Natural Language Processing

The Need for a Yoruba Corpus

Olamilekan Wahab

PYCON
NIGERIA 2018

The Agenda

Me

Motivation

Work Done/To be Done

Things That Can Go Wrong

Lessons

Questions

PYCON NIGERIA 2018

Who am I?

1. Machine Learning Engineer and Aspiring Machine Learning Researcher
2. Open Source Contributor
3. Huge fan of Saheed Osupa
4. Advocate for Open Machine Learning www.openml.org

Motivation

What led to this project and what is driving it?

NLP

Everybody is doing it

NLP

A little Story

How did I get drawn into this mess?

NLP Pipeline

STEP 1

Choose a corpus

STEP 2

Choose annotations(labels) to use

STEP 3

Choose or implement an NLP algorithm

```
import nltk
```

http://www.nltk.org/nltk_data/

Problem Statement

How do we build an extensive and standard corpus for the Yoruba Language?

Corpus

A source of language use example.

Electronic

Annotated

Balanced

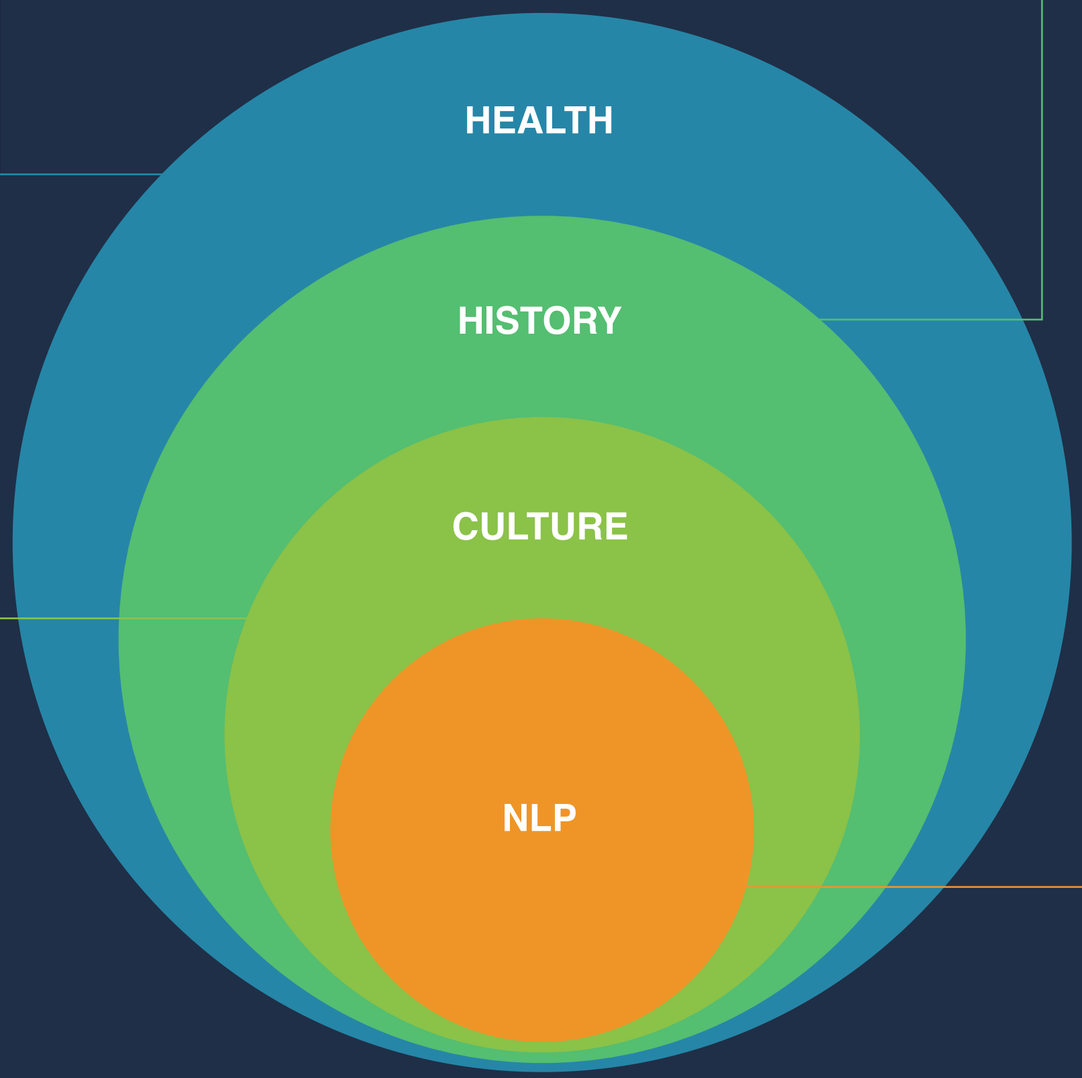
Unified

Work Done or To be Done

What has been done and what is yet to be?

7

10



18

2

Google

yoruba corpus

All Videos Images News Maps More Settings Tools

About 369,000 results (0.35 seconds)

Not Free

Yoruba text corpora – Sketch Engine
https://www.sketchengine.eu/user-guide/user-manual/corpora/.../yoruba-text-corpora/
Yoruba is one of the many languages whose text corpora are included in Sketch ... their own **Yoruba corpus** using the Sketch Engine's intuitive built-in tool.

