07/10/20

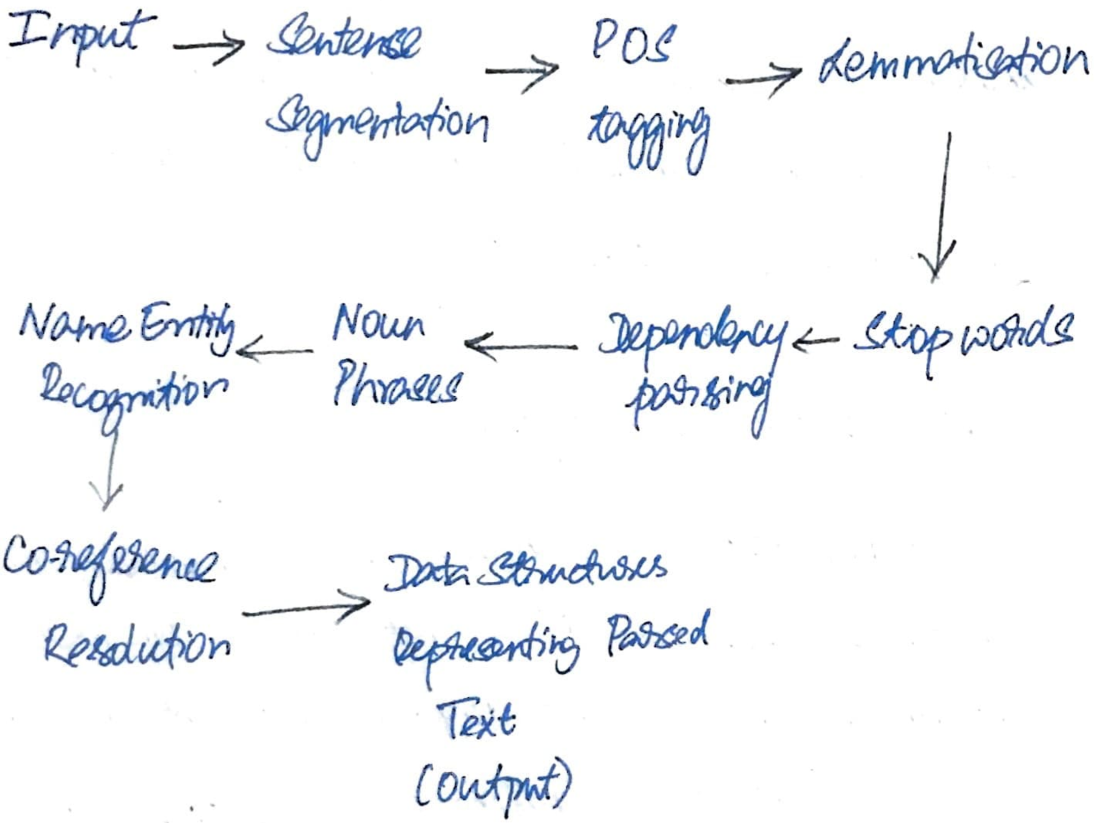
# CSE4022 - Natural Language Processing

18BCE0715
Sanjit CKS
Page: 1

## Digital Assignment

### Developing an NLP Pipeline

Steps involved:



Input text Taken: (Text Document)

~~The~~ Erode is the seventh largest urban agglomeration in Tamil Nadu. It is also the administrative headquarters of Erode district. Erode has a hilly terrain with undulating topography and semi-arid climate. River Kaveri flows through the city and an abundance of limestone is found in its beds. It is located centrally in the south-Indian Peninsula. It is located at 80 km from Coimbatore and 50 km from Tiruppur.

Being extremely popular for the textile industry, a lot of cotton spinning, weaving and knitting industries can be found in the region. Historically, it was part of the Kongu Nadu region in the Sangam age and was ruled by the Cheras who were later ousted by the Pandyas in 590 CE. It was later a prominent British trading point till independance was gained in 1947.

## Step 1: Sentence segmentation

The first step in the pipeline is to breakdown the text into individual sentences. Each unit can be imagined to be a separate idea. This dividing of text into meaningful units is achieved by using punctuations (?, ., !) as delimiters in English. With complex text processing techniques this can be achieved with high accuracy and efficiency.

By breaking down the input text we get the following sentences (first 4 are listed).

1. Erode is the seventh largest urban agglomeration in Tamil Nadu

2. It is also the administrative headquarters of the Erode district

3. Erode has a hilly terrain with undulating topography and semi-arid climate.

4. River Kaveri flows through the city and an abundance of limestone is found at its rocks.

## Step 2: Word Tokenisation

Each sentence further has units of meaning, that are words. By breaking down sentences into seperate words (tokens). This is important for classifying and counting them for a particular sentiment. Tokenisation in English is easy because words are separated by spaces. Punctuations will also be tokenised because they contain semantic meaning too.

from the 1st sentence in the previous section, after tokenisation.

[ "Erode", "is", "the", "seventh", "largest", "urban", "agglomeration", "in", "Tamil", "Nadu" ]

## Step 3: Parts-of-Speech Tagging for each Token.

To further understand the meaning of the sentence, we need to know the role each word is playing. Parts-of-speech tagging does this by tagging tokens with grammatical parts of speech labels like – Noun, verb, adverb etc. This is done by either lexical based, rule based, probabilistic or deep learning methods. The deep learning method is statistical and provides maximum efficiency.

After processing the previous sentence & tagging,

Erode, is , the , seventh , largest, urban,

Proper Noun verb Determiner Adjective

agglomeration , in , Tamil, Nadu

noun Preposition Noun

Step 4: Text Lemmatisation

In natural language words appear in different inflections. It helps to map these different inflections to the same root word to make the computer consider them as the same word. Finding the base forms of nouns and verbs could mean removing plurality and tense to get the "lemma". For example 'car' and 'cars' have a very minute difference but the computer classifies them as different words.

By lemmitizing the previous sentence.

Erode is the seventh largest urban agglomeration in Tamil Nadu
( Erode be the seventh largest urban agglomeration in Tamil Nadu)

The 'is' is changed to lemma 'be'. Note that 'be' (a state of being) is the lemmatised word of 'is', 'are', 'been', 'was' etc.

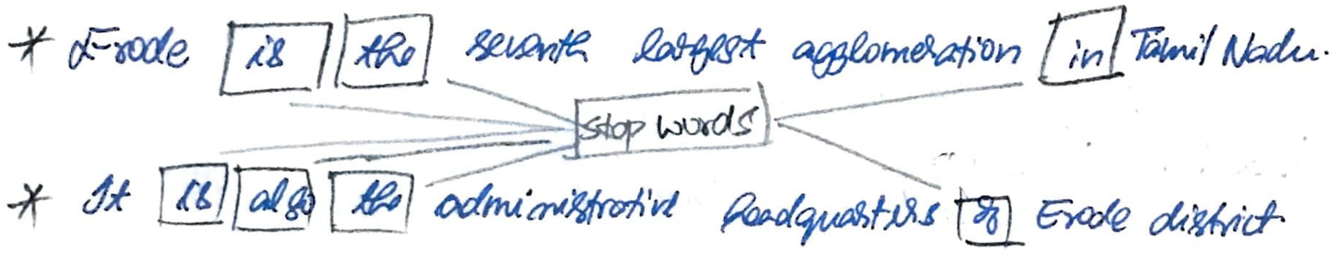
It is the administrative headquarters of Erode district
(It be the administrative headquarters of Erode district)

## Step 5: Identifying stop words

We speak with a lot of repetion and therefore natural languages have a lot of noise in the form of filler words like "and", "the", "a" etc. They occurs in higher frequency and can reduce the efficiency of machine learning models used in the following steps so they can be flagged as STOPWORDS and filtered out before statistical analysis.

A pre written list of stop words are loaded up during parsing to identify them. Depending on the domain, the stop words differ.

By filtering out the previous sentenance.



* Erode is the seventh largest agglomeration in Tamil Nadu.

