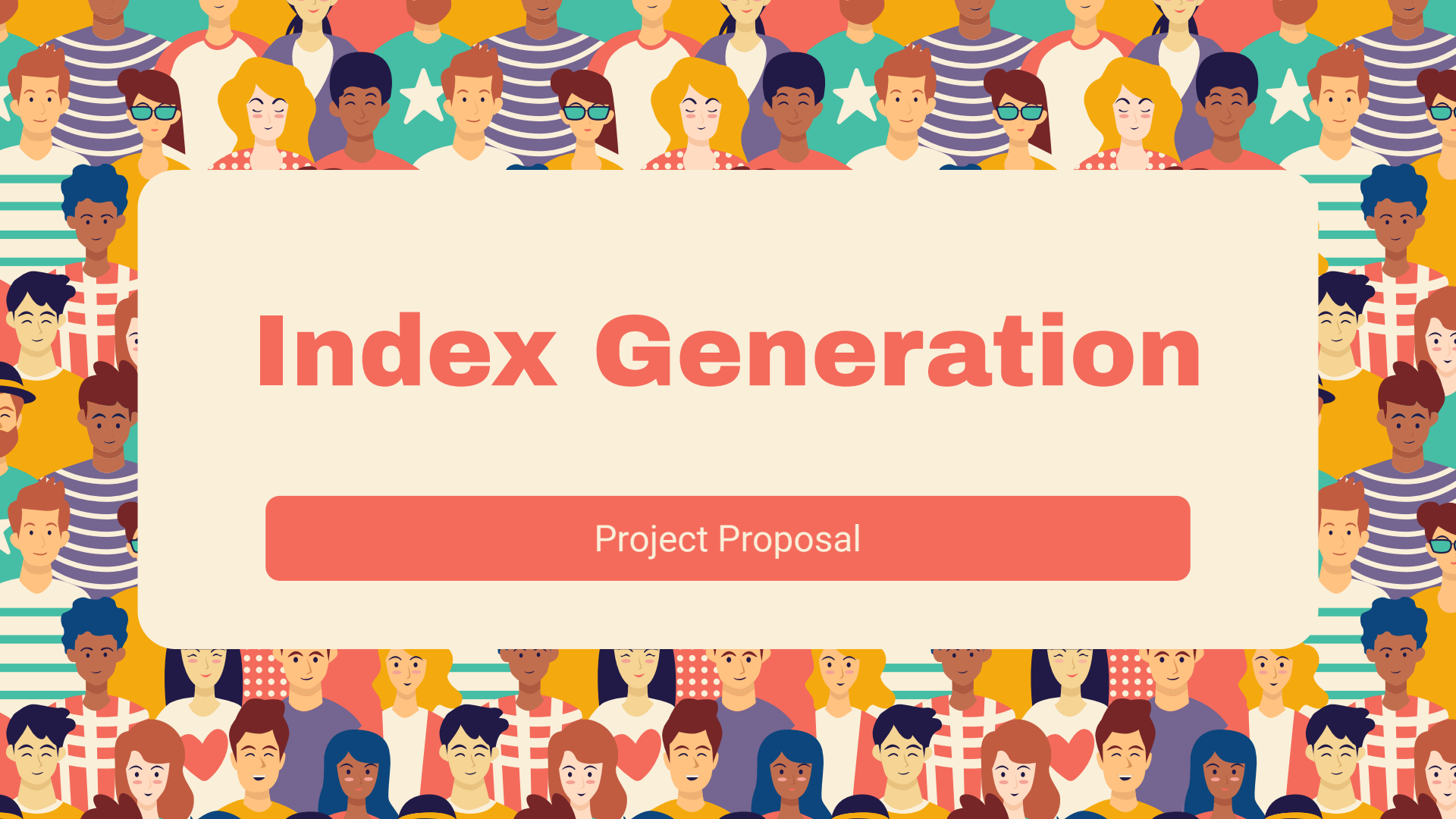

Index Generation
Project Proposal

Our team

201306 - Akash S P
201316 - Dinesh C
201320 - Harishankar S
201345 - Srinivasan K

Table of contents

Problem Statement

Index Generation for Venmurasu

Venmurasu (http://venmurasu.in) novel series has about 3.4 million words used. This project will create an index of the words and create a reference. Index all words and find stem of the indexed words.

Proposed Solution

1. Get sitemap of venmurasu website.
2. Scrape all pages mentioned in sitemap.
3. Use regular expressions to remove non-tamil unicode characters.
4. Index words using Tensorflow Tokenizer.
5. Stem all Tokenized words.

Tools used:

- Python 3
- Google's Tensorflow
- Google Colab
- BeautifulSoup scraping library

Technical challenges

1. **Getting URL of all pages of venmurasu site.**

   Solution: Obtain sitemap.xml file with contains all URL of all available pages.

2. **Extracting tamil-only words from the page.**

   Solution: Scrape individual pages using BeautifulSoup and remove all non-tamil character using regular expression.

3. **Getting unique words from extracted words.**

   Solution: Usage of Tensorflow.Keras Tokenizer.

4. **Unavailability of tamil stemmer in python nltk.**

   Solution: Implemented hand made functionality from scratch using rules of tamil grammar (prototype stage).

Technical challenges

1. **Scrapping all pages takes so much time.**

   Solution: Local file-level memoization implemented.

2. **Prioritizing functionality within stemming.**

   Solution: Working on it.

Proposal evolution

Stage 1 : Scraping

Stage 2: Extracting

Stage 3: Tokenizing and Indexing

Stage 4: Stemming

OBSERVATION AND RESULTS TILL NOW

- Scrapped 1936 pages of venmurasu site.
- Successfully indexed 4,49,239 unique tamil words.
- Implemented prototype functionality for stemming

Upcoming plans

1. Improvise Stemming algorithm.

2. Indexing resources other than Venmurasu site (e.g . Any Novel) to improve vocabulary.

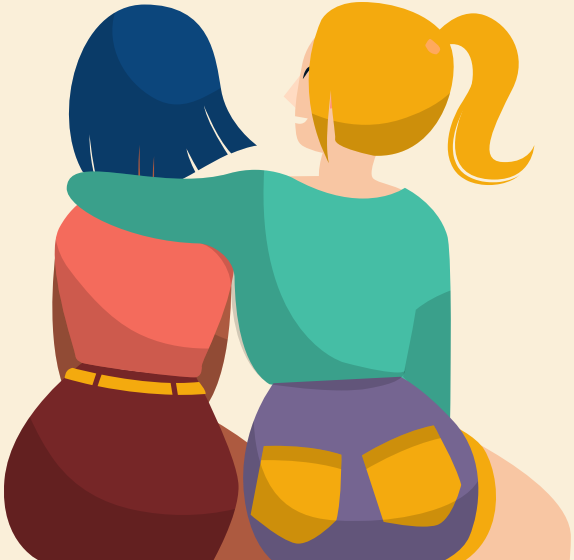
3. REST API based Website for live experimenting.

Resources

1. Venmurasu Website : venmurasu.in

2. Dr. Vairaprakash Gurusamy's research paper : www.ijcset.com/docs/IJCSET17-08-06-023.pdf

Thanks!



Credits: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik