Twittercorp. A python script for creating and distributing location-aware Twitter corpora.

Tom Ruette[*]

Manuscript of November 14, 2013[†]

Abstract

Corpora of Computer-mediated Communication allow for a perspective on written public language production. Also, there is a (questionable?) tendency to consider such corpora also as indicative of characteristics of spontaneous language use, due to its communicative and immediate character. With the institutionalization of online communication – as an example, many blogs are professionally written with the goal of making money and many twitter feeds are produced by companies or advertisers – the status of such corpora may need to be redefined. The youngest member on the list of Computer-mediated Communication protocols is Twitter. This whitepaper describes a python implementation for collecting a location-aware Twitter corpus. The possibility to automatically annotate tweets with the location affiliation of the Twitter user makes such collections a valuable tool for dialectologists. Also, given the strict Terms of Service of Twitter, a method was implemented to distribute Twitter corpora without violating these terms. Finally, the script contains aa search method that supports regular expressions.

1 Introduction

Corpora of Computer-mediated Communication allow for a perspective on written public language production. Also, there is a (questionable?) tension to consider such corpora also as indicative of characteristics of spontaneous language use, due to its communicative and immediate character. With the institutionalization of online communication – as an example, many blogs are professionally written with the goal of making money and many twitter feeds are produced by companies or advertisers – the status of such corpora may need to be redefined.

In order to investigate the language use of Computer-mediated Communication, it is necessary to have access to the linguistic production of this medium. Therefore, I implemented an easy to use python script to compile a Twitter corpus in a highly controlled way, so that it meets the highest corpus linguistic standards. Moreover, since science is a business of verification and falsification, a method is provided to distribute the compiled corpora (or compile a comparable one) without violating the Twitter Terms of Service. As such, research

[*] KU Leuven

[†] The development of this script has been a collaborative effort. A list of contributing developers can be found at GitHub (https://github.com/ruettet/twittercorp/graphs/contributors).

on the basis of a Twitter corpus that is compiled with the here described script should also comply with perhaps the most important requierement of science, i.e. data transparancy.

After running the script – which may take several weeks – a controlled corpus of Twitter messages is presented in an XML format, which contains per tweet the ID of the tweet, its Twitter user, the reported location of the Twitter user, a normalized location, the GPS coordinates of this normalized location, the timestamp, and obviously the text of the tweet itself. How to get to such a corpus is explained in detail in the following sections.

2 Quick user guide

3 Methodological steps

3.1 Recursive finding of users

3.2 Location normalization

3.3 XML creation

4 Distributing the corpus

5 Searching the corpus

6 Obtaining the script