NLP Project Toolbox

October 8, 2021

Agenda

1. Introduction
2. Problem Identification
3. Data exploration
4. Feature Engineering
5. Modelling
6. Conclusion

Introduction

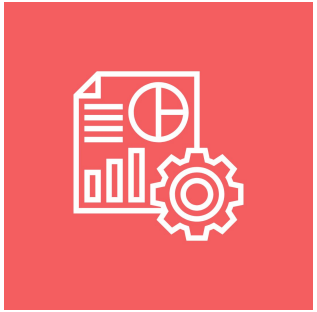
About Me

- My ML journey started in Marketing - I am now an MSc. Social Statistics candidate studying the Kenyan informal sector
- I love learning - I enjoy being either the teacher or the student
- Life can be an exciting adventure if you make it one - I hike and cycle

The NLP project toolbox

How is stuff made?

Get a tool



Use it to build something

Share the thing with loads of people to benefit from it

What is NLP?

Natural Language Processing
is a branch of Machine Learning
concerned with teaching computers
how to recognize patterns
in communications data
generated by humans
and/or use patterns learned
to generate responses for human audiences.

What is the NLP toolbox?

The NLP toolbox is a collection of concepts, tools and ideas available for building applications that can handle real-world challenges around understanding content from all over the world.

It can be used for any NLP task such as sentiment analysis, translation, transcription, audio synthesis etc.

