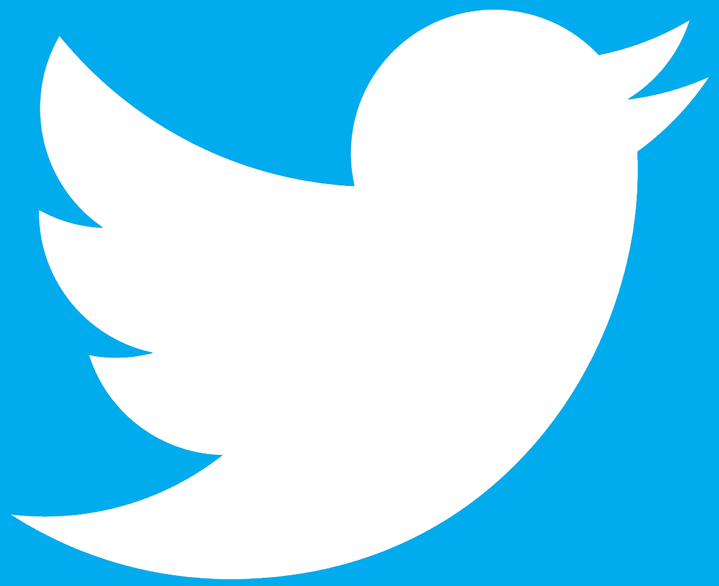
NATURAL LANGUAGE PROCESSING

(for the impatient)

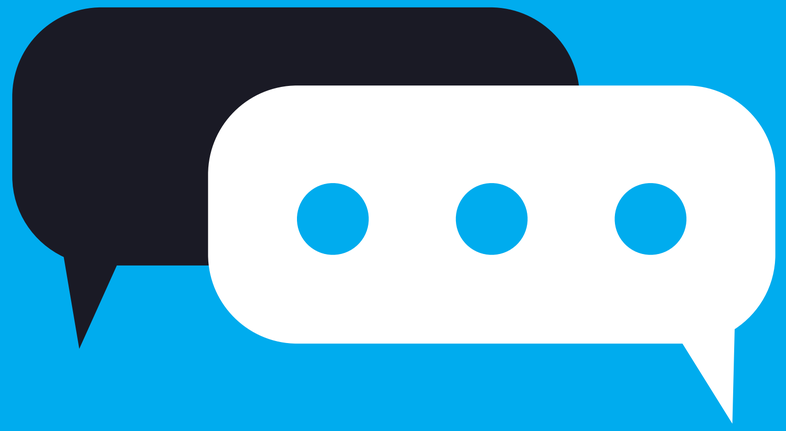
INTRODUCTION

Sebastian Dziadzio

cliqz.com
ghostery.com



@sebadzia



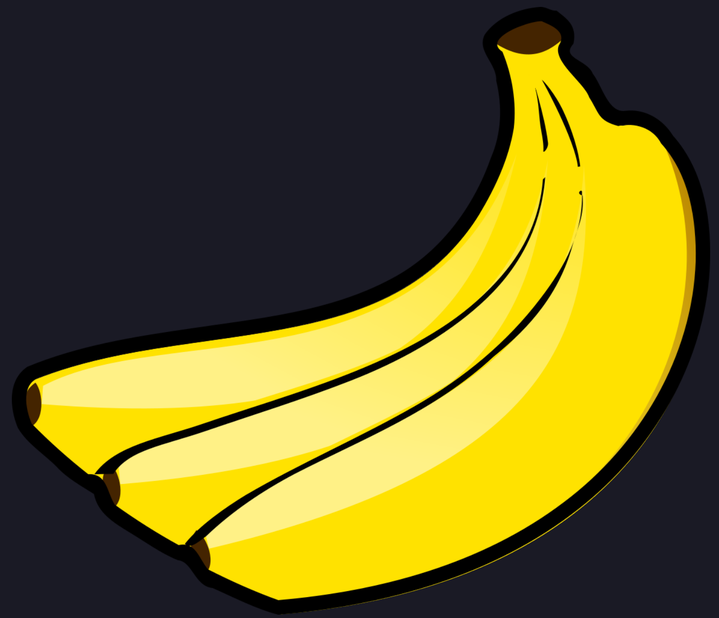
Natural Language Processing

Why is Natural Language Processing important?

Why is Natural Language Processing hard?

Time flies like an arrow.

