Sophie Chen

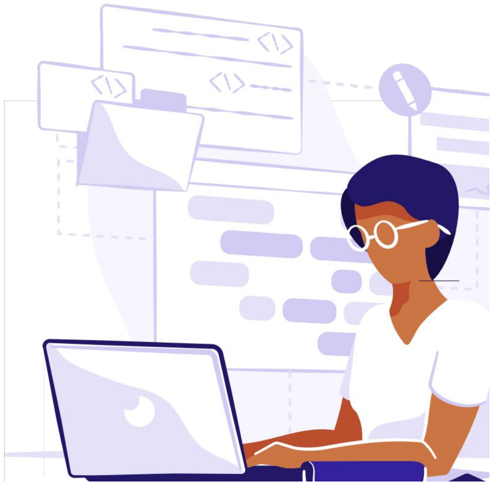
AI Folder Genie

Download Files where they Belong!

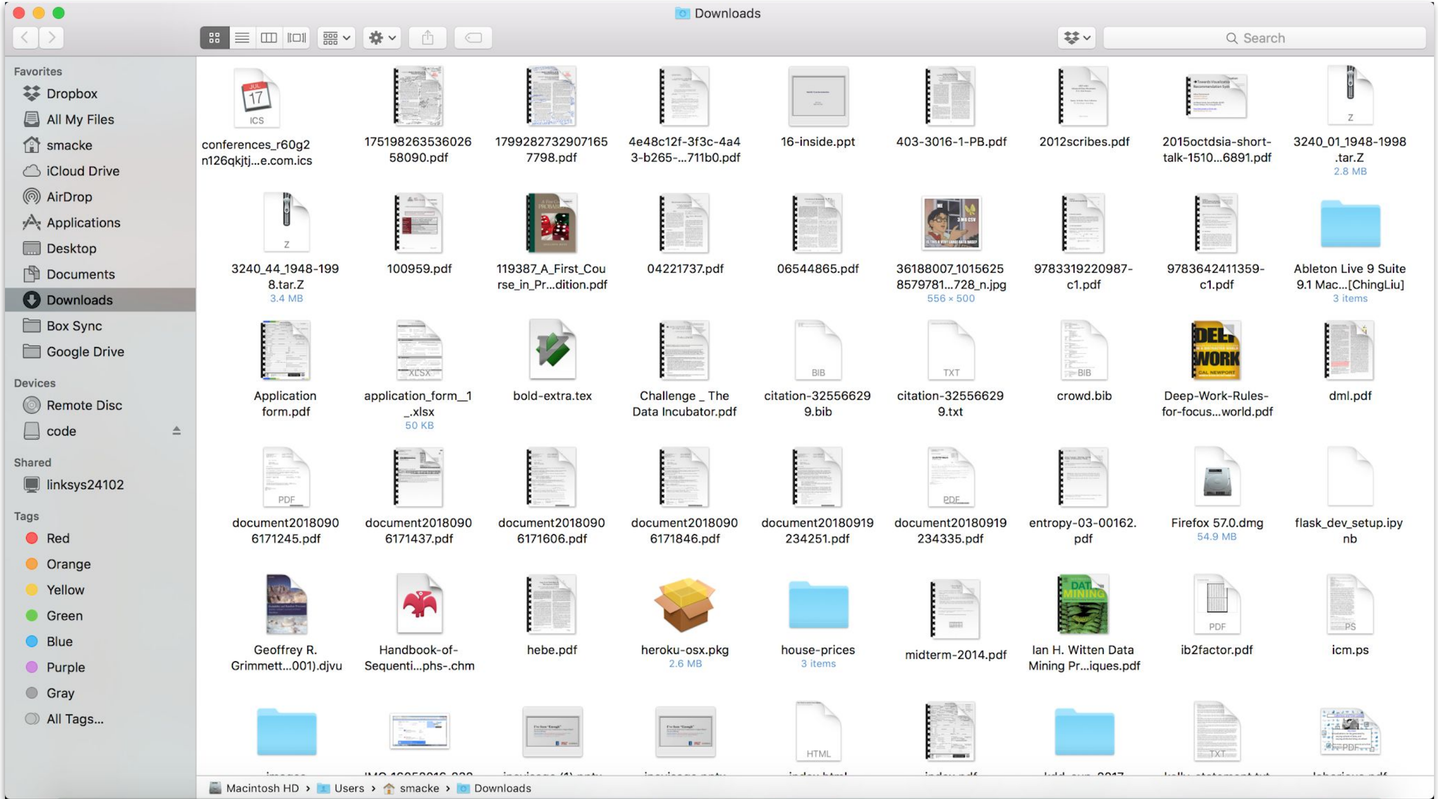
Download Every Day

Downloads :

- *Email Attachments*
- *Social Media Photos*
- *Cloud Shared Files*
- *Github Codes*
- *Digital Music*
- *...*

