Natural Language Processing of a Low-Resource Language

(Igbo, one of the West African Languages)

Student: Atajan Abdyyev, Harrisburg University
aabdyev@my.harrisburgu.edu
Principal Investigator: Stanley Nwoji PhD, Harrisburg University
SNwoji@harrisburgu.edu
Mentor: Iheb Abdellatif PhD, Harrisburg University
IAbdellatif@harrisburgu.edu

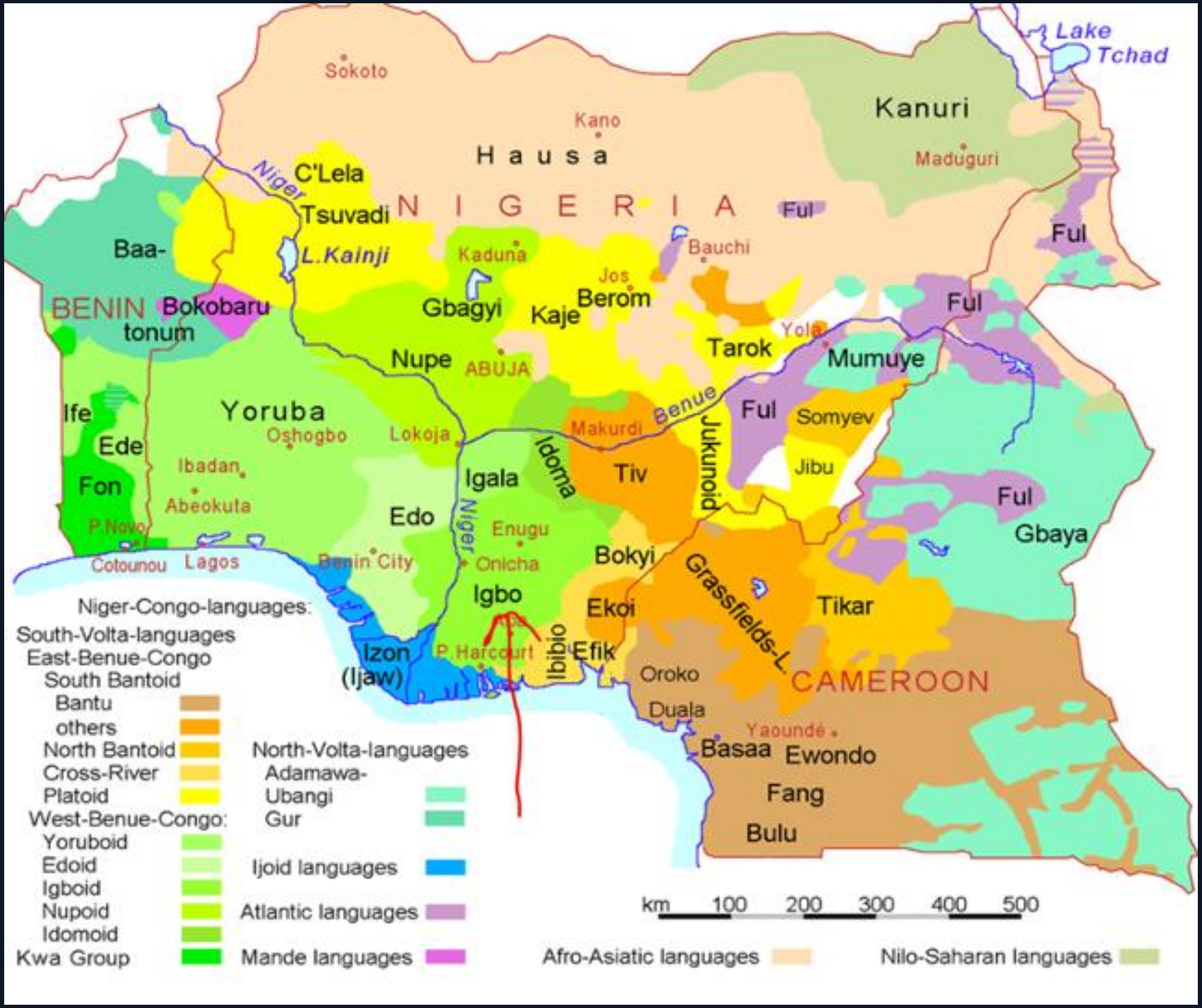
CAREERS CYBERTEAM

Date 4/10/2024

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

- To narrow down the NLP gap for Igbo a category language.



NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

- Timeframe
  - –September 2023
  - –April 2024

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

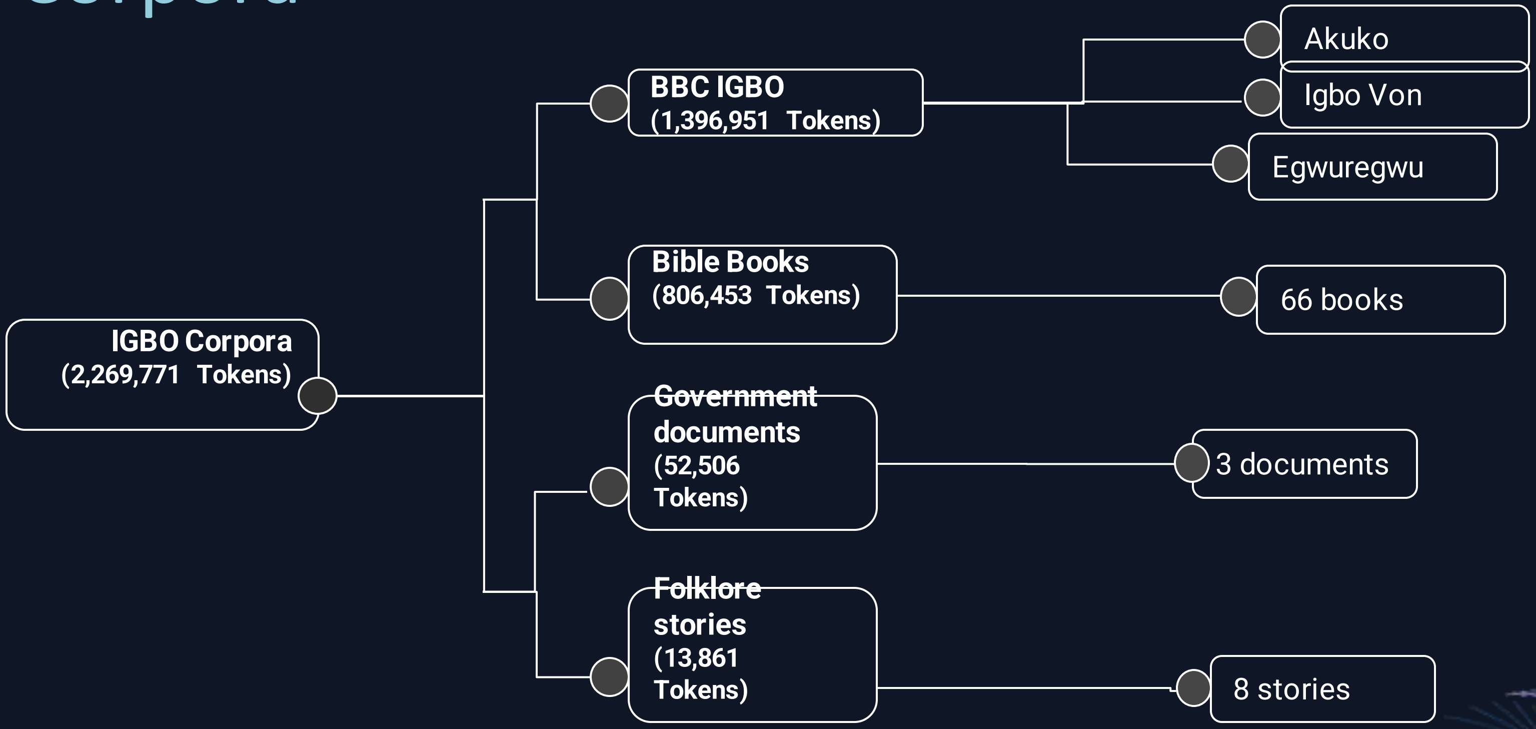
- Goals/Milestones
  - –Develop Igbo Corpora
  - –Cleaning Data
  - –Perform NLP Analysis (Statistical, ML, DL on POS and NER)
  - –Perform Text Categorization
  - –Perform Information Extraction
  - –Perform Translation