Fruit flies like a banana.



worstest
hatemonger
post-truth
pogonophobia

opinion size age shape colour origin material purpose noun

ugly large old brown English tweed hunting jacket

big fat Greek wedding

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

PROCESS & TOOLS

Get data

kaggle™

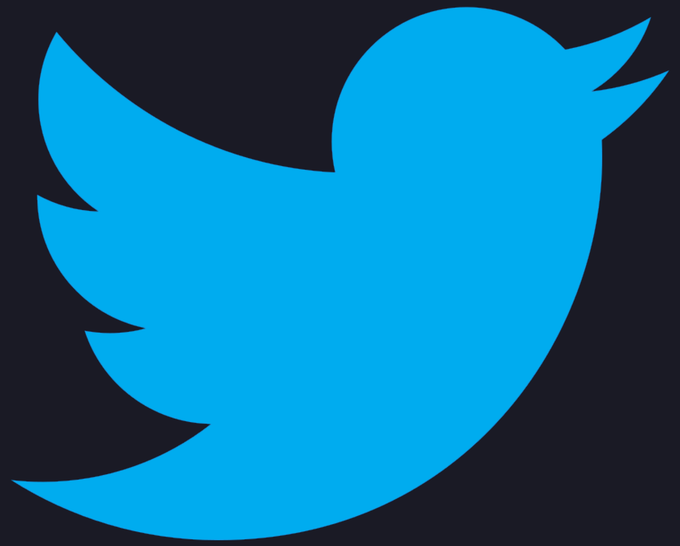
data.world

niderhoff/nlp-datasets

nlp.stanford.edu

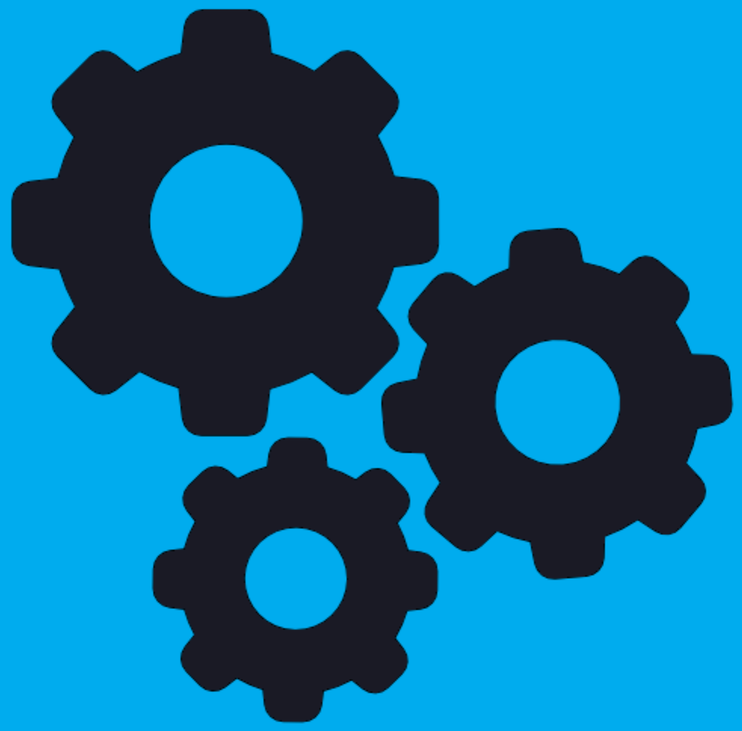
zil.ipipan.waw.pl

keon/awesome-nlp



Get data → Transform → Encode → Visualize → Model

Transform

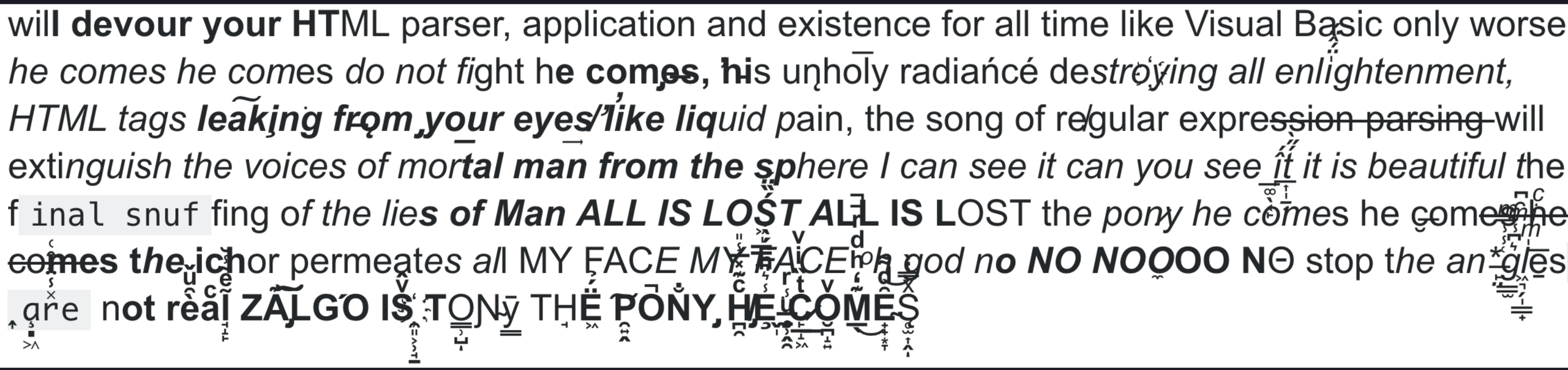


normalization
stemming
lemmatization
POS tagging

Plain Python
Unix
NLTK
SpaCy

@TRINArockstarr Trina flow, working on my abs & legs … So ""when sb ask ""what's yo REAL name"" I tell him Da Baddest Bitch 😂😂😱💪 #gym

trina flow working on my abs and legs so when somebody ask what is your real name I tell him da baddest bitch



