First year project
Mini-project: Demographics and Sociolinguistics

1. Goals of the project

With this project, you will perform a segmentation of texts from many authors while sharpening your programming, statistics, NLP and management skills. The segmentation will be based on external factors, like author age, or the sentiment expressed in the text. Sub-parts of this project will include:

- Choosing and collecting an adequate text collection for the task
- Use your existing NLP skills and new stastical / machine learning skills to split the data

2. Requirements

In addition to the requirement stated in the course description, you must work in ITU github, and you must provide a notebook that runs in Google Colab.

3. Assignment: Demographic Classification

You must find out who the major populations are that have written a set of texts. In extracting and describing this information, you might like to consider some of the following dimensions. These might be thought of as scalar (i.e. on a sliding scale) or class-based (i.e. a fixed number of distinct options) – that's up to you. You must have at least three different factors.

(a) Age
(b) location
(c) gender
(d) sentiment

In your analysis, you should go one step further. You should be able to describe, based on your data alone:

(a) what terms (e.g. unigrams and bigrams) are strong identifiers of each factor
(b) what the major groups are across these dimensions (e.g. negative sentiment/old/men)

Methodologically, you must:

(a) find adequate text sources to use for this (Kaggle might be a good starting point)
(b) pre-process the data to allow analysis
(c) extract relevant terms and analyse the data
(d) create a clearly and cleanly communicated presentation of the major findings

4. Hand-in and presentation

You must hand in:
- The github log (text or PDF)
- Python or Notebook result – it should be tested and run in Google Colab
- A project report

The project report is specified as follows:
Project write-up. Your hand-in should be no longer than 3 pages (with 1.5 cm margins and

11 pt font size), and should consist of precisely the following sections.

1. Introduction: Here you provide the context for the problem and re-state the brief.
2. Methodology: Here you define and describe your methods, with precise mathematics where applicable.
3. Data: Here you describe your data sources, and briefly summarize how you obtained them without referring to any code.
4. Results: Here you provide the technical results of your method over the data. Any tables with numerical results should be put in this section.
5. Interpretation: Here you interpret the results back into the original question's setting.
6. Error analysis: Here you give an account on some major short-comings of your methodology/data. (This does not include anything to do with group dynamics. This is inherent to the data and methods.)
7. Concluding remarks and future work: Here you provide a couple of sentences summarizing the results of the project and indicating how the methods/data could be improved. The improvements have to do with information/method considerations beyond your own group dynamics.
8. Disclosure statement: Here you may state if there were any serious unequal workloads among group members.

An important final note: Your hand-in should be self-contained. You are required to master your code for the oral exam. Your 10 minute oral presentation should correspond to the structure of your write-up. However, you are encouraged to have slide headings that are more communicative.