Yoruba corpus (yoWaC) | Sketch Engine
https://www.sketchengine.eu/yowac-yoruba-corpus/
Search yoWaC, the 2.8-million-word **Yoruba corpus** of texts from the Yoruba national domain. Texts were cleaned and deduplicated. The corpus contains ...

Yoruba | Corpus linguistics
https://corplinguistics.wordpress.com/tag/yoruba/
Feb 8, 2012 - There are over two thousand African languages, spoken (in situ) by 15% of the world's population. In density of linguistic diversity it is rivaled ...

Resourceful

[PDF] **Digital Yorùbá Corpus - IJISET**
ijiset.com/vol2/v2s8/IJISET_V2_I8_122.pdf
by O Fagbolu - Cited by 2 - Related articles

https://corplinguistics.wordpress.com/tag/yoruba/

Yoruba Wikipedia

LDC Lexicon Database

ASP corpus

Google Internalization Corpus Crawler

Babatunde Obalalu

Kola Tunbosun

PYCON NIGERIA 2018

What We Have Done

KJV

A combination of manually translated and scrapped bible verses of the Pentateuch chapters.

Bible

Ife, Ibadan and Ede

Manual translations of orikis of my home town (Ede), Ibadan and Ijebu Ode

Oriki

Yoruba names

Scrapped Yoruba names from Kola Tubosun's project.

Yoruba Names

Saheed Osupa, Pasuma

Collaborated with kiosk disc sellers to work on getting some songs by SO and Pasuma written and manually translated to English.

Fuji Music

Usage Words

An average of 2000 words/ sentences on usage of Yoruba in different contexts.