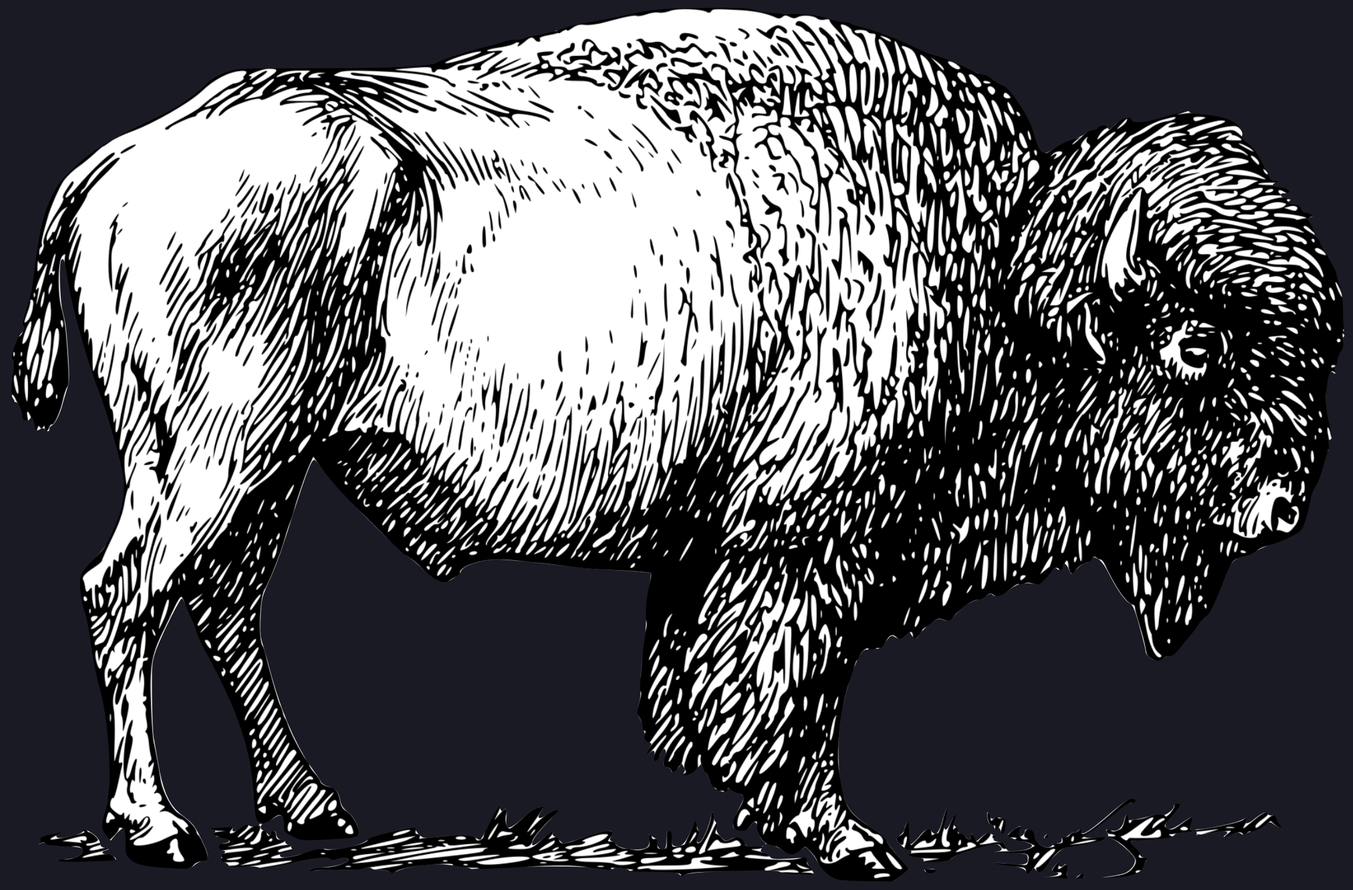
Listen, strange women lying in ponds distributing swords is no basis for a system of government. Supreme executive power derives from a mandate from the masses, not from some farcical aquatic ceremony.

Listen, strange woman lie in pond distribute sword be no basis for a system of government. Supreme executive power derive from a mandate from the mass, not from some farcical aquatic ceremony.