CAREERS CYBERTEAM

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Corpora



IGBO Corpora (2,269,771 Tokens)

BBC IGBO (1,396,951 Tokens)

Akuko

Igbo Von

Egwuregwu

Bible Books (806,453 Tokens)

66 books

Government documents (52,506 Tokens)

3 documents

Folklore stories (13,861 Tokens)

8 stories

CAREERS CYBERTEAM

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Corpora Cleaning

Excel:

Removing special characters, recurring instances of texts (ex: Ebe foto si, BBC, Names of articles), removing numbers where possible, sorting sentences to be by rows based on characters, removing any duplicates

Notepad

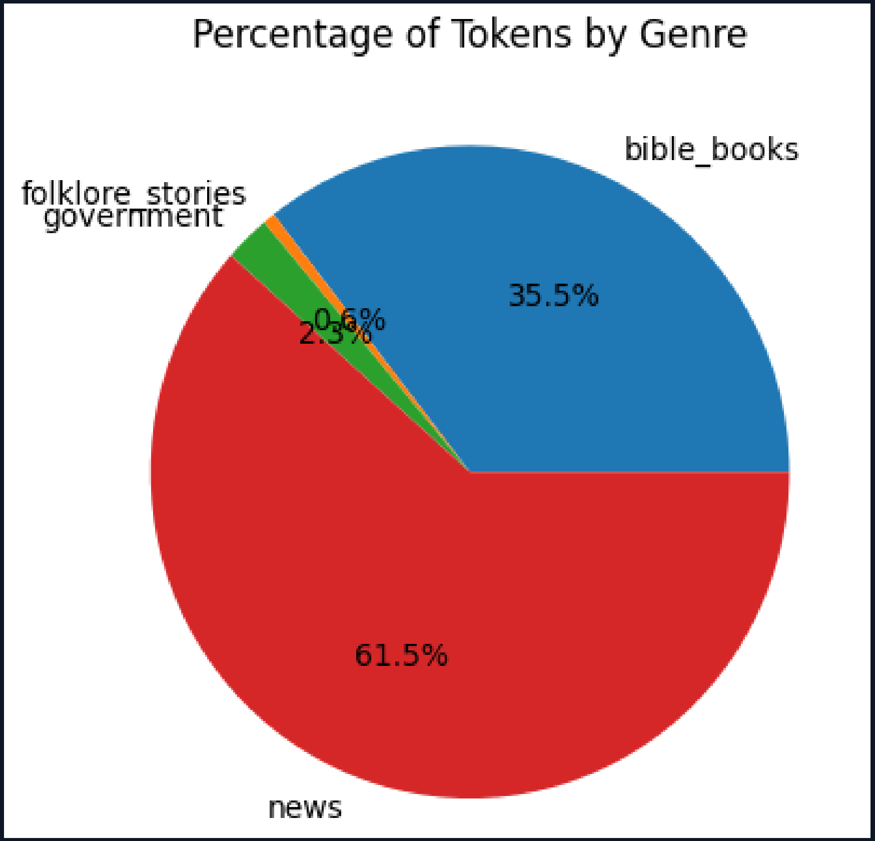
Storing Corpus files as .TXT

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Corpora

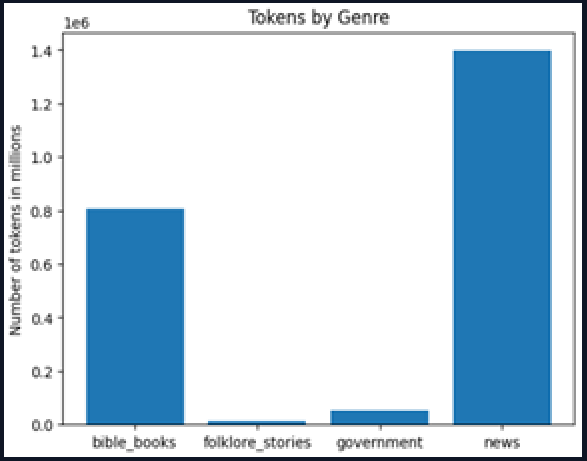
- Total Tokens : 2,269,771
- Total Sentences: 75,883
- Stored on Github

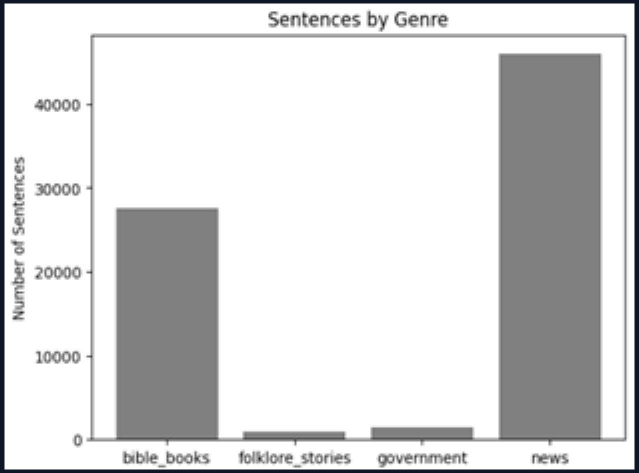


NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Tokens and Sentences by Genres





CAREERS CYBERTEAM

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Corpora (Tokens insights - Akuko vs Reuters)

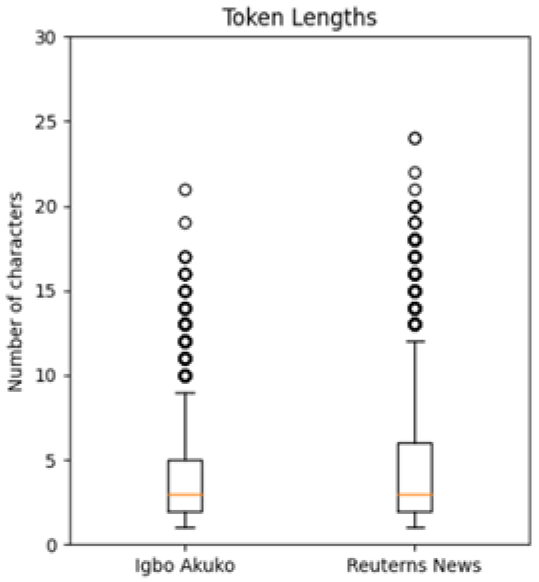
Tokens per Sentence: 32.99

Median token length: 1.00 characters

Igbo Akuko (news)