* It is also the administrative headquarters of Erode district

## Step 6: Dependancy Parsing

In in this step the relationships of words in a sentence are mapped. Related words are mapped hierarchically that results in a tree with root to leaves. The root is the main verb in the sentence.

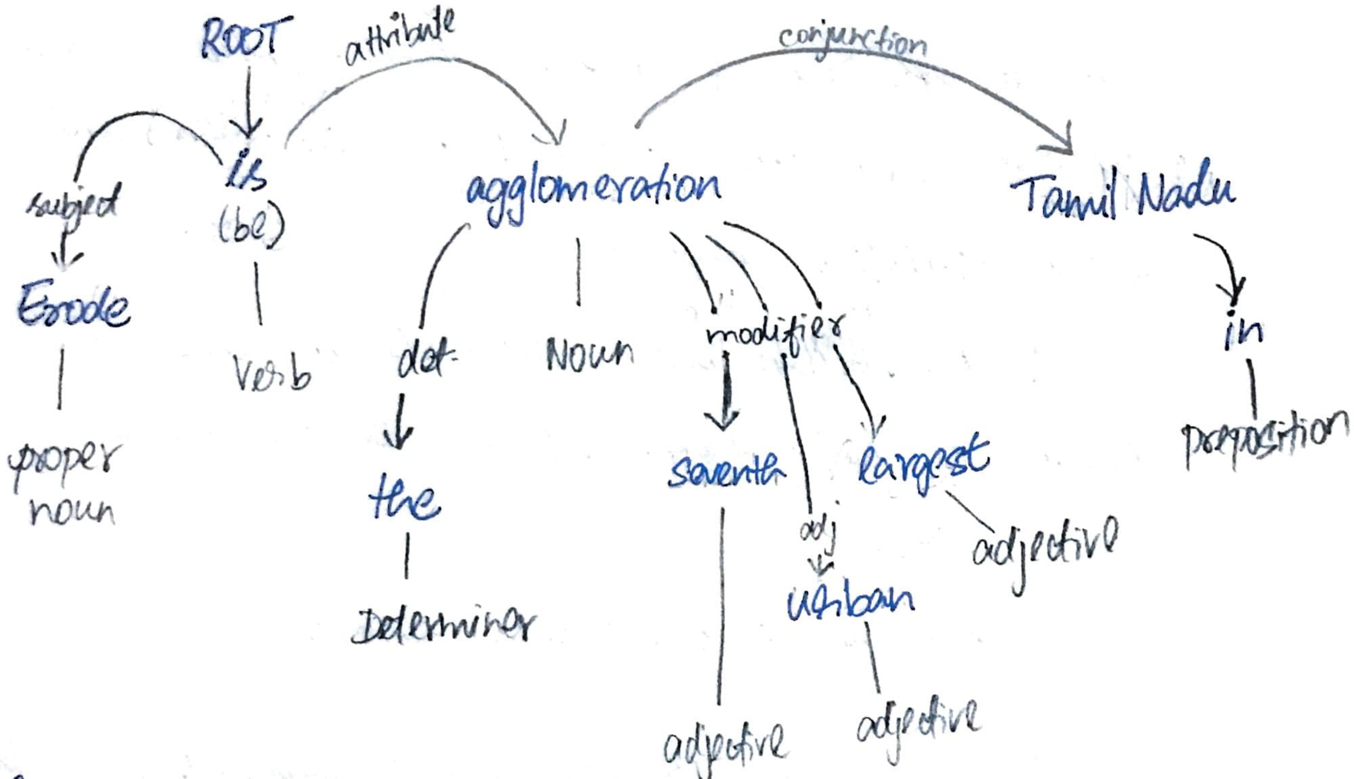
→ Parent word to each word is identified & mapped

→ The type of relationship could also be mapped.

→ Find and group noun phrases.

finding/grouping noun phrases that talk about the same thing, makes statistical further analysis easier.

For the first sentense in the text,



Grouping of Nounds

Tamil Nadu (Tamil + Nadu)

urban agglomeration (urban + agglomeration).

## Step 7: Named Entity Recognition

Some of the words in the processed text contain proper nouns that represent real world entities. In this step of Named Entity Recognition these actual physical real world entities are detected and tagged.

This is achieved by a combination of statistical models and context processing. The NER system detects objects like names of people, companies, physical locations, dates, names of events etc.

From the input text, the first line becomes.

Erode is the seventh largest urban agglomeration in Tamil Nadu.

↳ Physical location

physical location

Other examples include: Alice went to the market.

Cochella was amazing.

Step 8: Coreference Resolution

At this point in the pipeline, most of the semantic information from the sentence itself is obtained. But this brings another problem of pronoun usage in the 2nd, 3rd and following sentences 'It' is used to refer to Erode. This information requires passing accross one sentence.

With co-reference and named entity recognition using deep learning more data can be extracted. From the input text, the first 3 sentences,

Erode is the seventh largest urban agglomeration in Tamil Nadu. It is the administrative headquarters of Erode district. It has a hilly terrain with undulating topography and semi arid climate

It ⟷ Erode

# Implementation and Screenshots

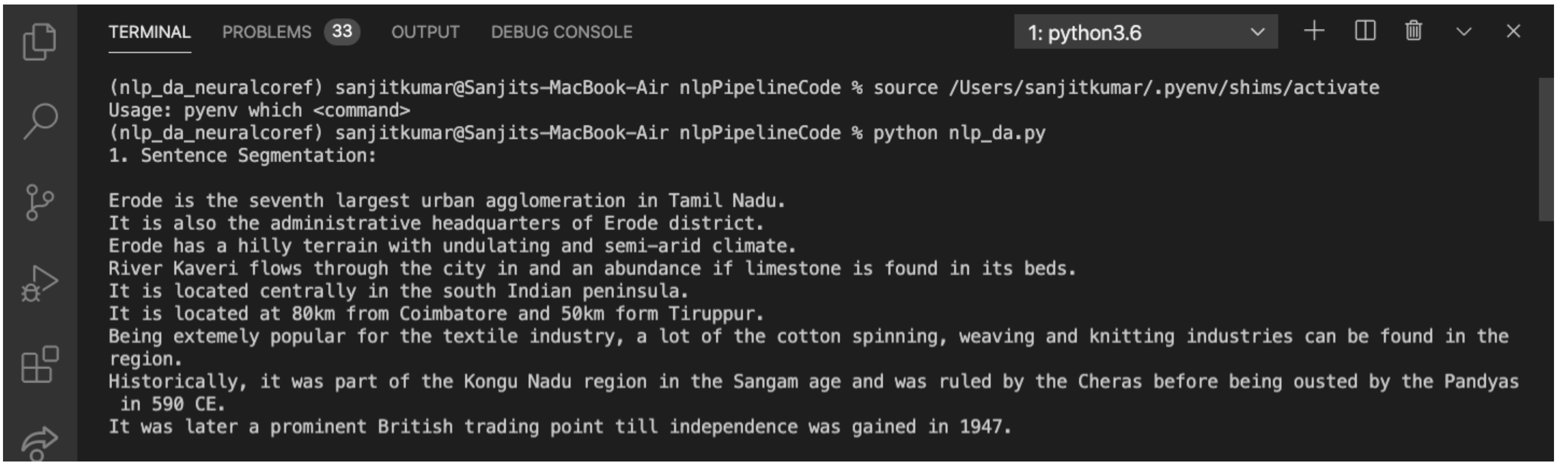
## Source Code (Steps and Numbers in Comments)

```
import spacy
import nltk
from nltk.stem.porter import *
from spacy.lang.en import English

# NLP PIPELINE - STEP WISE - SANJIT C K S - 18BCE0715
# Load the large English NLP model
nlp = spacy.load('en_core_web_lg')

# Input Text
text = u"""Erode is the seventh largest urban agglomeration in Tamil
Nadu. It is also the administrative headquarters of Erode district.
Erode has a hilly terrain with undulating and semi-arid climate. River
Kaveri flows through the city in and an abundance if limestone is found
in its beds. It is located centrally in the south Indian peninsula. It
is located at 80km from Coimbatore and 50km form Tiruppur. Being
extemely popular for the textile industry, a lot of the cotton spinning,
weaving and knitting industries can be found in the region.
Historically, it was part of the Kongu Nadu region in the Sangam age and
was ruled by the Cheras before being ousted by the Pandyas in 590 CE. It
was later a prominent British trading point till independence was gained
in 1947."""

# Parse the text with spaCy. This runs the entire pipeline.
doc = nlp(text)

# Segment Sentences
print("1. Sentence Segmentation:\n")
for sent in doc.sents:
    print(sent)

# Word Tokenization
print("\n\n")
print("2. Word Tokenization:\n")
word_tokens = []
for word in doc:
    word_tokens.append(word.text)
    print(word.text,end=",")

# POS Tagging
print("\n\n")
print("3. POS Tagging:\n")
for word in doc:
    print(word.text,  word.pos_, end=", ")

# Lemmatisation
print("\n\n")
print("4. Lemmatisation:\n")
lemmatised = []
```

```
for word in doc:
    lemmatised.append(word.lemma_)
    print(word.text + '  ===>', word.lemma_)

# Remove Stop Words
print("\n\n")
print("5. Remove Stop Words:\n")
from spacy.lang.en.stop_words import STOP_WORDS
filtered_sentence =[]
for word in word_tokens:
    lexeme = nlp.vocab[word]
    if lexeme.is_stop == False:
        filtered_sentence.append(word)
print("Before Removal: ",word_tokens)
print("After Removal: ",filtered_sentence)

# Dependency Parser
print("\n\n")
print("6. Dependency Parser:\n")
from spacy.pipeline import DependencyParser
from spacy import displacy
displacy.serve(doc, style='dep')
# parser = DependencyParser(nlp.vocab)
# processed = parser(doc)
# print(processed)

# NER
print("\n\n")
print("7. Name Entity Recognition:\n")
for entity in doc.ents:
    print(f"{entity.text} ({entity.label_})")

# Co-reference Resolution
import neuralcoref
print("\n\n")
print("8. Co-reference resolution:\n")
nlp = en_coref_md.load()
doc = nlp(test_sent)
print(doc._.has_coref)
print(doc._.coref_clusters)
```