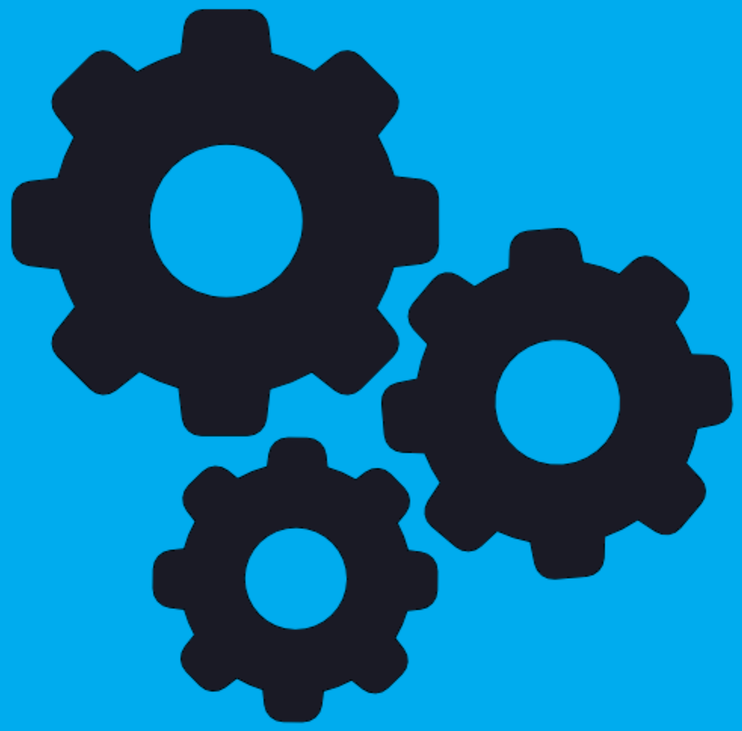
Listen, strang women lie in pond distribut sword is no basi for a system of government. Suprem execut power deriv from a mandat from the masses, not from some farcic aquat ceremon.

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo.

