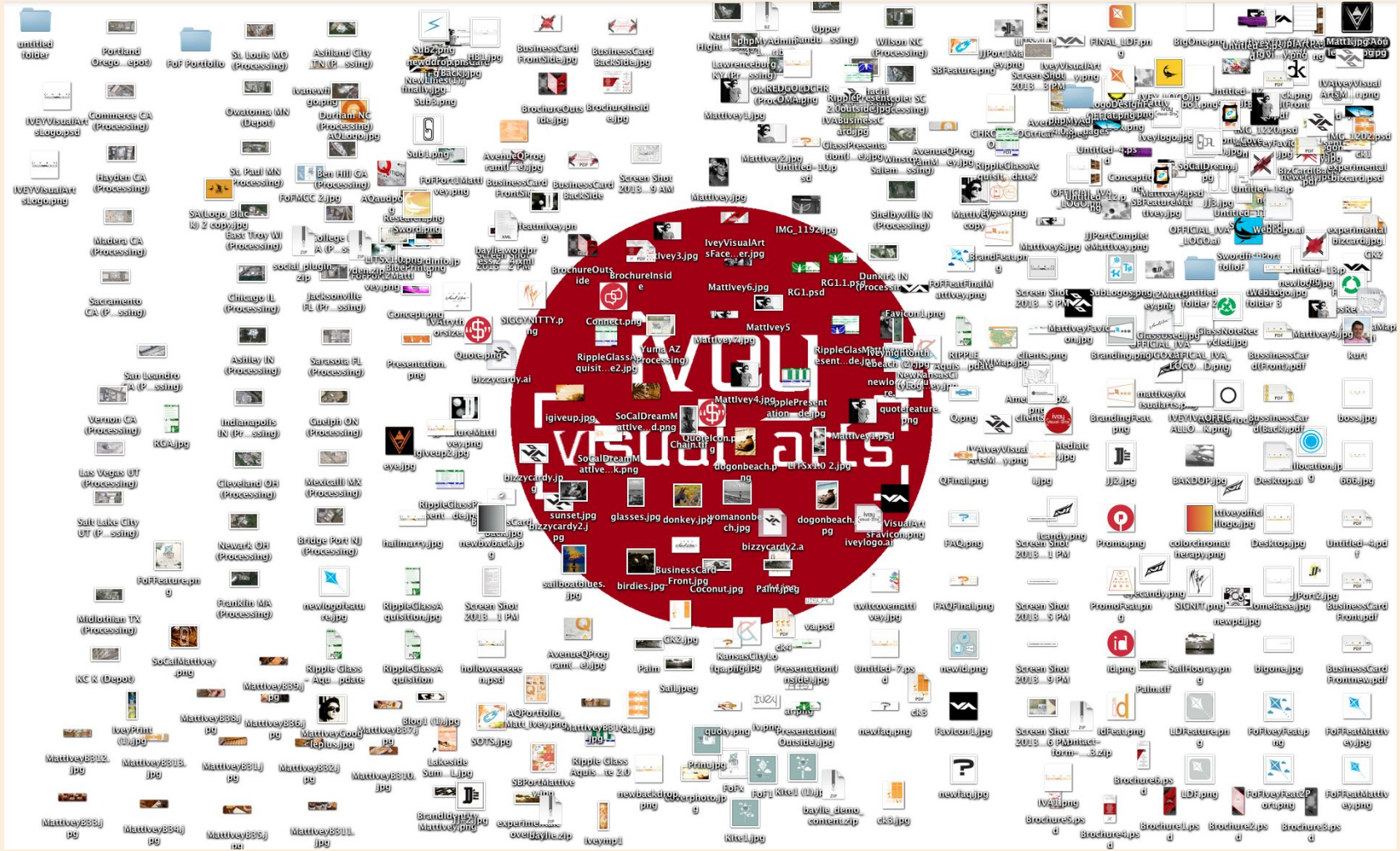
CONTROLLED CHAOS

A tool for taming digital clutter

by **Faustina Maria Giaquinta Caggiati**


[vdv]
visual arts

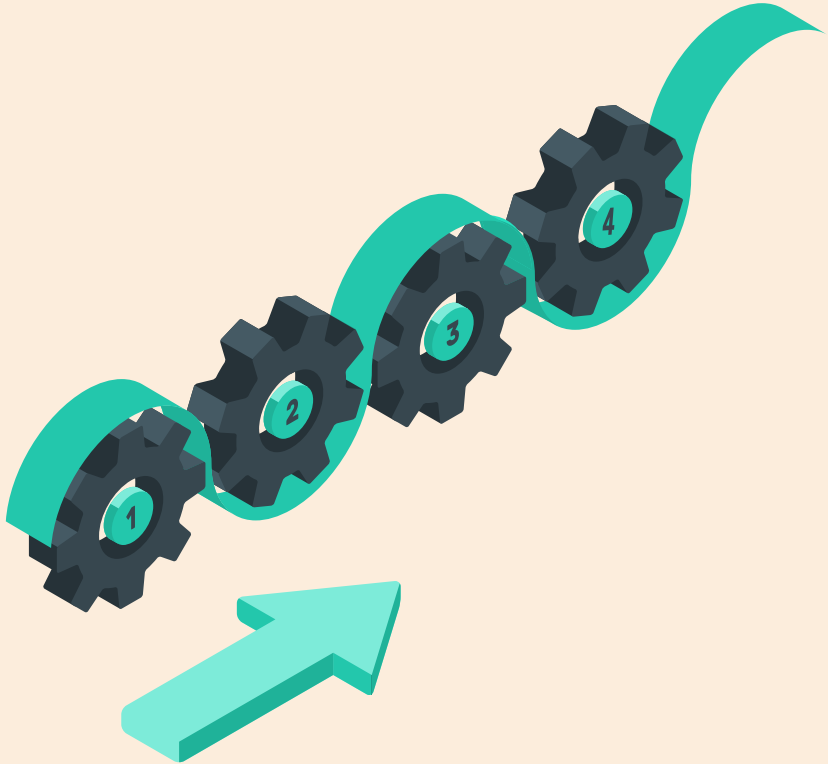
CAN MY DESKTOP NOT BE A METAPHOR FOR MY LIFE PLEASE?!

CONTROLLED CHAOS

a Machine Learning-based solution

UNDER THE HOOD

APPLICATION



STEP I - SCANNING FOLDERS

Select folders to scan archive files

STEP 2 - ARBITRARY CLASSIFICATION

Fetch names and absolute paths of files
Classify folders, media, compressed files, disk images, programs & system files, fonts, images, web/data/email-related files

STEP 3 - TEXT CLASSIFICATION

Extract text from text files