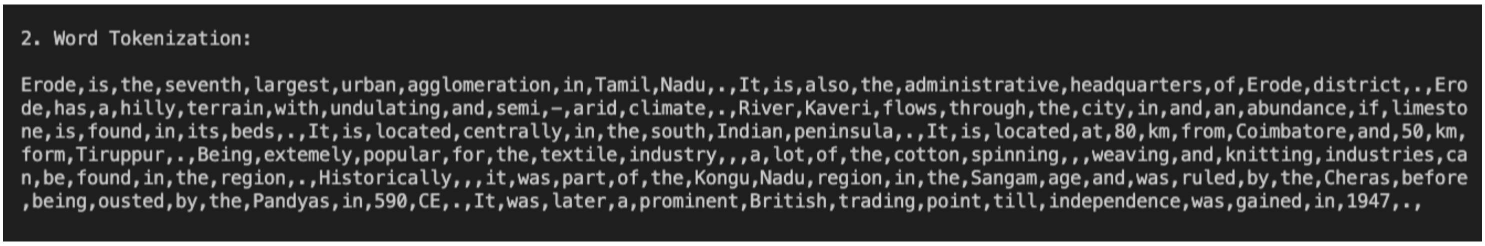
# Output

Part 1 - Sentence Segmentation



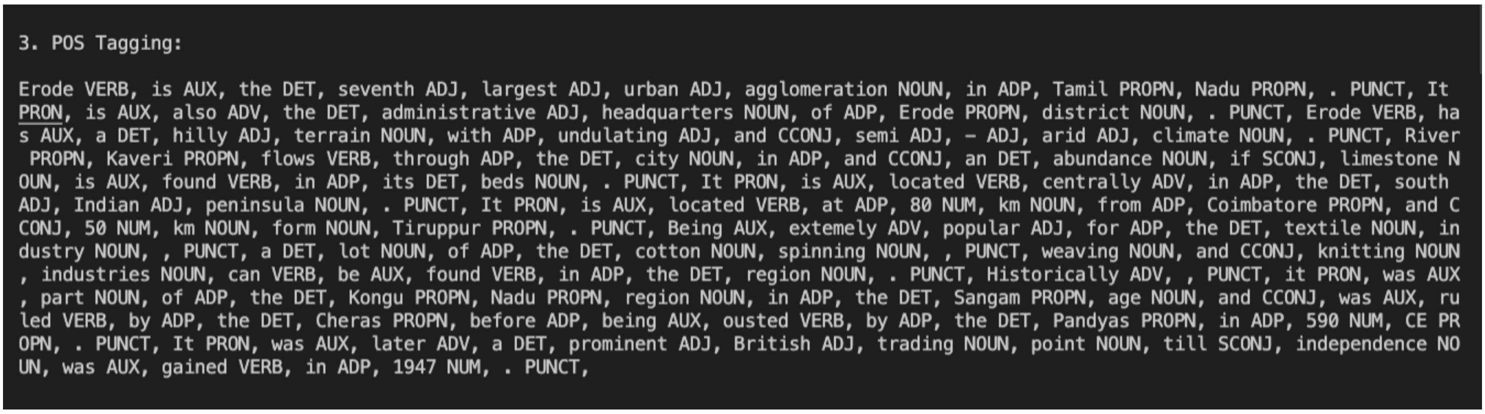
```
(nlp_da_neuralcoref) sanjitkumar@Sanjits-MacBook-Air nlpPipelineCode % source /Users/sanjitkumar/.pyenv/shims/activate
Usage: pyenv which <command>
(nlp_da_neuralcoref) sanjitkumar@Sanjits-MacBook-Air nlpPipelineCode % python nlp_da.py
1. Sentence Segmentation:

Erode is the seventh largest urban agglomeration in Tamil Nadu.
It is also the administrative headquarters of Erode district.
Erode has a hilly terrain with undulating and semi-arid climate.
River Kaveri flows through the city in and an abundance if limestone is found in its beds.
It is located centrally in the south Indian peninsula.
It is located at 80km from Coimbatore and 50km form Tiruppur.
Being extemely popular for the textile industry, a lot of the cotton spinning, weaving and knitting industries can be found in the
region.
Historically, it was part of the Kongu Nadu region in the Sangam age and was ruled by the Cheras before being ousted by the Pandyas
 in 590 CE.
It was later a prominent British trading point till independence was gained in 1947.
```

