# Text Preparation (NLP)

## Lecture 13

## Introduction


### 1) Lowercasing

→ Whenever you get text, in first step you bring it to lower case e.g.

"Condition of the car is very good, provided that the condition of such old cars".

→ Python language is case-sensitive

```
df[...].lower()
```

### 2). Remove Unnecessary (e.g. HTML tags, punctuation).

→ By using regex (import re)

→ Online tools available

### 3). Remove URL's

→ if not needed.

https: & http, www:

### 4). Remove Punctuations

" ! $ " % ------

→ We remove these for tokenization.

Buddy! What's up? ⇒ Buddy: Whats up

→ In python, import string or for big data use translate.

## Chat Word treatment

→ New trend, where words are being shortened.

- → Asap
- → Lol
- → Imag
- → FYI

## Spelling Correction

→ By using different libraries such as textblob.

## Removing Stop Words

→ From sentiment analysis & document classification perspective. However you don't do it in Pos tagging.

is, are, the, a, my, yours, which

These sentence formation words are removed.

→ These are libraries, one such is ~~NLTK~~ NLTK.

## Handling Emoji's

- → Either Remove
- → or replace ( 🙂 == happy)

→ Each emoji has unique identifier (U0000024E2)

→ In python import emoji.

## Tokenization

→ is breaching of your text data, into smaller parts.

This is Sparta. You fall here

→ Word level tokenization

[This , is , Sparta ...]

→ Sentence level tokenization

[This is Sparta, You fall Here]

CamScanner

→ I am new in new delhi
1 2 3 4 3 5

but correct is

→ I am new in new delhi
1 2 3 4 5

→ So one should take care of,
Prefix, Suffix, Infix (co-operation), Exceptions (let's u.s.)
$20 3pm

→ In Python
1). split() → on word level
split(.) → on sentence level

Are you going to peshawar? I am not.
→ on full stop whole two sentences

2) Regular Expressions
(import re)

3) Libraries
↳ NLTK

4) The best is Spacy.

Stemming

→ In grammer, inflection is word modification to express different grammatical categories such as tense case, aspect, voice, person numbers, gender & mood.

Eg. Walk Walking
Walked
Walks

do → undoable

→ Stemming is process to reduce inflection in words to their roots. Important in IR (Information Retrieval) systems.

eg. In google search (fish, fishy, fishnet) will show images related to fish.

CamScanner

→ NLTK is porter stemmer.
  ↳ (linguistics make it).

→ Lemmatization
  ↳ if you want faster stemming.
  ↳ if you want to show to your user.

In lemmatization, root will be english word.

story → stori

CamScanner