Characterizing Real World Datasets into MIME Types

Kevin Khoi
Nam Le
Khanh Duy Le

Three Datasets

1. TripAdvisor Review Dataset
2. Yelp Review Dataset
3. Cats and Dogs Audio Dataset

TripAdvisor Review Dataset

A millions reviews from 4333 hotels and 14 millions restaurant crawled from TripAdvisor.

MIME TYPES		
application/json	**text/plain**	**image/jpeg**
One file 2GB	One file 1.3 GB	More than millions .jpg image files

TripAdvisor Review Dataset

MIME TYPE	VOLUME	VELOCITY	VARIETY	VERACITY	VALUE
application/json	High	Low	High	Low	High
text/plain	High	Low	High	High	High
image/jpeg	High	High	High	Low	High

Software: Microsoft Excel, Open Office, JSON Editor, Sublime, Image Editor

Yelp Review Dataset

- Subset of 5.2M Yelp Reviews from 11 cities released to the public for sentiment analysis and graph research

MIME TYPES
text/plain
One file 1.3 GB

Yelp Review Dataset

MIME TYPE	VOLUME	VELOCITY	VARIETY	VERACITY	VALUE
text/plain	High	High	Medium	High	High

Software - Microsoft Excel, Open Office, LibreOffice Calc

Cats and Dogs Dataset

- 200+ variations of cats meowing and dogs barking for audio classification purposes

MIME TYPES
audio/wav
200+ wav files

Cats and Dogs Dataset

MIME TYPE	VOLUME	VELOCITY	VARIETY	VERACITY	VALUE
audio/wav	Low	Low	Low	N/A	High

Software - Audacity, Wave Editor