Analyzing Cannabinoid Edibles Patent Document Dataset with Non-negative Matrix Factorization

Tom Plunkett

Disclaimers

- Everything in this project is my personal opinion and does not represent the opinion of my employer.
- This presentation does not constitute legal advice.

Agenda

- Project Overview
- Dataset Selection
- Data Cleaning
- Exploratory Data Analysis
- TF-IDF
- Unsupervised Learning NMF with Frobenius & KL-Divergence
- Comparison with Supervised Learning Logistic Regression
- Results and Conclusion

Overview

- Patent offices around the world issue millions of patents every year. It is difficult for companies with large patent portfolios to understand and manage their patent assets for a specific technology area.
- Machine learning can help automate tasks that have previously been entirely handled by human beings. This project is an attempt to demonstrate the ability of machine learning to classify patents as belonging in a particular category of asset, in this case whether a patent is related to Cannabinoid Edibles technology or not.
- I rely on patent metadata information (title, publication date, serial number, family id). While more accurate results would be achievable using the full text and drawings of a full patent document, that would require significantly more processing power.

Machine Learning and Unsupervised Learning

- Machine Learning, which uses statistical models and algorithms to make predictions, includes Supervised Learning, Unsupervised Learning, Reinforcement Learning, Deep Learning, and other techniques
- Unsupervised Learning, which doesn’t rely on labels, is not interested as much in prediction, but more focused on discovering interesting things about the data. It includes information visualization, clustering, finding subgroups, recommendations, dimensionality reduction, similarities, matrix factorization, collaborative filtering, etc.
- I focused on unsupervised learning techniques appropriate to patent document metadata (title, publication date, etc.). In particular, I selected Matrix Factorization and other techniques.

Goal: Analyze Patents with Unsupervised Learning

- I wanted to demonstrate the use of Matrix Factorization for analyzing textual information related to patents
- Patent Documents include the body of the patent document (specification text, claims text, and patent drawings pdfs) and patent metadata (text and numeric data)
- Analyzing the body of patent documents would require significant processing capabilities that I didn’t want to utilize on this project
- I decided to focus on patent metadata for a matrix factorization problem using NMF, etc.

Dataset: Cannabinoid Edibles Classification Gold Standard

- The next step was to find a dataset with labeled patent metadata that I could use for training data for my deep learning project.
- A search engine result led me to a site that then led me to a github site created by Steve Harris. This site had several data sets and I selected the Cannabis data set for this unsupervised learning project.
- https://github.com/swh/classification-gold-standard
- This Cannabinoid Edibles classification gold standard data set looked like a good choice for my unsupervised learning project
- The data set was created by Tony Trippe. Steve Harris, Tony Trippe, David Challis, and Nigel Swycher published an article about the dataset
- https://www.sciencedirect.com/science/article/pii/S0172219019300791

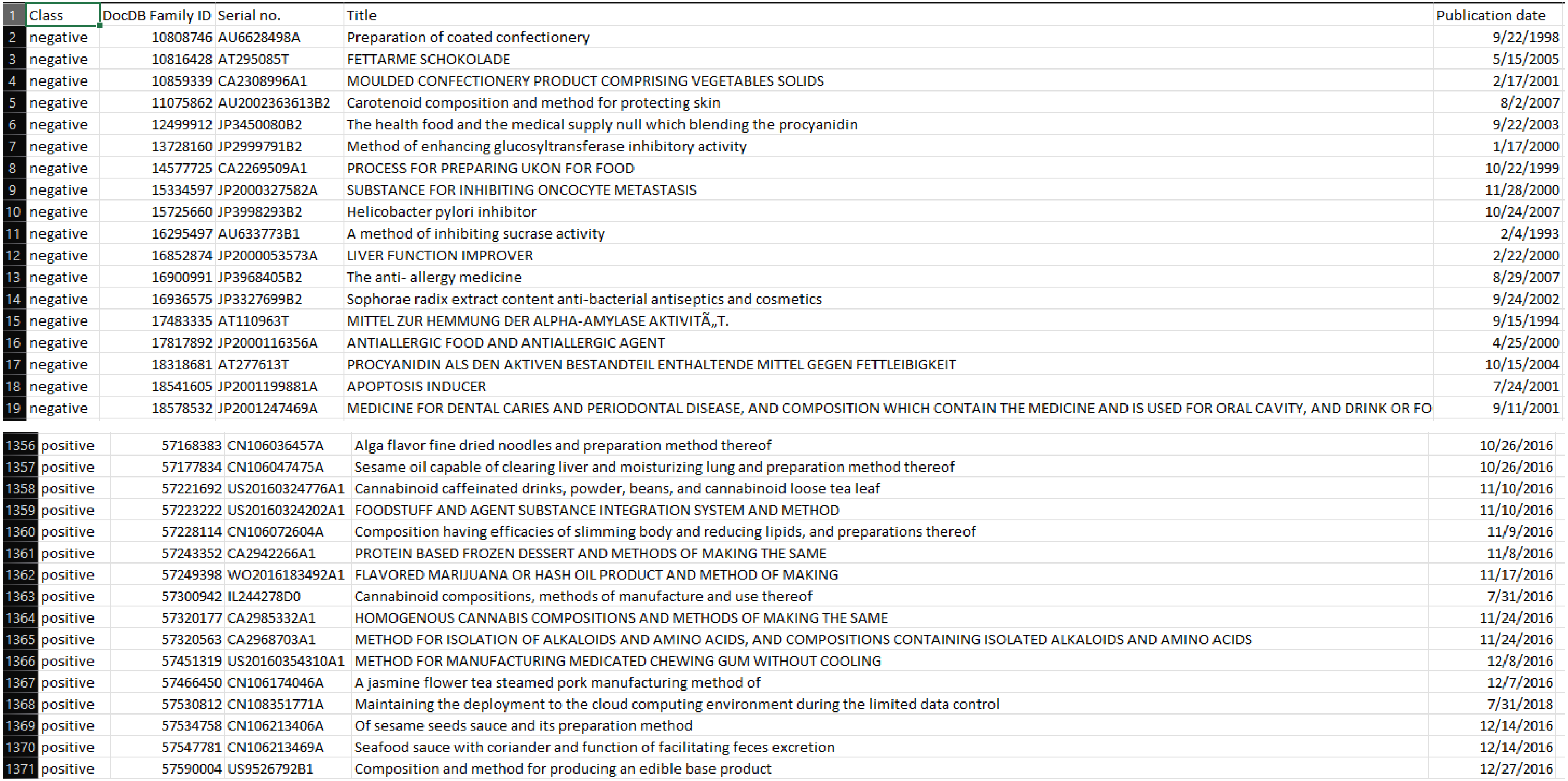
Cannabinoid Edibles Classification Description