PN N PN N V V PN N



Encode



word2vec
GloVe
doc2vec
sense2vec

gensim

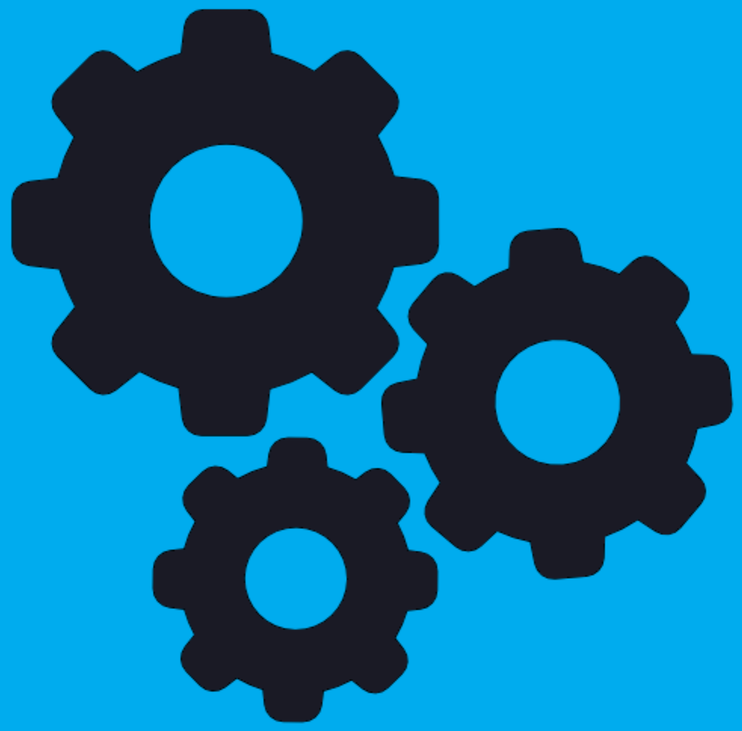
fastText

GloVe

EXAMPLE:
Information Retrieval

Get data → Transform → Encode → Visualize → Model

Visualize



t-SNE
PCA

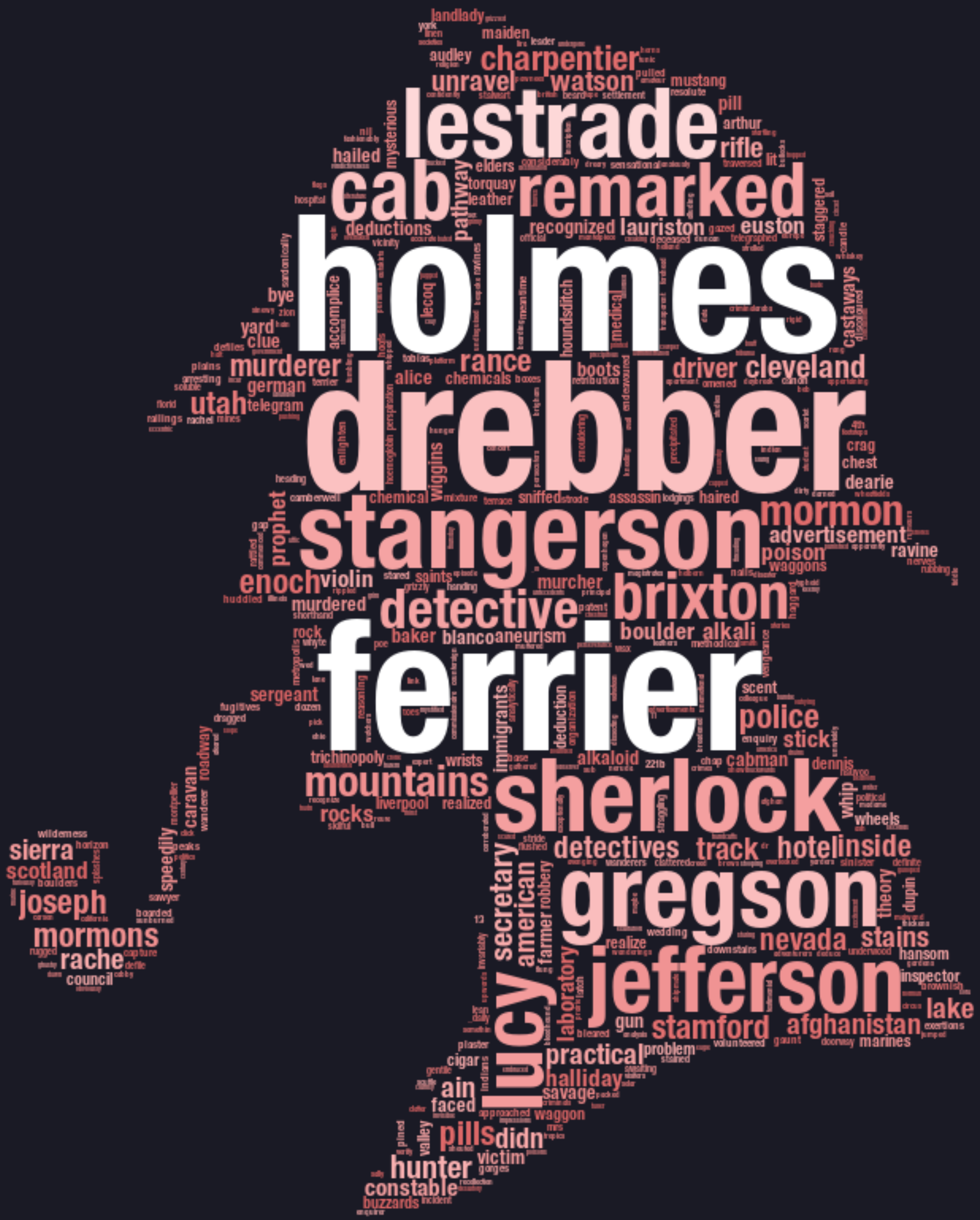
scikit-learn
matplotlib

EXAMPLE:
Project Gutenberg

darcy
bourgh
georgiana
pemberley
jane
forster
agreeable
catherine
behaviour
kitty
lydia
lucas
colonel
amiable
wickham
longbourn
netherfield
brighton
elizabeth
bennet
phillips
cousin
aunt
hertfordshire
mrs
hurst
civility
eliza
lucases
that
bingley
meryton
rosings
uncle
gardiner
hunsford
lizzy
charlotte
ladyship
fitzwilliam
collins
william
derbyshire
caroline
convinced
gratitude
resentment
compliment
mary
attentions
gracechurch
judgement
library
park
vanity
politeness
humour
anybody
neighbourhood


pig
pool
guinea
dinah
wasn
cook
lessons
turtle
rabbit
caterpillar
majesty
slates
pepper
hastily
croquet
wow
gardeners
soup
treacle
lory
hare
lizard
remarked
duchess
cheshire
pigs
hatter
footman
knave
gloves
bottle
pigeon
alice
lobster
doesn
dormouse
whiting
cats
mock
wouldn
sneezing
tea
baby
didn
jury
dodo
gryphon
timidly
mushroom
hedgehog
flamingo
mouse
cat
anxiously
serpent
angrily
fan
lobsters
twinkle
executioner
butter
hookah
hedgehogs
tarts
soo
puppy
haven
couldn
shouted
porpoise
tail
oop
kid
yer


baskerville
holmes
mortimer
moor
watson
stapleton
henry
barrymore
charles
sherlock
hound
grimpen
lestrade
baronet
convict
selden
merripit
stapletons
devonshire
baskervilles
coombe
frankland
hugo
cartwright
tracey
lyons
swear
moorland
hotel
gray
boulders
beard
alley
mire
wagonette
naturalist
tor
boot
huge
hut
mrs
legend
rocks
halloa
stick
james
granite
cabman
connection
dr
yew
charing
track
telegram
fog
laura
regent
butler
boots
cab
huts
baker
candle
heir
printed
dartmoor
devon
detective
hospital
postmaster
practitioner
county
tragedy
initials
slopes
hunt
naturalist
northumberland


holmes
drebber
ferrier
stangerson
lestrade
sherlock
gregson
jefferson
lucy
cab
remarked
brixton
detective
charpentier
watson
unravel
mormon
enoch
mountains
utah
murderer
cleveland
driver
rance
stamford
afghanistan
nevada
mormons
joseph
sierra
scotland
rache
hunter
constable
pills
police
stick
detectives
track
hotel
inside
secretary
american
practical
halliday
lauriston
euston
rifle
advertisement
poison
boulder
alkali
alkaloid
deductions
recognized
mysterious
hailed
laboratory
boots
accomplice
stains
lake
cabman

