Dealing with Big Data in the Humanities

Eileen Waegemaekers (Department of Linguistics)

1 Objective

Teach RPG students how to analyze and extract relevant data from big datasets to inform their own theories and hypotheses.

2 Content

Recent advances in technology have made huge amounts of data available and companies are using these to extract information valuable for their businesses (think, online clicking behaviour, suggested tunes on Spotify, prediction of medical conditions based on symptoms etc.). A large part of this data is linguistic data and with the right tools, it is possible to extract information that cannot only be used for commercial use but can also inform linguistic theories or test theoretically grounded hypotheses. Although students are usually aware of corpus data that can inform their research, they lack the tools to automatically extract information from these datasets. For example, if a linguistics student is interested in the acquisition of adjectives, it would be helpful to create a big dataset from several corpora with all occurrences of adjectives and their immediate context, timestamped for the age at which they occurred. Problematic is usually that formatting varies in different corpora which makes it hard to merge datasets. If the processing had to be done manually, it would take ages. In this two-day workshop, I am introducing the first steps for students to create a big dataset automatically, organize them according to their needs and perform analysis to identify clustering in the data. The two clustering techniques Principle Component Analysis and t-SNE will be introduced for this purpose. The main goal is to let students get acquainted with big data, tailored to their own research needs.

3 Learning outcomes

After this course, students will have basic tools to access and create large datasets, reorganize them for their purposes and do some simple analyses. They will be able to run python scripts and use the python package NLP to get linguistic information (e.g., frequency of words, parts of speech, constituents) from the data. They will also be able to perform Principle Component Analysis (PCA) and t-SNE analysis in R to find clusters in the datasets.

4 Activities

This is a two-day workshop of a total of 6 hours. Most activities are performed in pairs, although students are encouraged to individually run all scripts as it will enable them to reuse those same scripts for their own purposes. Discussion in class is highly encouraged and most of the activities will be hands-on, practical tasks.

day 1 (3 hours) In the first hour I will introduce applications of big data with a specific focus on scientific and linguistic uses. Students are invited to brainstorm about datasets that could inform their own research and think of possible patterns they would wish to identify. As the main analysis in the end carried out on the datasets is aimed at finding clusters in the data, the students' hypotheses should be centered around clustering. The second hour is used to let students in groups of two create their own big datasets from online corpora based on their own interests. A list of corpora and where they can be downloaded will be provided to the students. Additionally, scripts are provided to manipulate the data and let the students get

acquainted with running scripts. For example, students learn to remove empty lines from text files, to remove the first word from all lines in a text file and extract specific words or phrases from a text file. In the third hour we practice how to make small changes to the existing scripts and use the NLP package in python to get frequency information, tokenize and parse sentences. Again, preexisting scripts are provided, but they are all easily adaptable.

day 2 (3 hours) The second day is concerned with finding patterns and clusters in the data. The first hour is dedicated to explaining PCA and t-SNE. The following website 'http://distill.pub /2016/misread-tsne/' will be used to introduce t-SNE, as it clearly describes the advantages and disadvantages of the method and uses video animations that will enhance understanding. Students are encouraged to discuss the differences between the two methods. In the second hour we transform the datasets to csv files that can be imported into R and make sure the data is clean and ready to be analyzed. Students will learn how to remove missing values from their datasets and will learn how to plot their data using labels. Labeling of datapoints in a plot is a great way to explore your datasets. Finally, we will perform a PCA and t-SNE and interpret the results for the different datasets.

5 Equipment and programme requirements

Students should bring their own laptops and have RStudio installed. For students with Windows laptops, please also install **git bash**.

RStudio: *https://www.rstudio.com/products/rstudio/download/*
Git Bash: *https://git-for-windows.github.io/*