- Positive Patent Documents: The positive collection discuss edible items, which can include lozenges, beverages, or powders containing a cannabinoid substance that can be used directly by oral absorption, or by formulating into a foodstuff for oral consumption. Cannabinoid substances include products from Cannabis sativa, ruderalis, or indica as well as products coming from the processing of hemp including hemp seeds, fibers, or oils.
- Negative Patent Documents: All of the records in the negative collection mention an edible item of one sort or another, and specifically a foodstuff.

Data Cleaning

- Original dataset has 10,795 rows
- Only 1600 unique DocDB Family IDs
- I removed the second and subsequent documents from every patent document family, just keeping the first document in each family. This reduced dataset to 1600 documents. Patent family members are highly correlated (similar titles, etc.) and add little information.
- I converted the textual classification column “negative, positive” to numeric Boolean values “0,1”.

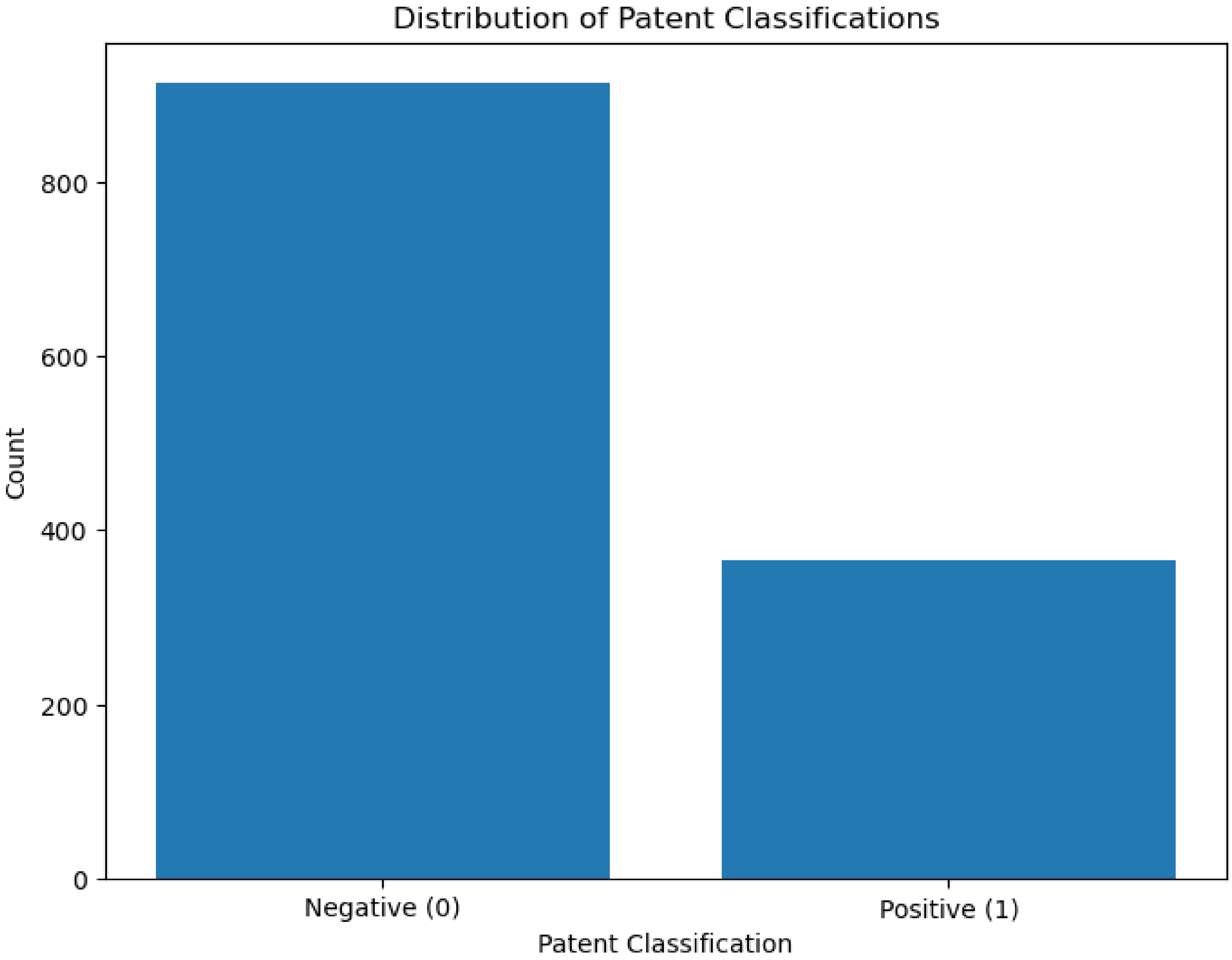
Cannabinoid Edibles Data Set Excerpt



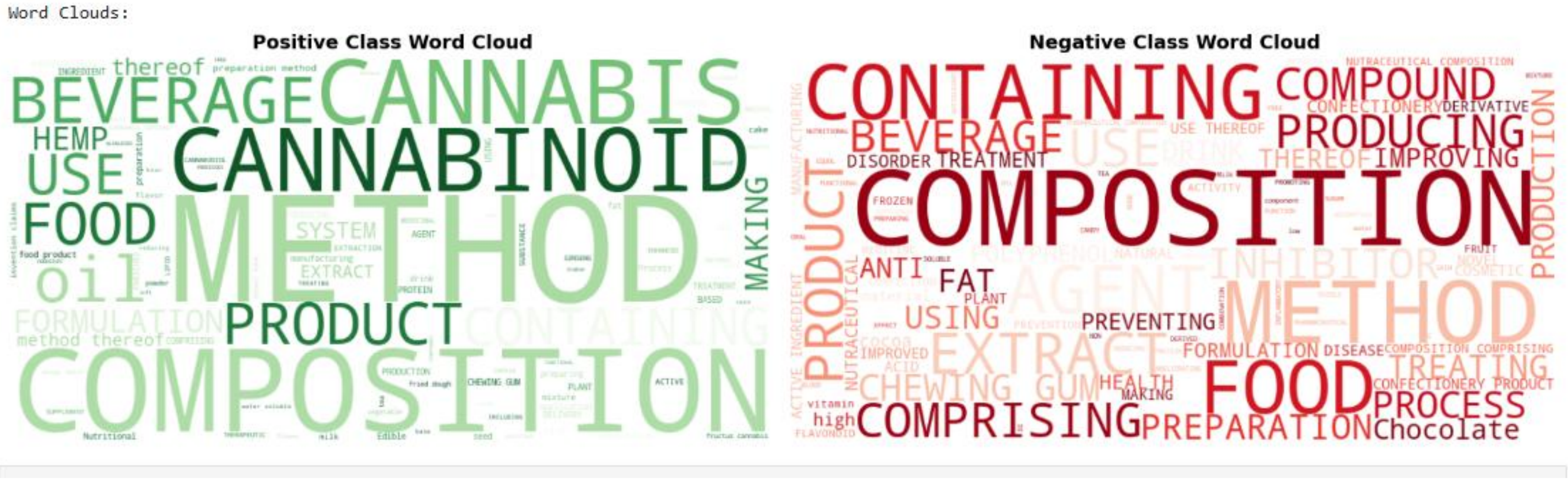
1	Class	DocDB Family ID	Serial no.	Title	Publication date
2	negative	10808746	AU6628498A	Preparation of coated confectionery	9/22/1998
3	negative	10816428	AT295085T	FETTARME SCHOKOLADE	5/15/2005
4	negative	10859339	CA2308996A1	MOULDED CONFECTIONERY PRODUCT COMPRISING VEGETABLES SOLIDS	2/17/2001
5	negative	11075862	AU2002363613B2	Carotenoid composition and method for protecting skin	8/2/2007
6	negative	12499912	JP3450080B2	The health food and the medical supply null which blending the procyanidin	9/22/2003
7	negative	13728160	JP2999791B2	Method of enhancing glucosyltransferase inhibitory activity	1/17/2000
8	negative	14577725	CA2269509A1	PROCESS FOR PREPARING UKON FOR FOOD	10/22/1999
9	negative	15334597	JP2000327582A	SUBSTANCE FOR INHIBITING ONCOCYTE METASTASIS	11/28/2000
10	negative	15725660	JP3998293B2	Helicobacter pylori inhibitor	10/24/2007
11	negative	16295497	AU633773B1	A method of inhibiting sucrase activity	2/4/1993
12	negative	16852874	JP2000053573A	LIVER FUNCTION IMPROVER	2/22/2000
13	negative	16900991	JP3968405B2	The anti- allergy medicine	8/29/2007
14	negative	16936575	JP3327699B2	Sophorae radix extract content anti-bacterial antiseptics and cosmetics	9/24/2002