Part 2 - Word Tokenisation



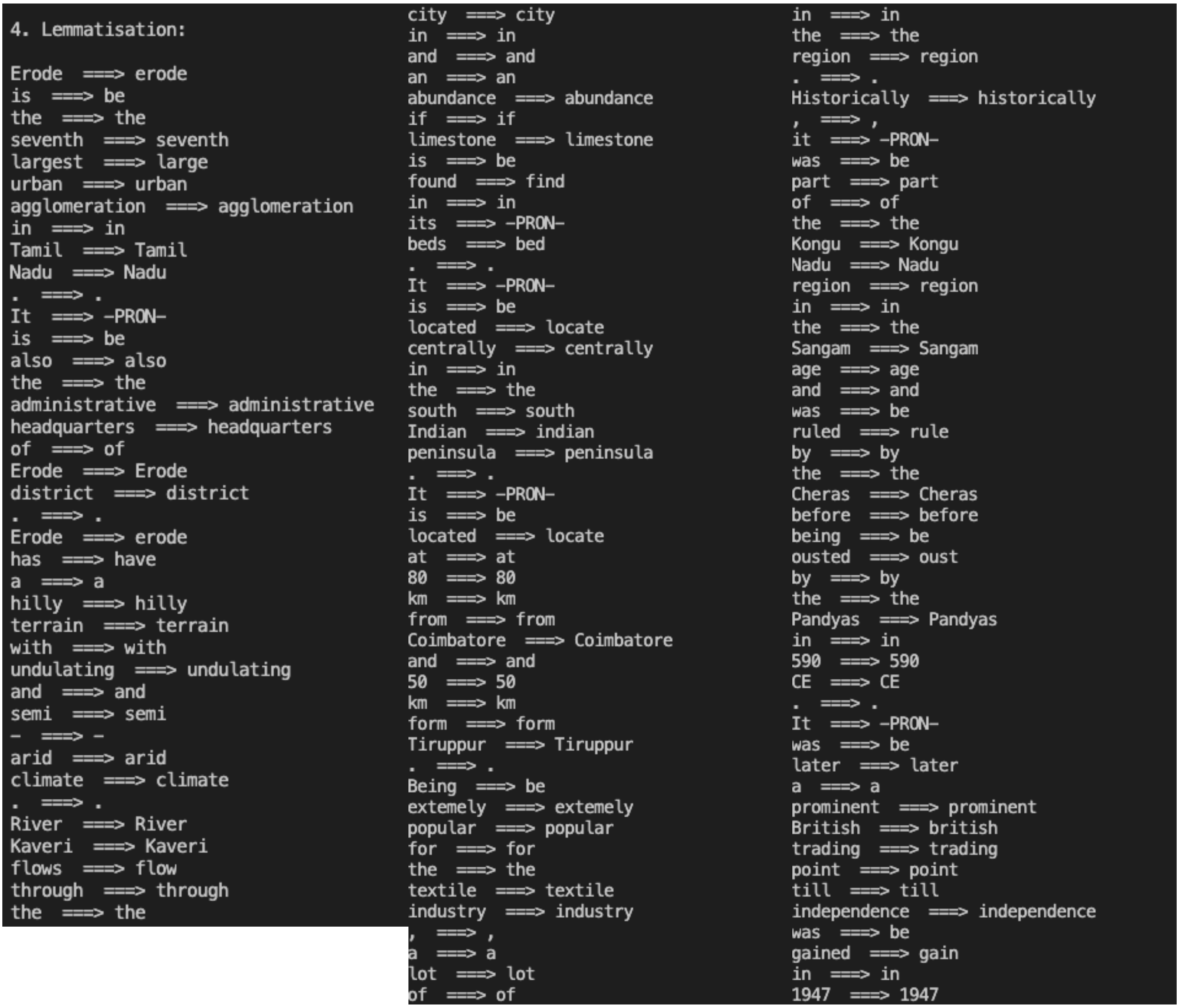
```
2. Word Tokenization:

Erode,is,the,seventh,largest,urban,agglomeration,in,Tamil,Nadu,.,It,is,also,the,administrative,headquarters,of,Erode,district,.,Ero
de,has,a,hilly,terrain,with,undulating,and,semi,-,arid,climate,.,River,Kaveri,flows,through,the,city,in,and,an,abundance,if,limesto
ne,is,found,in,its,beds,.,It,is,located,centrally,in,the,south,Indian,peninsula,.,It,is,located,at,80,km,from,Coimbatore,and,50,km,
form,Tiruppur,.,Being,extemely,popular,for,the,textile,industry,,,a,lot,of,the,cotton,spinning,,,weaving,and,knitting,industries,ca
n,be,found,in,the,region,.,Historically,,,it,was,part,of,the,Kongu,Nadu,region,in,the,Sangam,age,and,was,ruled,by,the,Cheras,before
,being,ousted,by,the,Pandyas,in,590,CE,.,It,was,later,a,prominent,British,trading,point,till,independence,was,gained,in,1947,.,
```

Part 3 - POS Tagging



```
3. POS Tagging:

Erode VERB, is AUX, the DET, seventh ADJ, largest ADJ, urban ADJ, agglomeration NOUN, in ADP, Tamil PROPN, Nadu PROPN, . PUNCT, It
PRON, is AUX, also ADV, the DET, administrative ADJ, headquarters NOUN, of ADP, Erode PROPN, district NOUN, . PUNCT, Erode VERB, ha
s AUX, a DET, hilly ADJ, terrain NOUN, with ADP, undulating ADJ, and CCONJ, semi ADJ, - ADJ, arid ADJ, climate NOUN, . PUNCT, River
 PROPN, Kaveri PROPN, flows VERB, through ADP, the DET, city NOUN, in ADP, and CCONJ, an DET, abundance NOUN, if SCONJ, limestone N
OUN, is AUX, found VERB, in ADP, its DET, beds NOUN, . PUNCT, It PRON, is AUX, located VERB, centrally ADV, in ADP, the DET, south
ADJ, Indian ADJ, peninsula NOUN, . PUNCT, It PRON, is AUX, located VERB, at ADP, 80 NUM, km NOUN, from ADP, Coimbatore PROPN, and C
CONJ, 50 NUM, km NOUN, form NOUN, Tiruppur PROPN, . PUNCT, Being AUX, extemely ADV, popular ADJ, for ADP, the DET, textile NOUN, in
dustry NOUN, , PUNCT, a DET, lot NOUN, of ADP, the DET, cotton NOUN, spinning NOUN, , PUNCT, weaving NOUN, and CCONJ, knitting NOUN
, industries NOUN, can VERB, be AUX, found VERB, in ADP, the DET, region NOUN, . PUNCT, Historically ADV, , PUNCT, it PRON, was AUX
, part NOUN, of ADP, the DET, Kongu PROPN, Nadu PROPN, region NOUN, in ADP, the DET, Sangam PROPN, age NOUN, and CCONJ, was AUX, ru
led VERB, by ADP, the DET, Cheras PROPN, before ADP, being AUX, ousted VERB, by ADP, the DET, Pandyas PROPN, in ADP, 590 NUM, CE PR
OPN, . PUNCT, It PRON, was AUX, later ADV, a DET, prominent ADJ, British ADJ, trading NOUN, point NOUN, till SCONJ, independence NO
UN, was AUX, gained VERB, in ADP, 1947 NUM, . PUNCT,
```

