Cannabis Analysis: Applying Statistical and Linguistic Methods to Understand Strain Names

Henrique Schechter Vera[1]

[1]Princeton University

Abstract

This project details the extraction and analysis of cannabis strain data found in cannabis use and education website Leafly. The analysis applies regular statistical methods and methods drawn from the field of Natural Language Processing (NLP) to find patterns and relations between information such as strain names, ratings, effects, flavors, and genealogical origins. The analysis aims to carve a path to quantitatively investigate the factors involved in cannabis strain names and cannabis strain success, by organizing publicly available data and conducting an initial exploratory analysis through the lens of NLP.

Introduction

Cannabis strain names bizarre; some names of the most popular strains are “Martian Candy”, “Obama Kush”, “XJ-13", “Ewok”, “AK-47”, "Alaskan Thunder Fuck" and “Stardawg” [**?**]. Some strain names are even counter-intuitive, connoting negative or otherwise unappealing characteristics. For example, some denote inedible or otherwise unpleasant substances, such as "Original Glue" and "Sour Diesel", and some even evoke danger, such as "White Widow".

The naming process must have some order, and so finding this sense in the chaos of strain names is intrinsically interesting, and may result in insights related to cannabis' chemical properties, human behavior, or perhaps even something else. More importantly, finding patterns in cannabis names, especially when relating these to strain popularity metrics, is bound to yield important insights for coming up with new strain names. This product naming problem is especially interesting because the cannabis industry is actively transitioning from illicit to legal throughout the United States, with over 33 states having legalized medical cannabis and 18 of those having also legalized cannabis for recreational use [1]. Finally, this problem is more relevant than ever, since cannabis sales are growing: they hit $20 billion in 2020, were on track to overcome $26 billion in 2021, and are projected to leap to $45.9 billion in 2025 [1].

Data

We scraped our data from Leafly.com. We collected data on 5774 different cannabis, each with 71 different attributes. See an example below. The most meaningful attributes include: Aliases; average rating (and rating count); awards; cannabinoid percentages; category or phenotype (sativa, indica, hybrid, edible); chemotype; energizing score and highness score; terpene concentrations; children and parent strains; scores for each condition (e.g. anxiety), effect (e.g. relaxed), negative effect (e.g. sore throat), flavor (e.g. earthy), and symptom (e.g. lack of appetite); growth information; similar strains; total followers; and written description.

Figure 1: Representation of a strain object, with truncated information, for strain "Jet Fuel".

```
{
    'slug':              'jet-fuel',
    'id':                118799,
    'aka':               'Jet Fuel OG, G6, Jet Fuel G6,
                         Jet Fuel Kush, G6 Kush',
    'articleTotalCount': 5,
    'articlesAvailable': True,
    'averageRating':     4.523809523809524,
    'award':             {'blurb': None,
                         'imageUrl': None},
                  [...]
    'trending':          False,
    'videoUrl':          None
}
```

We also created relevant lists of categories (fruits and colors), scraped from Wikipedia, that may also prove useful for future analysis. All data we scraped and created is available at https://schechterh.github.io/cannabis-analysis-data/.

Approach

Three things are unique about our approach:

1. We use data from the most popular cannabis website.
   - Other datasets are smaller in amount of strains, and have fewer fields per strain.
   - No analyses have been performed on cannabis datasets for this goal.
2. We apply techniques from natural language processing (NLP) and computational linguistics (CL) to preprocess or analyze the data in more meaningful ways. We use sentiment analysis, semantic embedding, part-of-speech-tagging, and more.
3. We answer questions and hypotheses that have not yet been investigated, using strain effects, popularity, genealogy, reviews, etc.

Results

We summarize the statistically significant results we acquired. We found evidence supporting the ideas that:

1. Certain effects (‘sleepy’, ‘energetic’, ‘tingly’, ‘focused’, and ‘aroused’) significantly contribute to strain popularity. Contributions to review count were much larger than to average rating, but were often *negative* contributions (decreasing popularity).
2. Strains with names that contain fruits tend to be *less* popular than those that do not.
3. Strain names with verbs tend to have higher average ratings but lower review counts. Strain names with adjectives tend to have lower average ratings.
4. Strains with at least one parent that is popular (75th percentile or higher in popularity) tend to have much higher average ratings and review counts (on average, 0.55 points higher in rating with 107.771 more reviews).
5. Opposite to what was expected, strains that signal their popular parents in their names tend to differ in popularity from their parents more than those that do not (using review counts).

Questions

Here are the research questions we investigated using the data we collected, using a mix of regressions analyses, difference of means tests, and other statistical techniques:

- What strain attributes might help make them popular?
- To what extent are strain names attributable to strain consumption effects (or vice versa)?
- What strain consumption effects are the most conducive to popularity?
- Do name patterns with categories influence strain popularity?
- Do names of popular strains share common characteristics?
- Does the biological genealogy of strains account for their popularity?
  - Follow-up: Does ‘signaling’ parent’s name increase popularity retention?

Discussion

The results we obtained may influence how cannabis sellers breed, market, and name new strains.

The first three reveal ways in which to exploit tendencies for popularity. They may (1) breed strains to yield specific effects that tend to be related to popular strains, (2) advertise strains to be particularly potent at evoking these effects (perhaps through names reminiscent of effects, like "Northern Lights" being relaxing), or (3) avoid fruits and adjectives when naming strains.

The latter two reveal only potential for exploitation. Clearly, trains somehow benefit from popular parents. If this is from similarity in chemical profiles, then sellers should breed primarily popular strains. Perhaps, however, this is due to recognition of relatedness to popular strains through signals in names, in which case naming strains accordingly might be the best way to utilize this effect.

Overall, in seeking to investigate questions and hypotheses that have been thus far unexplored, we carried out several analyses and ran countless tests: this was a great initial exploratory analysis of cannabis strain names through a computational and linguistic lens, using NLP techniques as well as a new and more comprehensive dataset.

References

[1] A. Wallace, “Cannabis is one industry that’s actually coming out of covid even stronger,” *CNN Business*, 2021.

Acknowledgements

First and foremost, thank you to my advisors Professor Christiane Fellbaum and Dr. Jérémie Lumbroso for inviting me to pursue this project, for supervising it, and for their continuous advice and support. Thank you to all students in COS IW 02, as well as our Teaching Assistants Pranay Manocha, Alan Ding, and Gyoonho Kong, for the indispensable feedback and ideas they offered each week. Thank you to Leafly for originally collecting the data used in this project. Lastly, thank you to all my friends who supported me throughout this endeavor.