The Simple Science Of Unsupervised NLP

goo.gl/zN2a46

Anubhav Singh

i@xprilion.com

MLCC Facilitator | Founder - The Code Foundation | Human

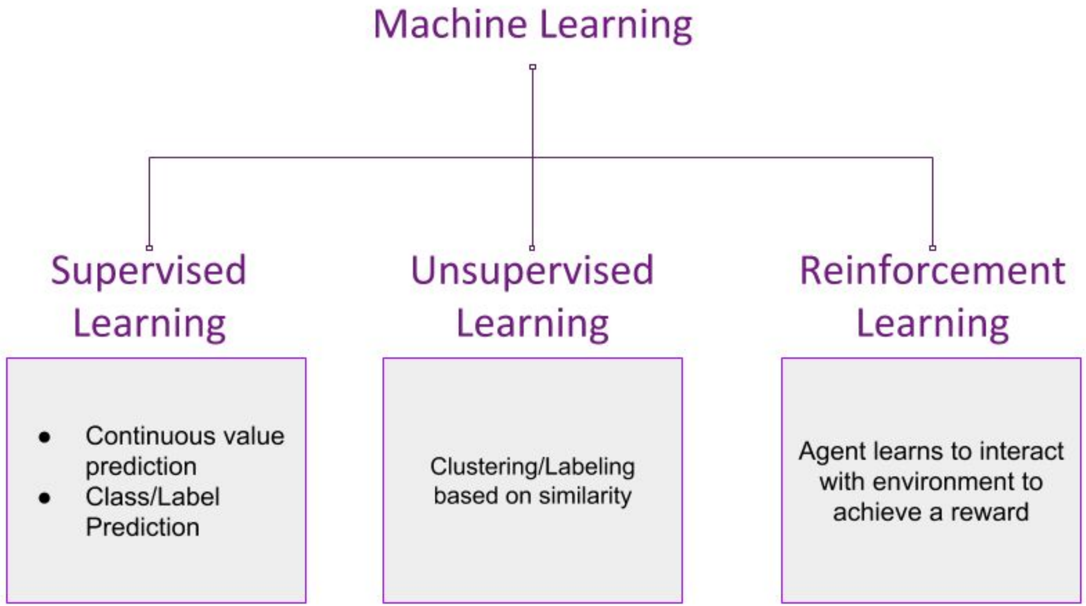
What exactly am I talking about?

What shall we be doing in the next 1 hour?

- Concept:	Unsupervised ML
- Concept:	NLP
- Concept:	Unsupervised ML for NLP
- Code:	The ~~Oil~~ Data we shall deal with
- Code:	Clustering

Unsupervised Machine Learning

Concept: Unsupervised ML



Machine Learning

Supervised Learning

- Continuous value prediction
- Class/Label Prediction

Unsupervised Learning

Clustering/Labeling based on similarity

Reinforcement Learning

Agent learns to interact with environment to achieve a reward

Unsupervised Machine Learning

Concept: Unsupervised ML

Clustering

Unknown data, put together in groups based on patterns exhibited (similar features)

Un-supervised ML

Classification

Unknown data, put together in groups based on training from labeled data previously