The NLP conceptual toolbox

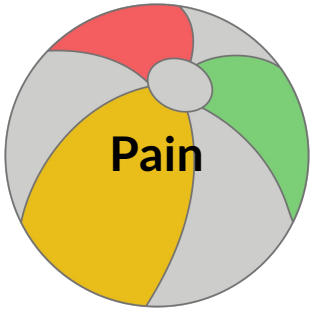
Problem

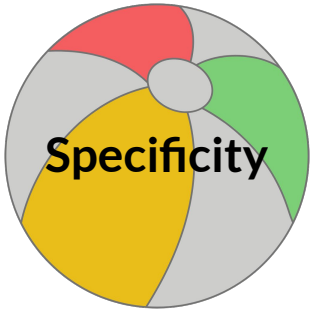
Exploration

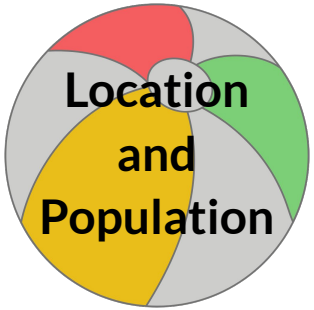
Shaping

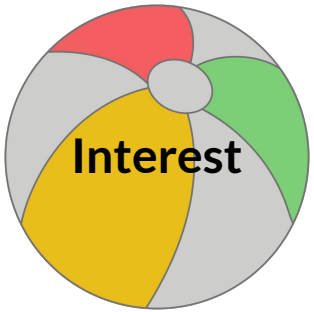
Experimentation

Problem Definition



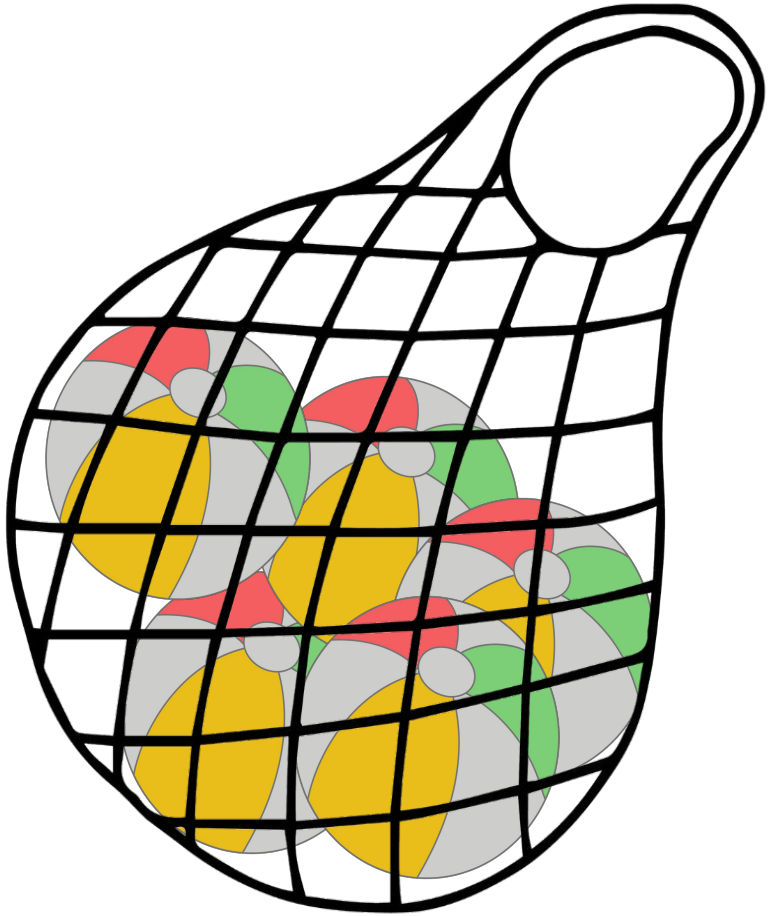






Problem Definition

- The problem has to address each of these 5 areas
- This makes it easier to pick a problem that you can actually solve



The NLP conceptual toolbox

Problem

Exploration

Shaping

Experimentation

Data Exploration

- What languages are spoken and is there slang or code-switching?
- What are the key topics and ideas?
- Which people/institutions that are deemed influential by the speakers/authors?