15	negative	17483335	AT110963T	MITTEL ZUR HEMMUNG DER ALPHA-AMYLASE AKTIVITÃ„T.	9/15/1994
16	negative	17817892	JP2000116356A	ANTIALLERGIC FOOD AND ANTIALLERGIC AGENT	4/25/2000
17	negative	18318681	AT277613T	PROCYANIDIN ALS DEN AKTIVEN BESTANDTEIL ENTHALTENDE MITTEL GEGEN FETTLEIBIGKEIT	10/15/2004
18	negative	18541605	JP2001199881A	APOPTOSIS INDUCER	7/24/2001
19	negative	18578532	JP2001247469A	MEDICINE FOR DENTAL CARIES AND PERIODONTAL DISEASE, AND COMPOSITION WHICH CONTAIN THE MEDICINE AND IS USED FOR ORAL CAVITY, AND DRINK OR FO	9/11/2001
1356	positive	57168383	CN106036457A	Alga flavor fine dried noodles and preparation method thereof	10/26/2016
1357	positive	57177834	CN106047475A	Sesame oil capable of clearing liver and moisturizing lung and preparation method thereof	10/26/2016
1358	positive	57221692	US20160324776A1	Cannabinoid caffeinated drinks, powder, beans, and cannabinoid loose tea leaf	11/10/2016
1359	positive	57223222	US20160324202A1	FOODSTUFF AND AGENT SUBSTANCE INTEGRATION SYSTEM AND METHOD	11/10/2016
1360	positive	57228114	CN106072604A	Composition having efficacies of slimming body and reducing lipids, and preparations thereof	11/9/2016
1361	positive	57243352	CA2942266A1	PROTEIN BASED FROZEN DESSERT AND METHODS OF MAKING THE SAME	11/8/2016
1362	positive	57249398	WO2016183492A1	FLAVORED MARIJUANA OR HASH OIL PRODUCT AND METHOD OF MAKING	11/17/2016
1363	positive	57300942	IL244278D0	Cannabinoid compositions, methods of manufacture and use thereof	7/31/2016
1364	positive	57320177	CA2985332A1	HOMOGENOUS CANNABIS COMPOSITIONS AND METHODS OF MAKING THE SAME	11/24/2016
1365	positive	57320563	CA2968703A1	METHOD FOR ISOLATION OF ALKALOIDS AND AMINO ACIDS, AND COMPOSITIONS CONTAINING ISOLATED ALKALOIDS AND AMINO ACIDS	11/24/2016
1366	positive	57451319	US20160354310A1	METHOD FOR MANUFACTURING MEDICATED CHEWING GUM WITHOUT COOLING	12/8/2016
1367	positive	57466450	CN106174046A	A jasmine flower tea steamed pork manufacturing method of	12/7/2016
1368	positive	57530812	CN108351771A	Maintaining the deployment to the cloud computing environment during the limited data control	7/31/2018
1369	positive	57534758	CN106213406A	Of sesame seeds sauce and its preparation method	12/14/2016
1370	positive	57547781	CN106213469A	Seafood sauce with coriander and function of facilitating feces excretion	12/14/2016
1371	positive	57590004	US9526792B1	Composition and method for producing an edible base product	12/27/2016