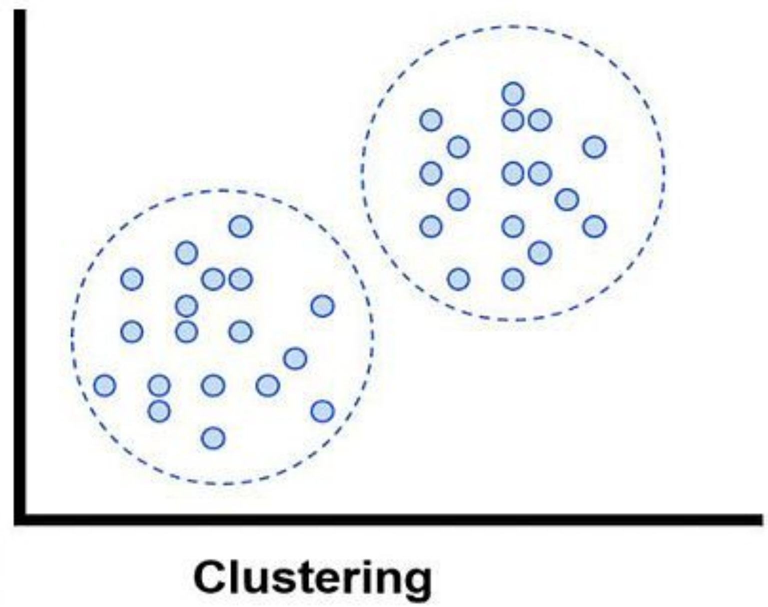
Supervised ML

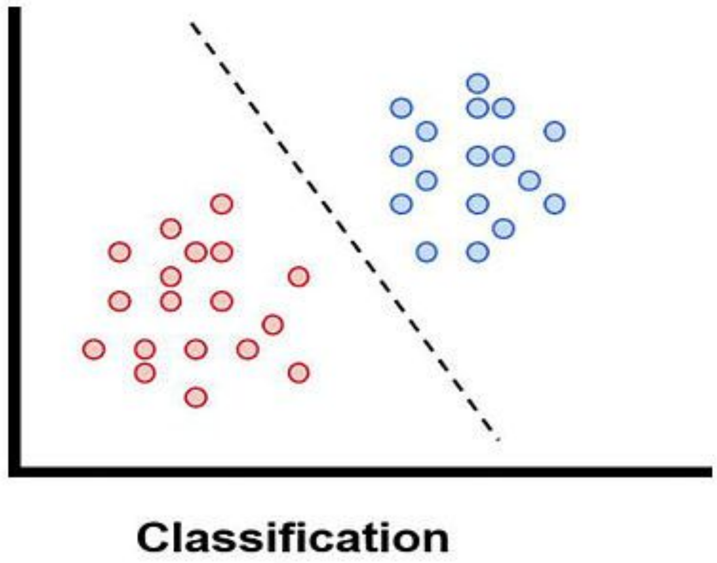
Unsupervised Machine Learning

Concept: Unsupervised ML

Clustering

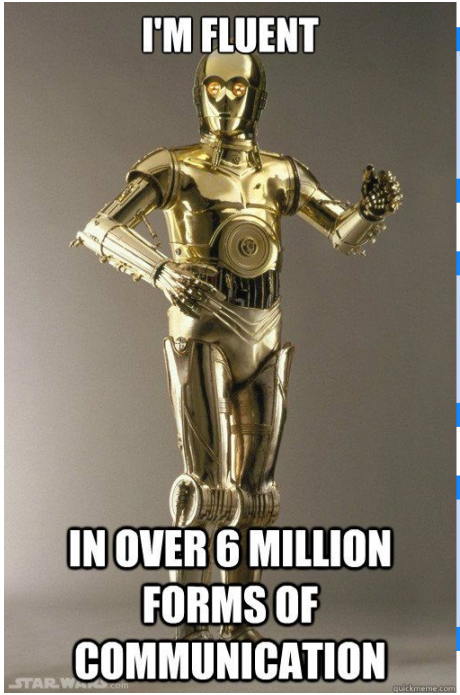
Classification





Natural Language Processing

Concept: NLP



Chatbots, Reply suggestions, Forum answer bots

Article Summarize, Auto essay evaluation, Topics

Sentiment, Natural Language Query, Assistant Apps

Natural Language Processing

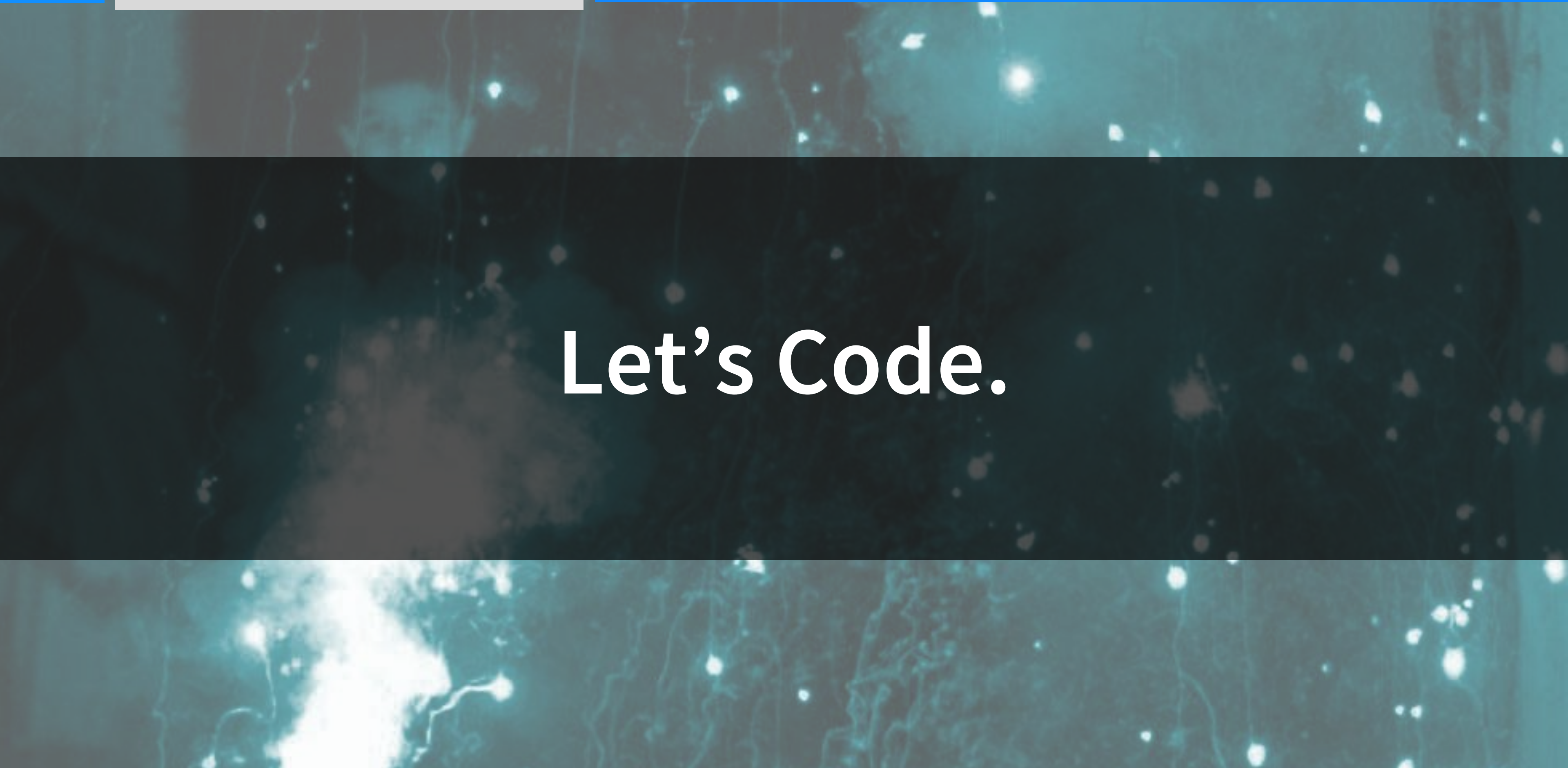
Concept: Unsupervised ML for NLP

Unsupervised NLP

- Group similar articles together
- Recommend articles based on previously read
- Extract Named Entities

Supervised NLP

- Categorize articles
- Spam filtering
- Chatbots



Let’s Code.

Environment Setup

Environment Setup

Things to install -

- Anaconda (Python)
  - Download from - goo.gl/EnYY9V
- NLTK
  - conda install -c anaconda nltk

NLTK Setup

In a (Anaconda) command prompt / terminal ---

```
user@ubuntu:~$ python

Python 3.7.0 (default, Jun 28 2018, 13:15:42)
[GCC 7.2.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>  import nltk
>>>  nltk.download()
```

Setup Working Directory

```
user@ubuntu:~$  git clone https://github.com/xprilion/df18.git
```

Knowing our data

Things already done:

- Imported data from WikiPedia and IMDB for 100 movies
- Data is in the “data” folder.
- Imported data is:
  - Movie Titles
  - Movie Genres
  - Movie Synopsis from IMDB
  - Movie Synopsis from WikiPedia

“STOP”, “HALT!”, “TARRY!”