The NLP conceptual toolbox

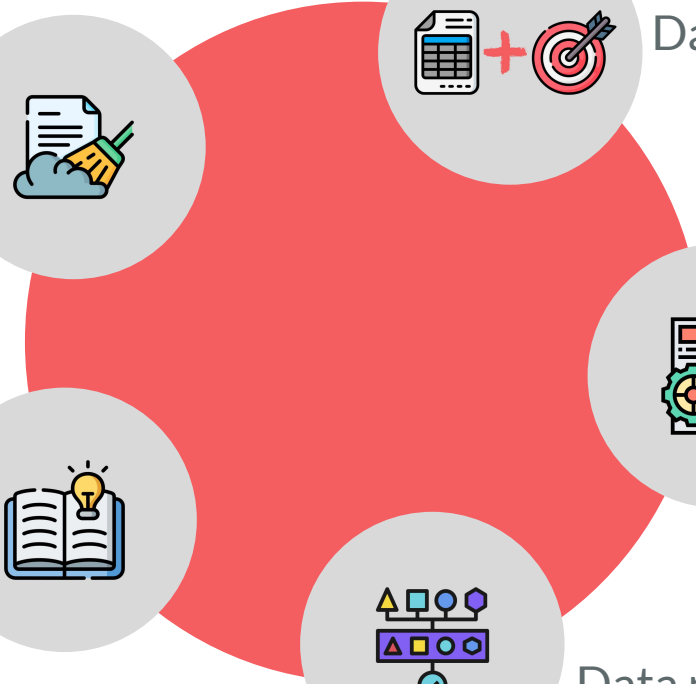
Problem

Exploration

Shaping

Experimentation

Feature Engineering



Data relevancy check

Data cleaning

Data validation

Domain knowledge:
Linguistics, Anthropology

Data pipelining

Pics c/o Freepik and Eucalyp and Smashicons

The NLP conceptual toolbox

Problem

Exploration

Shaping

Experimentation

Model Fitting & Evaluation

Algorithms

Corpus representation

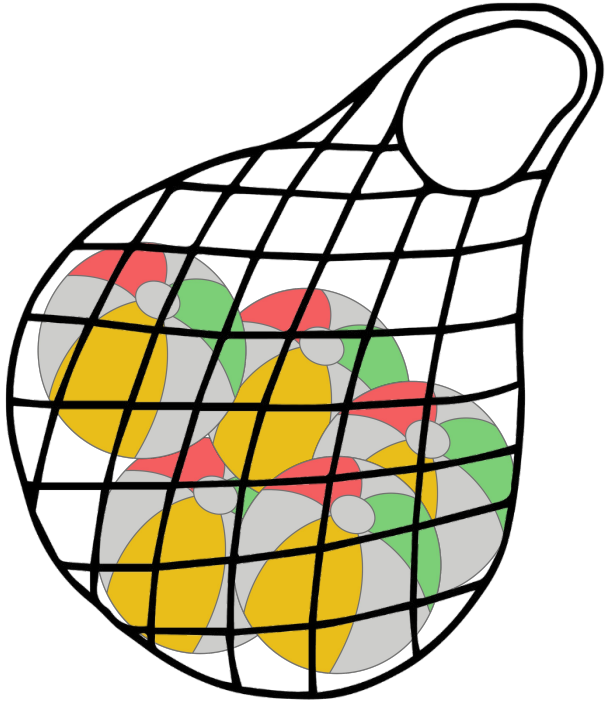
Modelling metrics

Modelling insights

Problem Identification

What is Problem Identification?

Problem Identification is a process of recognizing valuable problems that are hard for most people but easy to solve when technology and creativity are combined.



Problem

Size of Data

Structure of Data

Data Pain

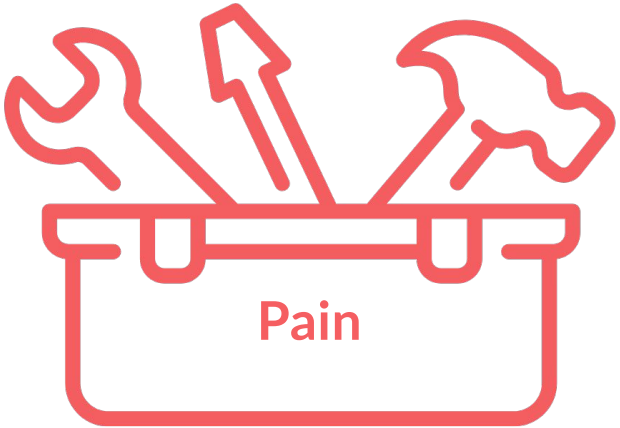
Number of Languages

Structure of Language

Number of Dialects

Short forms vs. misspellings

Linguistics Pain



Problem

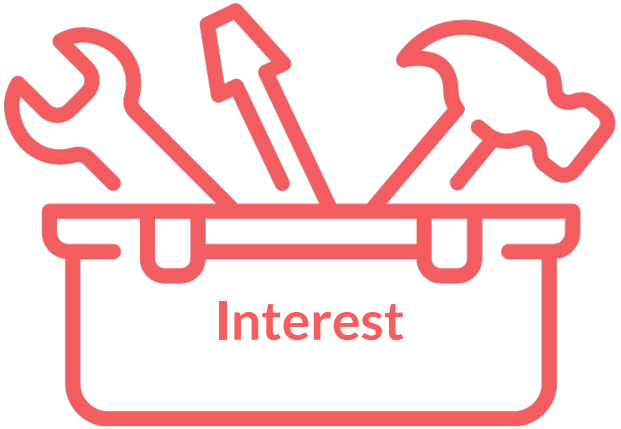
What do you read?

What do you watch?

What do you listen to?

What places do you want to travel to?

Who inspires you?



Problem

Data access credentials

Data privacy

Creator consent

Computational resources

Problem

Where is the creator located?

What is the creator's culture of origin?

What is the creator's current culture?

Problem

Generalizable results

Use cases

Project goals and objectives

Data exploration

What is Data Exploration?

Data Exploration

is a process

of analyzing the data

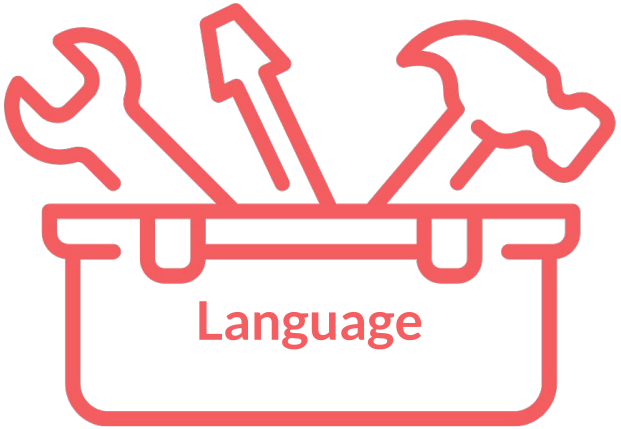
to find patterns within it and

summarize key characteristics

of the data.