Maximum token length: 21 characters

Average token length: 3.54 characters



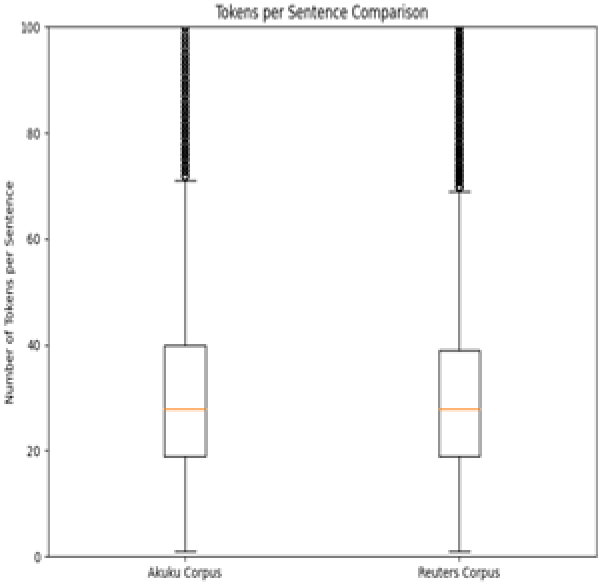
Tokens per Sentence: 33.85

Median token length: 3.00 characters

Reuters

Maximum token length: 42 characters

Average token length: 4.00 characters



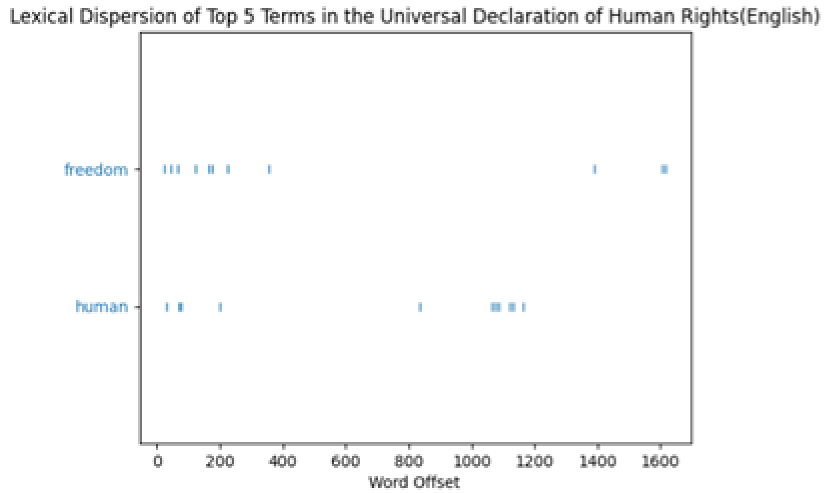
Corpora Tokens insights (lexical diversity)

Corpora (Tokens insights - UDHR)

Universal Declaration of Human Rights

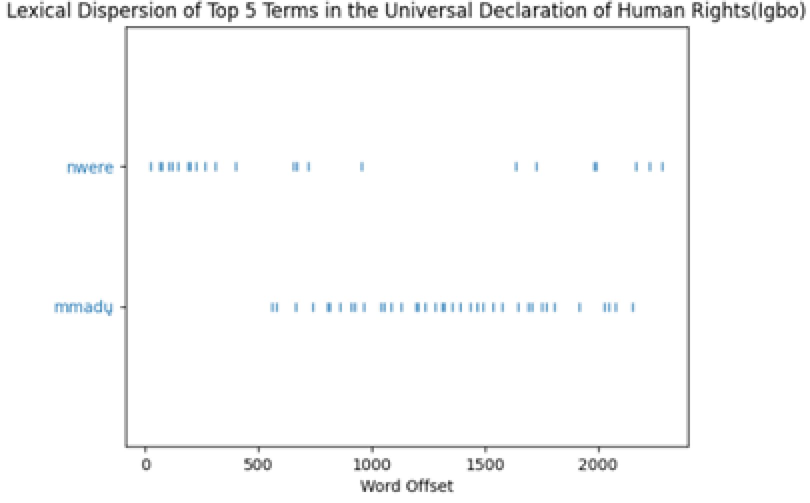
English

Total words in English version: 1907
Vocabulary size (unique words): 556
Lexical diversity (type-token ratio): 0.2916



Igbo

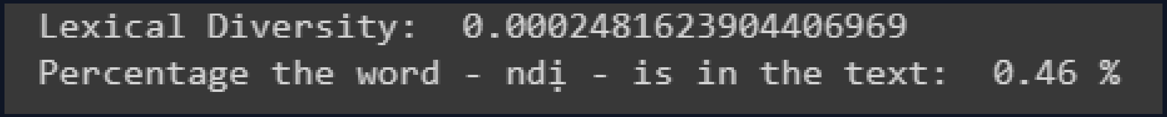
Total words in Igbo version: 2313
Vocabulary size (unique words): 573
Lexical diversity (type-token ratio): 0.2477