Cannabinoid Edible Patent Title Class Distributions

```
Class Distribution:
```



Cannabinoid Edible Patent Title Word Clouds



```
Word Clouds:
```

Positive Class Word Cloud

Negative Class Word Cloud

Term Frequency-Inverse Document Frequency (TF-IDF)

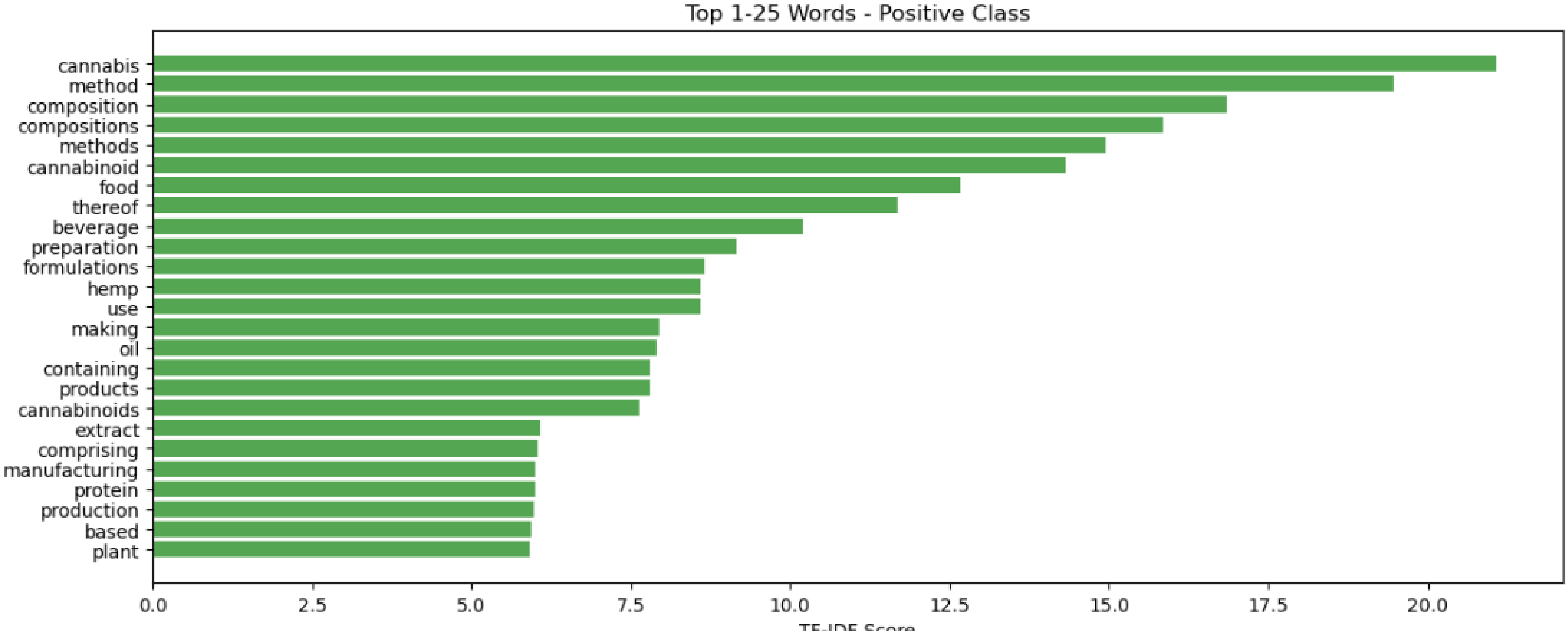
- Evaluates the importance of a term in a document in a corpus of documents
- Combines Term Frequency with Inverse Document Frequency
- TF(t,d) = Number of times t appears in document/number of terms in document
- Term frequency doesn’t deal with global document corpus and may also include stop words that lack meaning (the, and, etc.)
- Inverse Document Frequency puts more weight on rare words and less weight on common words
- TF-IDF helps us identify words that are important to specific documents

