NLP Pipeline:

1. Text Preprocessing:
2. Text Cleaning:
3. Lowercasing
4. Special character removal
5. Punctuation removal etc.
6. Tokenization on cleaned text
7. Normalization on Tokens
   1. Stemming
   2. Lemmatization (Please try to find out their difference like which is better in which case)
8. Stop words removal (Unimportant words in our text like a, an, the, preposition etc.)
9. Keyword Extraction
   1. TF-IDF
   2. Gensim
   3. Spacy, NLTK, YAKE, RAKE, KeyBert

Extracted keywords will be used in case of labeling our clustering

1. Featured Engineering/ Feature Extraction (Converting text data to numerical):
2. One Hot Encoding
3. TF-IDF
4. Word Embedding (Word2Vec, Glove, FastText)
5. Sentence Embedding (Averaging the Word2Vec/Glove/FastText generated vector, Transformer based Models)
6. Model Building

Clustering Algorithms:

1. K-Means
2. DBSCAN
3. Mini-Batch K-means etc.
4. Model Evaluation

Resources:

1. <https://www.analyticsvidhya.com/blog/2021/06/pre-processing-of-text-data-in-nlp/>
2. <https://www.analyticsvidhya.com/blog/2021/06/part-4-step-by-step-guide-to-master-natural-language-processing-in-python/>
3. <https://goodboychan.github.io/python/datacamp/natural_language_processing/2020/07/17/02-Text-preprocessing-POS-tagging-and-NER.html>
4. <https://www.analyticsvidhya.com/blog/2022/06/an-end-to-end-guide-on-nlp-pipeline/#:~:text=NLP%20Pipeline%20is%20a%20set,and%20Pipeline%20is%20non%2Dlinear>.
5. <https://towardsdatascience.com/text-preprocessing-in-natural-language-processing-using-python-6113ff5decd8>
6. <https://www.youtube.com/watch?v=R-AG4-qZs1A&list=PLeo1K3hjS3uuvuAXhYjV2lMEShq2UYSwX&ab_channel=codebasics>
7. <https://www.analyticsvidhya.com/blog/2022/01/four-of-the-easiest-and-most-effective-methods-of-keyword-extraction-from-a-single-text-using-python/>
8. <https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04>
9. <https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/?utm_source=blog&utm_medium=top_5_sentence_embedding>