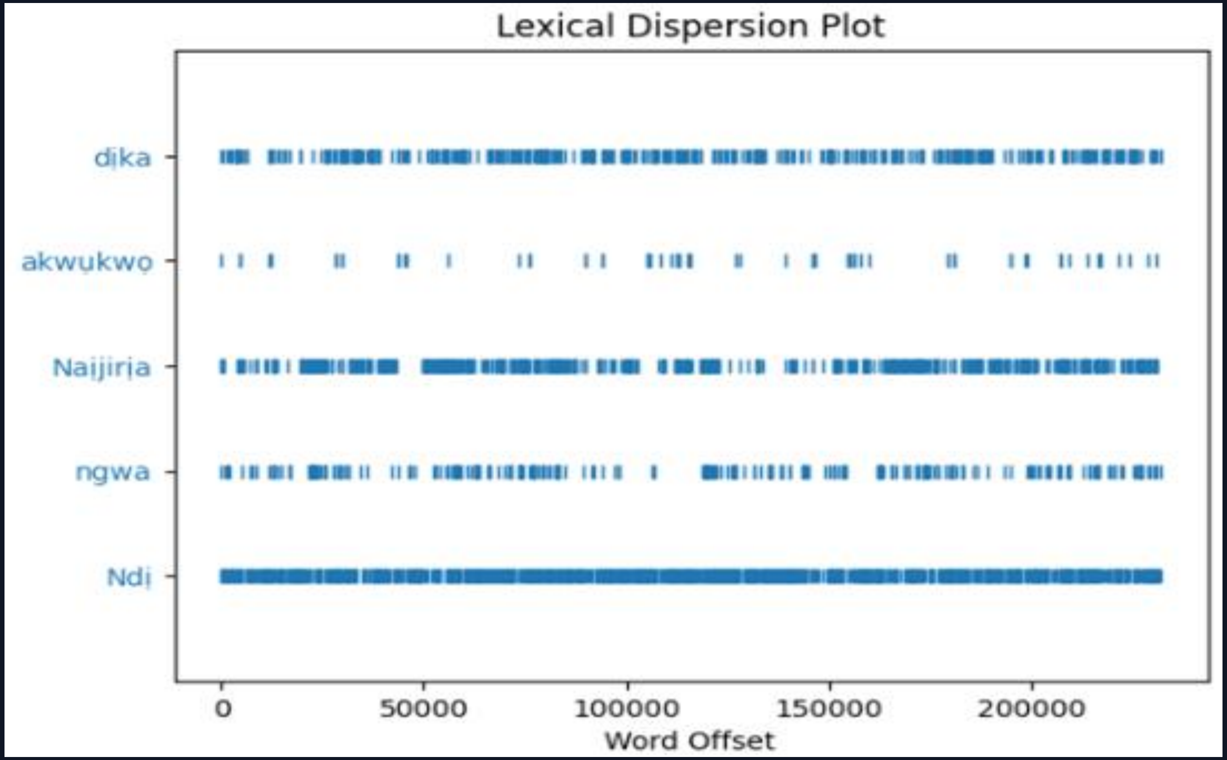
NLP of a Low-Resource Language

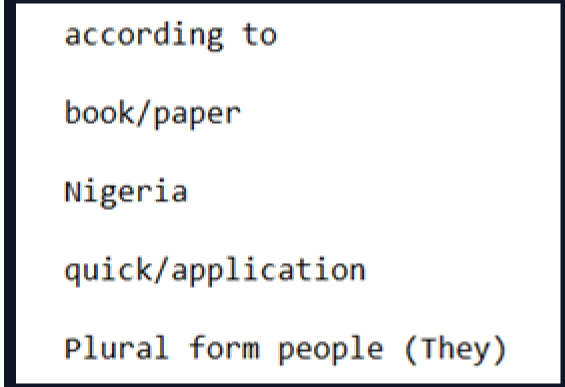
(Igbo, one of the West African Languages)

Corpora (Tokens insights - Akuko)



```
Lexical Diversity:  0.0002481623904406969
Percentage the word - ndị - is in the text:  0.46 %
```





according to

book/paper

Nigeria

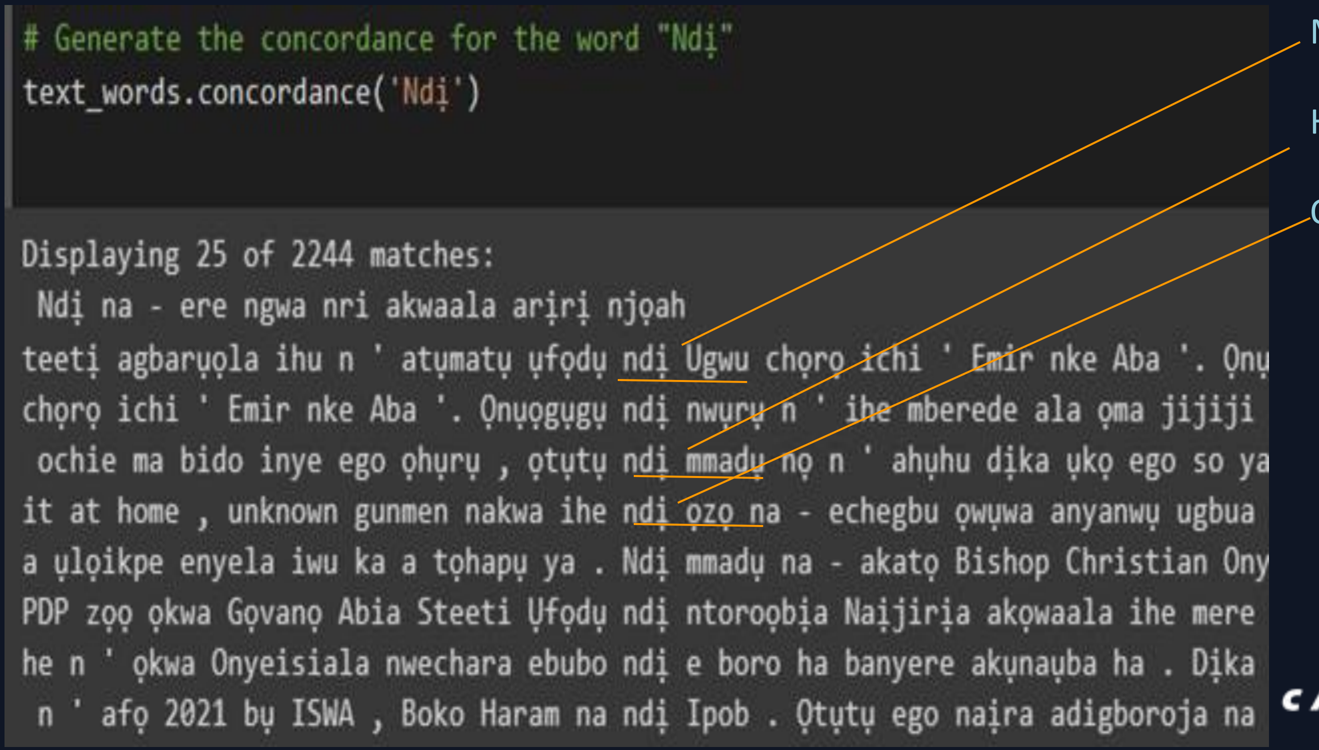
quick/application

Plural form people (They)

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Corpora (Tokens insights - Akuko)



```python
# Generate the concordance for the word "Ndị"
text_words.concordance('Ndị')
```

```
Displaying 25 of 2244 matches:
 Ndị na - ere ngwa nri akwaala arịrị njọah
teetị agbaruọla ihu n ' atụmatụ ụfọdụ ndị Ugwu chọrọ ichi ' Emir nke Aba '. Ọnụ
chọrọ ichi ' Emir nke Aba '. Ọnụọgụgụ ndị nwụrụ n ' ihe mberede ala ọma jijiji
 ochie ma bido inye ego ọhụrụ , ọtụtụ ndị mmadụ nọ n ' ahụhu dịka ụkọ ego so ya
it at home , unknown gunmen nakwa ihe ndị ọzọ na - echegbu ọwụwa anyanwụ ugbua
a ụlọikpe enyela iwu ka a tọhapụ ya . Ndị mmadụ na - akatọ Bishop Christian Ony
PDP zọọ ọkwa Gọvanọ Abia Steeti Ụfọdụ ndị ntoroọbịa Naịjirịa akọwaala ihe mere
he n ' ọkwa Onyeisiala nwechara ebubo ndị e boro ha banyere akụnaụba ha . Dịka
 n ' afọ 2021 bụ ISWA , Boko Haram na ndị Ipob . Ọtụtụ ego naịra adigboroja na
```

