Demographics and Sociolinguistics

Astrid Machholm, astma@itu.dk
Christoffer Ebbe Sommerlund, csom@itu.dk
Gustav Bakhauge, guba@itu.dk
Kasper Grønbæk, kgro@itu.dk
Sarah Dueholm Ramezanpour, sarr@itu.dk

28th of February 2020

Contents

1 Introduction

This report present an analysis performed on a data set containing both text material from many different authors along with their demographic information. By applying NLP (Natural Language Processing) to our text data, we can say something about sociolinguistics, which is also supported by statistic calculations.

2 Methodology

Cleaning data
Shuffling the data by rows
Size of test and train
Preparation of data for specific model or general (vocabulary, vectorization, n-grams)
Choice of model (nltk or SKlearn or something else) The math behind it all

3 Data

Our data set contains tweets about new year resolutions from the dates 31/12/2014 and 01/01/2015. The authors of the tweets are both female and male and from all over USA. Their location is split into three categories (city, state, and region) from which we have chosen [the most frequent states/region].

4 Results

5 Interpretation

6 Error analysis

7 Concluding remarks and future work

8 Disclosure statement