ASP Corpus

Lexicology

A database containing lexical and morphological usage of the Yoruba language

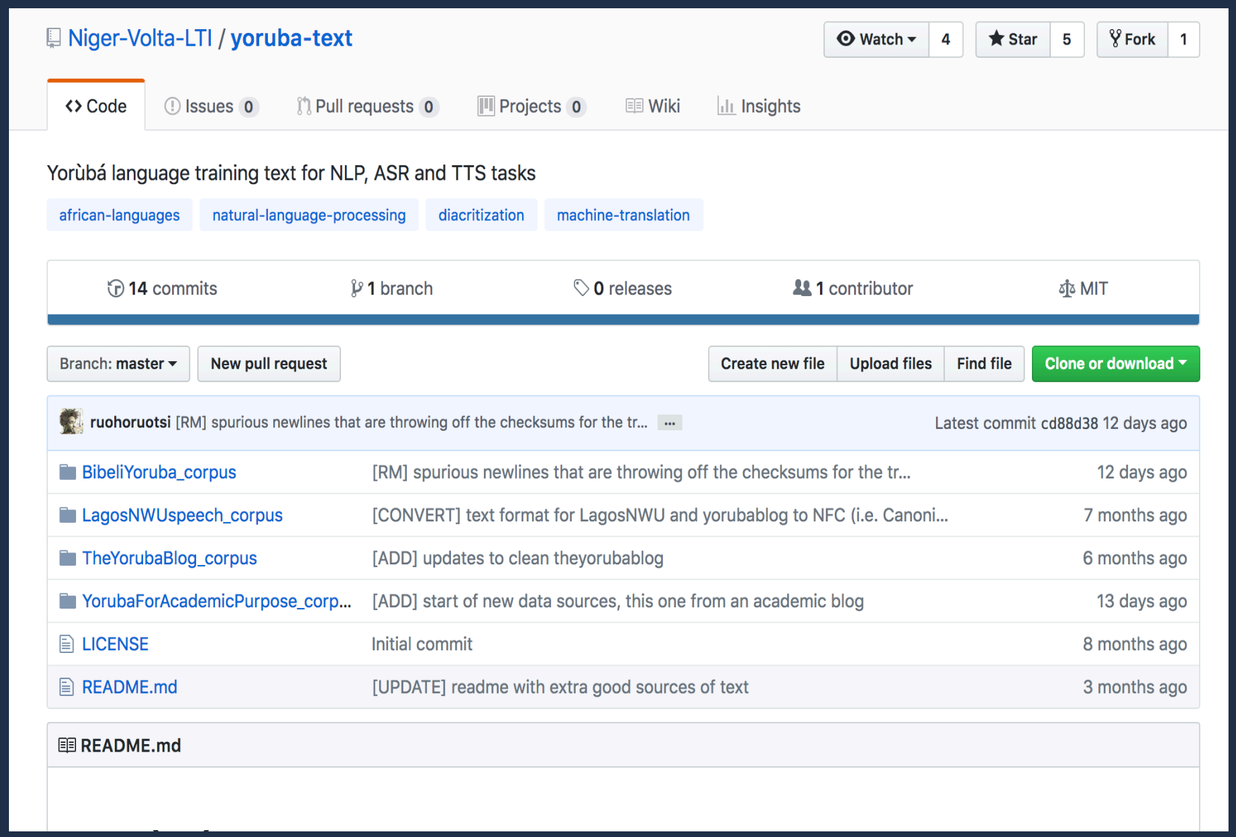
LDC Lexical Database

1. Talk to more linguists about the work, ask for advise and generally go more into academia for a solution.

2. Do more for open source. Try to bring more interested developers into the work and generally be more open.

3. Look more into existing (new) solutions.

4. More language support



PYCON NIGERIA 2018

What Can Go Wrong ?

It’s not an undocumented feature. It’s a bug

I. Scheduling
II. Adding new feeds
III. Synchronising feeds, finding duplicates
IV. Parsing different feeds/entries into a standard form
V. Monitoring

PYCON
NIGERIA 2018

Production-grade ingestion: Baleen

Configuration
+ logging
+ database
+ flags

Utilities
+ timez

Admin
+ ingest_feeds()
+ ingest_opml()
+ summary()
+ run()
+ export()

Logging
+ logger
+ mongolog

Ingest
+ feeds()
+ started()
+ finished()
+ process()
+ ingest()

Feed Sync
+ parse()
+ sync()
+ entries()

Post Wrangler
+ wrangle()
+ fetch()

<OPML />

OPML Reader
+ categories()
+ counts()
+ __iter__()
+ __len__()

Post
+ title
+ url
+ content
+ hash()
+ htmlize()