Mountain People

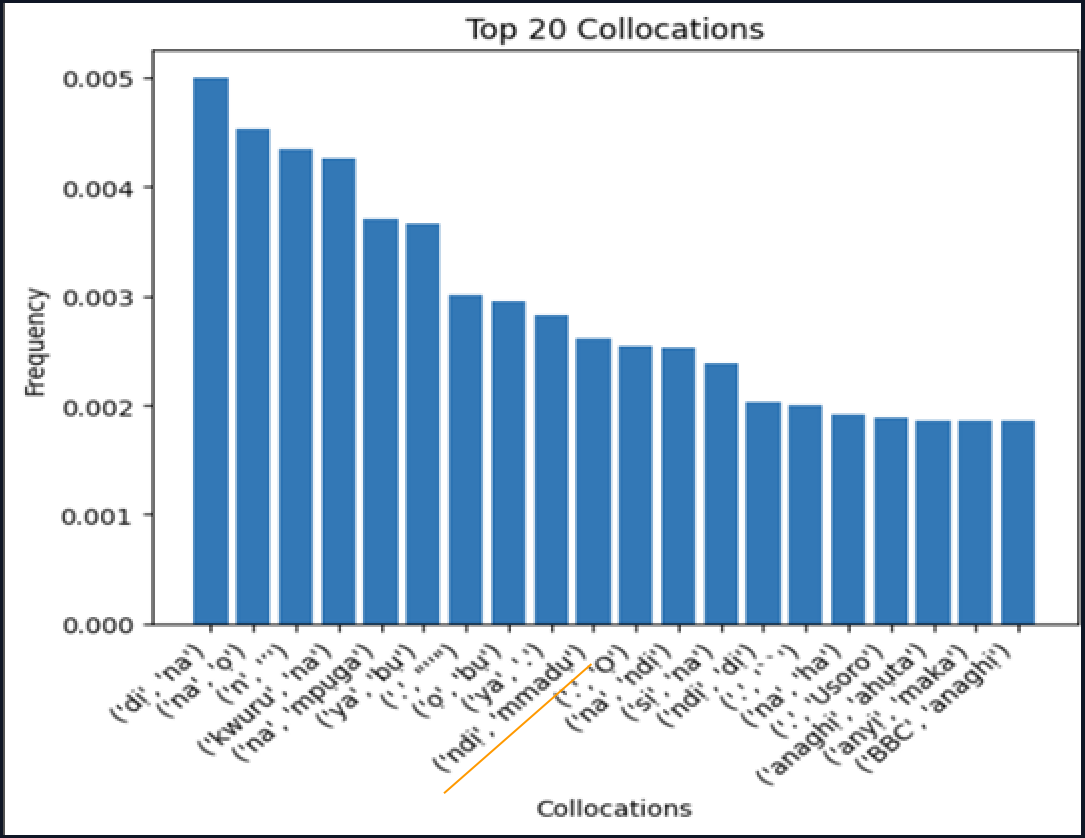
Human (people)

Others

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Corpora (Tokens insights (semantic information) - Akuko)

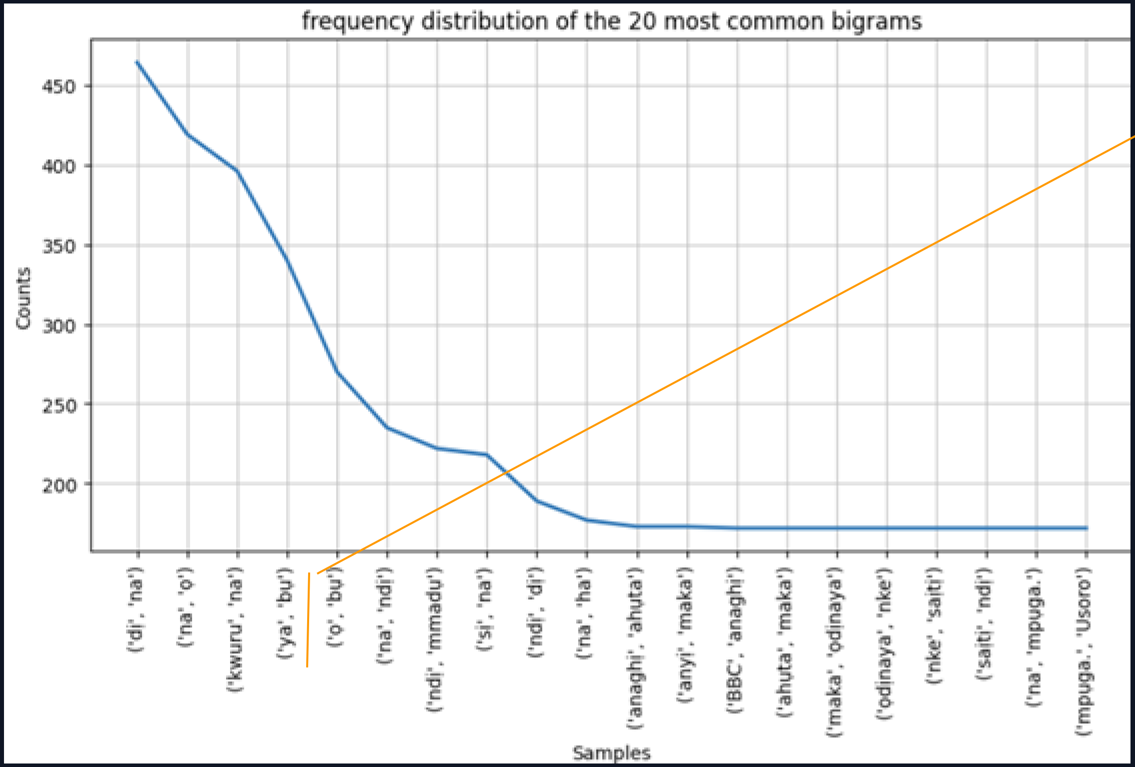


CAREERS CYBERTEAM

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Corpora (Tokens insights (phrase identification) - Akuko)



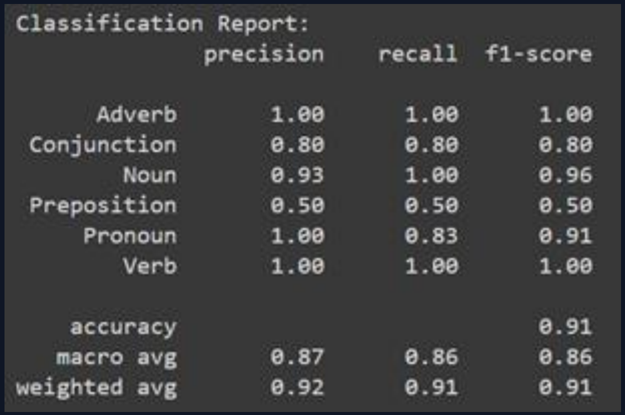
That is …
bu = is

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

NLP Analysis (POS - Bible/Jon 1)