Patent Title TF-IDF example

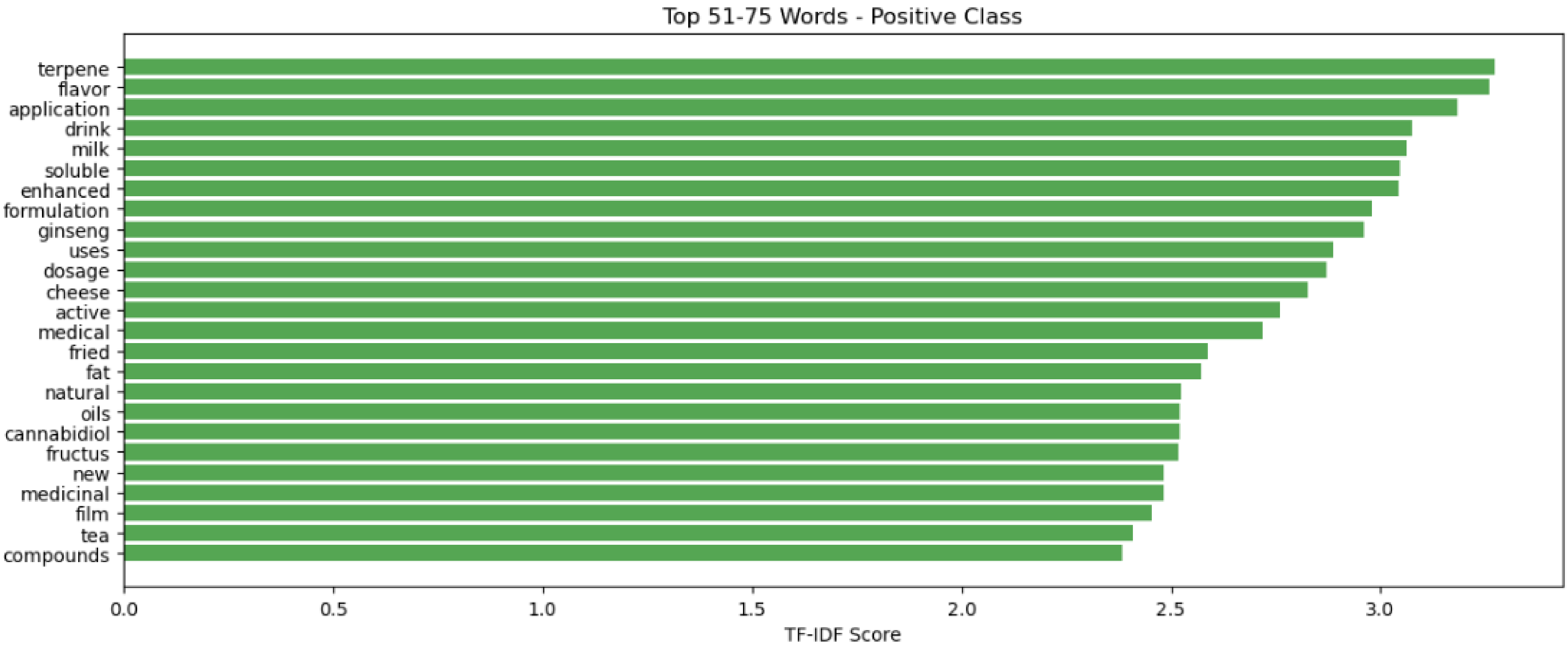
- Document 1: “Method for Cannabinoid Beverage Wellness product"
- Document 2: “Cannabis wellness process"
- Document 3: "Cannabinoid Edible manufacturing process and system“
- TF(cannabinoid, doc 1) = 1/6
- TF(cannabinoid, doc 2) = 0
- TF(cannabinoid, doc 3)=1/6
- IDF (cannabinoid, all docs) = log (3 docs/2 docs have cannabinoid)=.176
- TF-IDF is TF*IDF. Score is for a specific document. docs 1 and 3 have TF-IDF scores of .029 for a cannabinoid, so the term is somewhat important for those docs. Doc 2 has a TF-IDF score of 0.

TF-IDF top 25 words

```
Text Feature Analysis:
```



TF-IDF words 51-75



Matrix Factorization

- Use Matrix Factorization to analyze document vocabulary
- Several different MF methods including Singular Value Decomposition, Non-negative Matrix Factorization (NMF), and Approximation methods.
- I chose NMF using sklearn.decomposition.NMF
- I will compare Frobenius NMF and KL-Divergence loss. For KL-Divergence Loss, I will also vary the Number of Components hyperparameter and graph the impact to accuracy and reconstruction error.