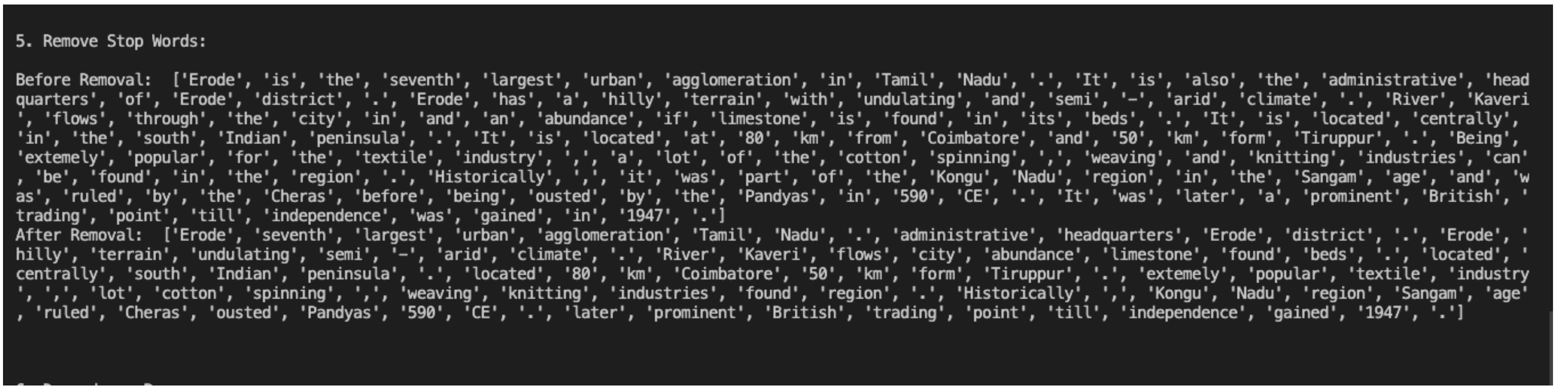
Part 4 - Lemmatisation



```
4. Lemmatisation:

Erode  ===> erode
is  ===> be
the  ===> the
seventh  ===> seventh
largest  ===> large
urban  ===> urban
agglomeration  ===> agglomeration
in  ===> in
Tamil  ===> Tamil
Nadu  ===> Nadu
.  ===> .
It  ===> -PRON-
is  ===> be
also  ===> also
the  ===> the
administrative  ===> administrative
headquarters  ===> headquarters
of  ===> of
Erode  ===> Erode
district  ===> district
.  ===> .
Erode  ===> erode
has  ===> have
a  ===> a
hilly  ===> hilly
terrain  ===> terrain
with  ===> with
undulating  ===> undulating
and  ===> and
semi  ===> semi
-  ===> -
arid  ===> arid
climate  ===> climate
.  ===> .
River  ===> River
Kaveri  ===> Kaveri
flows  ===> flow
through  ===> through
the  ===> the
city  ===> city
in  ===> in
and  ===> and
an  ===> an
abundance  ===> abundance
if  ===> if
limestone  ===> limestone
is  ===> be
found  ===> find
in  ===> in
its  ===> -PRON-
beds  ===> bed
.  ===> .
It  ===> -PRON-
is  ===> be
located  ===> locate
centrally  ===> centrally
in  ===> in
the  ===> the
south  ===> south
Indian  ===> indian
peninsula  ===> peninsula
.  ===> .
It  ===> -PRON-
is  ===> be
located  ===> locate
at  ===> at
80  ===> 80
km  ===> km
from  ===> from
Coimbatore  ===> Coimbatore
and  ===> and
50  ===> 50
km  ===> km
form  ===> form
Tiruppur  ===> Tiruppur
.  ===> .
Being  ===> be
extemely  ===> extemely
popular  ===> popular
for  ===> for
the  ===> the
textile  ===> textile
industry  ===> industry
,  ===> ,
a  ===> a
lot  ===> lot
of  ===> of
in  ===> in
the  ===> the
region  ===> region
.  ===> .
Historically  ===> historically
,  ===> ,
it  ===> -PRON-
was  ===> be
part  ===> part
of  ===> of
the  ===> the
Kongu  ===> Kongu
Nadu  ===> Nadu
region  ===> region
in  ===> in
the  ===> the
Sangam  ===> Sangam
age  ===> age
and  ===> and
was  ===> be
ruled  ===> rule
by  ===> by
the  ===> the
Cheras  ===> Cheras
before  ===> before
being  ===> be
ousted  ===> oust
by  ===> by
the  ===> the
Pandyas  ===> Pandyas
in  ===> in
590  ===> 590
CE  ===> CE
.  ===> .
It  ===> -PRON-
was  ===> be
later  ===> later
a  ===> a
prominent  ===> prominent
British  ===> british
trading  ===> trading
point  ===> point
till  ===> till
independence  ===> independence
was  ===> be
gained  ===> gain
in  ===> in
1947  ===> 1947
```

