Report and Analysis:

English:

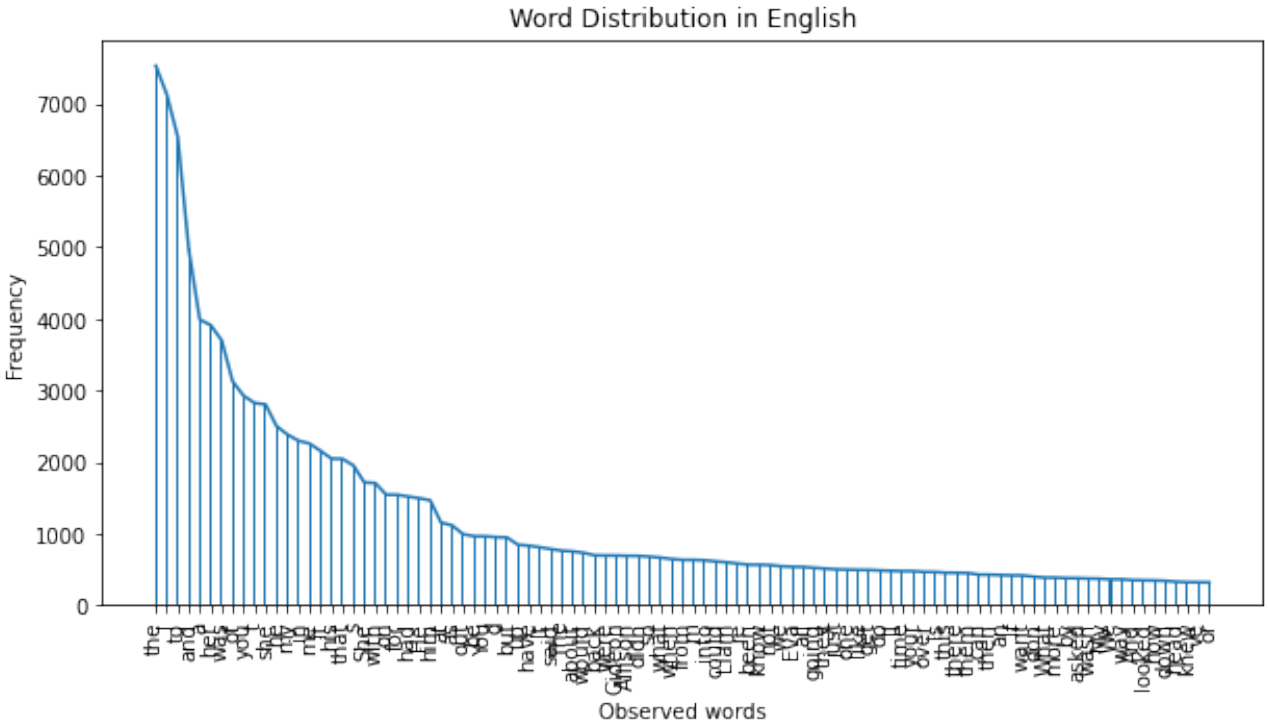
The scraping of the English Text has been done from https://allnovel.net

The found links are html files and scraping has been done via internal links recursively for the entire corpus.

The ‘p’ tags have been useful in scraping the text from the sites and the ‘a’ tags for the ‘href’ links to other pages and books.

Three Graphs have been plotted for the data:

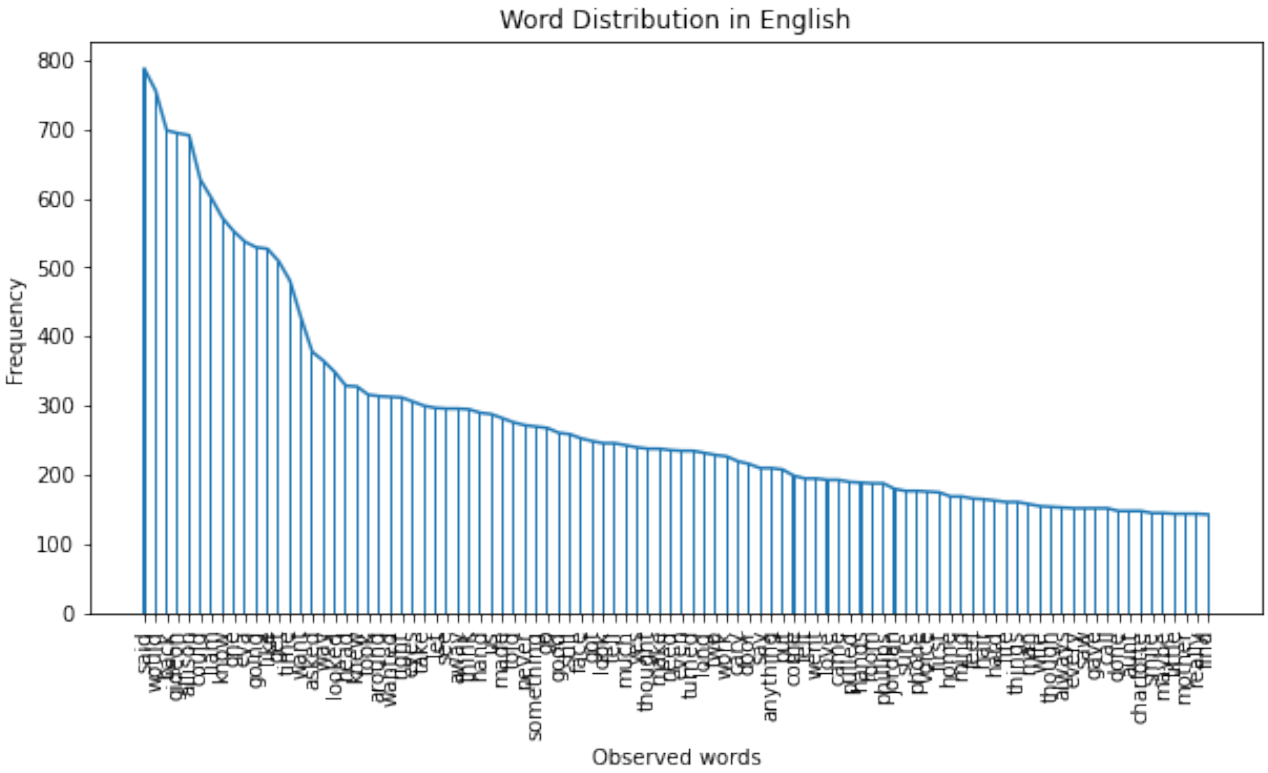
1. Before Removing StopWords:



The most common word in ‘Pre’ Stop word removed text is ‘the’ and it being the definite article, it is expected to Top the list as well

