Data Scraping: Lexical graduation in the word 'mito' in YouTube Users Comments

Rodrigo Esteves de Lima-Lopes
State University of Campinas | rll307@unicamp.br
Carolina Palma de Sousa Arruda
State University of Campinas | carolpalma203@gmail.com

Contents

1 Introduction

This script was developed for the analysis of Portuguese. We hope it helps colleagues in the LC area and popularize the use of R. It is part of our research project developed with CNPQ. Please drop us a line if you have any doubts or need any help.

1.1 Purpose of this repository

This repository brings the scripts used for lexical analysis of Brazilian Portuguese YouTube comments we wrote for our paper:

> Lima-Lopes, R. E. de and Arruda, C. P. S. (2020) Strategies for Gradation in YouTube comments. (under review)

This script is specifically about data scraping.

2 Packages

We are going to use the tuber package for data scraping (YouTube). In order to scrape data from YouTube you will probably need a Goggle developer account. It has more privileges and allows you to scrape all data necessary. However, dealing with any API is not a bed of roses and some configurations are not that obvious. This tutorial might bring a bit of help, but be prepared for some try and error.

3 Getting the comments

3.1 Setting the environment

The chunk bellow shows how to setup the environment to log on YouTube's API and scrape all comments.

```
library(tuber)
app_id <- "your app_if"
app_secret <- "your app_secret"
API_KEY <- "your app_secret"
yt_oauth(app_id, app_secret,token = "")
```

Important to observe that each of these variables will be given to you as you configure your Goggle developer account.

3.2 Downloading the data

Two commands will be necessary to download the comments:

```
statistics <- get_stats(video_id = "YOWJSpPcCqc")
comments <- get_all_comments(video_id="YOWJSpPcCqc")
```

- `get_stats` will download general data regarding the videos
- `get_all_comments` will download the video comments and save them in a data frame. Some other information is also made available, such as *author URL*, *ID*, *Display name* and many others. Please note that `YOWJSpPcCqc` is the actual ID for the video we studied.