Natural Language Processing (NLP)

1. Search Engine

Data crawling (Nudge)
Spider, Robot.txt
DFS → Unstructured

Preprocessing
- Tags, - Removed
- Case -
- Special character
- Punction
- Stop words
- emoji
- emoticons
- Lematization, Stemming
- Tokenization

Notebook Function Framework

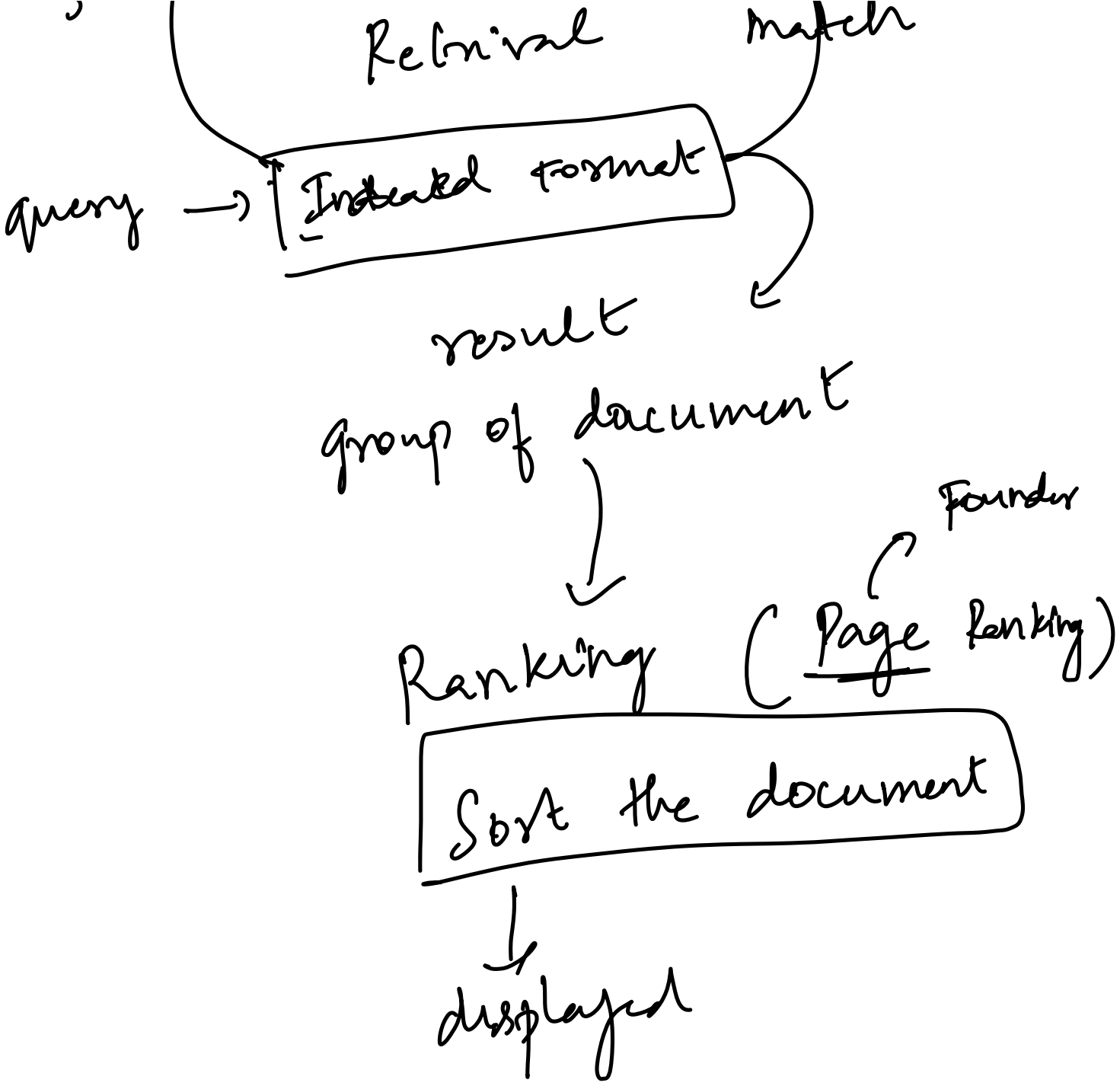
Structured Format

Indexing
- Unique word
- How many
- where it comes

n-grams
uni
bi
tri

NoSql (Cassandra, Hbase) RDMS

Scrapy


Retrival
match
query
Indexed format
result
group of document
Ranking
Founder
(Page Ranking)
Sort the document
displayed

