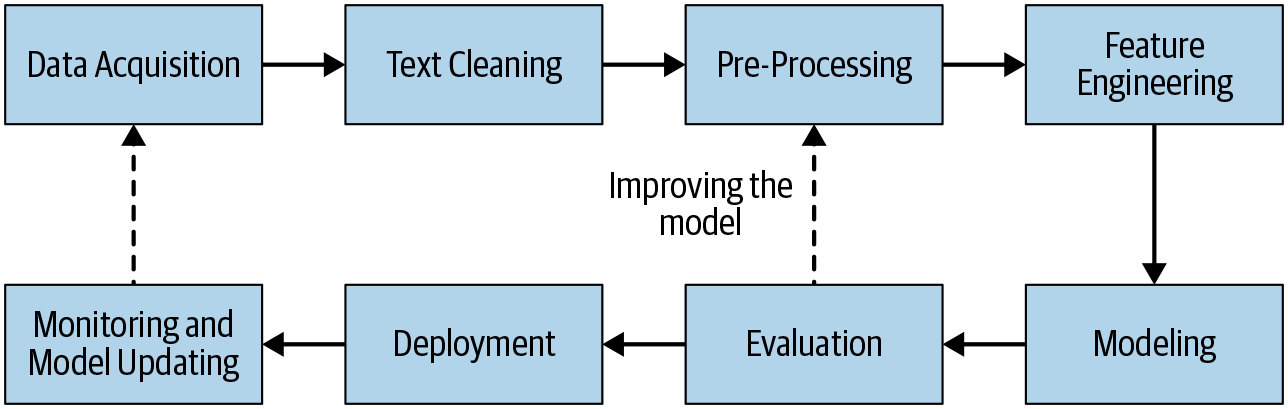
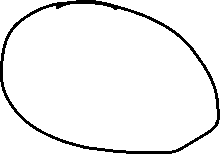
**NLP Pipeline**

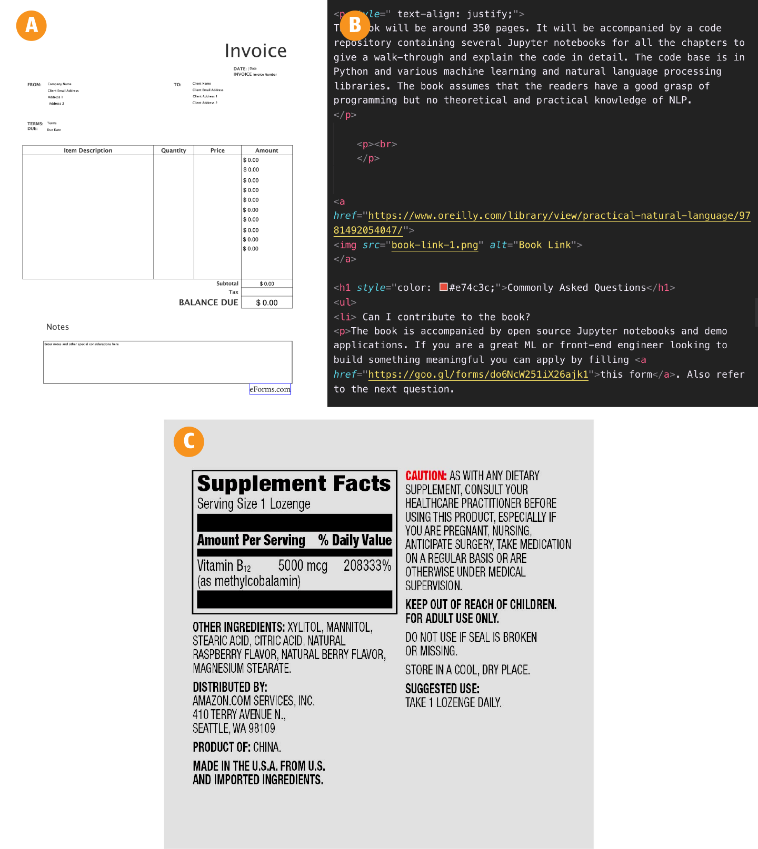




# Data Acquisition

* *Use a public dataset, private datasets(Domain- specific Banks, orgs)*
* *Scrape data*
* *Google Datasets*

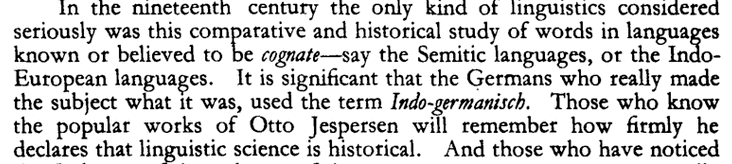
# Text Extraction and Cleanup



parse web pages- <https://www.crummy.com/software/BeautifulSoup/>

**OCR- Optical Character Recognition**

PDF-to-text extraction



# Pre-Processing

String Operations- converting to lower or upper cases, removing punctuations(.,!)

Tokenization- Sentence, Word

stop words- Most frequently used words- a,an,the,his, etc

Stemming- Removing the affixational inflections

Lemmatization- Base form(lemma) with semantic meaning

POS tagging – Noun, Adjective

# Feature Engineering

The goal of feature engineering is to capture the characteristics of the text into a numeric vector that can be understood by the ML algorithms

1. Bag Of Words
2. TF-IDF
3. Word 2Vec

# Modeling-

Machine Learning- Naïve Based, SVM, Logistic Regression

Deep Learning- RNN, LSTM

# Evaluation

Accuracy, Precision, Recall, F1 Score

# Deployment

**Web services**