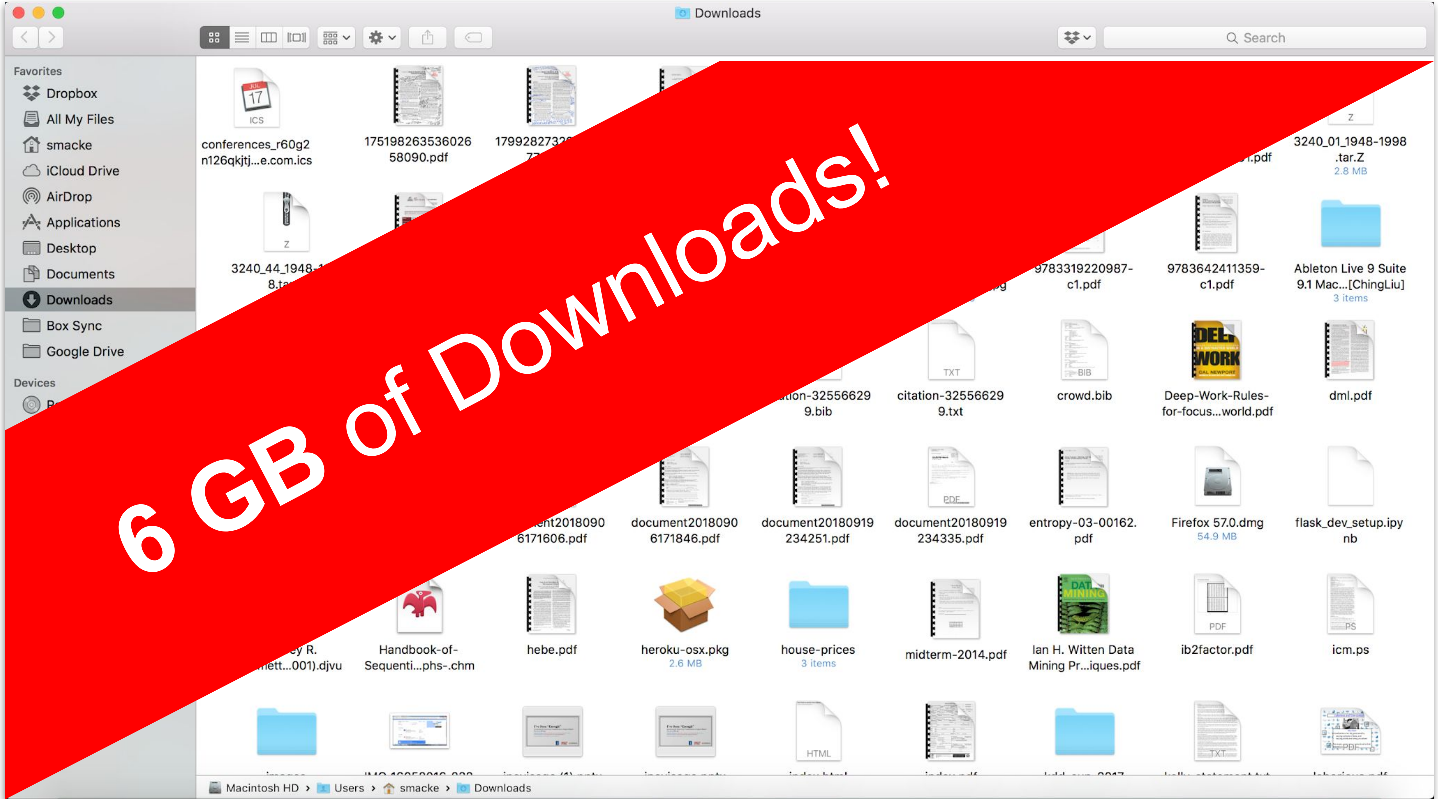


All go to *The Downloads Folder*



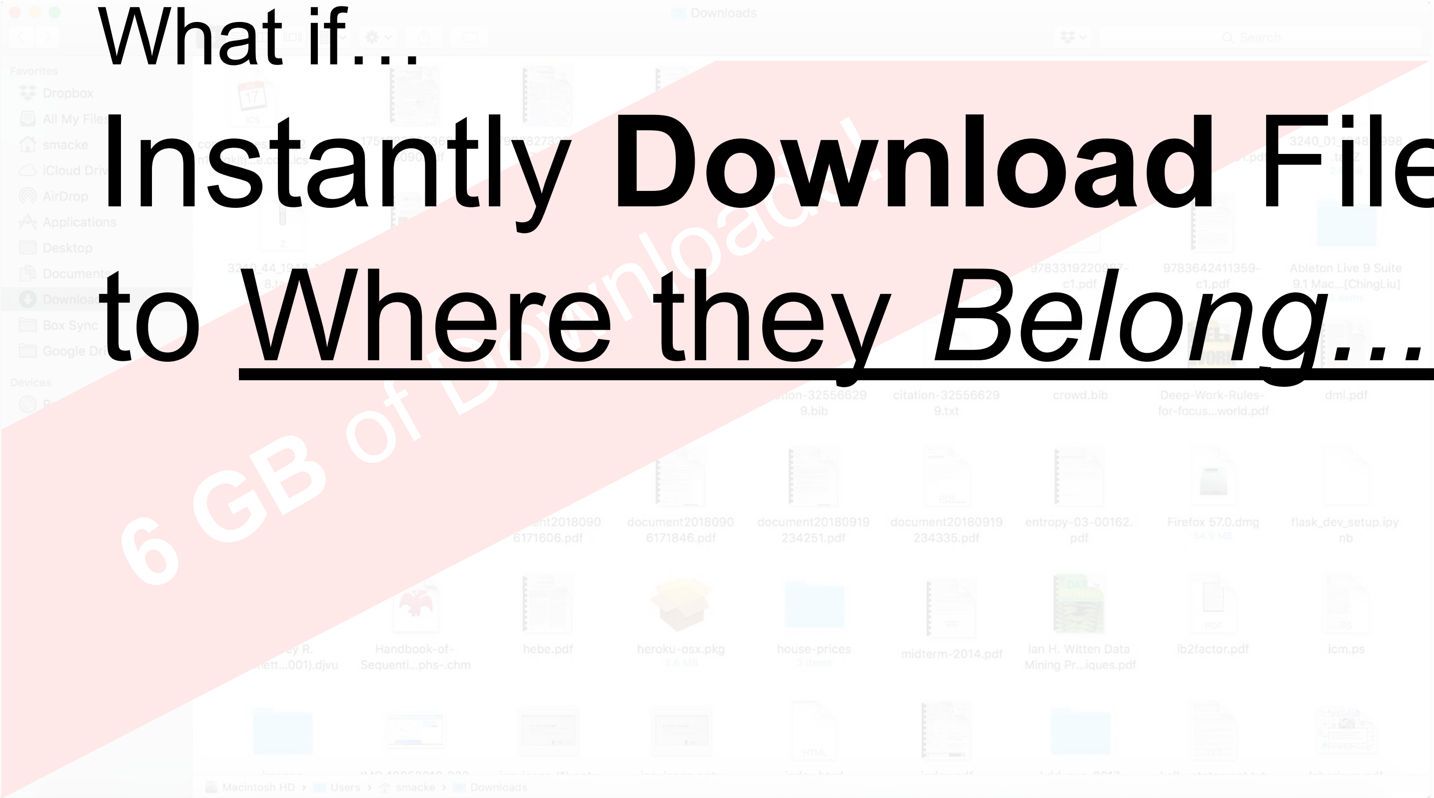
Disorganized File Structure!

Too many files in Downloads Folder



Disorganized File Structure!

Too many files in Downloads Folder



What if…

Instantly **Download** Files to <u>Where they *Belong*…</u>

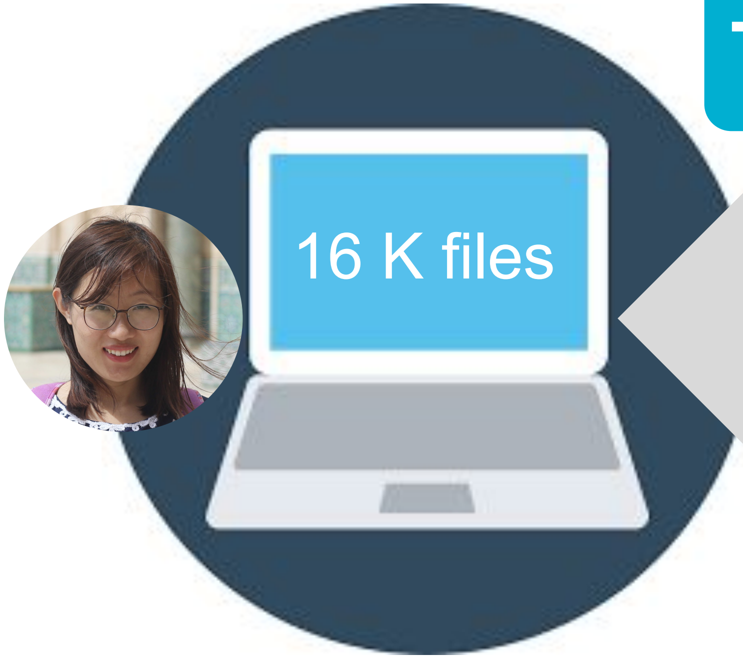
AI Folder Genie

available in the chrome web store*

Predict the ***Best*** Folder/Sub-Folder

for *YOUR* **File Download**,

for *YOUR* **File System**!