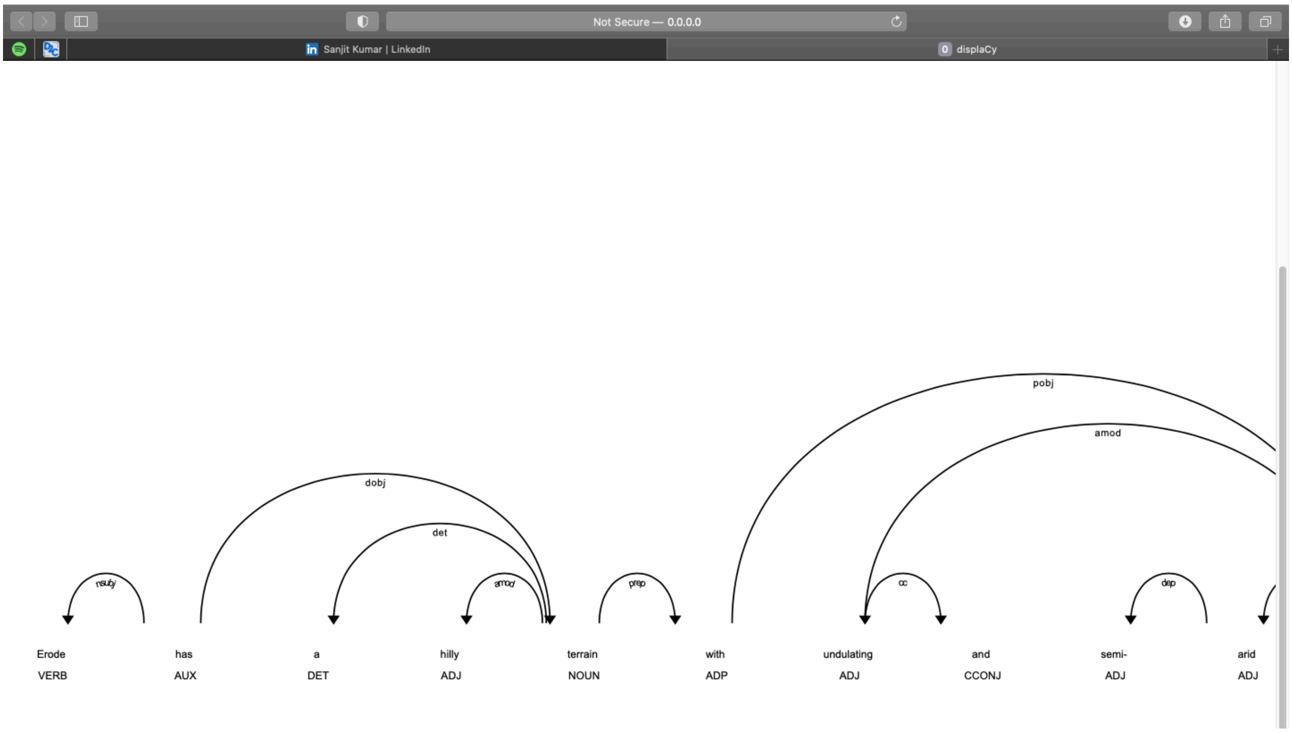
Part 5 - Remove Stop Words

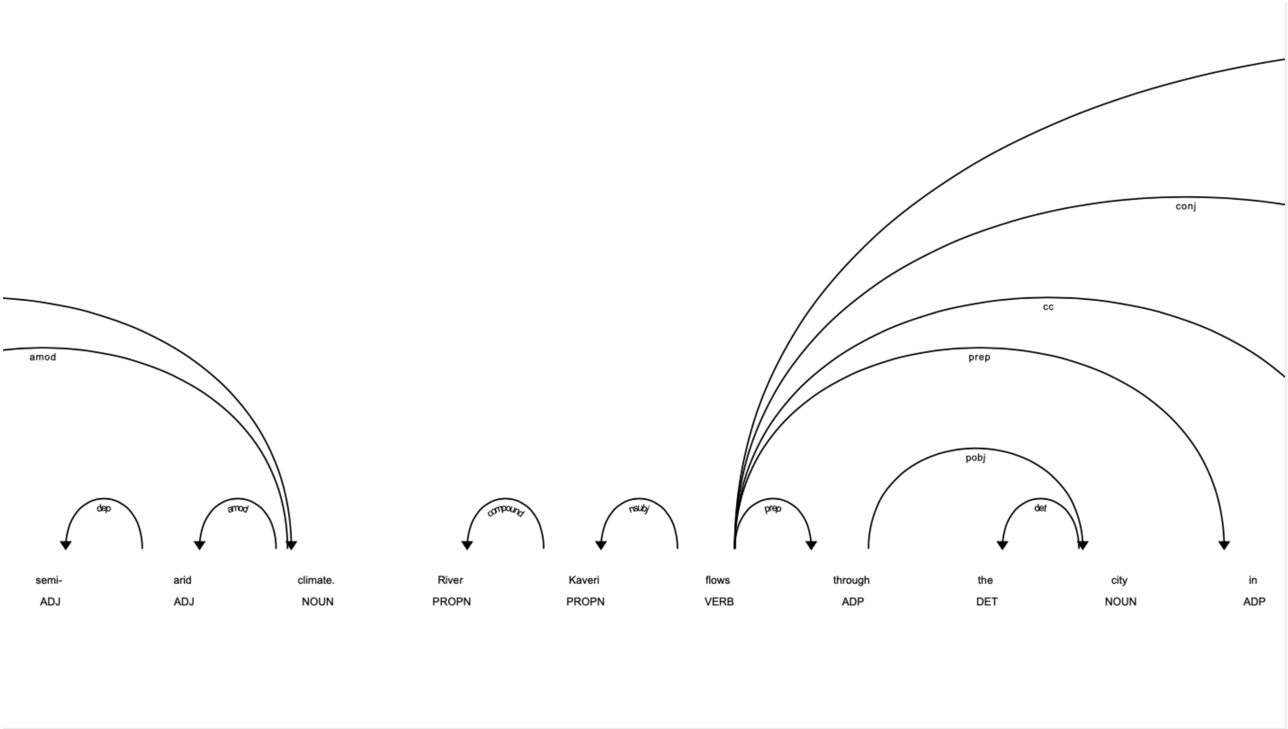
Before and After Removal

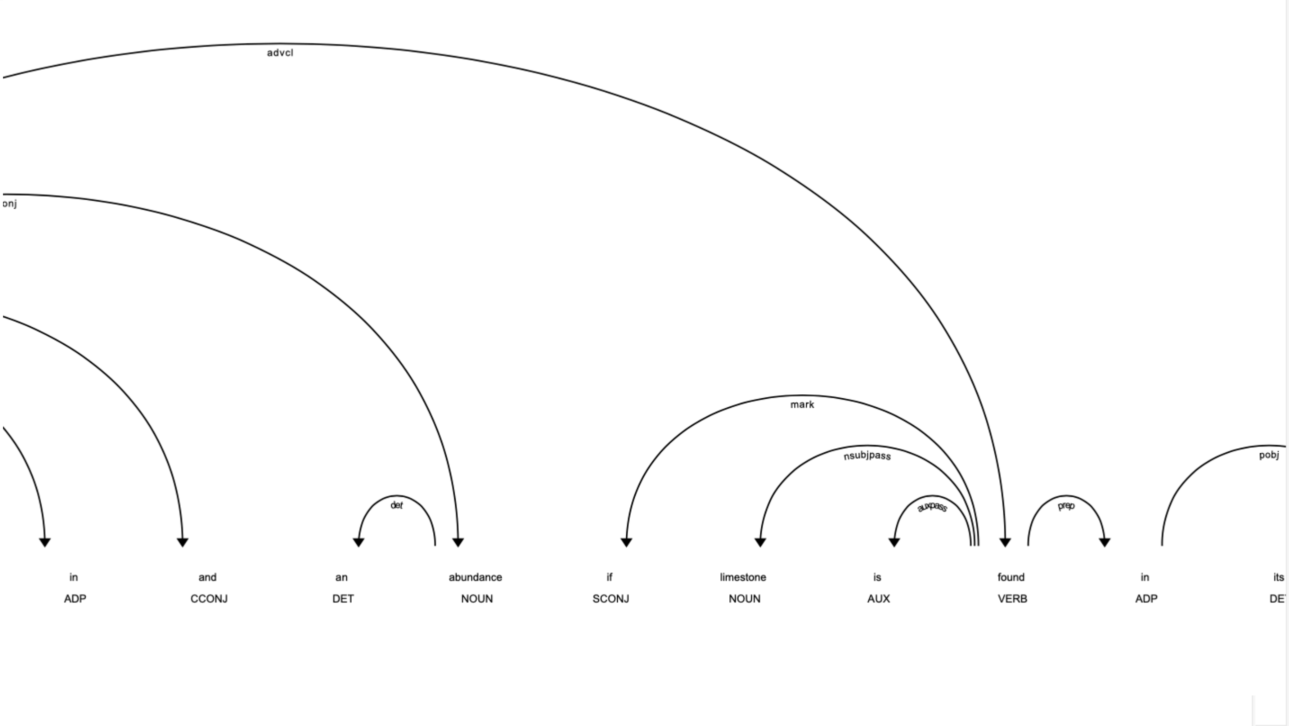


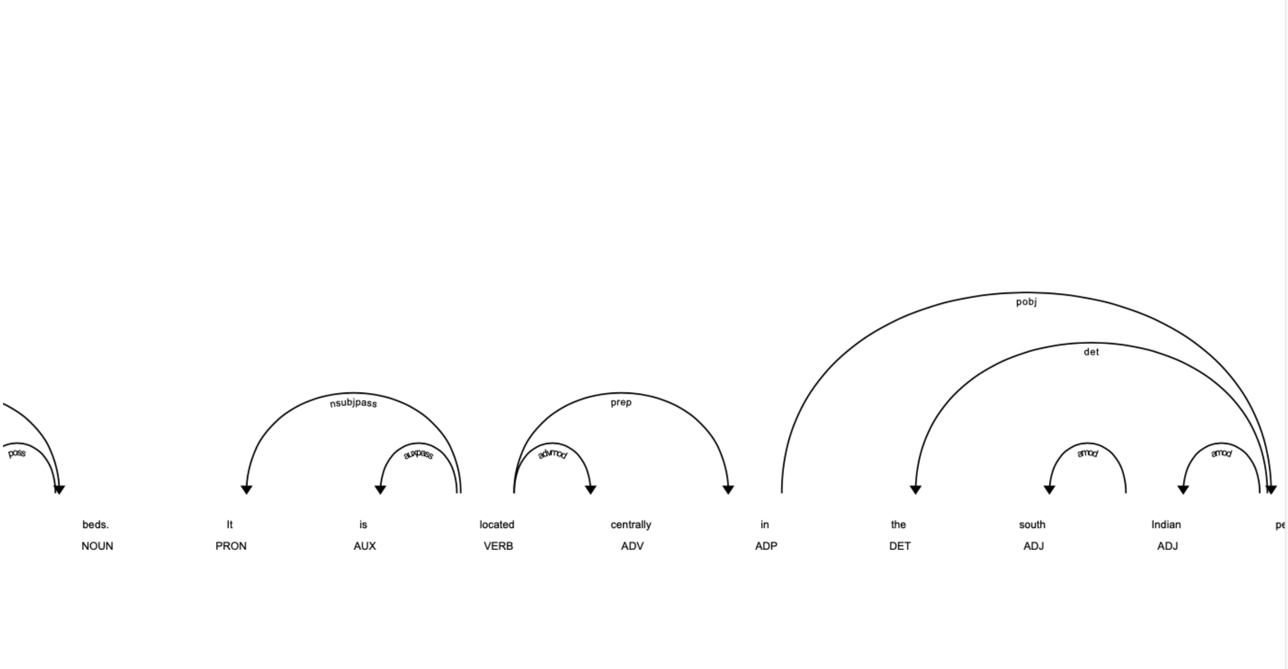
```
5. Remove Stop Words:

Before Removal:  ['Erode', 'is', 'the', 'seventh', 'largest', 'urban', 'agglomeration', 'in', 'Tamil', 'Nadu', '.', 'It', 'is', 'also', 'the', 'administrative', 'head
quarters', 'of', 'Erode', 'district', '.', 'Erode', 'has', 'a', 'hilly', 'terrain', 'with', 'undulating', 'and', 'semi', '-', 'arid', 'climate', '.', 'River', 'Kaveri
', 'flows', 'through', 'the', 'city', 'in', 'and', 'an', 'abundance', 'if', 'limestone', 'is', 'found', 'in', 'its', 'beds', '.', 'It', 'is', 'located', 'centrally',
'in', 'the', 'south', 'Indian', 'peninsula', '.', 'It', 'is', 'located', 'at', '80', 'km', 'from', 'Coimbatore', 'and', '50', 'km', 'form', 'Tiruppur', '.', 'Being',
'extemely', 'popular', 'for', 'the', 'textile', 'industry', ',', 'a', 'lot', 'of', 'the', 'cotton', 'spinning', ',', 'weaving', 'and', 'knitting', 'industries', 'can'
, 'be', 'found', 'in', 'the', 'region', '.', 'Historically', ',', 'it', 'was', 'part', 'of', 'the', 'Kongu', 'Nadu', 'region', 'in', 'the', 'Sangam', 'age', 'and', 'w
as', 'ruled', 'by', 'the', 'Cheras', 'before', 'being', 'ousted', 'by', 'the', 'Pandyas', 'in', '590', 'CE', '.', 'It', 'was', 'later', 'a', 'prominent', 'British', '
trading', 'point', 'till', 'independence', 'was', 'gained', 'in', '1947', '.']
After Removal:  ['Erode', 'seventh', 'largest', 'urban', 'agglomeration', 'Tamil', 'Nadu', '.', 'administrative', 'headquarters', 'Erode', 'district', '.', 'Erode', '
hilly', 'terrain', 'undulating', 'semi', '-', 'arid', 'climate', '.', 'River', 'Kaveri', 'flows', 'city', 'abundance', 'limestone', 'found', 'beds', '.', 'located', '
centrally', 'south', 'Indian', 'peninsula', '.', 'located', '80', 'km', 'Coimbatore', '50', 'km', 'form', 'Tiruppur', '.', 'extemely', 'popular', 'textile', 'industry
', ',', 'lot', 'cotton', 'spinning', ',', 'weaving', 'knitting', 'industries', 'found', 'region', '.', 'Historically', ',', 'Kongu', 'Nadu', 'region', 'Sangam', 'age'
, 'ruled', 'Cheras', 'ousted', 'Pandyas', '590', 'CE', '.', 'later', 'prominent', 'British', 'trading', 'point', 'till', 'independence', 'gained', '1947', '.']
```