StopWords

- stopwords = nltk.corpus.stopwords.words('english')
- stemmer = SnowballStemmer("english")

Breaking down the text

Tokenize

- Documents into Sentences
  - Done for you
- Sentences into words
  - words = [word for word in nltk.word_tokenize(sents[0])]

To the roots

Stemming

- To bring any text into its root word
- stemmer.stem(word)
- `stems = [stemmer.stem(t) for t in filtered_words]`

TF - IDF

“Term Frequency - Inverse Document Frequency”

Given a collection of documents (or corpus), how important a word is in to a particular document.

TF - IDF

Consider a document containing 100 words wherein the word sun appears 3 times. The term frequency (i.e., tf) for sun is then (3 / 100) = 0.03.

Now, assume we have 10 million documents and the word sun appears in one thousand of these.

Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4.

Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12

Grams...Kilos?

n-grams

“We welcome you to GDG Kolkata DevFest 2018”

2 - gram: [“We welcome”, “welcome you”, “you to”, ….]

3 - gram: [“We welcome you”, “welcome you to”, “you to GDG”, ….]

4 - gram: [“We welcome you to”, “welcome you to GDG”, ….]

KMeans

KMeans

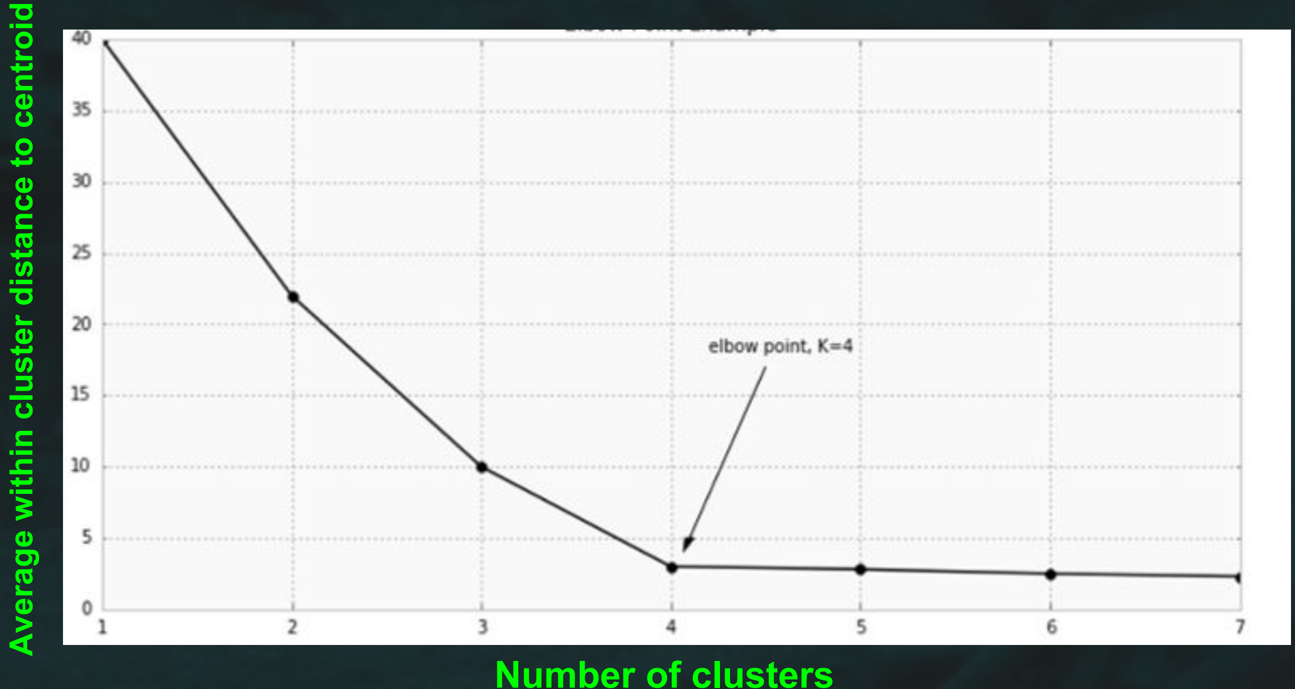
To cluster data points into ‘k’ clusters.