Non-Negative Matrix Factorization (NMF)

NMF is used to break down a large dataset into smaller meaningful parts.

For a matrix A of dimensions $m \times n$ where each element is ≥ 0, NMF factorizes it into two matrices W and H with dimensions $m \times k$ and $k \times n$ respectively, where both matrices contain only non-negative elements:

$Am\times n\approx(Wm\times k)(Hk\times n)$

where:

A → Original input matrix (a linear combination of W and H)

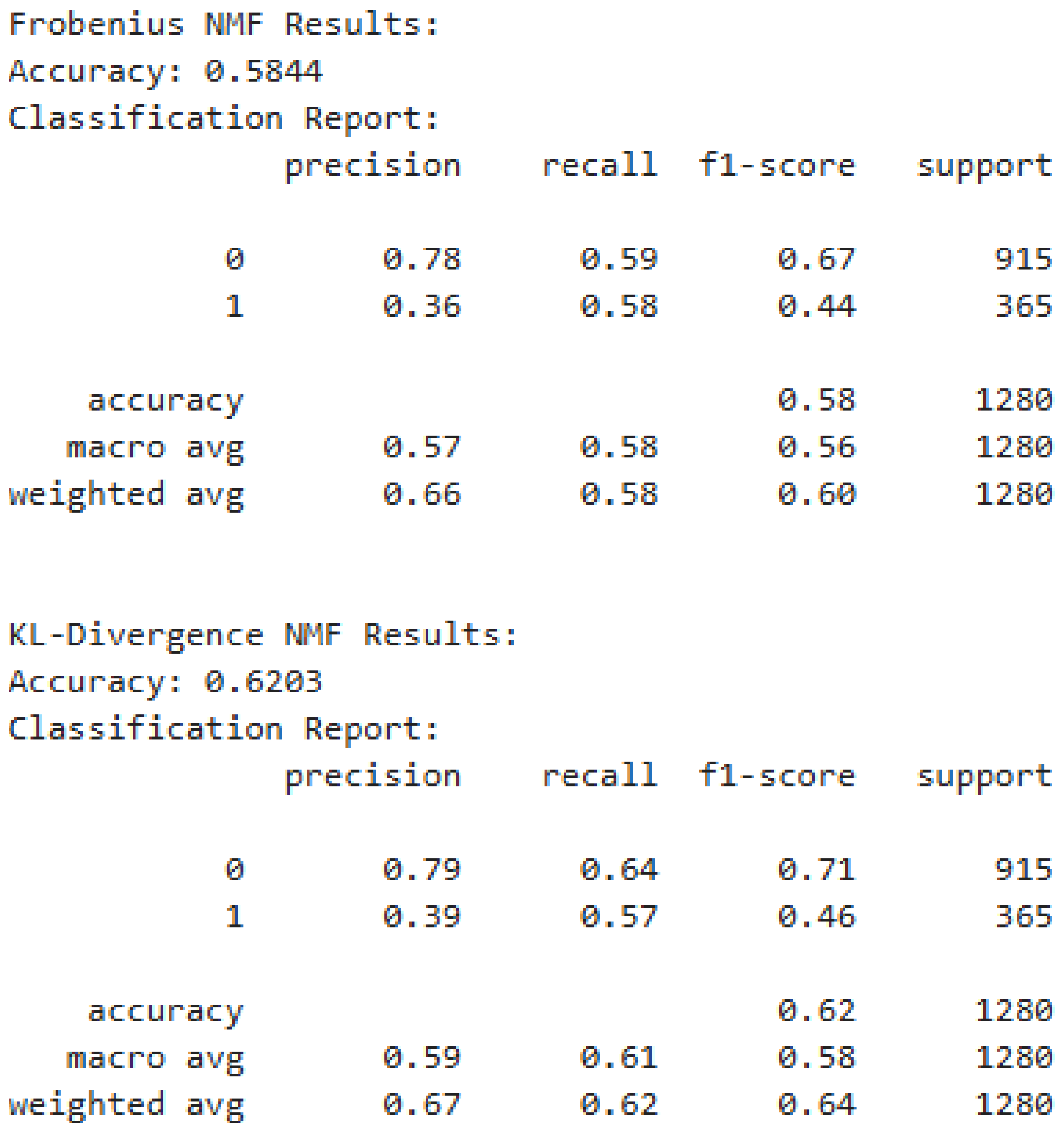
W → Feature matrix (basis components)

H → Coefficient matrix (weights associated with W)

k → Rank (dimensionality of the reduced representation where $k \leq \min(m, n)$)

NMF helps to identify hidden patterns in data by assuming that each data point can be represented as a combination of fundamental features found in W.