persuasion

pride

emma

eyre

sense

shirley

villette

categories

politics

poetics

constitution

caesar

leaves

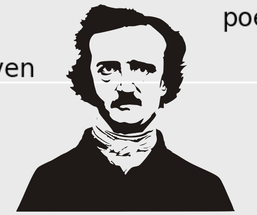
paradise

macbeth

poems

hamlet

raven



crime

karamazov

idiot

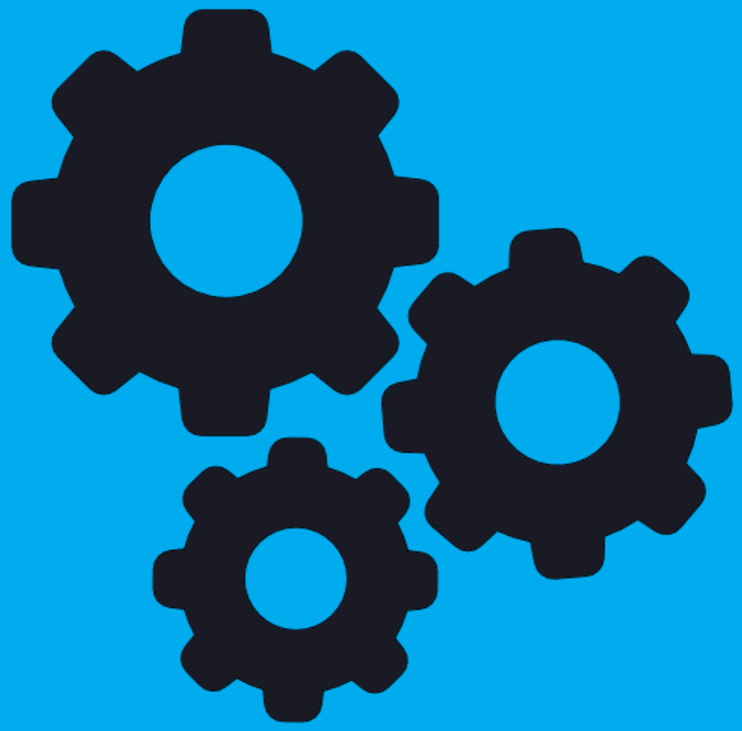
scarlet

four

baskervilles

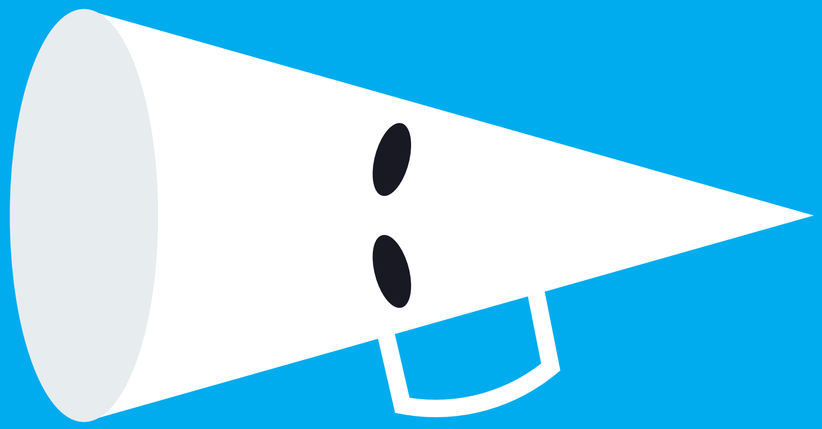
Get data → Transform → Encode → Visualize → Model