Choice of ‘k’ is up to the user.

How to choose the right value of ‘k’?

The Elbow

KMeans



Cosine Similarity

Cosine Similarity

1. Julie loves me more than Linda loves me
2. Jane likes me more than Julie loves me

Words: “me Julie loves Linda than more likes Jane”

Cosine Similarity

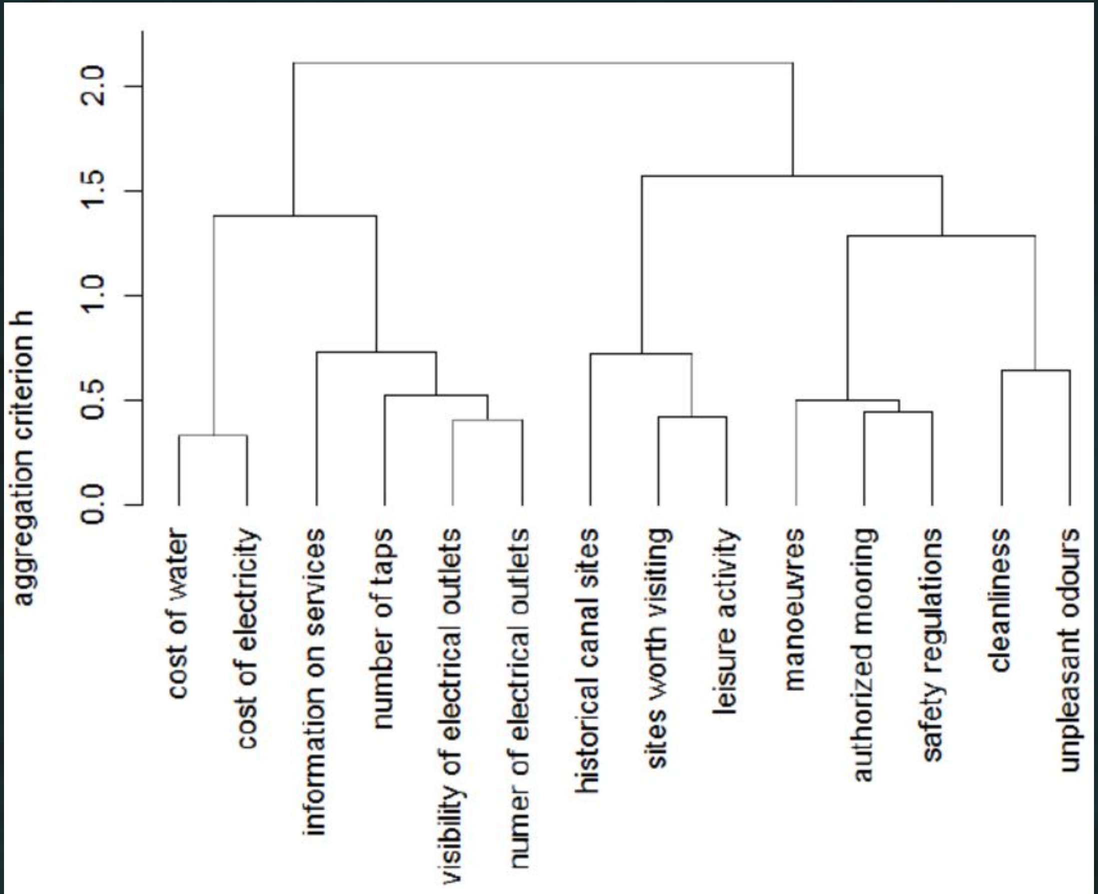
me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1

a: [2, 0, 1, 1, 0, 2, 1, 1]

b: [2, 1, 1, 0, 1, 1, 1, 1]

