MSSS Project Repository

Paul Hunt

February 17, 2022

1 Introduction

This is the repository for all documents pertaining to my MSSS research project. Unfortunately it was built for my own personal use, so I apologize for any lack of organization.

2 Repository Contents

2.1 Important/Currently in-use

- **MSSS_Project.Rmd** is the living up-to-date version of the project.
- **Aaj_tak1** is a directory containing text from articles scraped from Aaj Tak in December 2021.
- **BBC1** is a directory containing text from articles scraped from BBC Hindi in December 2021.
- **Getting AajTak Text.ipynb** is a python script for scraping text from articles listed on the front page of Aaj Tak and saving them by date.
- **Getting BBC Text.ipynb** is a python script for scraping text from articles listed on the front page of BBC and saving them by date.
- **perso-arabic_2008.pdf** A dictionary of loanwords in the Hindustani language (both Hindi and Urdu) I am working on using entries from this dictionary to identify Arabic loanwords.

2.2 Might be useful later

- **Aaj_tak** Contains text scraped from Aaj Tak after December 2021.
- **BBC** Contains text scraped from BBC after December 2021.
- **Jang** Contains articles from the Urdu-language newspaper Daily Jang. I might use these to compare Urdu to Hindi in terms of number of Arabic loanwords.

- **Getting Jang Text.ipynb** is a python script for scraping text from articles listed on the front page of Daily Jang and saving them by date.

2.3 Outdated stuff

- **S626_Project.Rmd** an early iteration of this project for the S626 course.

- **Dictionary.ipynb** A first attempt at using python to format my dictionary of Arabic words. It was not successful, and I have switched to R.

- **perso_arabic_loanwords.pdf** The first few pages of the dictionary used for testing my scripts.