Cannabis Analysis: Understanding Strain Names

Henrique Schechter Vera

COS IW 02: Natural Language Processing
Professor Christiane Fellbaum

Motivation: Cannabis strain names

- Particularly bizarre. Some popular examples: “Martian Candy”, “Grape Ape”, “Obama”, “Original Glue”, “Strawberry Cough”
- Counterintuitive: “White Widow”, “Sour Diesel”

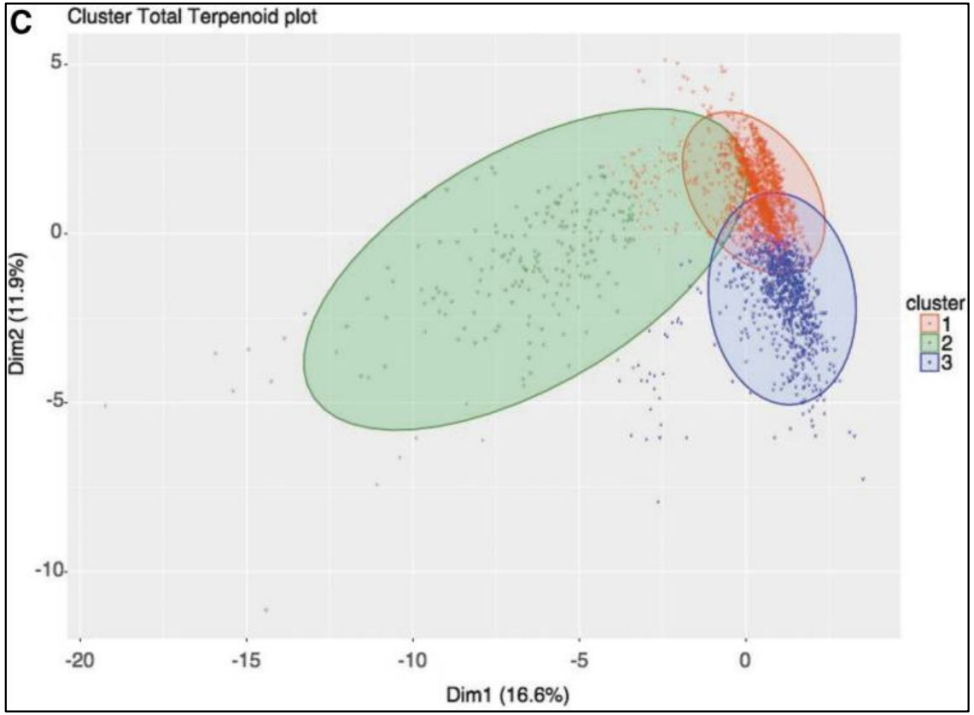
- Intrinsically interesting
- Product naming
  - Illicit market → legal

Goal

1. Find patterns in the chaos
2. Find insights useful for creating new strain names
    - Encompassed in the above

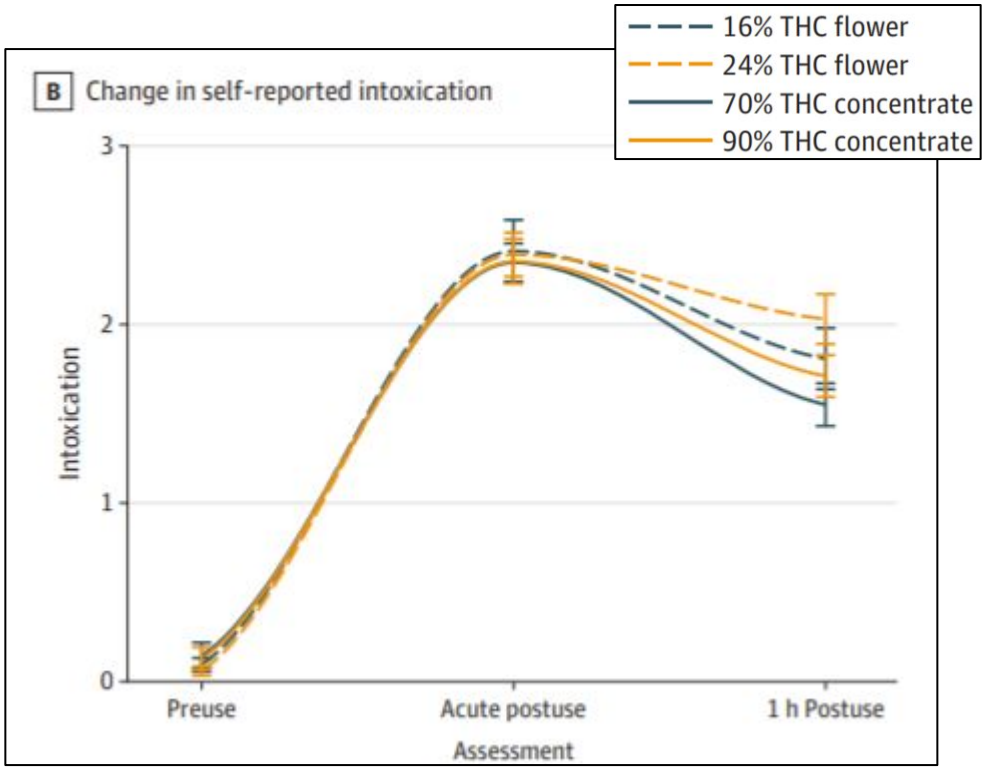
(1) Names “overestimate [chemical/genetic] diversity