Cluster documents

STEP 4 - INDEXER

Present clusters and associated files with symbolic links

TEXT CLASSIFICATION

TEXT EXTRACTION

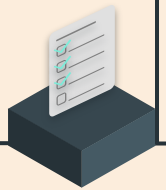
Extraction with Apache Tika

Storage in MongoDB

PREPROCESSING

Lemmatization with Spacy large model

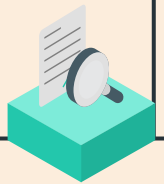
Snowball Stemming with NLTK



FEATURE EXTRACTION

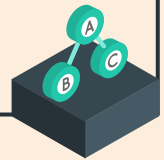
Term Frequencies Matrix (**TF-IDF**) with NLTK

SVD with Sklearn to decompose term matrix and get optimal **k** clusters



CLUSTERING

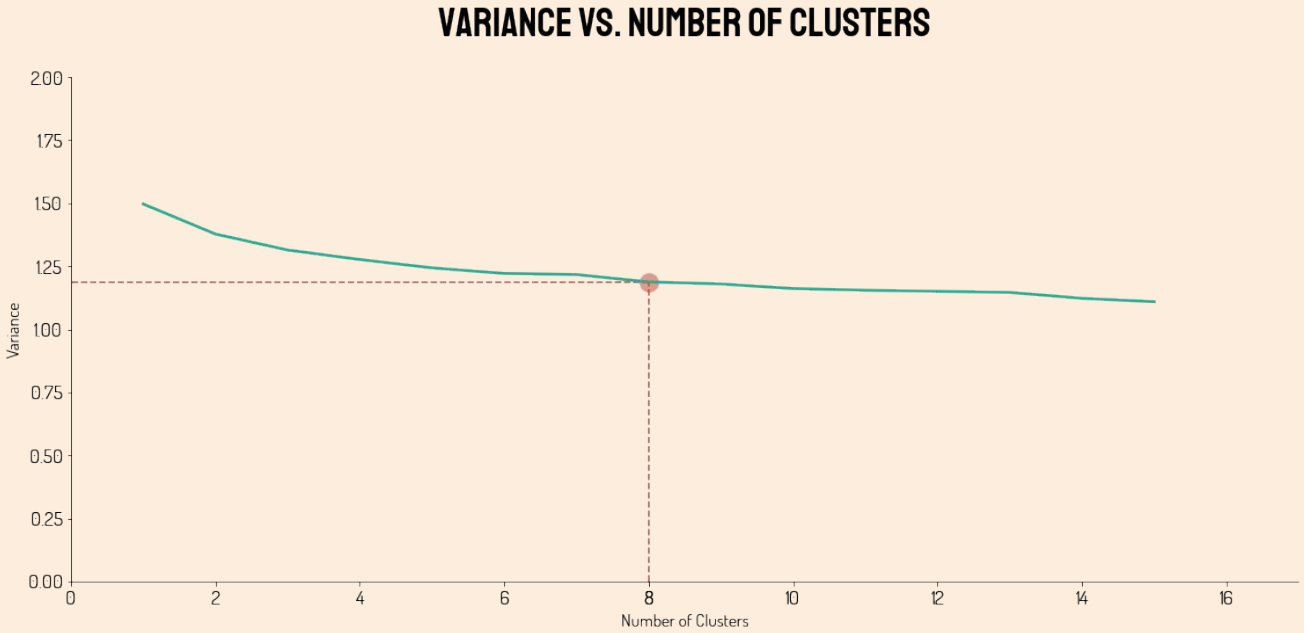
KMeans (Sklearn) with optimal **k clusters** for files analyzed