Exploration

- Direction of writing
- Type of script
- Code-switching
- Number of languages & dialects
- Summary Statistics for Words & Categories
- Stylistic devices



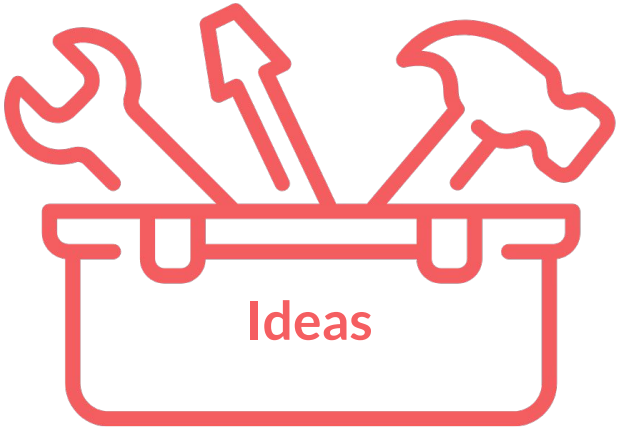
Exploration

Topic indicators

Word frequency

Word distribution

Group By operations & Contingency Tables



Exploration

Cross-referencing

Named Entity Recognition Frequency

Feature Engineering

What is Feature Engineering?

Feature Engineering

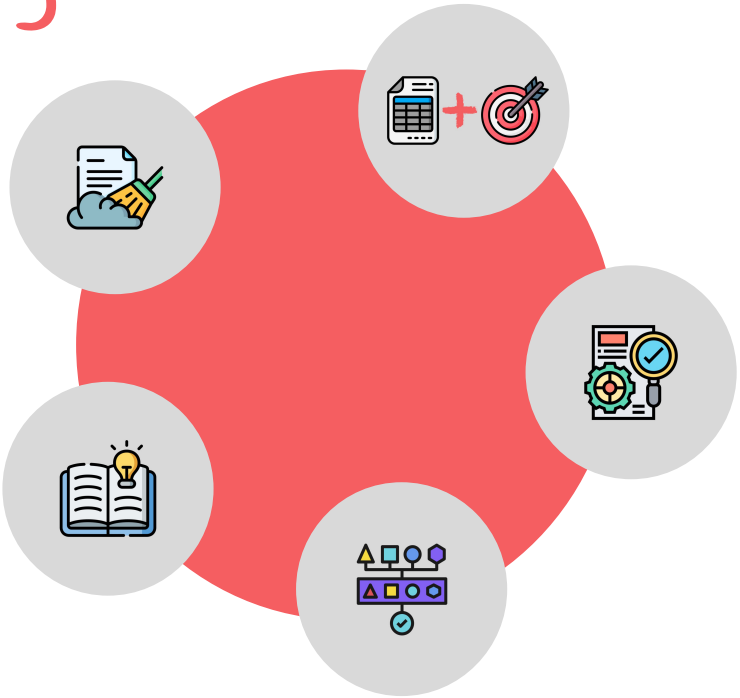
is a process

of transforming the data

to make it easier for the computer to

“understand” and produce relevant results

during modelling

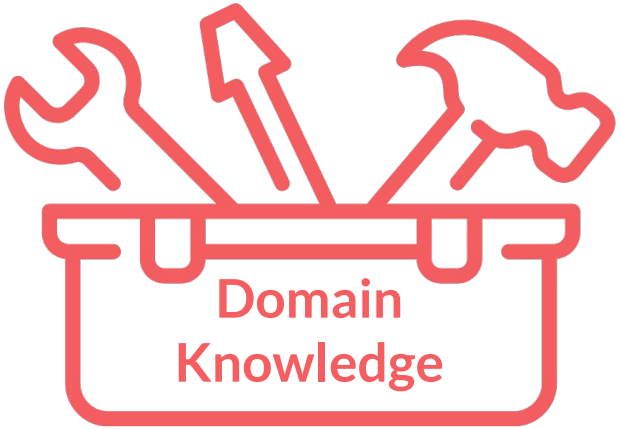


Shaping

- Stemming & Lemmatization
- Data privacy
- N-grams
- Tokenization
- Regex patterns
- Encoding