- *Cannabis chemovar nomenclature misrepresents chemical and genetic diversity*
- Thousands of samples drawn from 396 differently named strains cluster best into 2-3 well-defined groups



(2) THC concentration unrelated to user intoxication

- Widely thought that THC content increases intoxication, high THC considered recreational
- Recent study did not find this to be the case



Related Work: Insights

- Differences in strain names not accounted for by chemical makeup
- Differences in consumption effects not accounted for by chemical makeup
- Users still report very different consumption experiences for each strain

Related Work: Insights

1. Need lens other than chemical to understand (differences in) strain properties like consumption effects, names: use linguistics
2. If certain strain properties are not attributable to chemical makeup, maybe they are placebo or psychologically induced by strain names

Approach

1. Use data from most popular cannabis website
    - Other datasets are smaller, have fewer fields
    - No analyses performed on cannabis datasets for this goal
2. Apply techniques from natural language processing (NLP) and computational linguistics (CL) to preprocess or analyze the data in more meaningful way
3. Answer questions/hypotheses that have not yet been investigated, using strain effects, popularity, genealogy, reviews, etc.

Implementation: Scraping Data

1. Collect individual strain URLs
2. Scrape data for each strain
3. Handle special cases
    a. Different page layout for some strains
    b. Data unavailable for some strains, maintain a blacklist

Implementation: Scraping Data (Ex.)

```
{
    '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
}
```

Implementation: Scraping Reviews

More problematic

- Required rendering JavaScript: unable to be run in Colab
- Used pydrive and GoogleAuth to upload to cloud
  - Still in progress
  - Incompatible with Princeton account: will run out of space

What factors are correlated to ratings?

- Ran **linear regression tests:** strain ratings vs. (1) THC concentration, (2) effect scores (happy, energetic, etc.), (3) flavor scores (pungent, pine, etc.), (4) sentiment value of name
- Also regression test of sentiment of name vs. sentiment of top effect
- Variance of individual strains is too high for regression analysis
  - All negative results (not statistically significant)
  - $R^2 \approx 0.001$

What effect is the most conducive to higher ratings?

- For each effect, separate strains with given top effect from all other strains. Run T-test for difference in means of ratings
- Significant but small differences. Biggest improvement: *aroused*, +0.1 (p-value < 0.0002; positive!)

Do name patterns with categories influence ratings?

- Manually created list of fruit names, color names
  - Grouped shades into broad colors
- Other categories will be done programmatically using named-entity recognition (NER) models (tagging words with categories)
- Do strains with category names have greater ratings? Negative
- Do strains with (fruit/color) names have greater ratings? Negative

Do names of popular strains share common (linguistic, maybe lexical) characteristics?

- Do strains with ratings above 4.7 have different sentiment values?

  Negative

  - High threshold accounts for negatively skewed distribution
  - Sentiment computed using VADER library
- Do strains with ratings above 4.7 have different part-of-speech distributions? Negative
  - POS tagging using NLTK

Does the biological genealogy of strains account for their popularity

- Do strains which have a popular (rating > 4.7) parent have higher ratings? Positive
  - 3.60 vs. 4.15 average rating
  - P-value < 0.00007
- Intend to investigate if and how popularity dwindles as you go down the descendant line (for ex., linearly?)

Follow-up: Does ‘signaling’ parent’s name increase popularity retention?

- See if popularity retention is attributable to names
- Do strains which share a word with their parent (e.g. “OG Kush” and “Bubba Kush”) have more similar ratings to those parents? Negative
- Intend to investigate using not just exact word matches but also strong semantic similarity to parents (using word embeddings)
- Intend to test whether, for strains that have a child *with* a shared word and a child *without* one, the former children have more similar ratings

Conclusion

- Conducted extensive initial/exploratory analysis of cannabis strain names using a novel dataset and lens (computational linguistics)
- Most impactful result: genealogy has *strong* consequences for strain popularity
  - If not attributable to genetics/chemistry, could be name recognition: multiple tests planned
  - Could determine how cannabis producers breed in the future

Future Work

- Already mentioned extensions, revisions, and related tests
- Use amount of reviews instead of average rating as ‘popularity’ heuristic
  - Rating is more vulnerable to human variance, amount of reviews is more reflective of the market
- Dataset opens the floor for more research
  - Newest, largest (in strains & fields)
  - 71 fields, many with sub-fields (e.g. ‘effects’ and ‘conditions’), plus user reviews

Acknowledgements:

Thank you for all your advice, ideas, and support:

- Christiane Fellbaum
- Jérémie Lumbroso
- Pranay Manocha, Alan Ding, and Gyoonho Kong
- And all the students in COS IW 02

And thank you Leafly.com for the scrapable data!

Questions?