# Generative AI Pipeline

1. Data Acquisition

* Available Data(csv,text,pdf, docx,xlsx or others)
* Other Data (DB, internet, API, web scrapping)
* No Data (create your own data) 🡪 LLM to generate Data  
  Note: if you have less data then you perform data augmentation,  
  ex: Replace with synonyms (I am Data Scientist 🡪 I am AI engineer)
* **BIGRAM FLIP**I am Bappy 🡪 Bappy is my name
* **BACK TRANSLATE**
* **ADDITIONAL DATA/NOISE**I am a data Scientist. I love my job.

1. Data Preparation/ Preprocessing

* CLEANUP: HTML,EMOJI, SPELLING CORRECTION
* BASIC PRE-PROCESSING:  
  Tokenization (a) SENTENCE b) WORD
* OPTIONAL PREPROCESSING –( Stop word removal,  
   steaming(less used today), Lemmatization(more used)  
  **steaming 🡪 play, played and playing 🡪 root form Play**
* Lemmatization
* PUNCTUATION REMOVAL
* LOWER CASE
* LANGUAGE DETECTION

ADVANCE PRE\_PROCESSING

* Part of Speech Tagging
* Parsing
* Coreference resolution

1. Feature Engineering

TEXT VECTORIZATION ( TFIDF, Bag of words, word2vec, one hot encoding, transformers models)

1. Modeling

Choose between open source and paid one

1. Evaluation

Intrinsive Evaluation (GEN AI Engineer)

Extrinsic Evaluation – After doing the deployment

1. Deployment
2. Monitoring and model updating/retraining

## Text Representation

1. What is feature extraction from text/images? Edges patterns
2. Why we need it? Numbers or vectors
3. Why it is so difficult ? (Test data)
4. What is the core idea?
5. Some techniques?

Example audio

A diagram of a frequency response

AI-generated content may be incorrect.

Audio and videos can have fixed data as table but for text data is very variable (that is input size is variable).

Ways to solve it

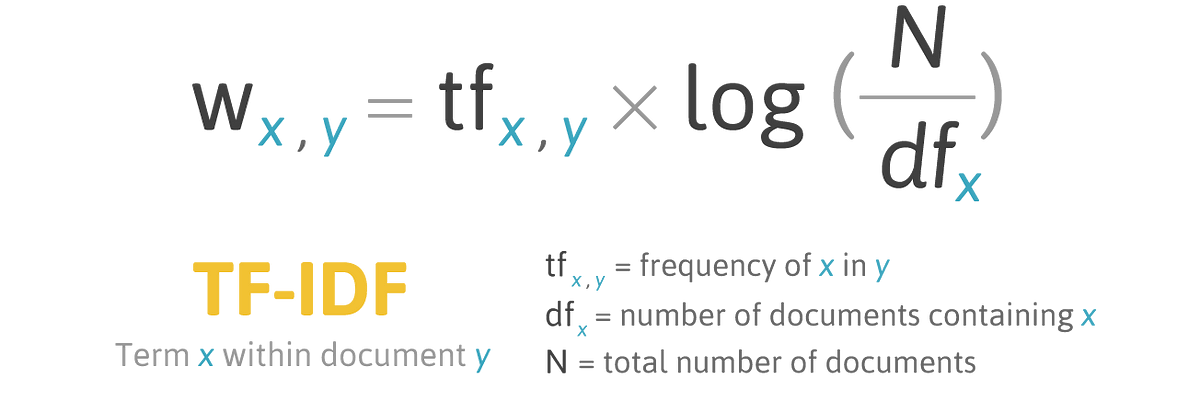
1. One- hot encoding
2. Bag of words

One-Hot Encoding 🡪 Drawbacks 🡪 Sparcity 🡪 No fixed size 🡪 out of vocabulary issue 🡪 Does not capture Semantic meaning 🡪 Don’t follow

Bag of Word  
Other methods TFIDF, Word2Vec, And Transformer models

N-Gram- It is a concept were you consider N-consecutive elements as 1 unit/token

TF-IDF- works on one statistical Equation



TF-IDF🡪 Drawback 🡪 Cannot contain semantic information

**Word2Vec**

Word2Vec uses 2 kinds of architecture CBOW and SKIPGRAM