2. Pre processing :

- Textblob
- NLTK
- Spacy
- Gensim
- Beautiful Soup

3. Similarity

Euclidean distance

flaw → Text

↳ magnitute

↳ match is having preference with Document word count

Cosine similarity :

Angle

angle 0° ⇒ 1 ⇒ 100% similar

angle 90° ⇒ 0 ⇒ 0% similar

4. Recommendation Engine

Content based Recommendation engine

Collaborative Filtering
- Item based
- User based

→ Apriori Algorithm

5. Text to Numbers.

1. One hot
2. Bag of Words
3. TFIDF
4. Word Embedding

One hot
1. No Frequency of word in Documents corpus
2. No context
3. No order

Bag of words:
1. No Frequency of word in corpus
2. No Context
3. No order

TF IDF:

1. No context
2. No order

Word Embedding

Word2Vec

Skipgram

Continuous Bag of words (CBOW)

disadvantage → gensim → Train

1. Static embedding

Word2Vec

Glove

FastText → HW

6. Text Classification

Text to Number

ML model

DL model

LSTM algo

CNN algorithm

7. Sentiment Analysis

Lexicon :

1. Textblob
2. Vader
3. Afinn

Text Classification

1. ML
2. DL
3. LSTM
4. Pretrained

8. Text Clustering

K-means → similar bucket

9. Topic Modelling

LDA → probabilistic Model

10. Seq2Seq

Encoder – Decoder

Machine Translation App

11. # Transformer :

Drawback of Seq2Seq :

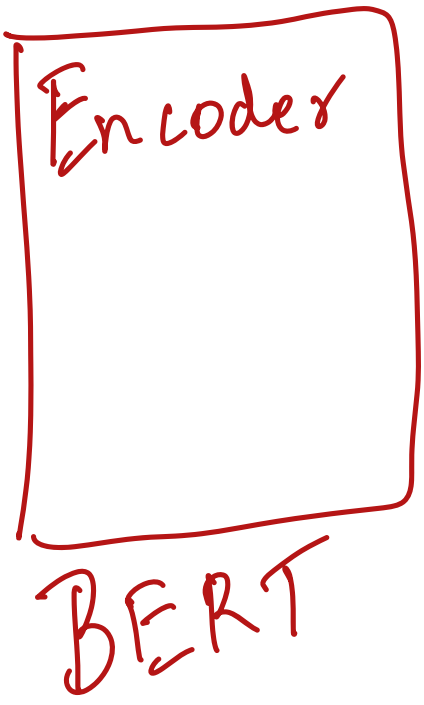
1. Input is taken Token by Token
2. The Decoder can only start after encoder completes the whole document
   The encoder takes time to create the Context vector based on the size of document
3. If we have large token size Context vector can miss the context of whole
4. Static embedding for token
   bank, bank
   track, track

Transformer architecture

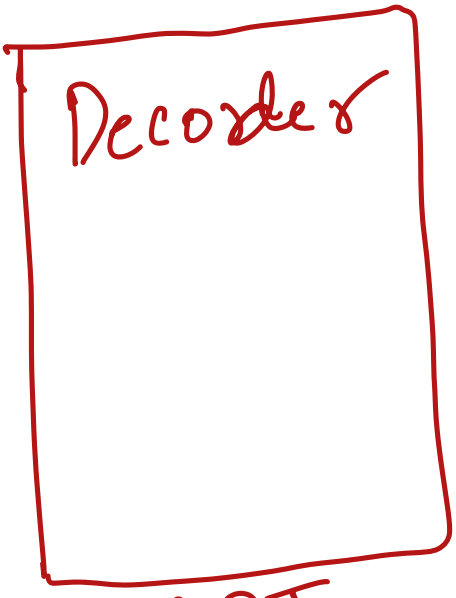
Attention is all you need

↳ Google Researcher

↳ Revolutionized



BERT



GPT

1 Decoder

GPT-1

GPT-2

GPT-3, 3.5

Creating the GPTs' model

Prompt Engineer

Langchain
RAG
Lama idex

Vector Stage
pinecone DB

MoE → Mixture of experts
Mistral

Lang Graph

Bedrock Agents

Tools to use on Agents

Text

Olama

Studs Agents

MCP

Optional based on time

Image Generation

GAN