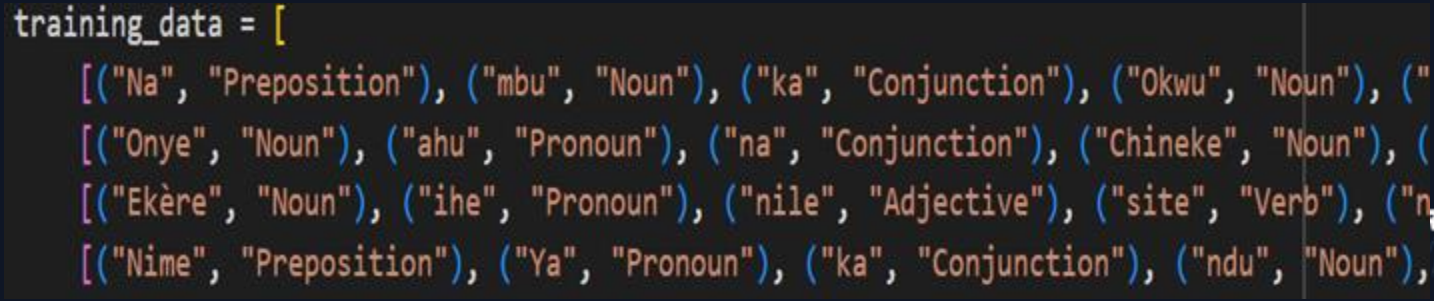
- 7 Unique Parts of speech used with couple hundred tokens in total
- Applied statistical NLTK tagger to get sample results
- Highlighting POS of interest for unigram tokens
- Achieved accuracy at 44% at Unigram level vs Machine Learning crfsuite library showed accuracy showed better results averaging 92%



```
Classification Report:
              precision    recall  f1-score

      Adverb       1.00      1.00      1.00
 Conjunction       0.80      0.80      0.80
        Noun       0.93      1.00      0.96
 Preposition       0.50      0.50      0.50
     Pronoun       1.00      0.83      0.91
        Verb       1.00      1.00      1.00

    accuracy                           0.91
   macro avg       0.87      0.86      0.86
weighted avg       0.92      0.91      0.91
```



```
training_data = [
    [("Na", "Preposition"), ("mbu", "Noun"), ("ka", "Conjunction"), ("Okwu", "Noun"), ("
    [("Onye", "Noun"), ("ahu", "Pronoun"), ("na", "Conjunction"), ("Chineke", "Noun"), (
    [("Ekère", "Noun"), ("ihe", "Pronoun"), ("nile", "Adjective"), ("site", "Verb"), ("n
    [("Nime", "Preposition"), ("Ya", "Pronoun"), ("ka", "Conjunction"), ("ndu", "Noun"),
```

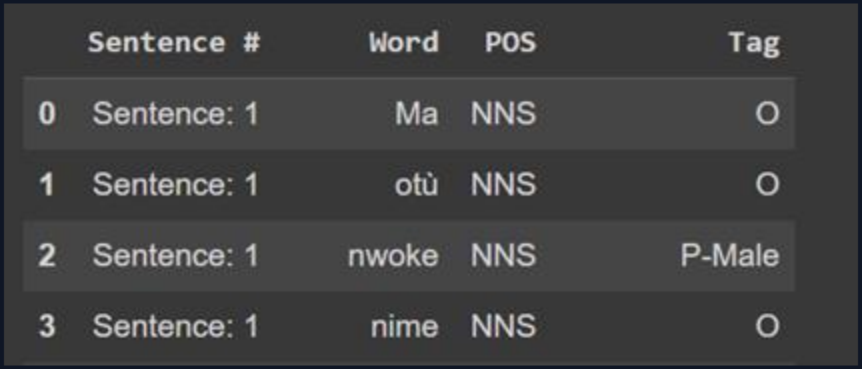
REERS
CYBERTEAM

NLP of a Low-Resource Language

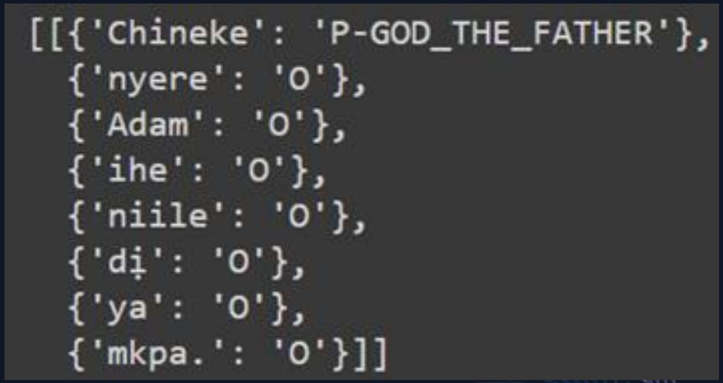
(Igbo, one of the West African Languages)

NLP Analysis (NER - Bible/Jon 1)

- Currently 29 tokens used for NER tags
- Applied deep learning Hugging Face tagger to get results
- Accuracy was achieved at eval loss:0.5 and F1 Score 0.3



	Sentence #	Word	POS	Tag
0	Sentence: 1	Ma	NNS	O
1	Sentence: 1	otù	NNS	O
2	Sentence: 1	nwoke	NNS	P-Male
3	Sentence: 1	nime	NNS	O



```
[[{'Chineke': 'P-GOD_THE_FATHER'},
  {'nyere': 'O'},
  {'Adam': 'O'},
  {'ihe': 'O'},
  {'niile': 'O'},
  {'dị': 'O'},
  {'ya': 'O'},
  {'mkpa.': 'O'}]]
```

