*Prepared by:* **Tanvir Ishraq Khan**

the supervised training preprocessing pipeline process serially:

## insert **intents** and corresponding **labels** into **python lists**

## *data cleaning*:

#remove punctuations, only keep text (use **regex.** Only keep a-z, A-Z)

# convert to lower case

# remove stop words (use **nltk stopwords**)

## *preprocessing start*:

# stemming (use **nltk stemmer**)

# Fit tokenizer(used **tensorflow tokenizer.**

**Fit** tokenizer on intents text**.** It creates **word\_index** I.e. number index for the vocabulary.)

# tokenization (turn intent **texts\_to\_sequences** of number using vocabulary word\_index)

# pad all intents to same length (use **tensorflow pad\_sequences**)

-- vectorization using tensorflow complete! --

Alternative option for vecotorization: Sklearn countvectorizer

Predict intent label from intent input text.