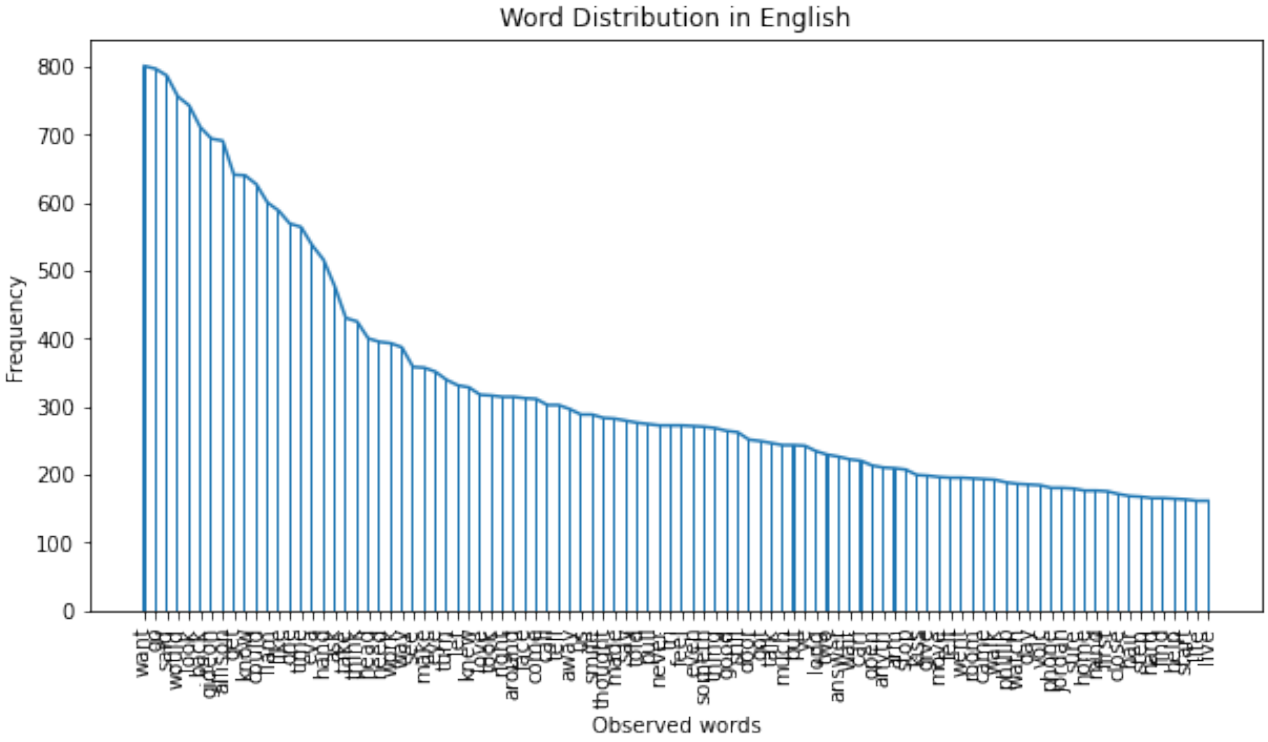
2. After Removing StopWords:

The word ‘want’ has had the highest frequency:



The word with highest frequency is ‘said’ and the Top 100 have been printed.

3. Word Frequency after Stemming:



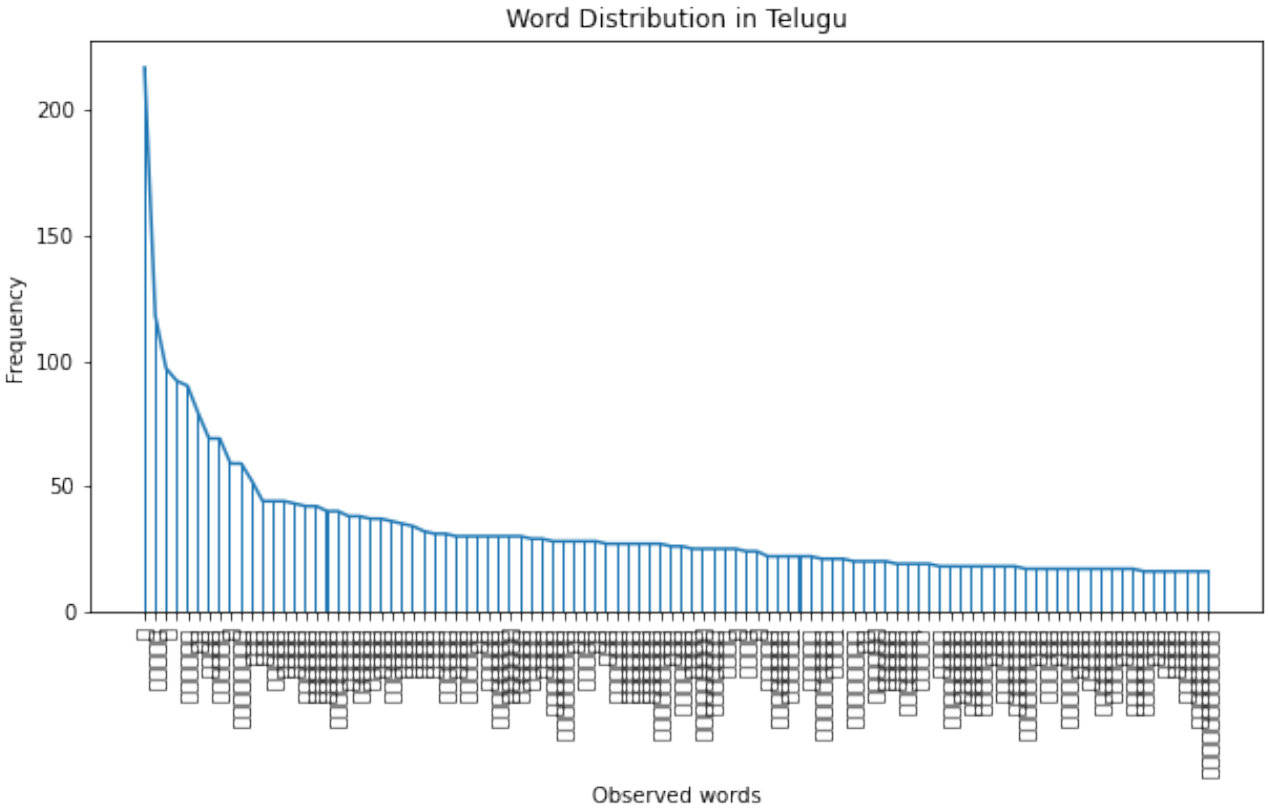
The word ‘want’ has the highest frequency in the data after Stemming.