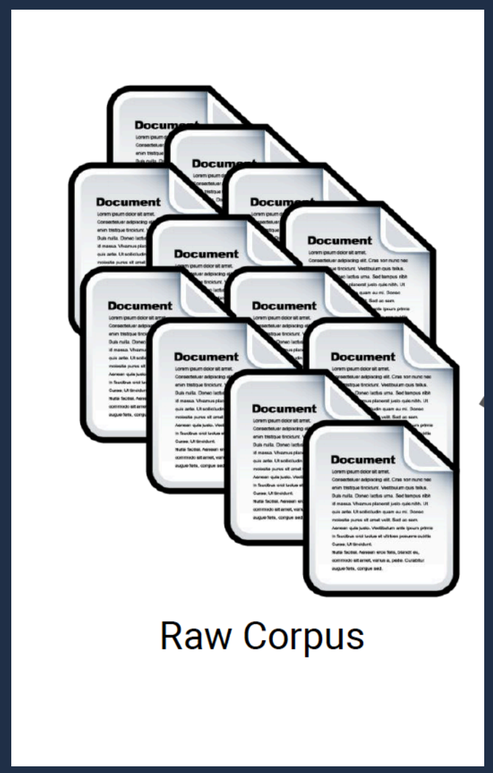
connect()

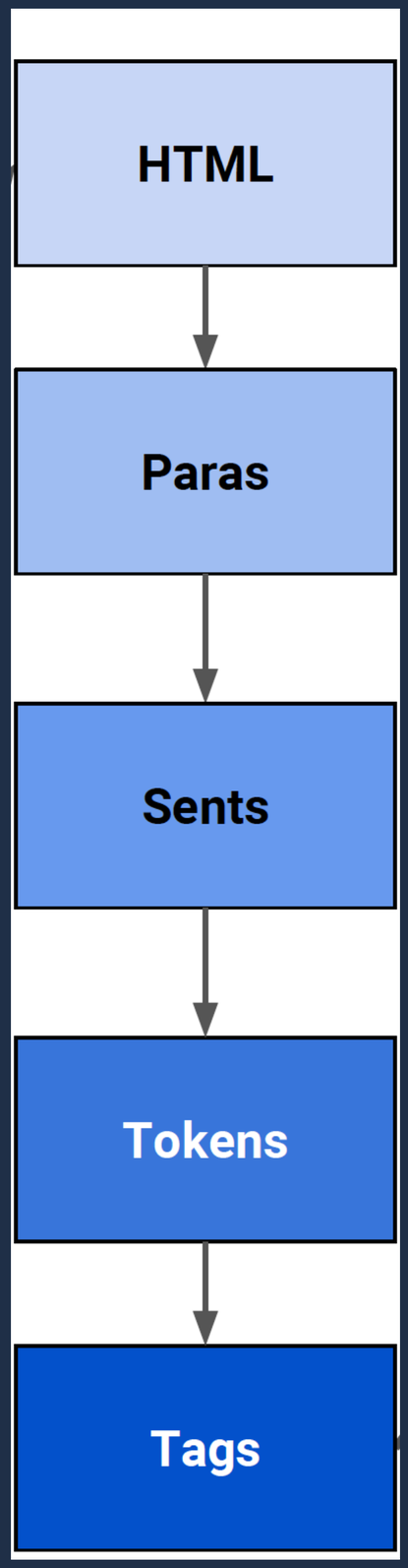
ingest()

Document

Exporter
+ export()
+ readme()