Text Classification

- **File** names (features)
- **Folder** names (labels)

Processed file names with Natural Language Analyses with NLTK and scikit learn

'CP 2006 Theoretical potential energy surfaces for excited mercury trimers.pdf'

'cp', '2006', 'theoret', 'potenti', 'energi', 'surfac', 'for', 'excit', 'mercuri', 'trimer', 'pdf'

Extract the Features

CS/Insight
hierarch
bootstrap

Pictures/Image
jpg
larg

videos/wild_china
wild
china

Hierarchical Labels

Files in:

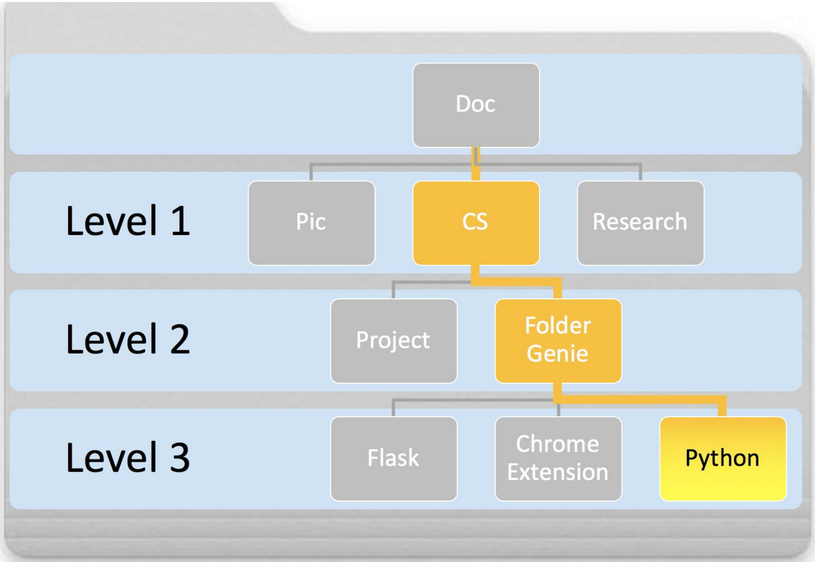
Doc/CS/FolderGenie/Python

Hierarchical labels:

Level 1: CS

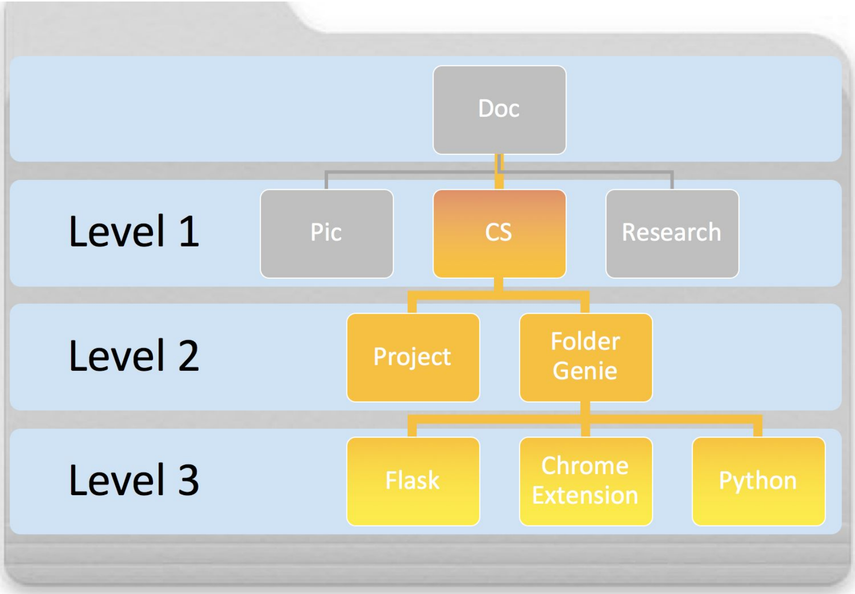
Level 2: CS/FolderGenie

Level 3: CS/FolderGenie/Python



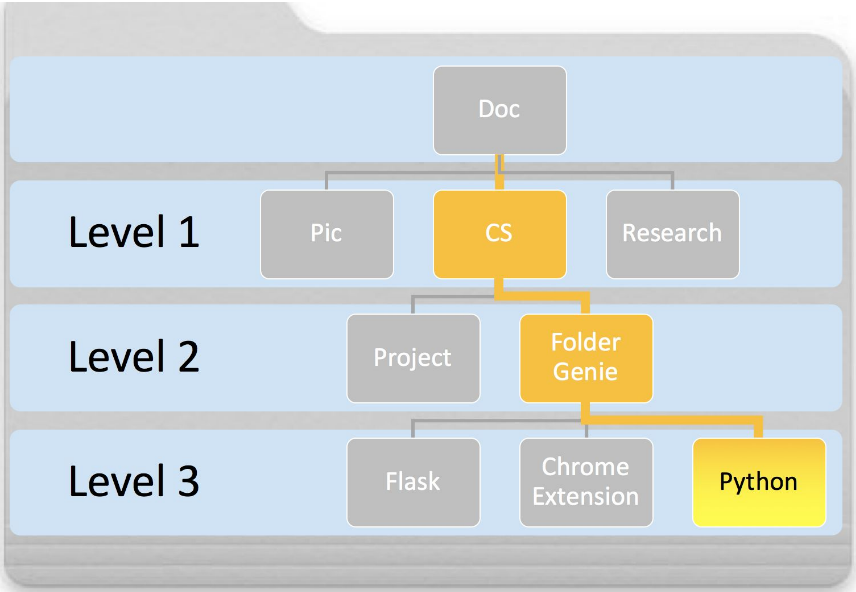
Hierarchical Classification

Softmax classifier trained at each folder

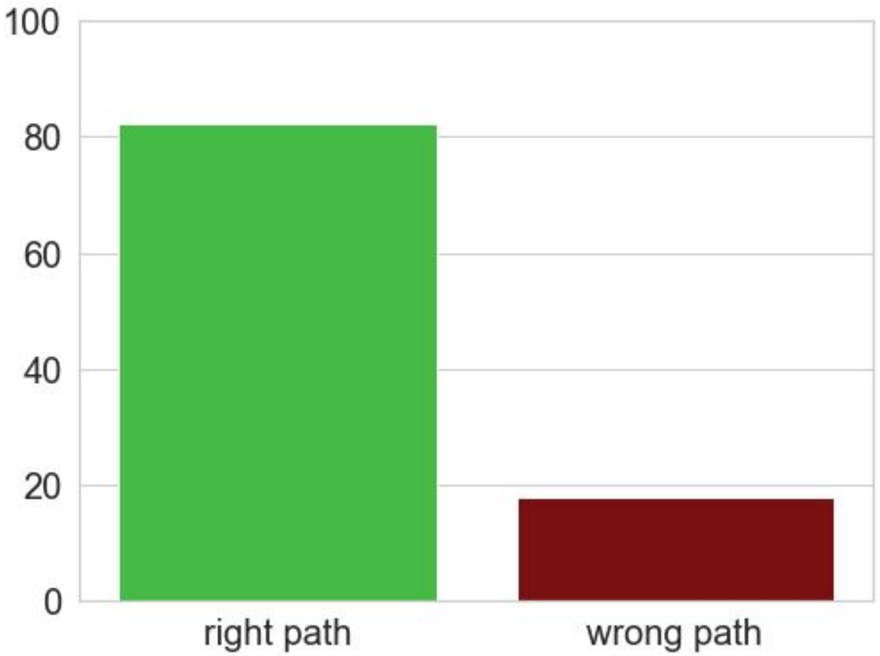


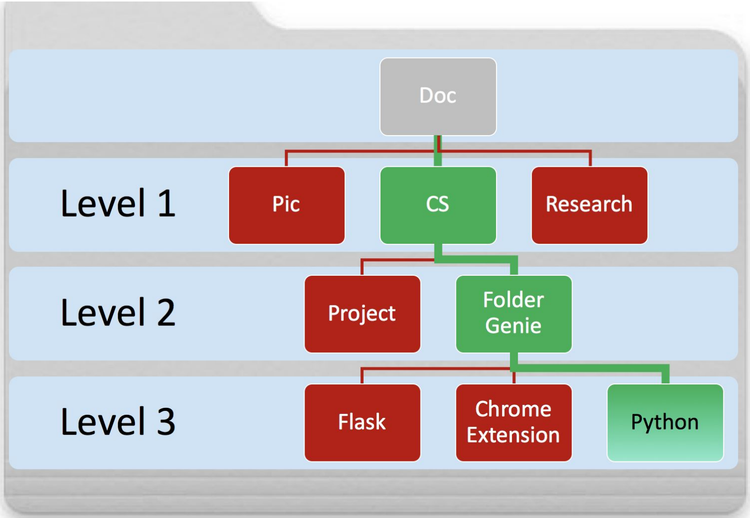
Hierarchical Classification

Predictions only advance deeper when confident

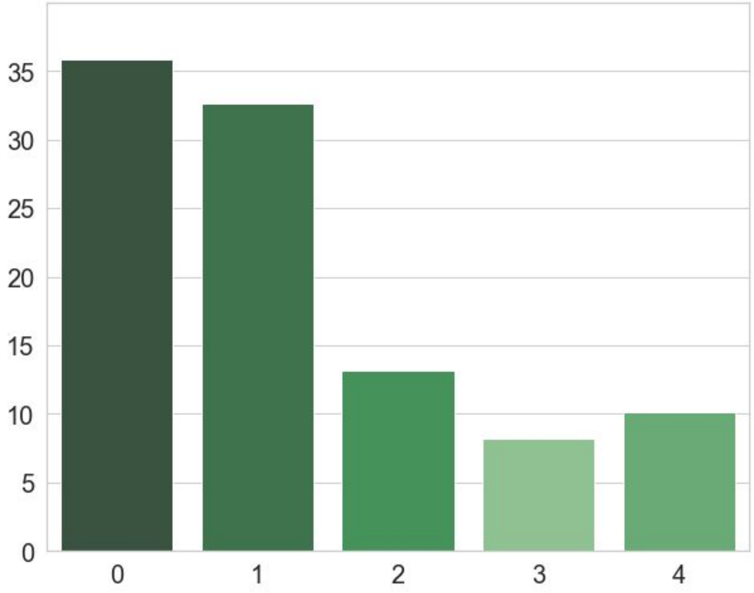


82% go to *Right Path!*

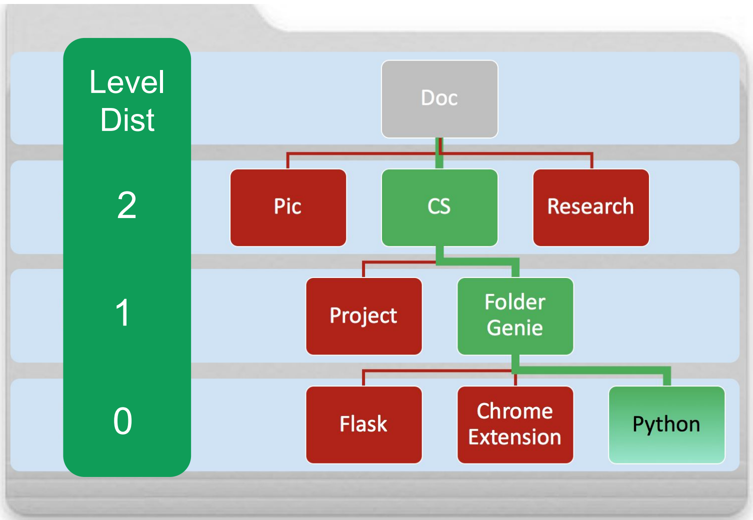




69% go to ***within 1* level distance!**



Level Distance = Actual - Predicted



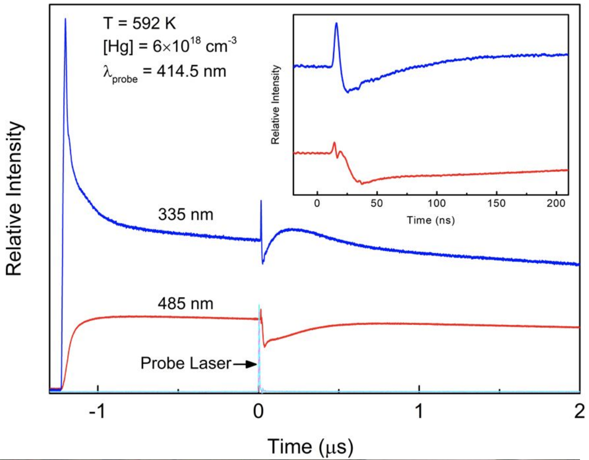
local file folder levels: 9 in total
first 5 considered for predictions

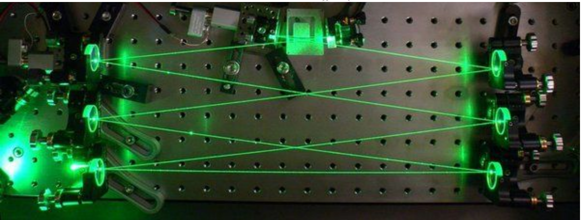
Sophie Chen

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PhD in Electrical & Computer Engineering, Spectral Analysis

Github: SophieGarden
Linkedin: Sophie-Chen-Data





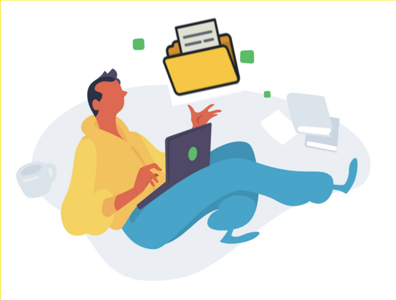
Thank you for attending my talk!