Non-Negative Matrix Factorization (NMF) Results

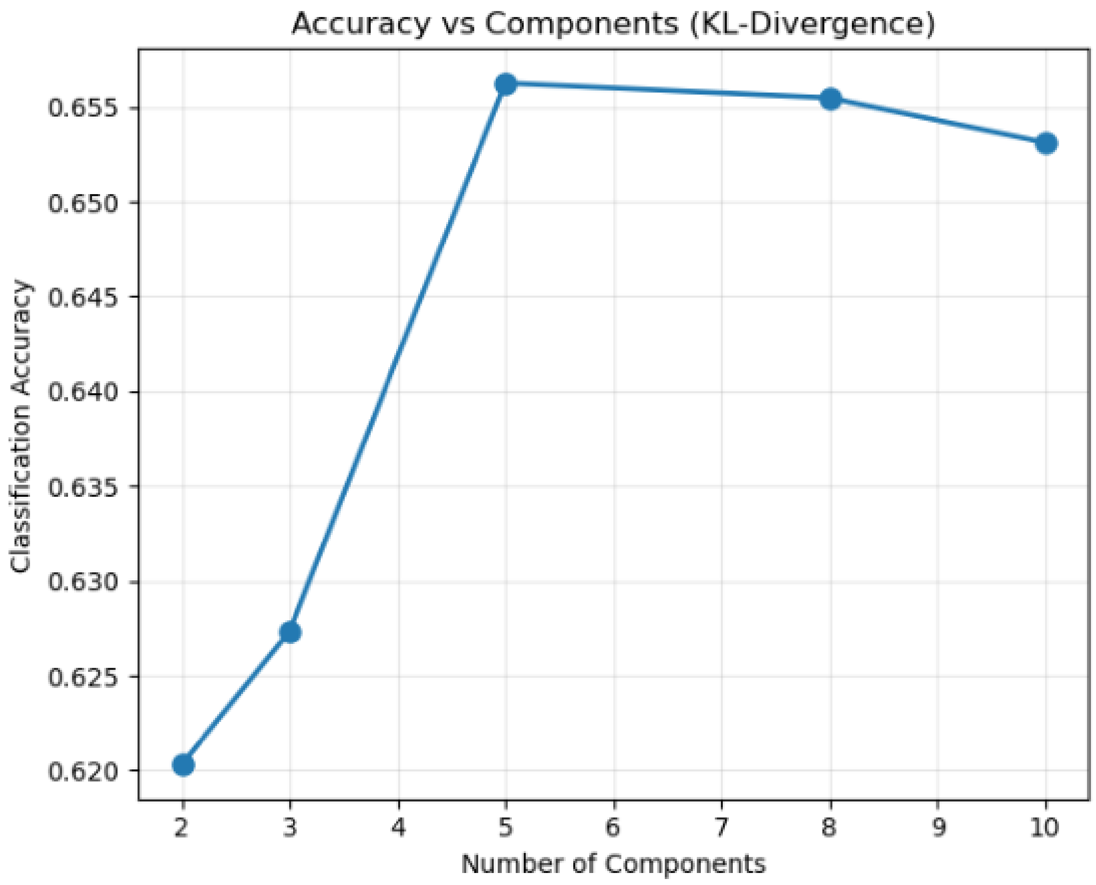


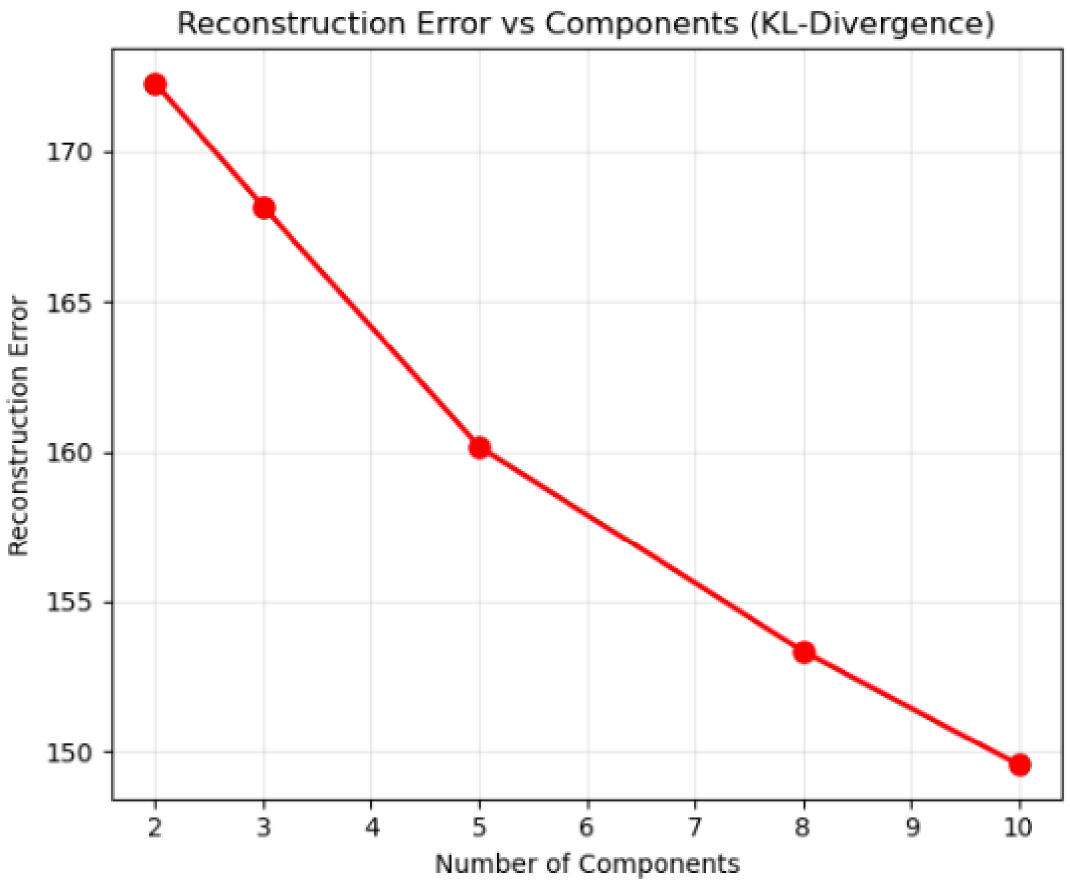
```
Frobenius NMF Results:
Accuracy: 0.5844
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.59      0.67       915
           1       0.36      0.58      0.44       365

    accuracy                           0.58      1280
   macro avg       0.57      0.58      0.56      1280
weighted avg       0.66      0.58      0.60      1280


KL-Divergence NMF Results:
Accuracy: 0.6203
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.64      0.71       915
           1       0.39      0.57      0.46       365

    accuracy                           0.62      1280
   macro avg       0.59      0.61      0.58      1280
weighted avg       0.67      0.62      0.64      1280
```

