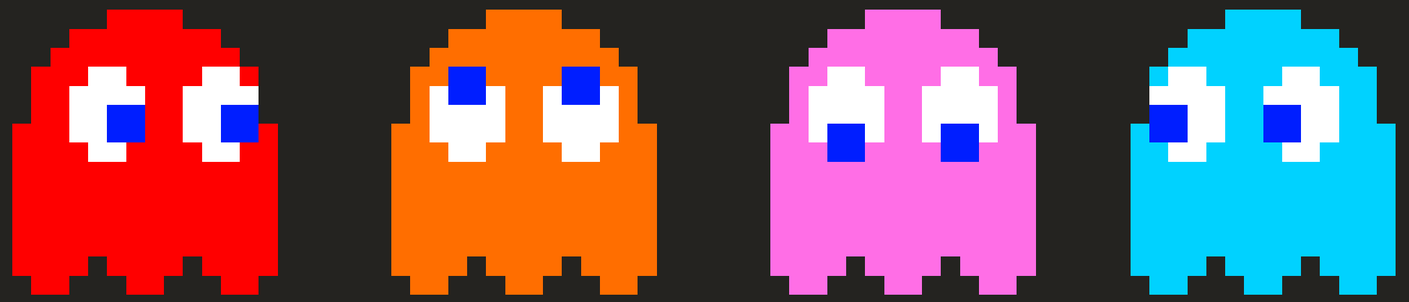
SICSS Sense

A GHOSTLY HERO'S JOURNEY:

A TEXT ANALYSIS ON EPP'S DISSERTATIONS METADATA FROM THE PAST 15 YEARS



May 23rd, 2025

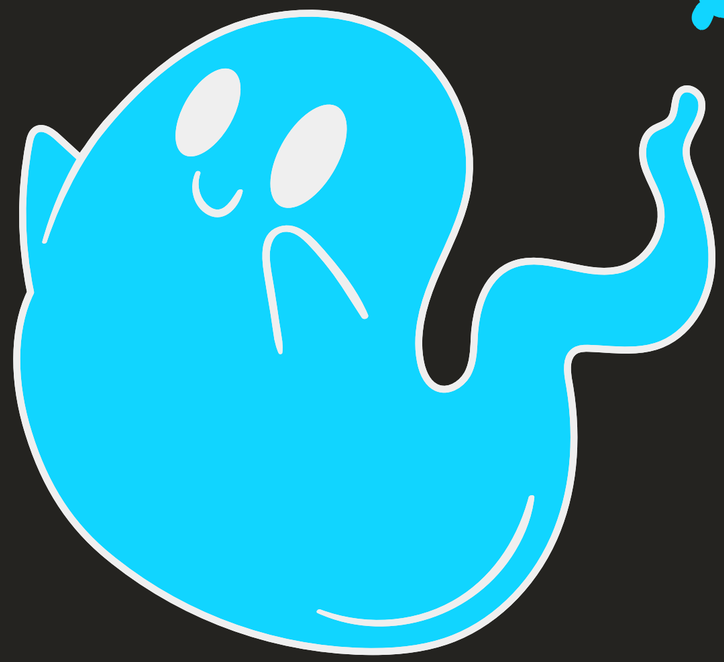
CONTENTS

THE CALL TO ADVENTURE

- The heroes received the call from the Engineering and Public Policy (EPP) Department at Carnegie Mellon University to analyse how research has changed over time:
  - Can we observe any trends through dissertations?
  - Impact of faculty additions / losses on dissertation content?
  - How have methodologies changed?
- Reason of the call: John, Hannah, and Stacy are selfish and made the project all about themselves

SUPERNATURAL AID

- The heroes got magical help from Group 6 to craft and download a dataset
- The dataset is comprised of EPP's PhD dissertations
- They are posted publicly as PDFs on KiltHub
- We could get abstracts and metadata through Zotero
- 87 Variables: Key (str), Publication date (str), Title (str), Author (str), Abstract (str)
- 197 rows, where each row is one dissertation



KiltHub

REFUSAL OF THE CALL

The heroes were worried the limited dataset would not give them enough information to draw significant conclusions:

- Dissertations are only available for recent years (2009-)
- For time trends, some years only have one or two dissertations
- Abstracts are occasionally sparse

THE CROSSING OF THE THRESHOLD

The heroes started their journey into the unknown by cleaning the dataset:

- Remove stopwords and a custom stopword dictionary
- Tokenize abstracts by words
- For time-series, drop years with less than 3 dissertations

```r
# Load the data

df <- read.csv("data/df.csv", stringsAsFactors = FALSE)

# Create a new df with year and abstract and key. Keep only necessary fields
# and filter NA
abstracts <- df %>%
  filter(!is.na(Abstract.Note), !is.na(Publication.Year)) %>% #remove na's
  select(Key, Abstract.Note, Publication.Year) #have a df with chosen columns

# Define custom stopwords to remove
custom_stopwords <- c("thesis", "chapter", "dissertation", "user",
                      "university", "research", "study", "studies",
                      "student", "develop")
#tokenize by word
token_stem <- abstracts %>%
  unnest_tokens(word, Abstract.Note) %>%
  filter(str_detect(word, "[a-z]")) %>%
  mutate(
    word = str_to_lower(word),
    word = str_replace_all(word, "(co\\s?-?2|co₂)", "carbon dioxide")) %>%
    #same word for co2
  filter(!word %in% custom_stopwords) %>%      # remove specific unwanted words
  anti_join(stop_words, by = "word")%>%       # remove standard stopwords
```

ROAD OF TRIALS

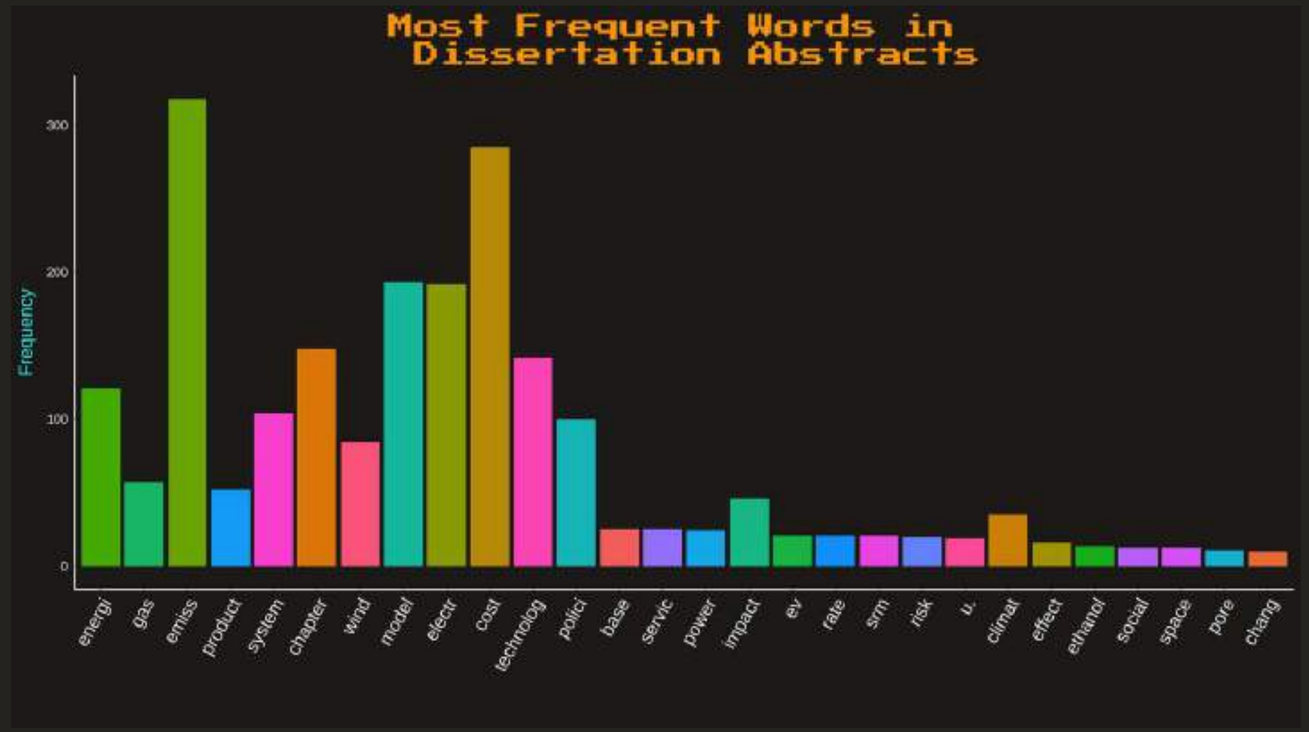
The exploratory analysis of the dataset came with technical and theoretical challenges that tested the heroes throughout the analysis of the dataset:

- Trends in AI terms in dissertations
- Proportion of keywords by year
- Energy Dictionary and Comparing Schools
- Co-concurrence of terms and stems in the same abstract

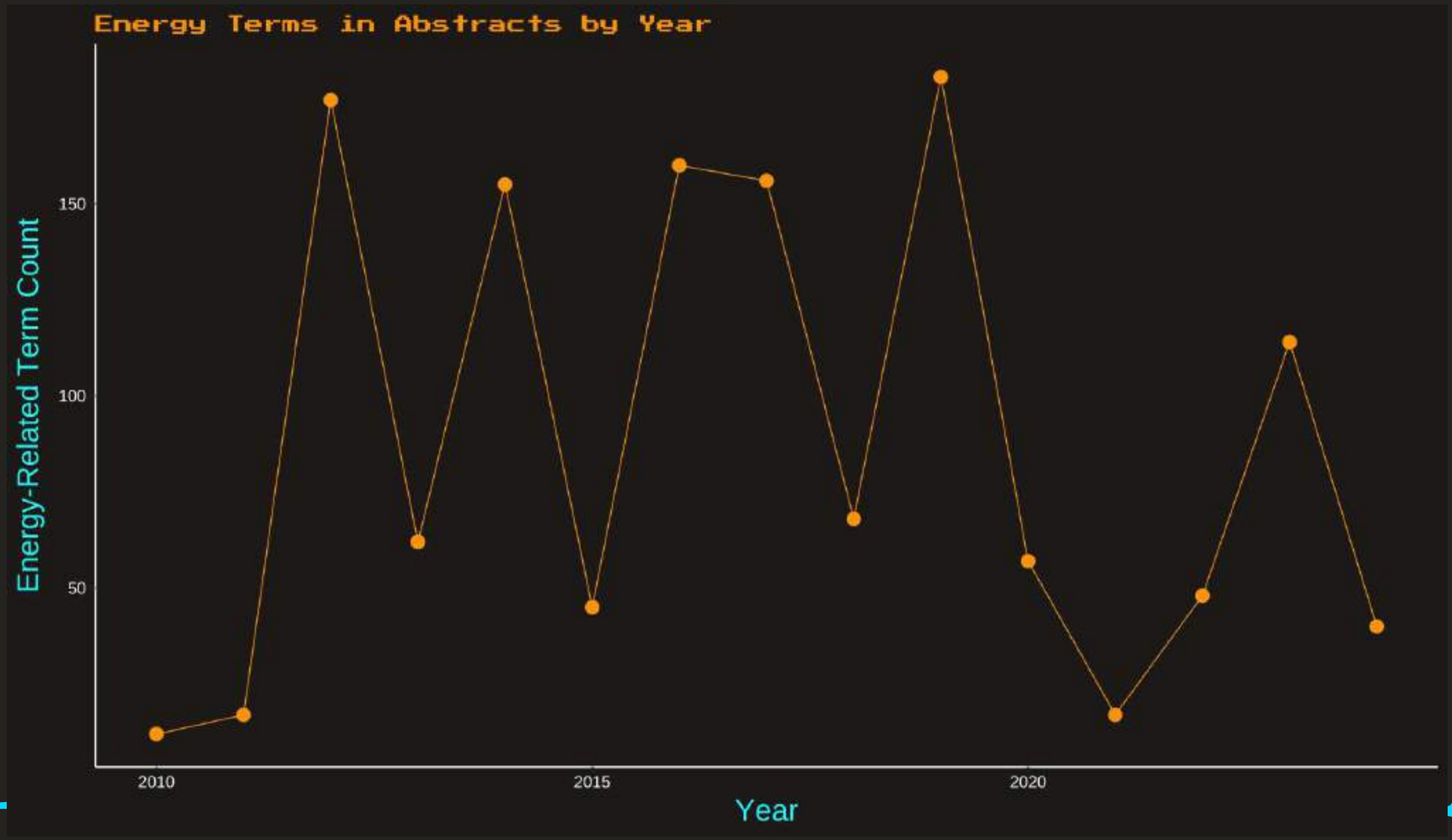
ROAD OF TRIALS

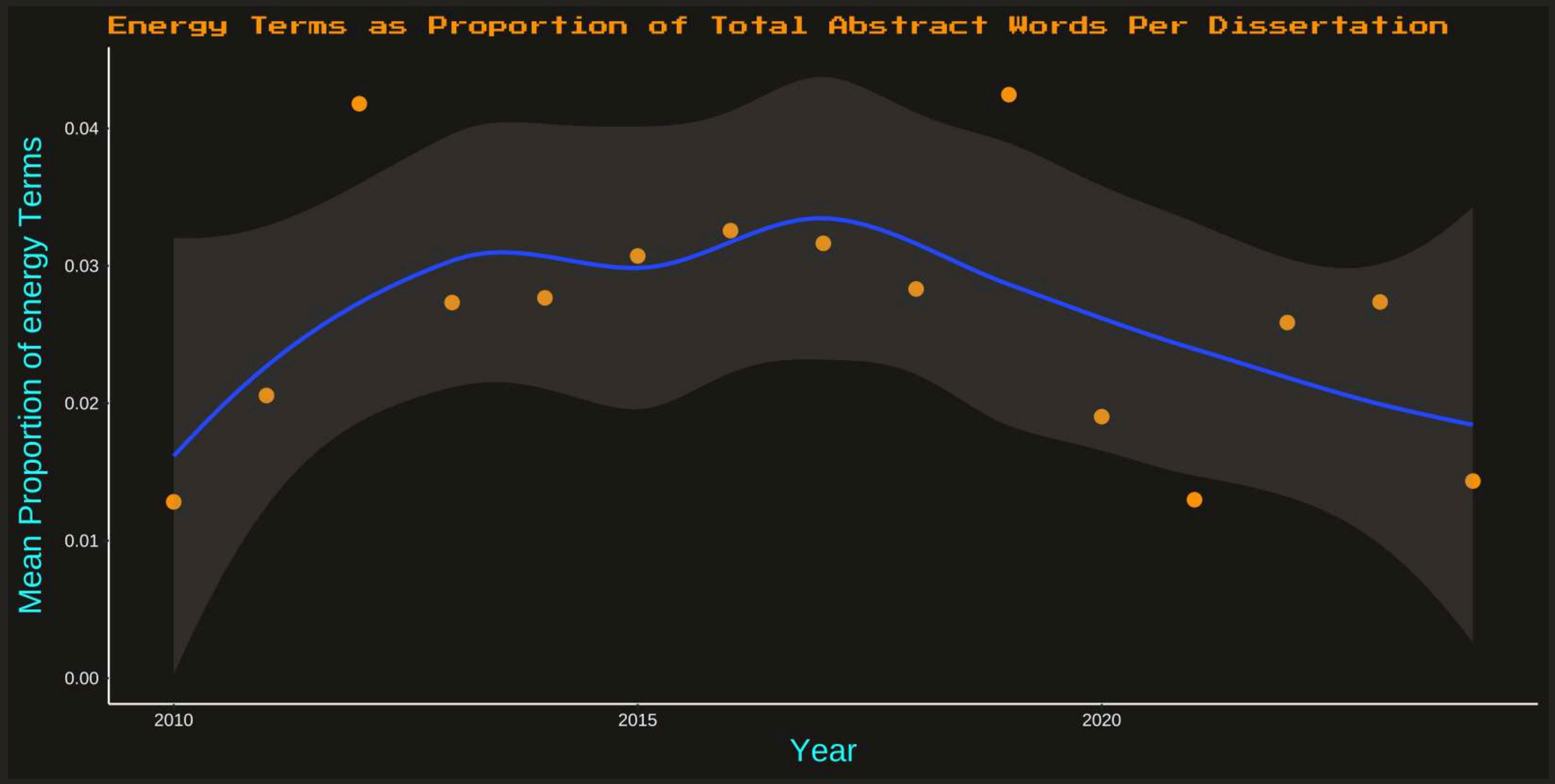
Description: Text analysis of dissertation abstracts (including tokenizing, removing stop words, and stemming). In addition to creation of an energy dictionary.

Challenge: Deciding additional stop words; Normalizing Data; Visual Creation



```r
energy_term <- c("energy", "electricity", "power", "grid", "utility", "smart grid",
  "transmission", "distribution", "electrification", "load",
  "outage", "demand response", "microgrid", "storage", "battery", "energy justice")
```




Energy Terms as Proportion of Total Abstract Words Per Dissertation
Mean Proportion of energy Terms
0.04
0.03
0.02
0.01
0.00
2010
2015
2020
Year

ROAD OF TRIALS

Description: Analysis of the proportions of thesis per year that adopt and/or reference artificial intelligence and machine learning

Challenge:

- Assumes accurate detection of “ML related” mentions based on word / text pattern matching
- Based on abstracts and not the full text thesis.

ML vs Non-ML Theses by Year