Welcome to try

Available soon

AI Folder Genie

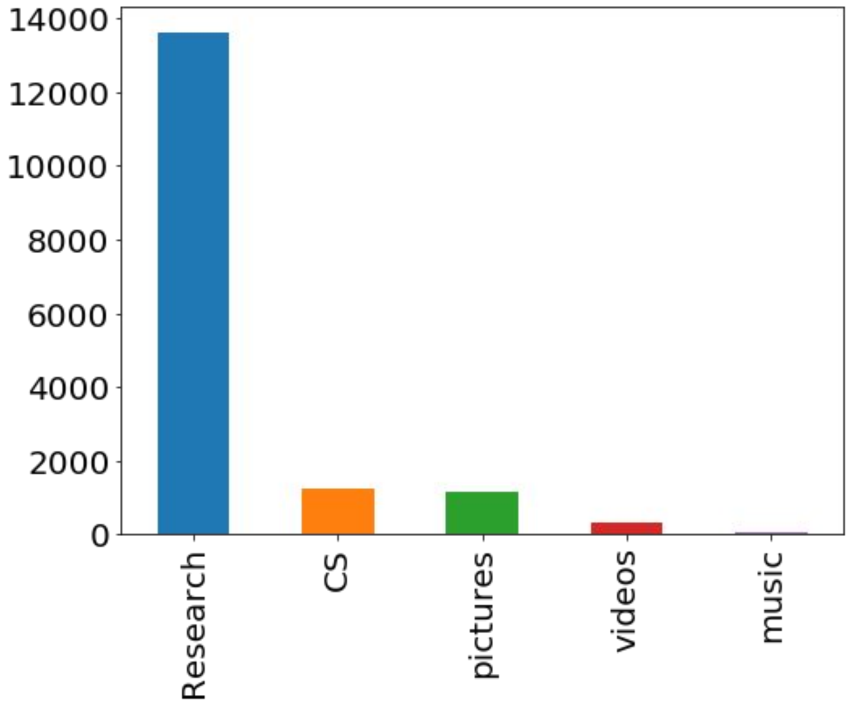
available in the chrome web store

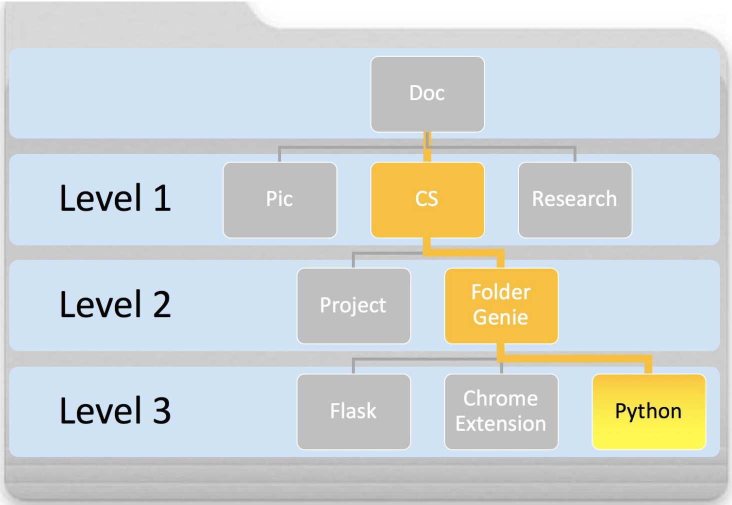


Future Plans

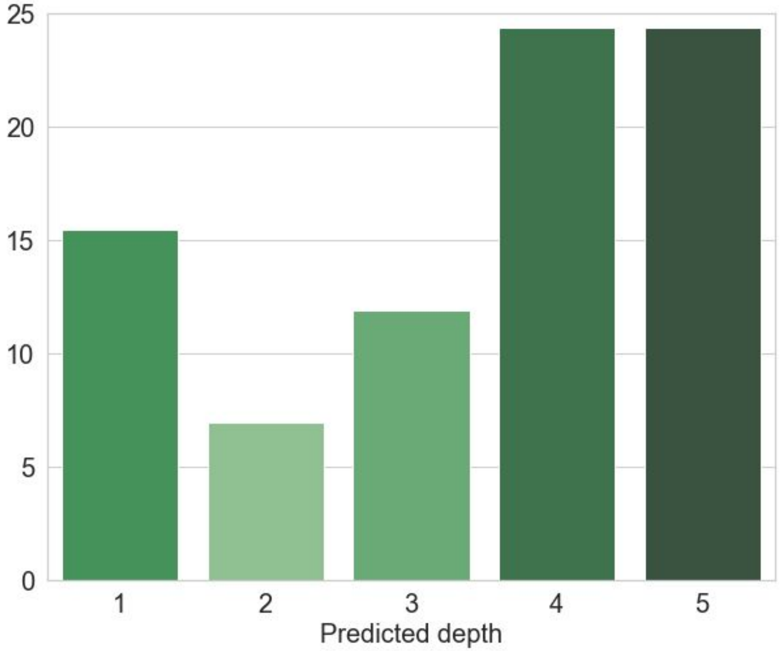
- Chrome Extension + Desktop App: sync between all devices, Dropbox…
- Output a list of choices
- Features: metadata, NLP folder names, included folder names as part of features

Metrics for ***Unbalanced Data***



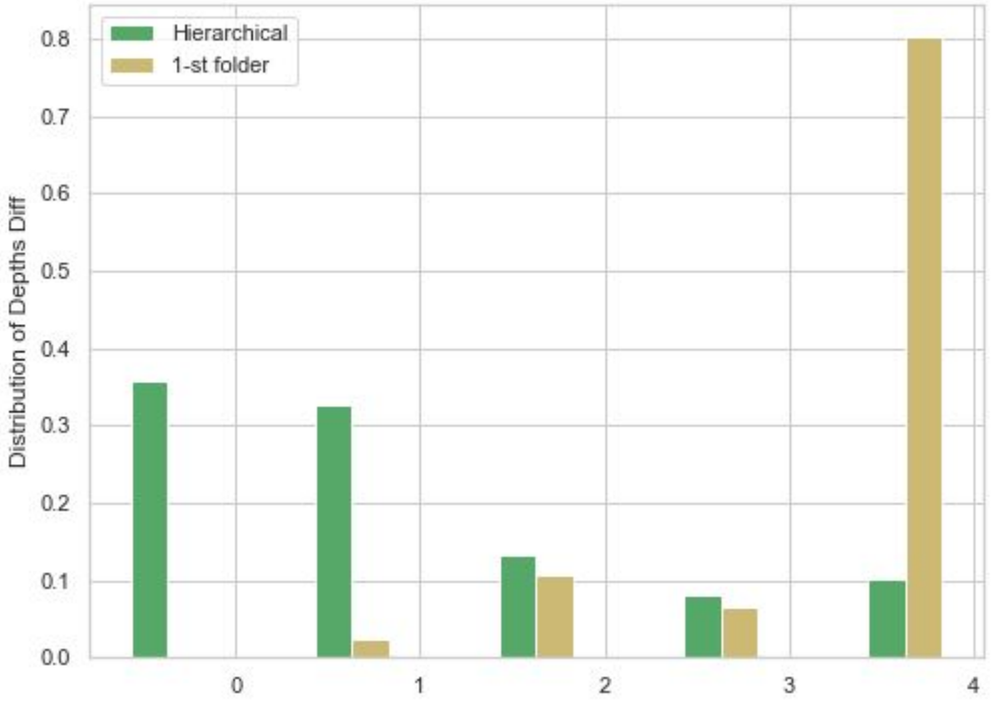


Predicted Depths



local file folder levels: 9 in total;
first 5 considered for predictions

Comparison of Depths Diff



'CP 2006 Theoretical potential energy surfaces for excited mercury trimers.pdf'

'cp', '2006', 'theoret', 'potenti', 'energi', 'surfac', 'for', 'excit', 'mercuri', 'trimer', 'pdf'

Natural Language
Analyses with NLTK

pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Original File Name

Remove - _ .