Frobenius NMF vs. KL-Divergence NMF

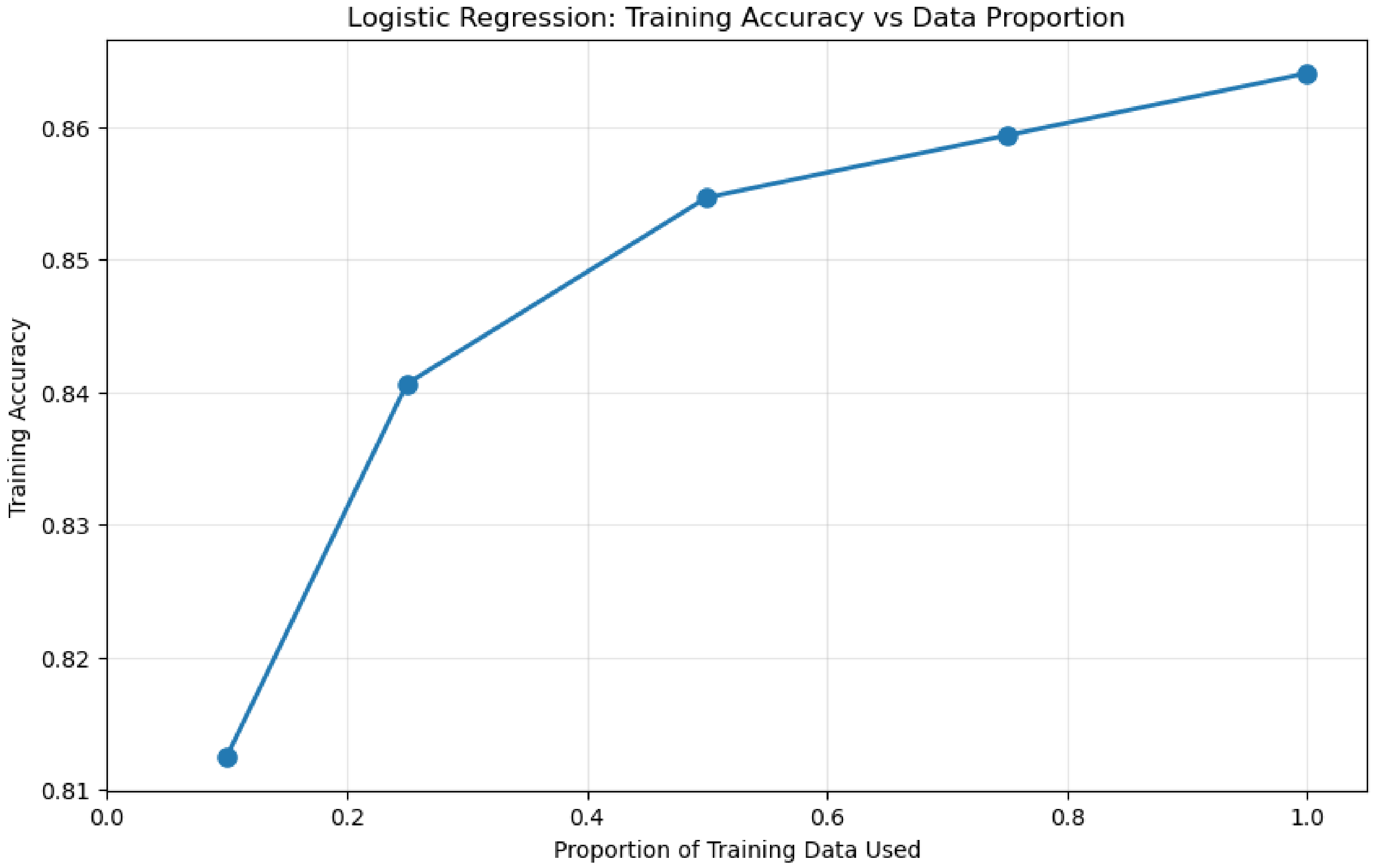
- Frobenius NMF penalizes large errors due to squaring. Frobenius NMF tends to produce more average solutions. Can be dominated by high frequency terms.
- KL Divergence is more sensitive to relative differences between values. Can better capture subtle topic distinctions.
- KL divergence probably performs better here because Patent titles have sparse TF-IDF representations, which leads to more distinct topic clusters. KL divergence is also better at handling outliers and patent titles might include rare technical terms.

Varying the Number of Components Hyperparameter for KL-Divergence





Supervised Learning Logistic Regression



Conclusions

- My thanks to Tony Trippe, Steve Harris, and the team that created the Cannabinoid Edibles Classification Gold Standard data set. I hope more groups release data sets of patent information for machine learning.
- I learned about using Non-negative Matrix Factorization, KL-Divergence, TF-IDF, etc. to process information with regards to Patent documents
- This project emphasized the value of fully understanding your data set and the relationships within your data set.