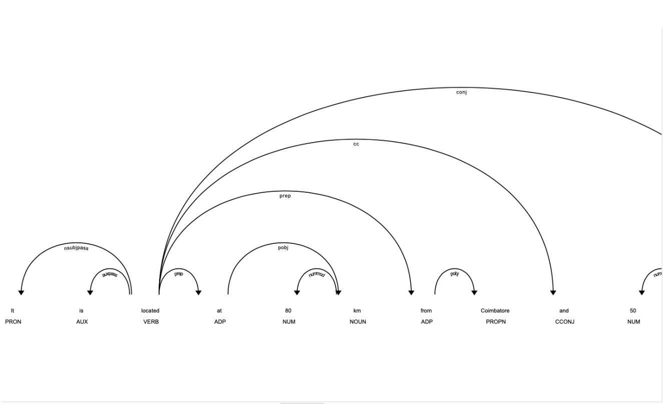
Part 6 - Dependancy Parsing and Graph

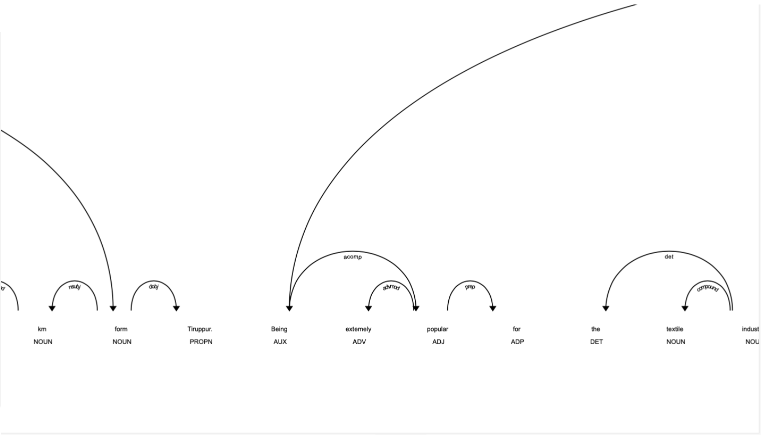


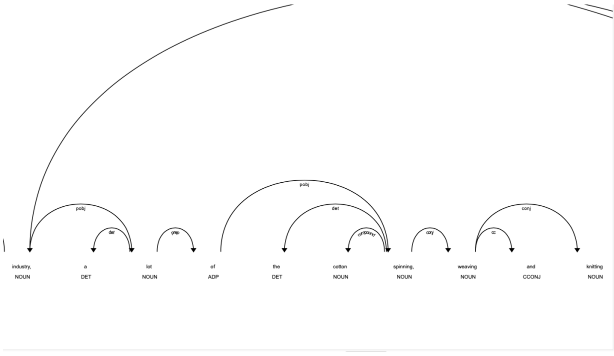

conj
cc
amod
prep
pobj
dep
amod
compound
nsubj
prep
det
semi-
ADJ
arid
ADJ
climate.
NOUN
River
PROPN
Kaveri
PROPN
flows
VERB
through
ADP
the
DET
city
NOUN
in
ADP

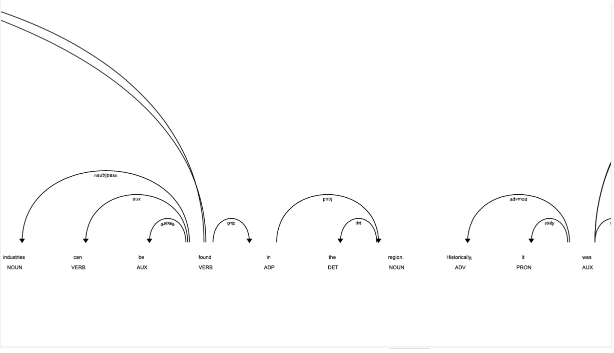

advcl
onj
mark
nsubjpass
pobj
det
auxpass
prep
in
ADP
and
CCONJ
an
DET
abundance
NOUN
if
SCONJ
limestone
NOUN
is
AUX
found
VERB
in
ADP
its
DE

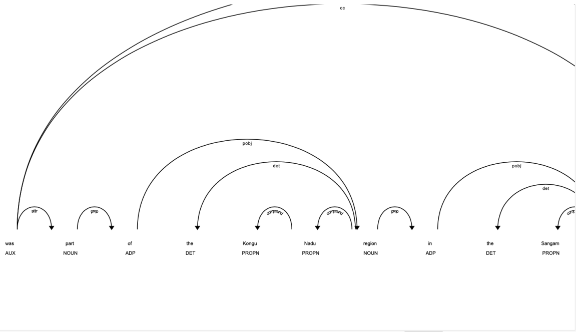

pobj
det
nsubjpass
prep
poss
auxpass
advmod
amod
amod
beds.
NOUN
It
PRON
is
AUX
located
VERB
centrally
ADV
in
ADP
the
DET
south
ADJ
Indian
ADJ
pe

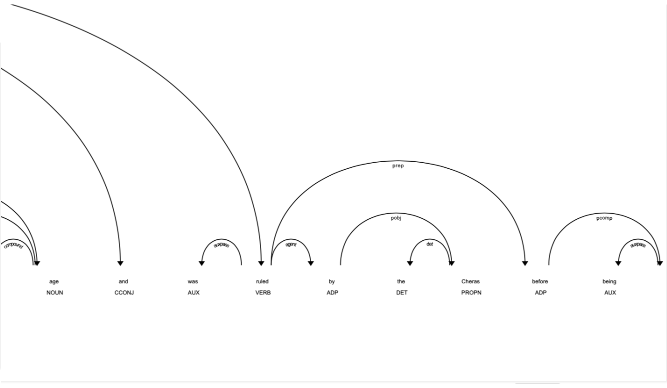

conj
cc
prep
nsubjpass
auxpass
prep
pobj
nummod
pobj
It PRON
is AUX
located VERB
at ADP
60 NUM
km NOUN
from ADP
Coimbatore PROPN
and CCONJ
50 NUM