Tokenize

Stem

Remove Stop Words

tf-idf

Feature Engineering (NLP)

['CP-2006-Theoretical-potential-energy-surfaces for excited mercury trimers.pdf']

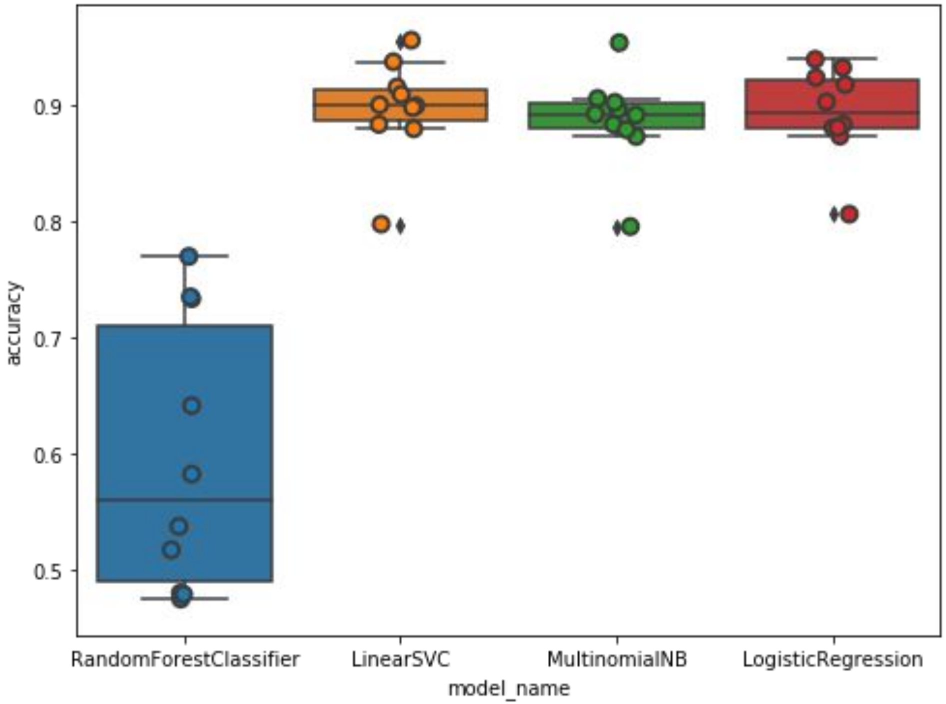
['cp-2006-theoretical-potential-energy-surfaces for excited mercury trimers.pdf']

[['cp', '2006', 'theoretical', 'potential', 'energy', 'surfaces', 'for', 'excited', 'mercury', 'trimers', 'pdf']]

[['cp', '2006', 'theoret', 'potenti', 'energi', 'surfac', 'for', 'excit', 'mercuri', 'trimer', 'pdf']]

['cp 2006 theoret potenti energi surfac for excit mercuri trimer pdf']

Text Classification: Softmax Regression

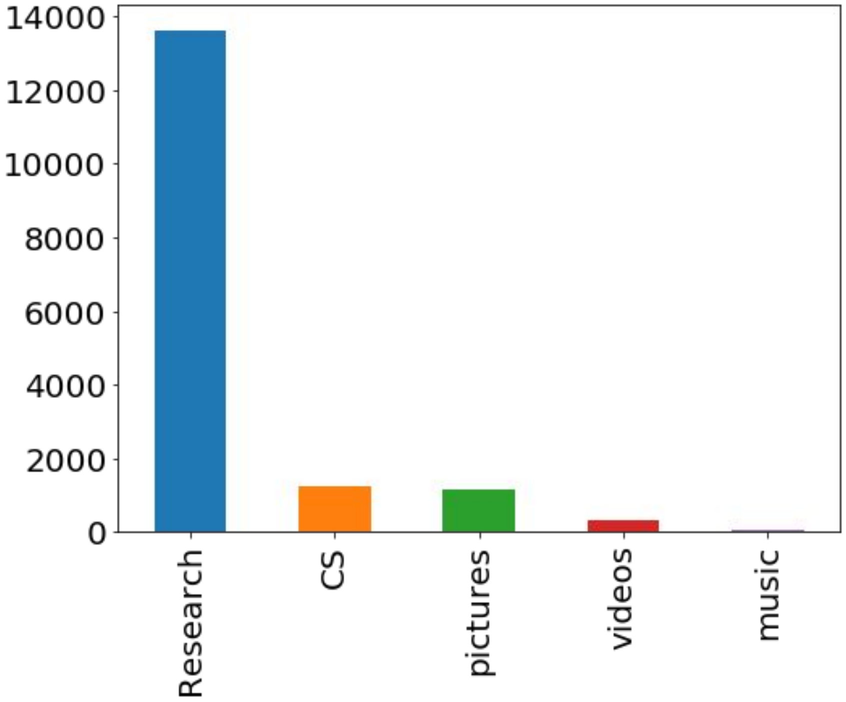


Logistic regression:

Fast, stable, with predict_prob

1st-depth AUC >0.9

Text Classification: Softmax Regression



Text Classification:

- Logistic Regression
- 1st-Folder AUC 90%

Challenges:

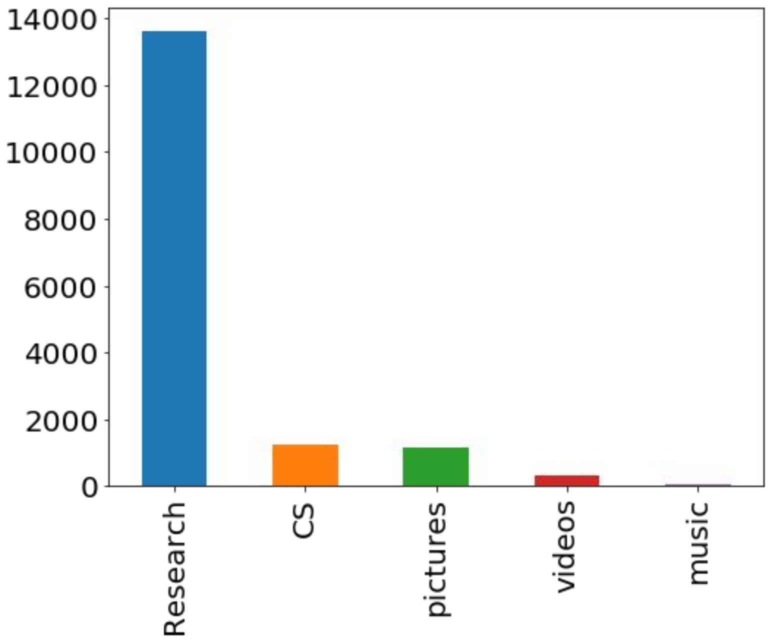
- Varied depth of folders
- 2st-Folder AUC 65%
- Files live in interior nodes