Feed
+ title
+ link
+ active

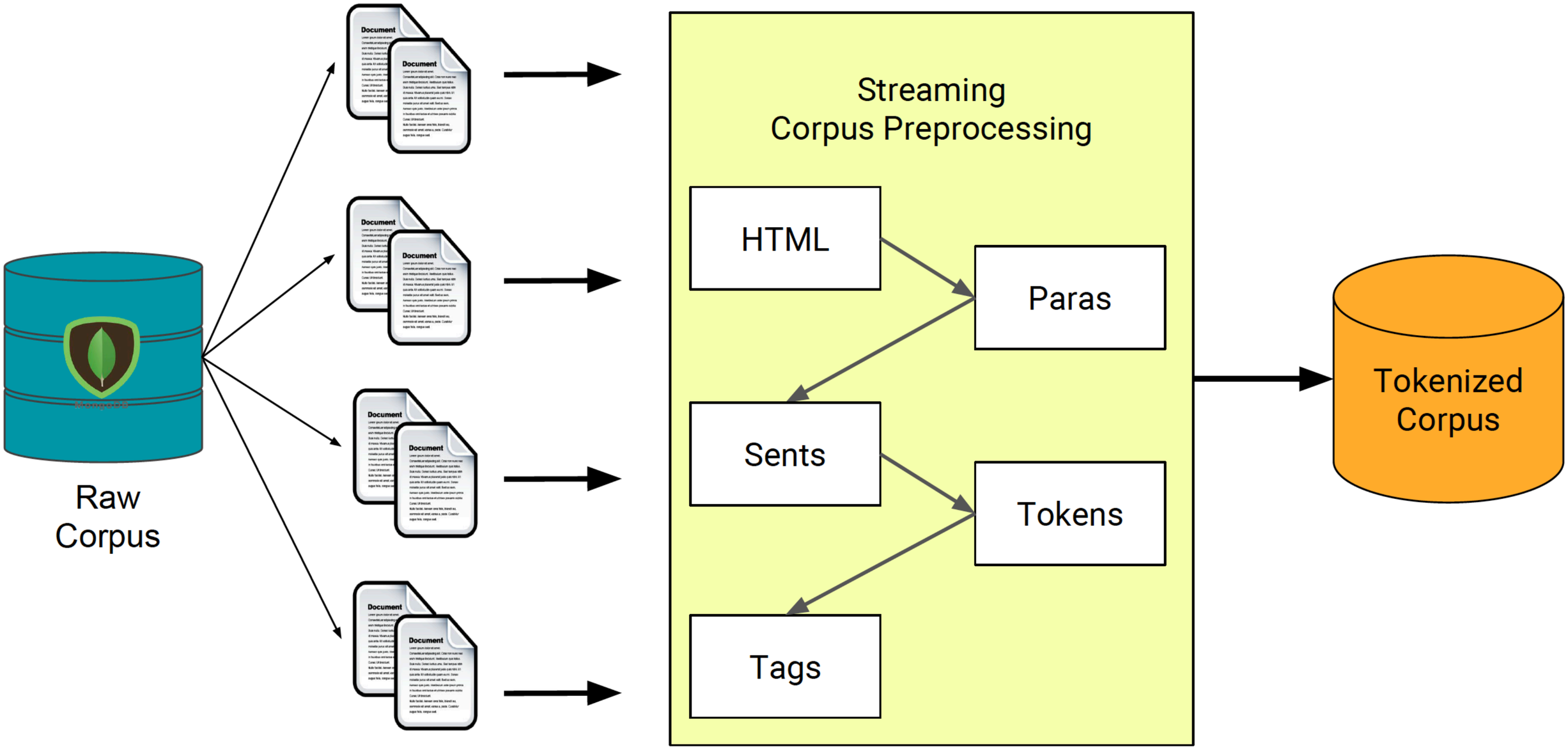


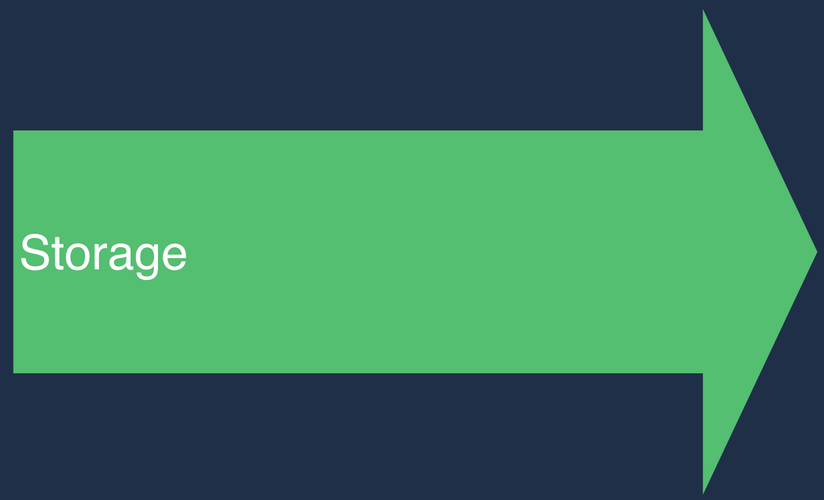


Corpus

Batch

PYCON NIGERIA 2018


Document
Document
Document
Document
Document
Document
Raw
Corpus
Streaming
Corpus Preprocessing
HTML
Paras
Sents
Tokens
Tags
Tokenized
Corpus



I. Database choice
II. Data representation, indexing, fetching
