**Gender of phrases dataset**

1. **Intro**

The objective of this dataset is the creation of a ground truth for NLP models. To do so the dataset is composed by a set of manipulated phrases where each one has either a female and a male subject or a all female/male universe (every subject/object of the phrase are female or male).

To chose the original phrases, they should be of similar length to avoid biases that may be created by the length (I ended up not considering this). The phrases should also be in the 3rd person of the singular. The original phrases were obtained from entries of Wikipedia related to the top 100 books of the day 17/3/2022 of the Gutemberg project. I collected the entire Wikipedia page of these books but only considered phrases related to the plot of the story.

1. **Getting the URLs**

These pages were obtained using the package Selenium.

* Firstly the [*https://www.gutenberg.org/browse/scores/top*](https://www.gutenberg.org/browse/scores/top)was acceced and the list of the books were collected.
* I saved (pickle) both the list of books with and without the author (the author sometimes is useful because the name of the book may be a movie for example).

1. **Getting the Wikipedia entries**

* the url used for the Wikipedia pages was a google search with the ‘https://www.google.com/search?q="en.wikipedia.org"+’ to which I added the name of the book+author separated by + (replaces the white spaces with +)
  + ex: [https://www.google.com/search?q="en.wikipedia.org"+Thus+Spake+Zarathustra+A+Book+for+All+and+None+by+Friedrich+Wilhelm+Nietzsche](https://www.google.com/search?q=%22en.wikipedia.org%22+Thus+Spake+Zarathustra+A+Book+for+All+and+None+by+Friedrich+Wilhelm+Nietzsche)'
  + these urls were saved in a list
* using these urls And selenium I collect all paragraphs of the Wikipedia pages (<p> on html). These pragraphs are then appended to a list.

Note: to access these urls firstly we need to get to google (and accept the cookies and sometimes it asks if I’m a robot, and I need to manually complete the reCAPTCHA)

I also deleted the wikipedia references that appear on the middle of the texts, before saving them into a pickle file -> raw\_texts.pkl.

1. **From Wikipedia paragraphs to SpaCy sentences**

This document has 5000+ paragraphs. And each of them is separated into phrases using the small model of the spacy library. From here I get 19500+ phrases (I also tried to split using ‘.’ But with the spacy model each token and phrase has more information).

This information can be quite useful. Some of these are:

* dep\_ - which I used to get the ROOT verbs of a phrase – the verb that the model considers the main verb in each phrase.
* morph – which I used to get the person and number of verbs in the phrases (I only considered phrases where the verb was in the third person of the singular).
* children – which I used to find the subject of related to the ROOT verbv (which would be the main subject of the phrase). – Nota: this returns a generator- turn into list.

1. **Selecting phrases which have the ROOT verb in the 3rd person singular**

After using these information to obtain only the phrases that have the ROOT verb in the 3rd person of the singular, I have 7800+ phrases.

1. **keeping only smaller sentences**

Some of these phrases are big and complex. Large sentences will be harder to learn for the computer (they are harder to understand for humans as well). Some guidelines indicate that phrases with 15-20 words are clearer.

https://techcomm.nz/Story?Action=View&Story\_id=106#:~:text=A%20common%20plain%20English%20guideline,2009%3B%20Vincent%2C%202014) -> (Cutts, 2009; Plain English Campaign, 2015; Plain Language Association InterNational, 2015)(Cutts, 2009; Vincent, 2014).

Because of this, I excluded phrases that had more than 30 tokens (some tokens are punctuation so I left a bit of a buffer). This step is optional, but it removes almost 3000 sentences.

1. **Deleting phrases that didn't end in '.'**

I also deleted some phrases that didn’t end with a ‘.’ Because these phrases would not finish a thought (some phrases are not well obtained by the spacy model – more complex models give more accurate phrase delimitation).

1. **Remove sentences where the subject is 'it'**

I then removed the phrases where the subject was ‘it’, since this is also a 3rd person/singular but it is not interesting for the dataset because it is neutral.

1. **Selecting "allowed" sujects (nouns, pronouns, and proper nouns)**

I consider that the only subjects allowed are:

* proper nouns – selected using token.pos\_ == 'PROPN'
* pronouns he and she – selected using token.morph.get('PronType')==['Prs] and (token.morph.get('Gender')==['Masc'] or token.morph.get('Gender')==['Fem']
* common names (some of which are useful such as father or sister, some are proper names that were misclassified and some are not useful, this needs to be checked by hand).

I did this selection printing the list of subjects in each group (9.1). I also deleted subjects that only appeared once (in particular in the proper names group) because most of the times these are not characters of the book but maybe a researcher that said something about the book (9.2).

Other proper names that I deleted were author’s names (this was done both in this section and in the selection of sentences (point 12)

1. **Deleting duplicated phrases**

There were some phrases duplicated that I couldn’t delete checking if the text was the same in two phrases and delete one of them.

1. **Highlighting of root verb and subject**

I also created a function that allow me to see the subject and verb selected by spacy so it is easier to see if the phrase is or not useful. This helped me chose by hand in each line the phrases that mattered.

1. **Selection of useful phrases**

This selection is done by hand and the selected sentences should:

* be part of the plot and not someone unrelated to the plot speaking about it
  + considerations about the author/book
* not contain citations ('this character said "this"')
  + some contain " "
  + some contain you/me
* phrases starting with -" Chapter x - ..."
* errors
  + in phrases (the Rachel, wrong words (81 from PROPN))
  + phrases starting with lower case (don't have the entire thought)
  + starting with numbers ("(number)")
* confusing phrases that usually have a '-'

In this part I added both the useful and the not useful phrases to lists so that they can both be retrieved.

1. **Saving the final phrases**

Lastly I pickled the useful phrases for later use.