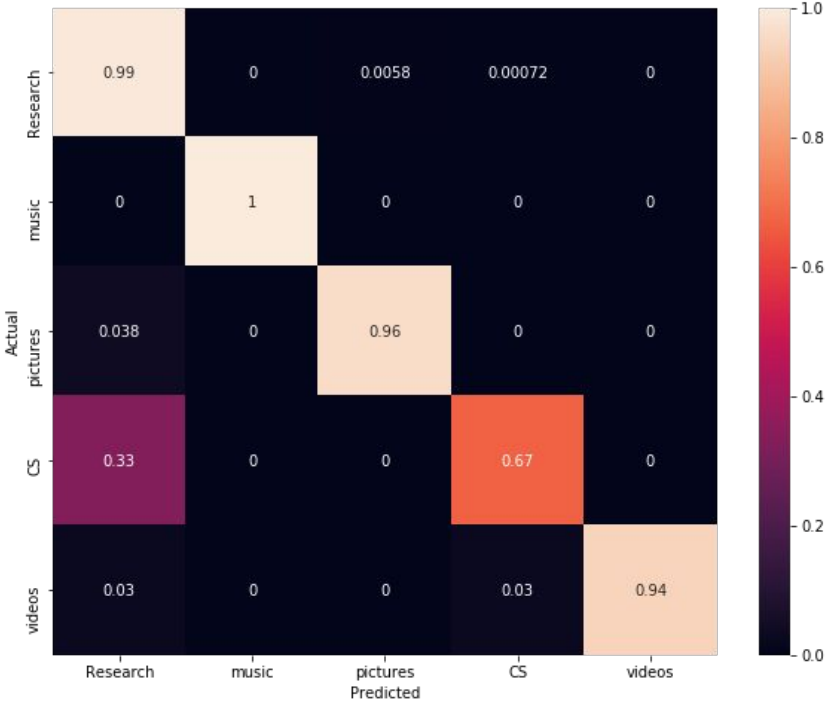
Solutions:

- **Hierarchical Classification**

Text Classification: Softmax Regression

- 16400+ files in total
- 1st split accuracy > 90%





Softmax regression (or multinomial logistic regression)

sklearn.linear_model.LogisticRegression implements **one-vs-rest** by default when given more than two classes.
One-vs.-res (or *one-vs.-all*, OvA or OvR, *one-against-all*, OAA) strategy involves training a single classifier per class, ***with the samples of that class as positive samples and all other samples as negatives.*** This strategy requires the base classifiers to produce a ***real-valued confidence score*** for its decision, rather than just a class label.

Sigmoid (Logistic)

$$h_\theta(x) = \frac{1}{1+\exp(-\theta^\top x)}$$

Softmax

$$\begin{aligned} P(y^{(i)} = k | x^{(i)}; \theta) &= \frac{\exp((\theta^{(k)} - \psi)^\top x^{(i)})}{\sum_{j=1}^K \exp((\theta^{(j)} - \psi)^\top x^{(i)})} \\ &= \frac{\exp(\theta^{(k)\top} x^{(i)}) \exp(-\psi^\top x^{(i)})}{\sum_{j=1}^K \exp(\theta^{(j)\top} x^{(i)}) \exp(-\psi^\top x^{(i)})} \\ &= \frac{\exp(\theta^{(k)\top} x^{(i)})}{\sum_{j=1}^K \exp(\theta^{(j)\top} x^{(i)})}. \end{aligned}$$

Softmax regression (or multinomial logistic regression)

Other Names

- Multinomial logistic regression
- Multivariate logistic regression
- Multiclass logistic regression
- Maximum entropy (MaxEnt) classifier
- Softmax regression

Relationship to Logistic Regression

In the special case where $K = 2$, one can show that softmax regression reduces to logistic regression. This shows that softmax regression is a generalization of logistic regression. Concretely, when $K = 2$, the softmax regression hypothesis outputs

$$h_\theta(x) = \frac{1}{\exp(\theta^{(1)\top} x) + \exp(\theta^{(2)\top} x^{(i)})} \begin{bmatrix} \exp(\theta^{(1)\top} x) \\ \exp(\theta^{(2)\top} x) \end{bmatrix}$$

Taking advantage of the fact that this hypothesis is overparameterized and setting $\psi = \theta^{(2)}$, we can subtract $\theta^{(2)}$ from each of the two parameters, giving us

$$\begin{aligned} h(x) &= \frac{1}{\exp((\theta^{(1)} - \theta^{(2)})^\top x^{(i)}) + \exp(\vec{0}^\top x)} \begin{bmatrix} \exp((\theta^{(1)} - \theta^{(2)})^\top x) \exp(\vec{0}^\top x) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{1+\exp((\theta^{(1)}-\theta^{(2)})^\top x^{(i)})} \\ \frac{\exp((\theta^{(1)}-\theta^{(2)})^\top x)}{1+\exp((\theta^{(1)}-\theta^{(2)})^\top x^{(i)})} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{1+\exp((\theta^{(1)}-\theta^{(2)})^\top x^{(i)})} \\ 1 - \frac{1}{1+\exp((\theta^{(1)}-\theta^{(2)})^\top x^{(i)})} \end{bmatrix} \end{aligned}$$

Thus, replacing $\theta^{(2)} - \theta^{(1)}$ with a single parameter vector θ', we find that softmax regression predicts the probability of one of the classes as $\frac{1}{1+\exp(-(\theta')^\top x^{(i)})}$, and that of the other class as $1 - \frac{1}{1+\exp(-(\theta')^\top x^{(i)})}$, same as logistic regression.

One-vs-All and One-vs-One

The difference is the number of classifiers you have to learn, which strongly correlates with the decision boundary they create.

Assume you have N different classes. One vs all will train one classifier per class in total N classifiers. For class ii it will assume ii-labels as positive and the rest as negative. This often leads to imbalanced datasets meaning generic SVM might not work, but still there are some workarounds.

In one vs one you have to train a separate classifier for each different pair of labels. This leads to $N(N-1)/2$ classifiers. This is much less sensitive to the problems of imbalanced datasets but is much more computationally expensive.

How does it work?