Model



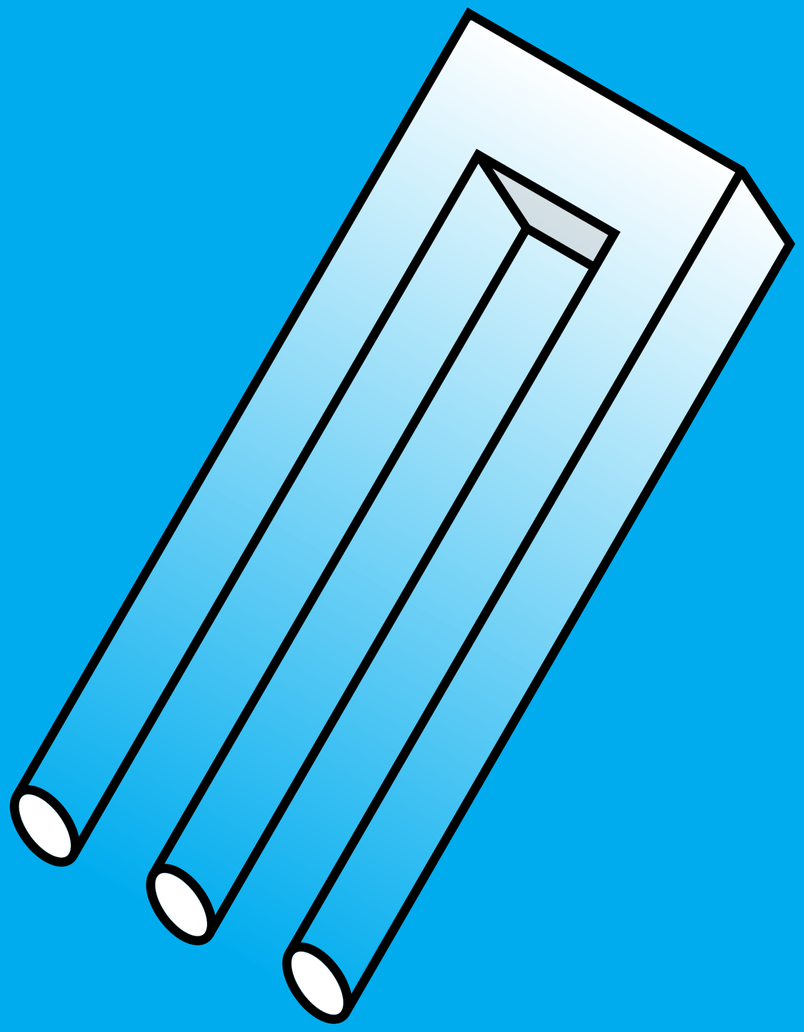
old school ML
(bi-)LSTM
attention

scikit-learn
Keras
TensorFlow
SpaCy

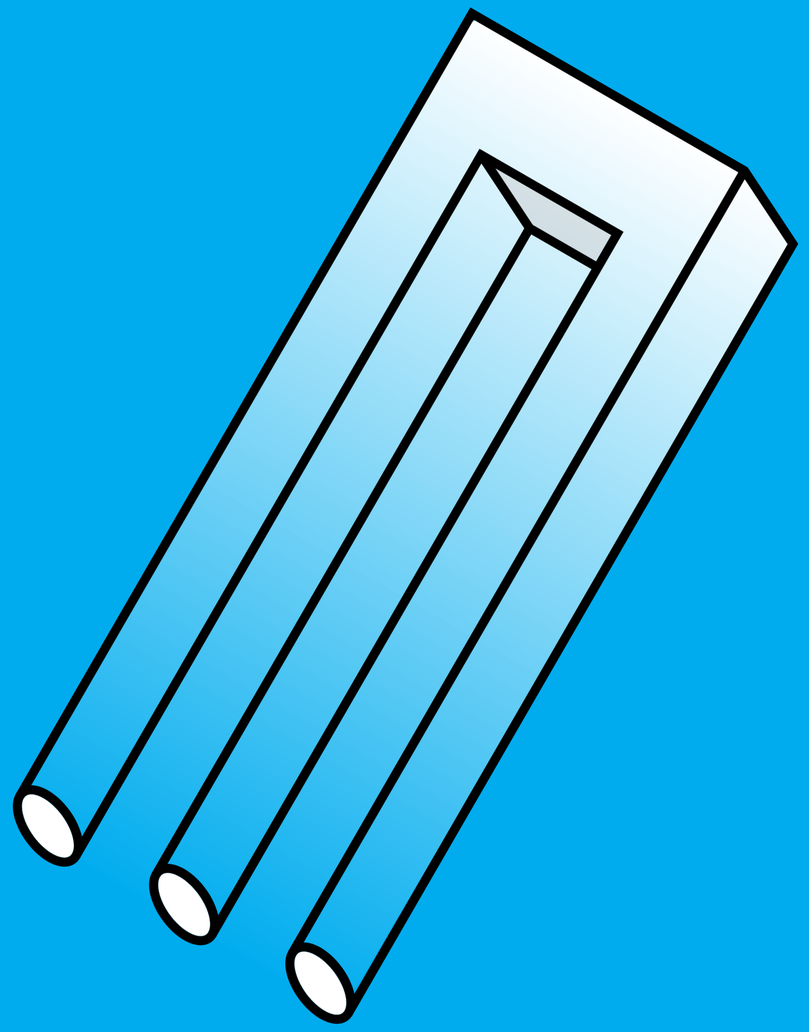


EXAMPLE:
Hate Speech Detection

BEST PRACTICES



DEFINE
the problem



task

data

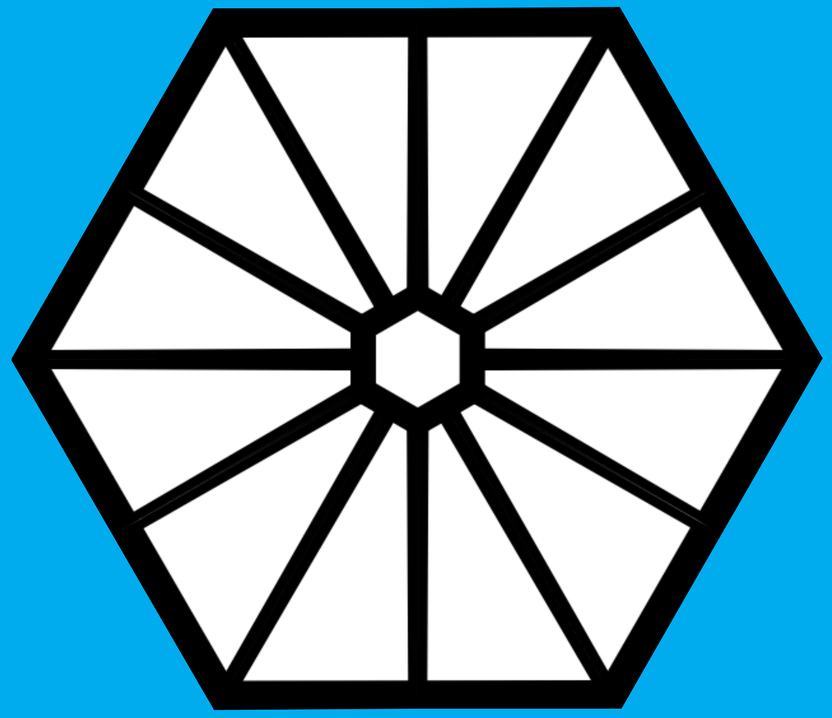
metric

assumptions

motivation

INSPECT
the data

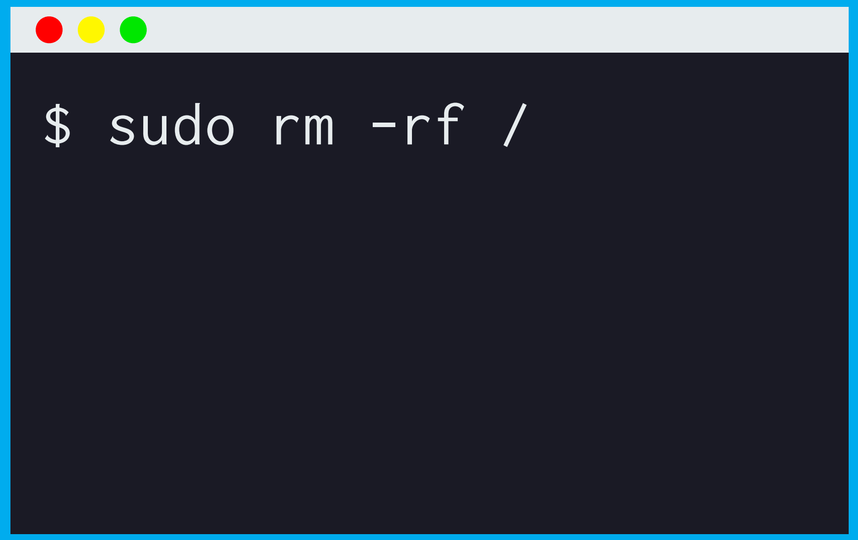
noise
order
imbalance
sparsity
encoding



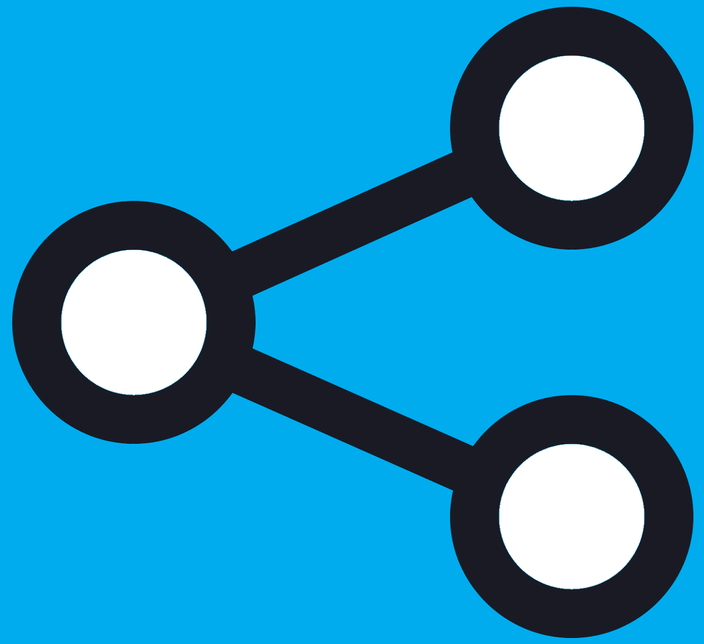
AVOID reinventing the wheel

CHOOSE
the simplest tool possible

OPEN
the black box



MASTER
the terminal



GET

your work out there

THANK YOU!

@sebadzia

IMAGES

Title image (Flickr Commons)

Vault Boy in a suit (Nukapedia, CC-BY-SA)

Twitter logo (Flaticon free license)

Banana clipart (CC0 1.0)

Pretzel clipart (Freepik free license)

Vault Boy with a wrench (Nukapedia, CC-BY-SA)

Tool icon (Flaticon free license)

Gear icon (public domain)

Jane Austen silhouette (Wikimedia commons)

Pride and Prejudice (public domain)

Shakespeare (CC BY 3.0)

Bison (CC0 1.0)

Thumb up Vault Boy (Nukapedia, CC-BY-SA)

Impossible trident (Wikipedia, public domain)

Black box (CC0 1.0)

Share icon (CC BY-ND 3.0)

Swiss army knife

Holmes silhouette (public domain)

IPI PAN logo (public domain)

SpaCy logo (CC BY-SA 4.0)