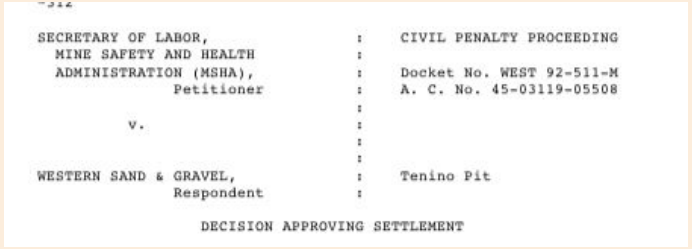
NUMBER OF CLUSTERS

FINDING MINIMUM OPTIMAL NUMBER OF CLUSTERS

Algorithm automatically detects optimal number of clusters generated

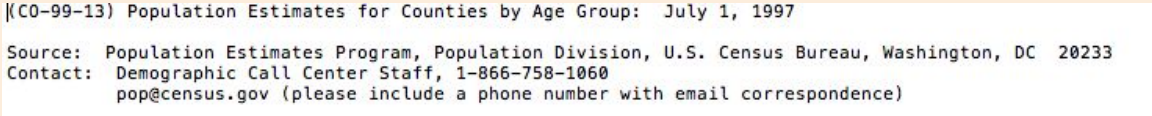


TEST CASE



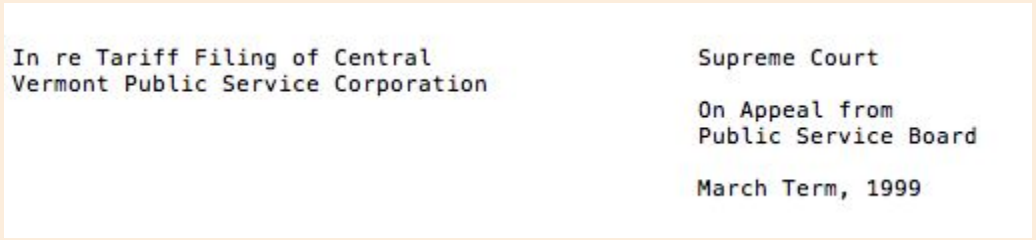
SECRETARY OF LABOR, : CIVIL PENALTY PROCEEDING
MINE SAFETY AND HEALTH :
ADMINISTRATION (MSHA), : Docket No. WEST 92-511-M
Petitioner : A. C. No. 45-03119-05508
:
v. :
:
:
WESTERN SAND & GRAVEL, : Tenino Pit
Respondent :

DECISION APPROVING SETTLEMENT



(CO-99-13) Population Estimates for Counties by Age Group: July 1, 1997

Source: Population Estimates Program, Population Division, U.S. Census Bureau, Washington, DC 20233
Contact: Demographic Call Center Staff, 1-866-758-1060
pop@census.gov (please include a phone number with email correspondence)

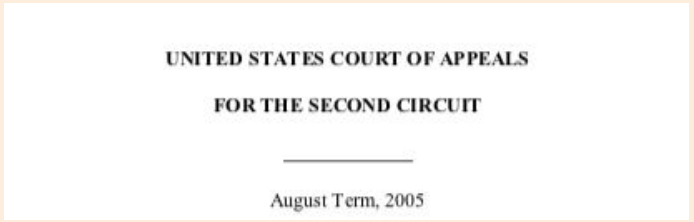


In re Tariff Filing of Central
Vermont Public Service Corporation

Supreme Court

On Appeal from
Public Service Board

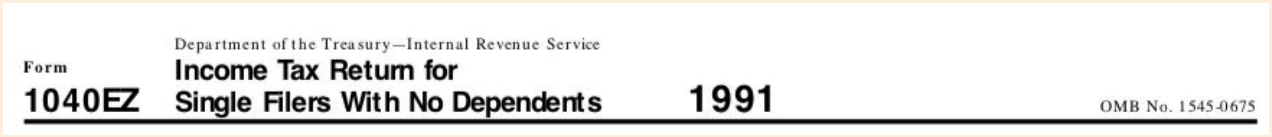
March Term, 1999



UNITED STATES COURT OF APPEALS

FOR THE SECOND CIRCUIT

August Term, 2005



Department of the Treasury—Internal Revenue Service

Form **1040EZ** **Income Tax Return for Single Filers With No Dependents** **1991**

OMB No. 1545-0675

THANKS

faustinamgiaquinta@gmail.com
+1 (917) 960 3057
faustinamaria.com
linkedin.com/in/faustinagiaquinta
twitter.com/miss_sizigia