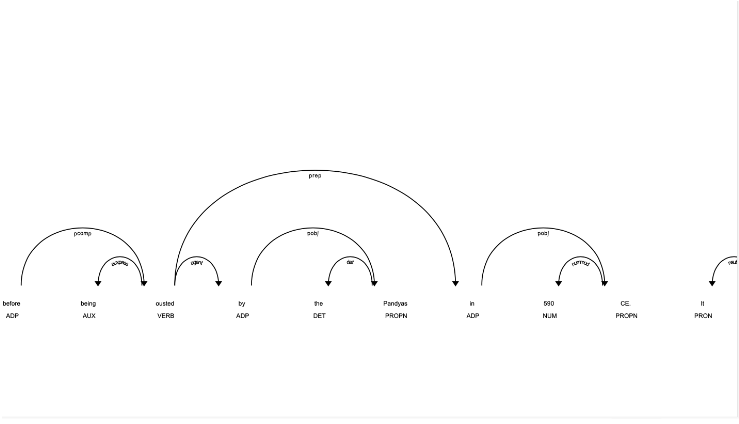

nsubj
dobj
acomp
advmod
prep
det
compound
km NOUN
form NOUN
Tiruppur. PROPN
Being AUX
extemely ADV
popular ADJ
for ADP
the DET
textile NOUN
indust NOU

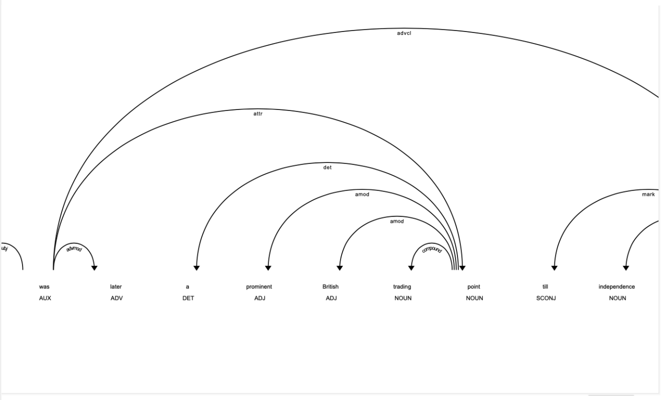

pobj
det
prep
pobj
det
compound
conj
cc
conj
industry, NOUN
a DET
lot NOUN
of ADP
the DET
cotton NOUN
spinning, NOUN
weaving NOUN
and CCONJ
knitting NOUN


nsubjpass
aux
auxpass
prep
pobj
det
advmod
nsubj
industries NOUN
can VERB
be AUX
found VERB
in ADP
the DET
region. NOUN
Historically, ADV
it PRON
was AUX

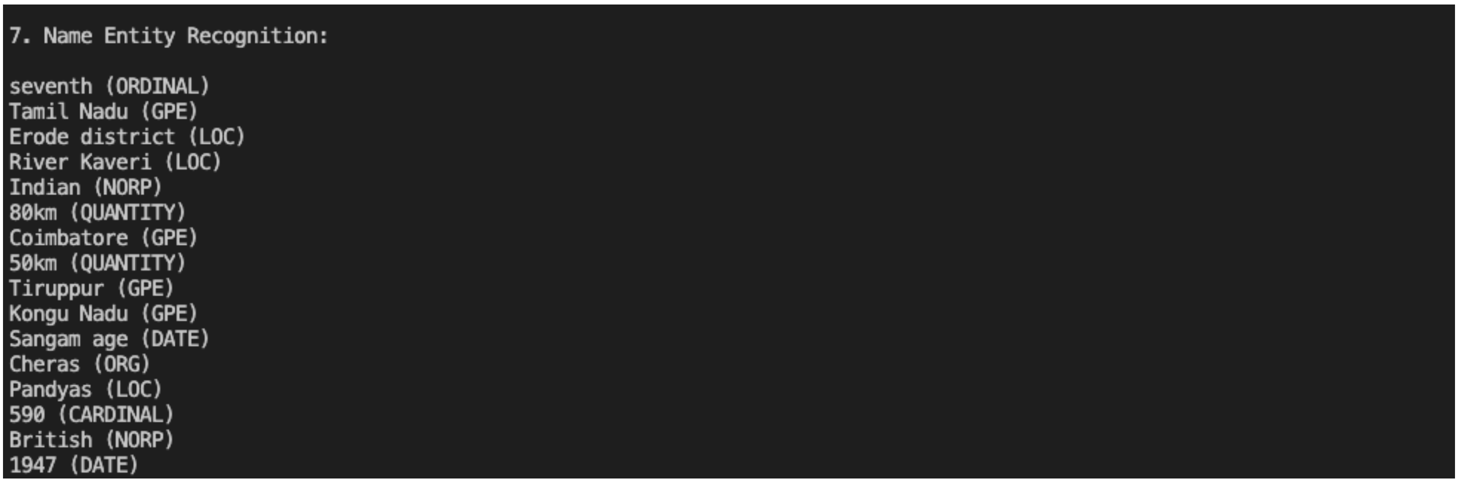

cc
attr
prep
pobj
det
compound
compound
prep
pobj
det
was AUX
part NOUN
of ADP
the DET
Kongu PROPN
Nadu PROPN
region NOUN
in ADP
the DET
Sangam PROPN


prep
pobj
det
auxpass
agent
pcomp
auxpass
age NOUN
and CCONJ
was AUX
ruled VERB
by ADP
the DET
Cheras PROPN
before ADP
being AUX


prep
pcomp
auxpass
agent
pobj
det
pobj
nummod
nsubj
before ADP
being AUX
ousted VERB
by ADP
the DET
Pandyas PROPN
in ADP
590 NUM
CE. PROPN
It PRON

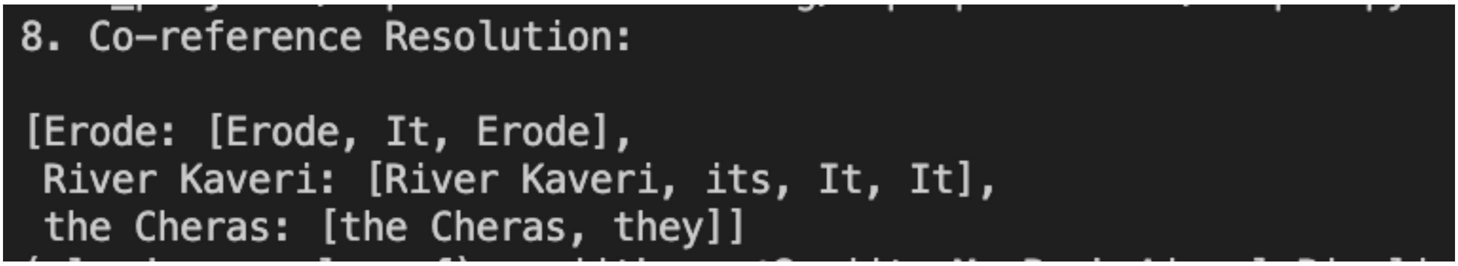

advcl
attr
det
amod
amod
compound
advmod
mark
was AUX
later ADV
a DET
prominent ADJ
British ADJ
trading NOUN
point NOUN
till SCONJ
independence NOUN

Part 7 - Name Entity Recognition



```
7. Name Entity Recognition:

seventh (ORDINAL)
Tamil Nadu (GPE)
Erode district (LOC)
River Kaveri (LOC)
Indian (NORP)
80km (QUANTITY)
Coimbatore (GPE)
50km (QUANTITY)
Tiruppur (GPE)
Kongu Nadu (GPE)
Sangam age (DATE)
Cheras (ORG)
Pandyas (LOC)
590 (CARDINAL)
British (NORP)
1947 (DATE)
```

Part 8 - Co-reference resolution



```
8. Co-reference Resolution:

[Erode: [Erode, It, Erode],
 River Kaveri: [River Kaveri, its, It, It],
 the Cheras: [the Cheras, they]]
```