# Stemming and Lemmitization

## Stemming

Process of reducing infected words to their word stem.

This may not give meaningful word after convert.

For example: history and historical. This will be converted to histori

Finally,final,finalized -> fina

Going, goes, gone -> go

Why?

* Use case like reviews, spam classifier. Word stem is important.

Application

* Sentiment Classifier
* Gmail spam classifier

Problem

* Produced intermediate representation of the word may not have any meaning

Eg: fina, intelligen

## Lemmatization

This is solution to stemming.

Process of converting infected words to meaningful words.

It takes time for processing than stemming.

For example: history, historical -> This is converted to history

Finally,final, finalized -> final

Going,goes,gone -> gone

These application gives response to human that must be meaningful.

Application

* Chatbots
* Question answer application

## Bag Of Words

Nltk count\_vectorizer is used

Stop keywords -> these, a, all, repetitive are stop and are not important in paragraph.

Sent 1 -> He is a good boy

Sent 2 -> She is a good girl

Sent 3 -> Boy and girl are good

NOTE: Lowering sentences is important.

After removing stop words

Sent1 -> good boy

Sent2 -> good girl

Sent3 -> Boy girl good

How do we derive sentence using Bag of word.

Now using histogram

What is histogram -> I will go and see for each and every word and I will just try to find out what is frequency of word.

For example here in all sentence after stop words of every new word

| Words | Frequency |
| --- | --- |
| good | 3 |
| boy | 2 |
| girl | 2 |

Sort in descending order. Once sorted BOW can be applied

Applying BOW

2 types

1. Binary BOW

We either have 1 and 0

Sent1 -> good boy

Sent2 -> good girl

Sent3 -> Boy girl good

Here vectors

|  | f1 | f2 | f3 |  |
| --- | --- | --- | --- | --- |
|  | good | boy | girl | Ooutput feature(dependent feature) |
| Sent1 | 1 | 1 | 0 |  |
| Sent2 | 1 | 0 | 1 |  |
| Sent3 | 1 | 1 | 1 |  |

Disadvantage:

* Values are either 1 and 0. Both are having 1 and 0. They are semantic it is treating all as same. For sentiment analysis we need to have more weightage.

Solution

TFIDF

For small dataset BOW is good but huge dataset needs to have using Word2Vec is best.

## 4. TFID (Term Freuquency - Inverse Document Frequency)

Term Frequency (TF) =

Inverse Document Frequency =log( )

TF \* IDF

After removing stop words

Sent1 -> good boy

Sent2 -> good girl

Sent3 -> Boy girl good

Disadvantage:

* Both Bow and TF-IDF approach semantic information is not stored. TF-IDF gives importance to uncommon words.
* There is definitely chance of over fitting.

Solution:

* Word2Vec model