III. Connection and configuration
IV. Error tracking and handling
V. Exporting

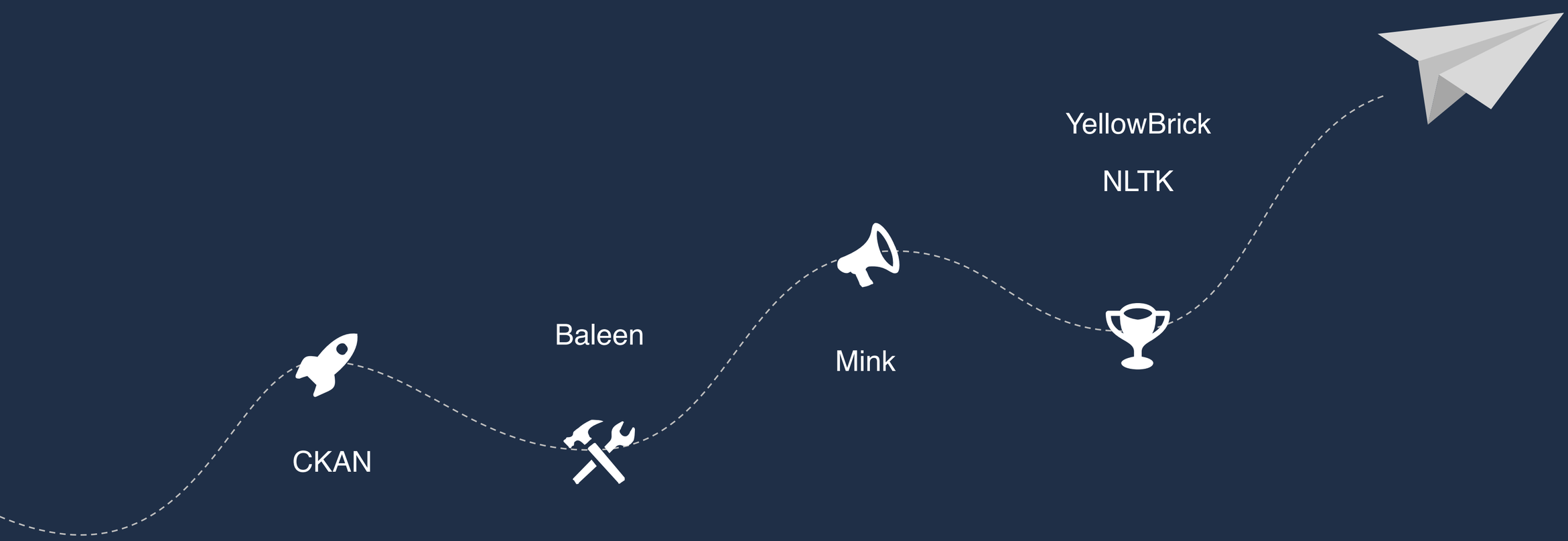
Lessons

What did we learn from all of this?

Lessons

1. Building a corpus is not an easy task.
2. Optimise for flexibility and easy iteration
3. Talk to people
4. Diversify

Tool set



thank you

Twitter : __olamilekan__

Github : olamyy

Email : olamyy53@gmail.com