CS 613 Assignment 2

Deadline: 30^{th} August 2021

1 Introduction

In the first assignment, each group was allocated a topic in the shared spreadsheet. Each member of the team was responsible for curating data from at least one subtopic. In this assignment, we shall annotate the dataset to conduct some meaningful ML experiments in the future. This paper from Google Research discusses in detail why good quality data creation (both curation and annotation) is important for progressing AI. The detailed description of the tasks for this assignment is as follows:

1.1 Sub-topic specific tasks

Each team member should select at least 100 random tweets (each Tweet should have at least five tokens after removing hashtags and mentions) from his/her curated corpus. Paste these random tweets in a spreadsheet and annotate the following for each tweet (create separate columns to answer the following questions in your spreadsheet).

1. For each selected tweet, mark if the Twitter assigned language tag is correct. If not, what is the correct language tag (see list of tags here)? In case the text mixes multiple languages, assign a combined tag by separating them using a hyphen. For example, if your Tweet text mixes Hindi (either in Devanagari or Roman) and English tokens, your first answer will be 'No' and the second answer would be 'Hi-En'. [5 points]

2. In case the Tweets are mixed ones. What is the main language (whose grammar is followed) and what is the embedded language/s (whose few tokens are embedded in the main language). For example, in the Tweet "items ko cart me daal ke app band kar dena is not funny", the main language is 'Hi' and embedded language is 'En'. [2 points]

1.2 Topic specific tasks

Now, the team has to combine the annotations from all teams members and answer the following questions:

1. What is the percentage of tags that were incorrect in question 1 in Section 1.1? What could be some of the possible reasons for this error? Any suggestions to improve the language identification? [3 points]

2. Do you think your data curation method (in assignment 1) was biased to some specific Indian languages? Why and why not? [2 points]

3. Can you give some examples of neologisms from your data? [2 points]

1.3 Statistical Analysis of Entire Curated Data

The next task is to conduct a statistical analysis of the entire curated data (not just the annotated data). The following are the tasks:

1. Create a frequency list of tokens present in the dataset in the decreasing order. What are the top-20 words? What is the percentage of stop-words in the top-20 rank list? [3 points]

2. Does the data follow heap's law? Show by plotting $|V|$ vs N. What are the values of curve fitting parameters K and β? [3 points]

2 Submission details

Please submit your answer script (PDF) here. The answer script should contain the answers to the above questions and links (from your Google Drive) for the spreadsheet files (one for each team member to answer Section 1.1).