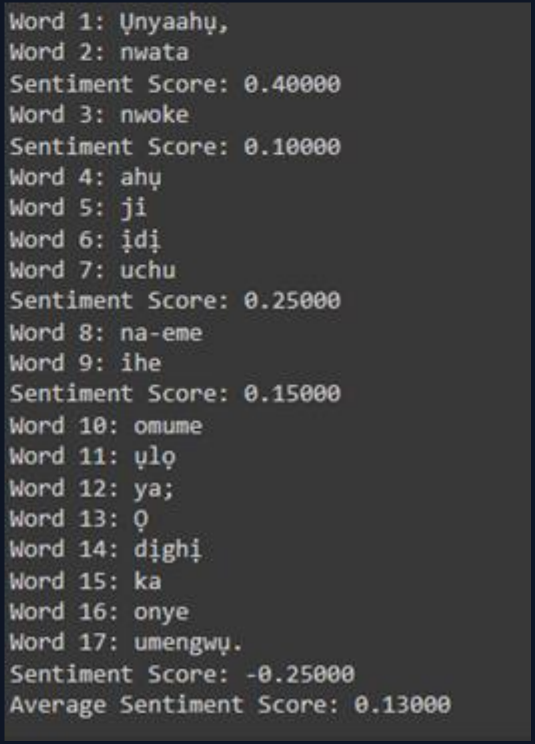
CAREERS CYBERTEAM

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

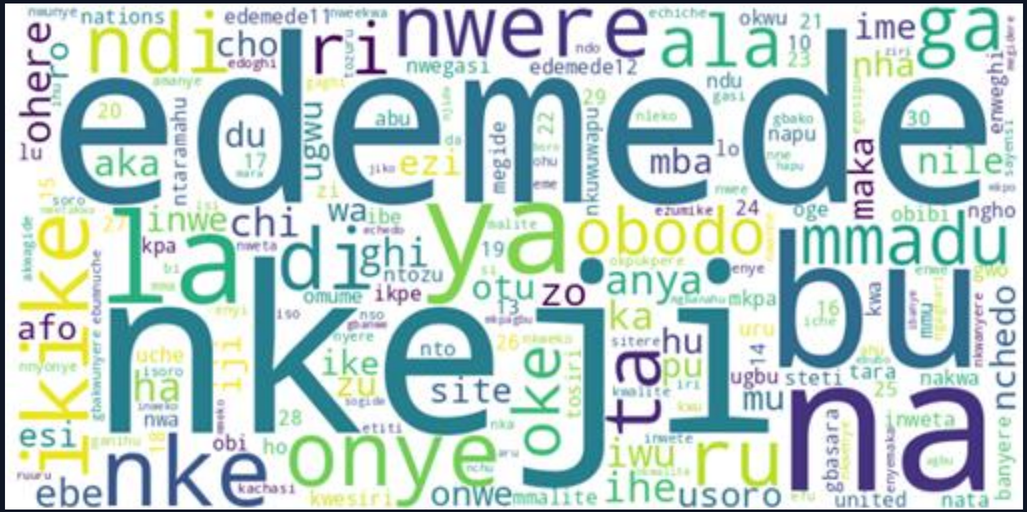
Information Extraction (Sentiment Analysis, Stopwords, WordCloud - UDHR)

Sentiment Analysis: `Yesterday, the boy diligently did his homework; He is not like a lazy person.`



```
Word 1: Ụnyaahụ,
Word 2: nwata
Sentiment Score: 0.40000
Word 3: nwoke
Sentiment Score: 0.10000
Word 4: ahụ
Word 5: ji
Word 6: ịdị
Word 7: uchu
Sentiment Score: 0.25000
Word 8: na-eme
Word 9: ihe
Sentiment Score: 0.15000
Word 10: omume
Word 11: ụlọ
Word 12: ya;
Word 13: Ọ
Word 14: dịghị
Word 15: ka
Word 16: onye
Word 17: umengwụ.
Sentiment Score: -0.25000
Average Sentiment Score: 0.13000
```

Word Cloud of UN Declaration of Human Rights



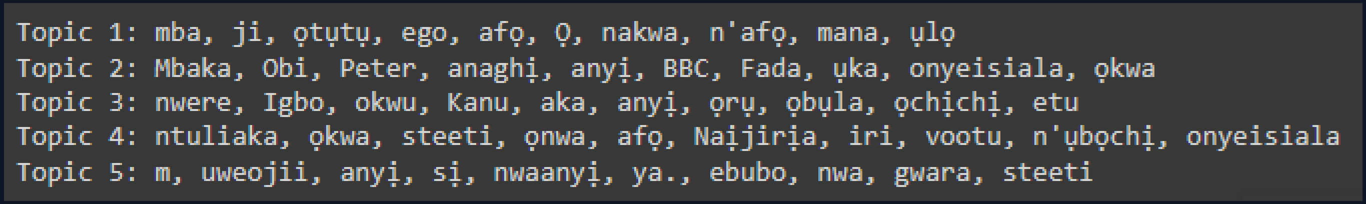
CAREERS CYBERTEAM

NLP of a Low-Resource Language

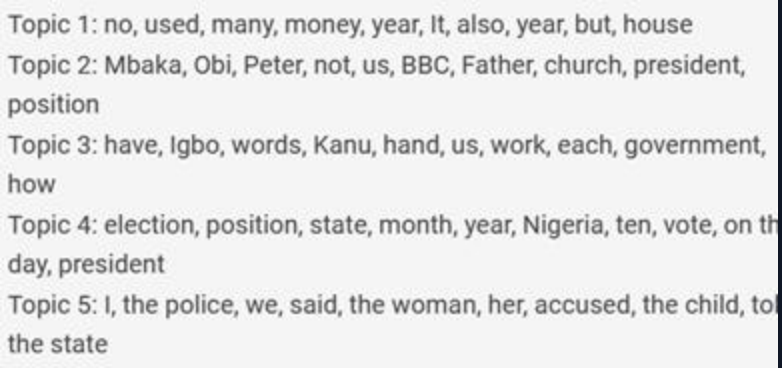
(Igbo, one of the West African Languages)

Text Categorization (Unsupervised ML Topic Modeling - Akuko)

Text Categorization (Topics)



```
Topic 1: mba, ji, ọtụtụ, ego, afọ, Ọ, nakwa, n'afọ, mana, ụlọ
Topic 2: Mbaka, Obi, Peter, anaghị, anyị, BBC, Fada, ụka, onyeisiala, ọkwa
Topic 3: nwere, Igbo, okwu, Kanu, aka, anyị, ọrụ, ọbụla, ọchịchị, etu
Topic 4: ntuliaka, ọkwa, steeti, ọnwa, afọ, Naịjirịa, iri, vootu, n'ụbọchị, onyeisiala
Topic 5: m, uweojii, anyị, sị, nwaanyị, ya., ebubo, nwa, gwara, steeti
```



Topic 1: no, used, many, money, year, It, also, year, but, house

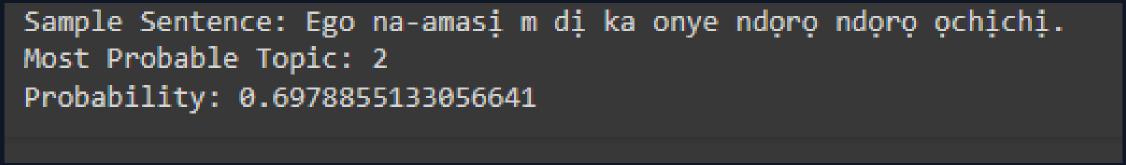
Topic 2: Mbaka, Obi, Peter, not, us, BBC, Father, church, president, position

Topic 3: have, Igbo, words, Kanu, hand, us, work, each, government, how

Topic 4: election, position, state, month, year, Nigeria, ten, vote, on th day, president

Topic 5: I, the police, we, said, the woman, her, accused, the child, to the state

I like money as a politician.