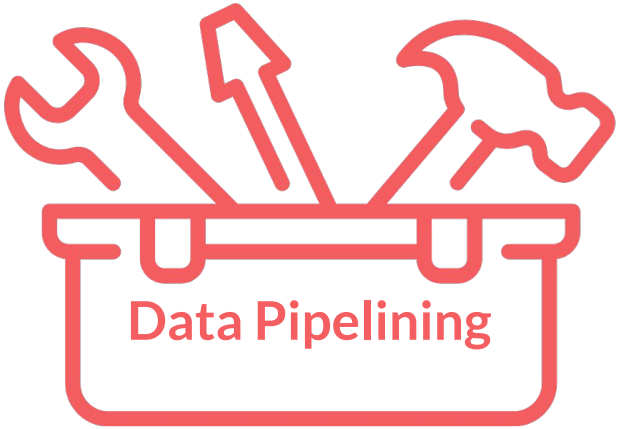
Shaping

- Academic Papers
- Subject Matter Expert
- Organizational History
- Language Speakers



Shaping

- Security
- Local Operating System
- Visualization Tools
- Cloud environment
- Data Lineage
- Jupyter vs. R Shiny



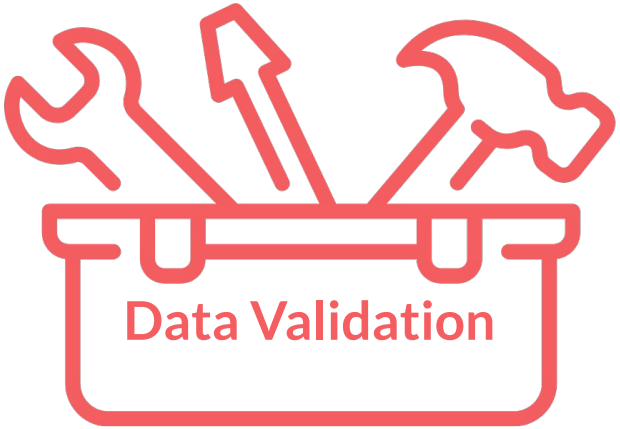
Shaping

Graphs

Data Exploration Results

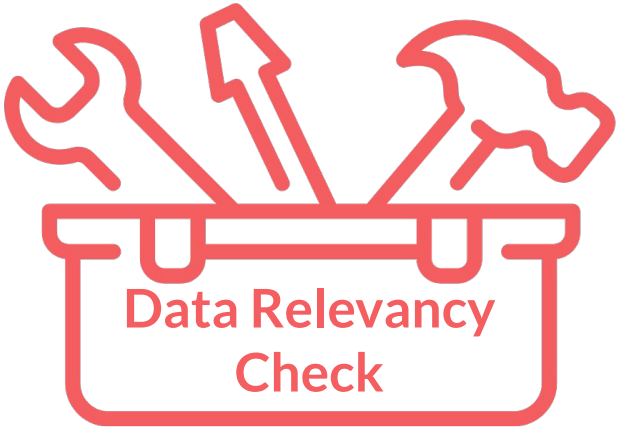
Algorithms

Project Objectives



Shaping

- Data Cleaning Insights
- Data Validation Insights
- Stakeholder Feedback
- Project Objectives



Modelling

What is Model Fitting & Evaluation?

