Team Mukabidndi

Data Augmentation
English and Kinyarwanda

Team Mukabindi

By Arnauld Kayonga &
Floris Nzabakira

Topic

We chose to use **Religion related sources**.

Why?

- The most parallel translated topic in the world
- Content used by a lot of different people with different beliefs

Sources - Holy Scriptures

Title	English Link	Kinyarwanda Link
Bible	https://servervideos.hopto.org/XMLBible/EnglishKJBible.xml	https://servervideos.hopto.org/XMLBible/Kinyarwanda2001Bible.xml
Quran	https://quranenc.com/en/home/download/csv/english_saheeh Reference: QuranEnc.com	https://quranenc.com/en/home/download/csv/kinyarwanda_assoc Reference: QuranEnc.com

Sources - Ellen G. White Books - egwwritings.org

Title	English Link	Kinyarwanda Link
abakurambere n'abahanuzi	https://m.egwwritings.org/en/book/84.68#68	https://m.egwwritings.org/rw/book/13880.54#54
abahanuzi n'abami	https://m.egwwritings.org/en/book/88.33#33	https://m.egwwritings.org/rw/book/13882.23#23
intambara ikomeye	https://m.egwwritings.org/en/book/132.21#21	https://m.egwwritings.org/rw/book/13878.29#29
inama ku mirire n'ibyokurya	https://m.egwwritings.org/en/book/384.84#84	https://m.egwwritings.org/rw/book/13891.51#51
ibyakozwe n'intumwa	https://m.egwwritings.org/en/book/127.22#22	https://m.egwwritings.org/rw/book/13874.21#21
imibereho yejejwe	https://m.egwwritings.org/en/book/138.19#19	https://m.egwwritings.org/rw/book/13883.21#21
inkuru itera ibyiringiro	https://m.egwwritings.org/en/book/14211.14#14	https://m.egwwritings.org/rw/book/14367.15#15
inyandiko z'ibanz	https://m.egwwritings.org/en/book/28.23#23	https://m.egwwritings.org/rw/book/13885.29#29
uburezi	https://m.egwwritings.org/en/book/29.29#29	https://m.egwwritings.org/rw/book/13996.32#32
uwifuzwa ibihe byose	https://m.egwwritings.org/en/book/130.21#21	https://m.egwwritings.org/rw/book/13875.24#24

Extraction

Extracting Bible

- Manual download script manual as xml
- Used CodeBeautify to change into json, https://codebeautify.org/jsonviewer
- Parsed the bible by verse
- 30,872 parallel sentences Extracted

Extracting Quran

- Manually downloaded using referral link
- 6,235 Parallel Sentences

Extracting Ellen G. White Books

- Extracted links using Beautifulsoup
- 21,752 sentences extracted from the links
- Split larger paragraphs into smaller sentences

Cleaning

Our Process

- Remove Emails
- Remove Urls
- Remove Hashtags
- Remove Phone numbers
- Remove Script Tags
- Remove Numbers
- Remove Punctuation
- Normalisation
- Remove unicode characters
- Spelling

Remove Emails

Filter Criteria

- “****@***.***

Examples of filtered emails:

- admin@gmail.com
- admin @gmail.com
- admin@ gmail.com
- admin@gmail. com

Remove URLS

Filter Criteria

- “****@***.*** “

Sample filtered URLS:

- http://www.google.com
- https://www.google.com
- https: / /www.google.com
- http://www. google.com
- http://www.google .com
- http;//www.google.com

Remove Hashtags

Filter Criteria

- “#*****“

Sample filtered Hashtags:

- #admin
- #admin123
- #admin-123

Remove Phone

Sample filtered Phone:

- (123)-456-7890
- 1234567890

Remove Script Tags

Sample filtered Script Tags:

- <html>admin</html>
- < html>admin</html>
- <html >admin</ html >

Normalisation

- Universal lower casing
- Removal of numbers digit
- Removal of stop words

Spell Check

Comments:

Brick & Mortar spell check which takes a while to go through the entire corpus and quite ineffective since all the corpus has been hand translated and thoroughly checked by competing organisations

Deliverables and Conclusion

Deliverables

All the scripts are done using reusable codes in python

- Extraction is done using functions
- Cleaning is done in a single object class

Conclusion

Given that the final objective is to augment and clean the english to kinyarwanda dataset, these are the conclusions:

- Total lines extracted and cleaned amount to 58,860 sentences/paragraph corpus.
- Spell checker for english sentences is time consuming and ineffective since