Co-sine angle: ~ 0.822

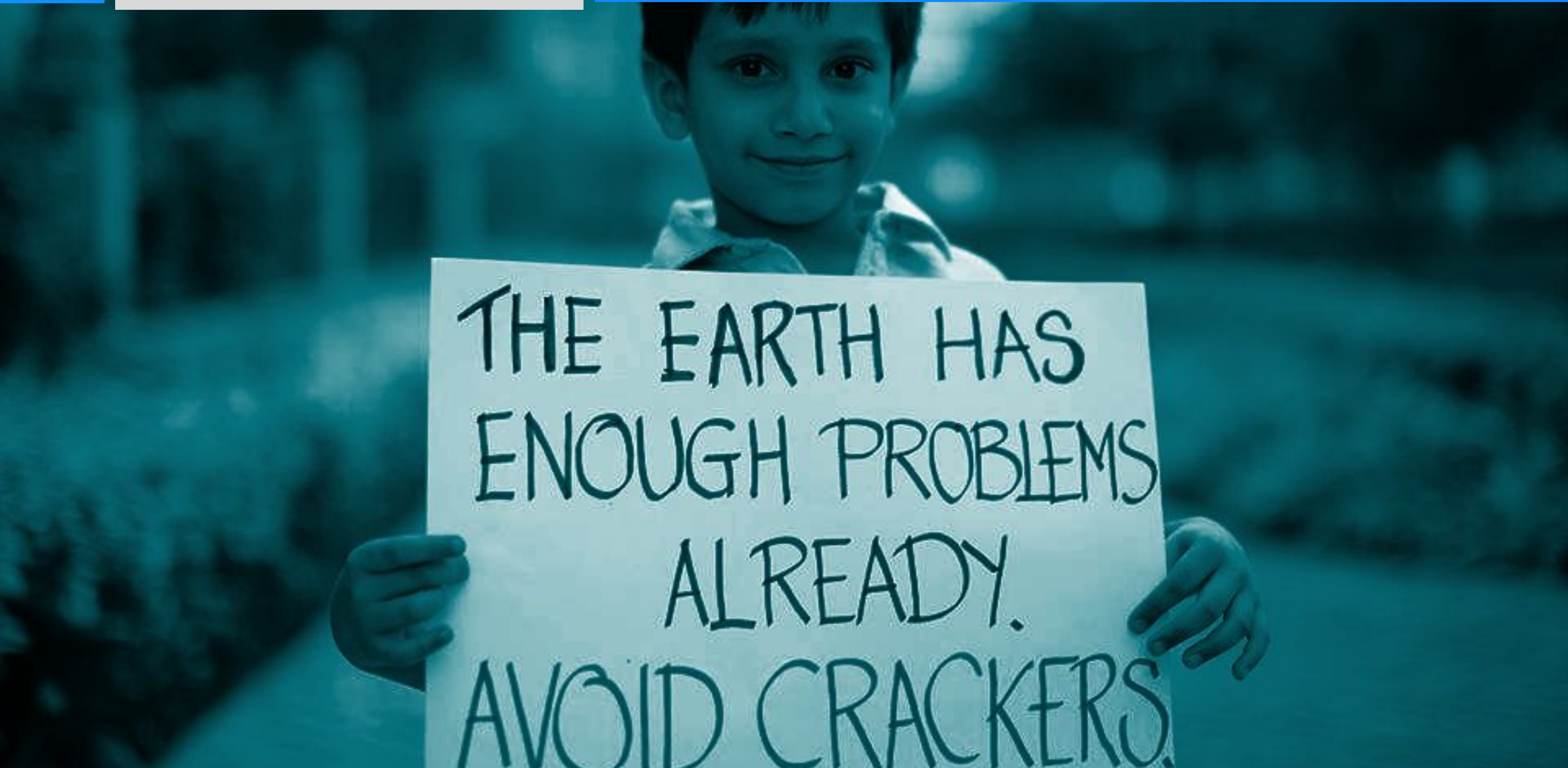
Dendrograms



Questions?

Are We Done?

Answers.



We wish you…

The Code Foundation is a venture to build meaningful open source software which are currently owned and kept close-sourced by companies.

We've just started. We need you! :) Happy Diwali!