Telugu:

A similar scraping technique has been used forn the Telugu text as well.
https://sakshi.com News Paper Agenmcy’s website has been used to scrap the Telugu Text.

This time the text not only had Ads and videos but also emojis in it’s corpus, which was eventually filtered for analysis.

Stanza a Stanford initiative has provided the necessary processors for Analysis, in which unfortunately removal of StopWords was not present and other iNLTK libraries proved to be inefficient systems.
However The Graph of the highest frequency of words is given below:



Printing the X – Axis in Telugu wasn't being possible
But the StopWord with Highest Frequency is ఈ.
And the Frequency of a non StopWord is: చేసారు.

Note: Removal of StopWords can be done just creating a list of all StopWords, looping the entire list of words over the loop and all the non-matching words gets printed.

Algorithm:

Words used in the Algorithm are around 100 to make up a proper dataset to analyse, as picking 100 words from 10000 sentence seemed legible.

The words have been sorted in Ascending order of their occuring frequencies and the size of appearance on the output will vary with the number of occurrences, which leads to a word cloud.


gideon
eva
back
one
would
could
like
know
want
said
time
going
get
looked
way
us
head
cary
eyes
see
face
hand
right
look
hard
told
good
well
husband
things
much
got
make
think
even
let
go
mother
asked
still
smile
took
say
wife
never
love
gaze
felt
angus
hands
something
knew
come
anything
around
need
feel
made
long
mouth
pulled
room
turned
away
hair
tell
wanted
take
lips
two
gave
together
taking
door
came
first
went
work
talk
dad
else
thought
stood
mine
cross
breath
man
arash
reached
put
life
yes
needed
left
behind
since
new
feet
onto
corinne
mind
god
better
okay
found
angel
find
held
really
walked
chair
office
fingers
thing
ready
yet
give
side
maybe
last
keep
night
shoulder
everything
saw
place
inside
voice
call
home
nothing
little
moment
living
bed
shot
making
best
words
table
raúl
set
sure
toward
loved
mom
desk
beneath
shook
weekend
dr
petersen
taken
deep
thinking
chris
enough
another
caught
across
monica
pushed
mark
glanced
lucky
arm
getting
wedding
woman
hot
day
open
smiled
photo
might
every
help
ever
fuck
married
body
dark
without
always
phone
sex
legs
family
done
started
slid
arms
sat
kiss
looking
someone