Modelling

is a process

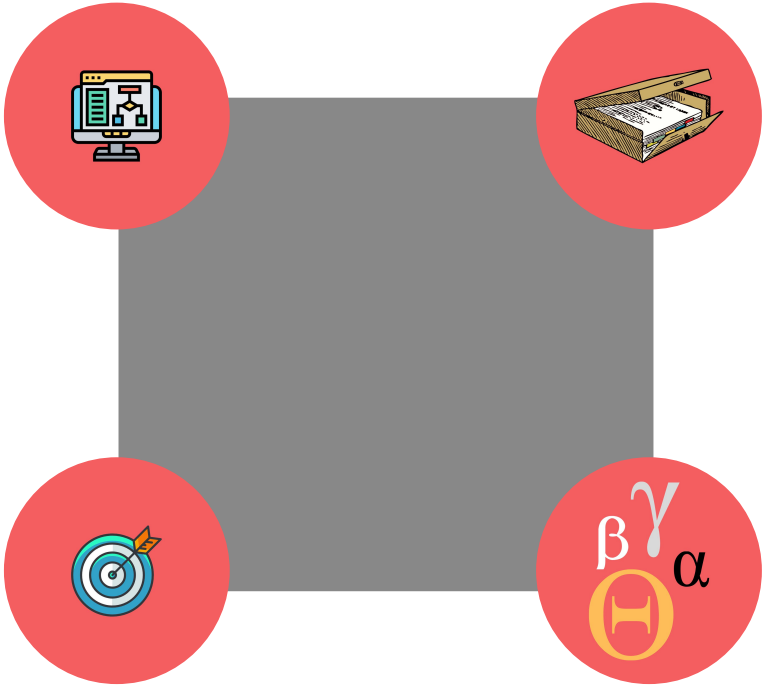
of training the computer to

look for patterns in data

and testing how well the model

predicted patterns in new data

and solved the real-world problem.



Experimentation

One-hot Encoding

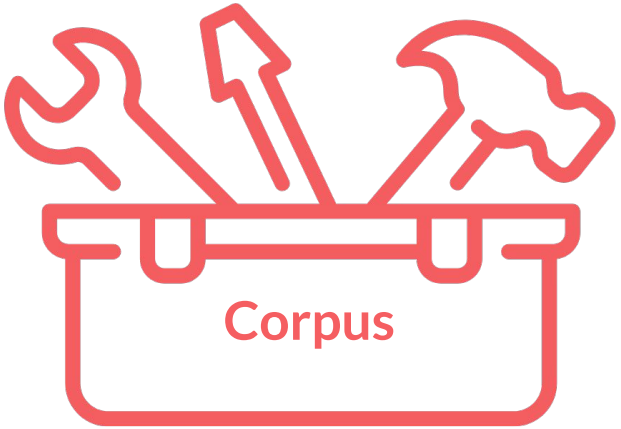
Vectorization

Cleaned Audio

Tokenization

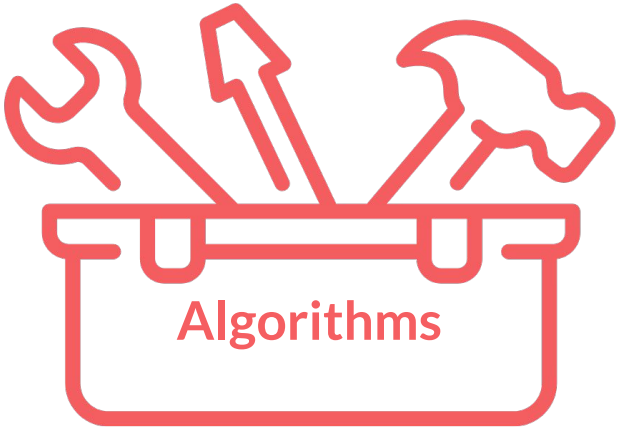
Cleaned Text

Cleaned Video



Experimentation

- Under The Hood Mathematics
- Process Visualization
- Supervised vs. Unsupervised
- Error Handling & Optimization



Experimentation

F1 score, Accuracy & Precision

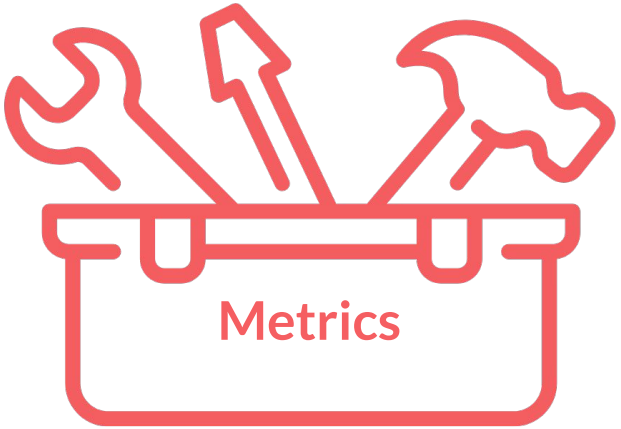
Mathematical Distances

Cosine

Error Analysis Results

Sample Correctness

Project Goal & Objectives



Experimentation

- Model Statistical Significance
- Parameter Interpretation
- Parameter Statistical Significance
- Unexpected Results