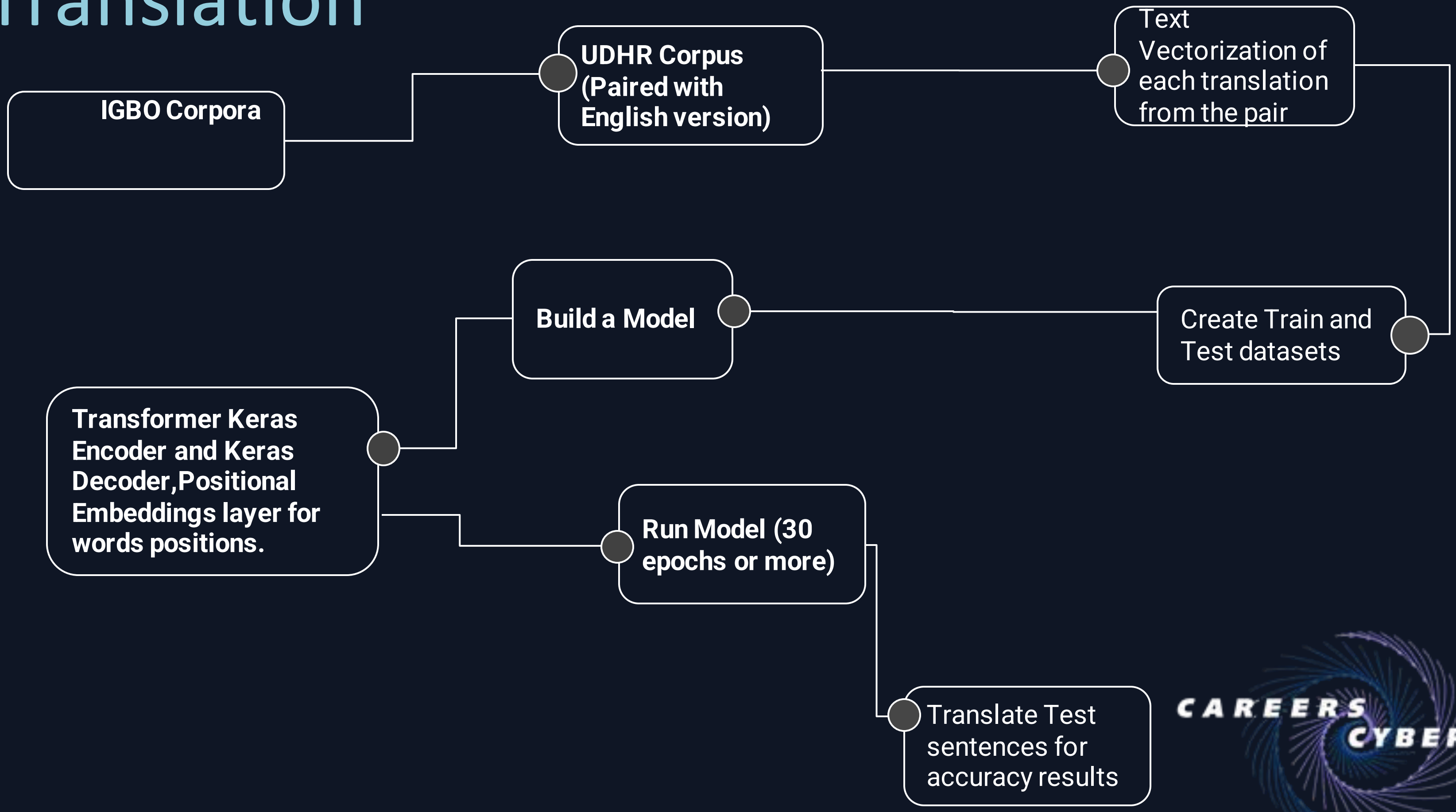
```
Sample Sentence: Ego na-amasị m dị ka onye ndọrọ ndọrọ ọchịchị.
Most Probable Topic: 2
Probability: 0.6978855133056641
```

CAREERS CYBERTEAM

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

Translation



IGBO Corpora

UDHR Corpus (Paired with English version)

Text Vectorization of each translation from the pair

Create Train and Test datasets

Build a Model

Transformer Keras Encoder and Keras Decoder,Positional Embeddings layer for words positions.

Run Model (30 epochs or more)

Translate Test sentences for accuracy results

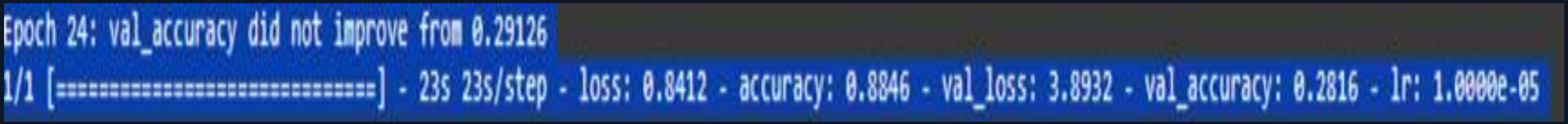
CAREERS CYBERTEAM

NLP of a Low-Resource Language

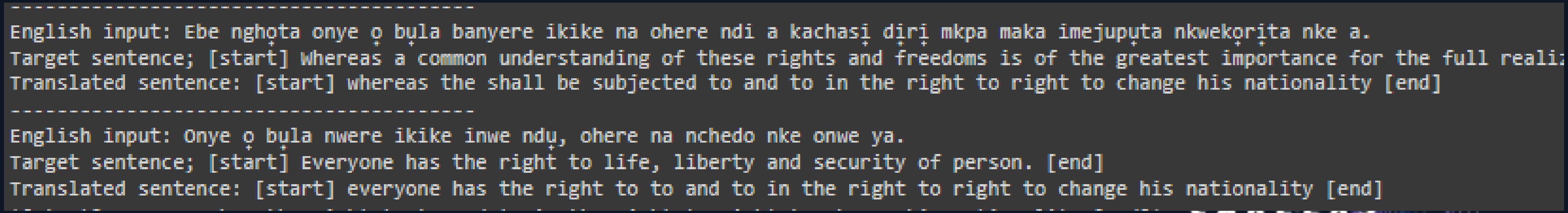
(Igbo, one of the West African Languages)

Translation

Read Paired Sentences Data (Igbo and English of UNDHR) -> Vectorize each pair -> Building Model -> Train Model -> Decode Sentences



```
Epoch 24: val_accuracy did not improve from 0.29126
1/1 [==============================] - 23s 23s/step - loss: 0.8412 - accuracy: 0.8846 - val_loss: 3.8932 - val_accuracy: 0.2816 - lr: 1.0000e-05
```



```
----------------------------------------
English input: Ebe nghọta onye ọ bụla banyere ikike na ohere ndi a kachasị dịrị mkpa maka imejupụta nkwekọrịta nke a.
Target sentence; [start] Whereas a common understanding of these rights and freedoms is of the greatest importance for the full reali:
Translated sentence: [start] whereas the shall be subjected to and to in the right to right to change his nationality [end]
----------------------------------------
English input: Onye ọ bụla nwere ikike inwe ndụ, ohere na nchedo nke onwe ya.
Target sentence; [start] Everyone has the right to life, liberty and security of person. [end]
Translated sentence: [start] everyone has the right to to and to in the right to right to change his nationality [end]
```

CYBERTEAM

NLP of a Low-Resource Language

(Igbo, one of the West African Languages)

- What we accomplished
  - –Developed diverse Corpora
  - –Executed Statistical, Machine Learning, Deep Learning approaches on various NLP applications
  - –Performed Translations
  - –Created Notebooks to easily run

Lessons Learned

- What went well?
  - –Developing Corpora
  - –Trying traditional NLP applications
  - –Understanding how a new to me language through NLP applications
  - –Get more tagged data

CAREERS CYBERTEAM

Lessons Learned

- What could we have done differently?
  - Have another student proficient in both English and Igbo for manual tagging of the corpus
  - Use a system that has higher computational power.

Publications/Contributions

Prepared for Publication:

Title: Building the Low-Resource Nigerian Igbo Language Corpora (2024)

- *Ready for Submission to IEEE*

CAREERS
CYBERTEAM

Publications/Contributions

1. Future Plan: Expand Corpora to Scraping Videos, more news, messages, songs,etc.
2. Papers plan to publish:
   - NLP applications
   - Translation

CAREERS CYBERTEAM

Thank you

CAREERS CYBERTEAM