Data

Feature Engineering

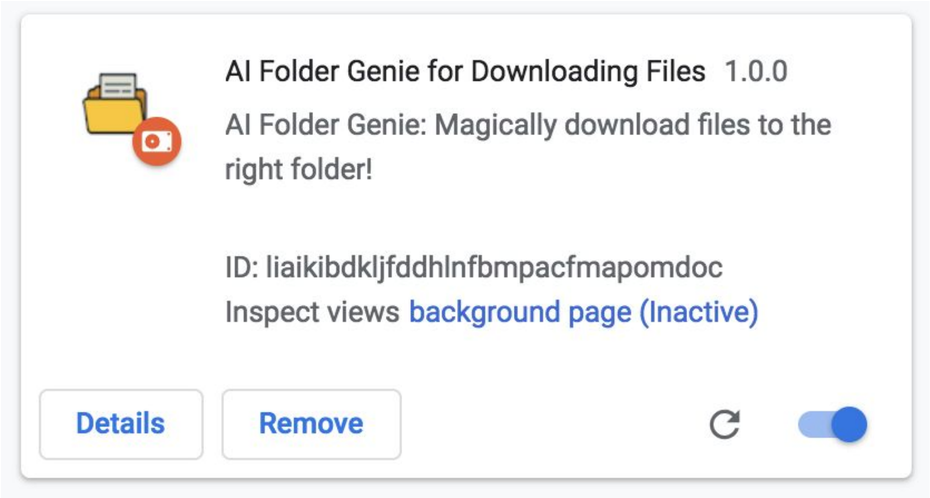
Label Engineering

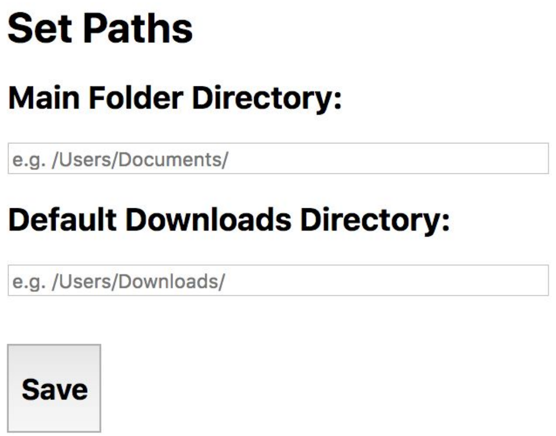
Classifier Picking

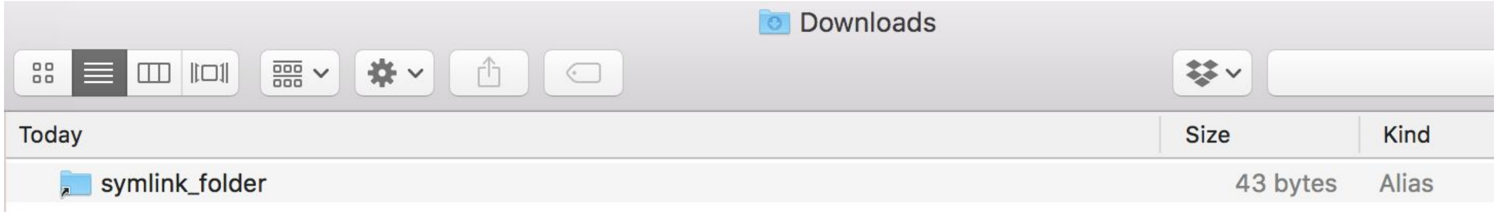
Hierarchical Classification (wrote my own algorithm)

Move files (symbolic links)

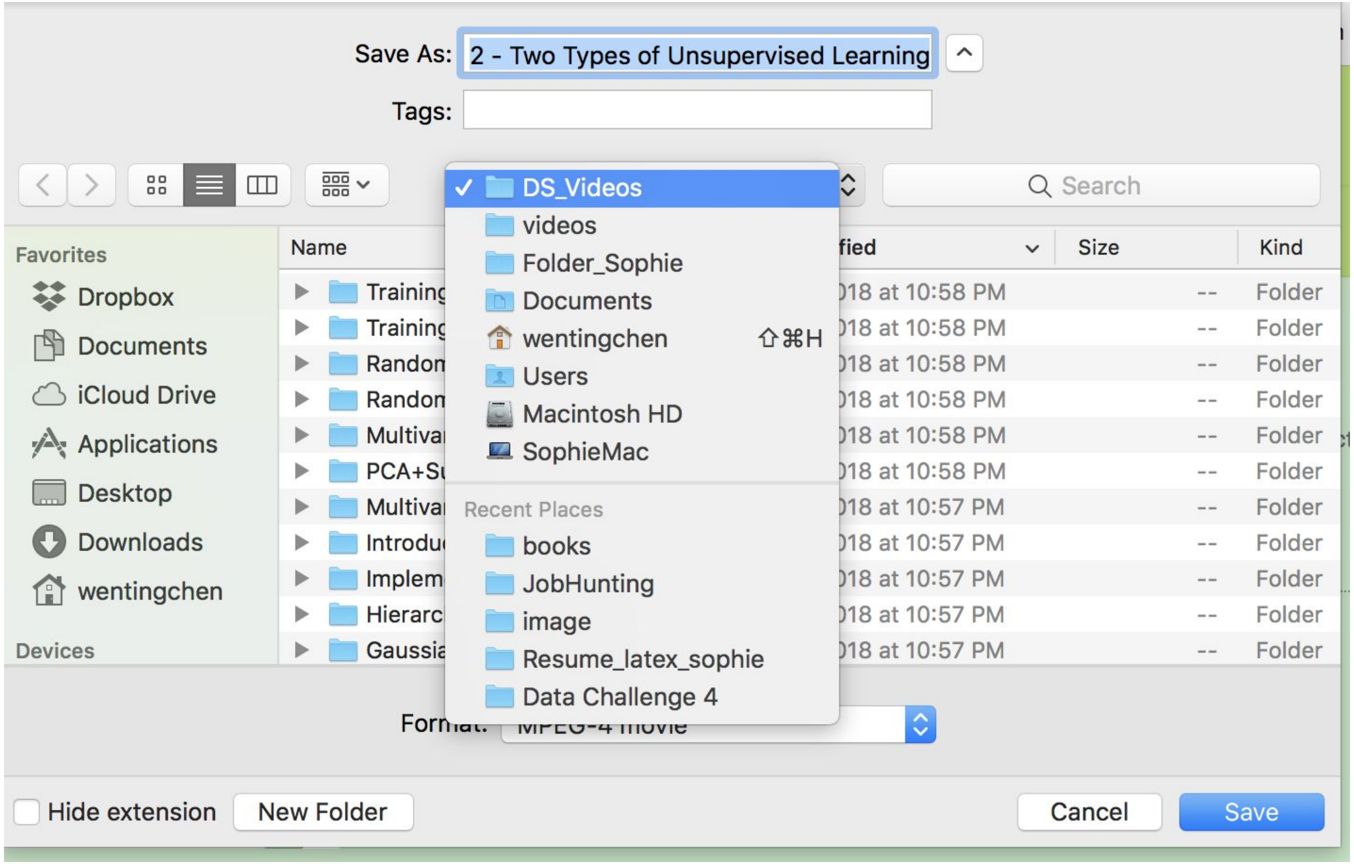
AI Folder Genie Chrome Extension







Prediction go to Nested Folder



Sophie's Folder Structure