Harvesting, Processing and Visualising Geo-Encoded Data from Social Media

Nikola Ljubešić
Department of Knowledge Technologies
“Jožef Stefan” Institute, Ljubljana

Aims of this tutorial

1. Understand APIs
2. Learn how to harvest data via APIs (hands-on)
3. Process harvested content (hands-on)
4. Perform downstream analyses, inferences and visualisations

Data harvesting from social media

All social media sites have open APIs

APIs (application programming interfaces) - interfaces for various processes / programs to interact

- Twitter
- Mobile app (Instagram) used for posting on Twitter

- Twitter
- Your program for data harvesting (TweetCat)

API examples

Twitter API

- Search API - enables your program to query for specific keywords
- Streaming API - sends part of the currently published content to your program

Facebook Graph API

- Enables updating or reading the “social graph” via object IDs
  - Nodes - User, Photo, Page, Post, Comment
  - Edges - connections between nodes (Post and Comment)
  - Fields - attributes of nodes, such as User’s birthday

Using APIs

Communication

1. Via HTTP with cURL (command line tool), urllib (Python library)
2. Libraries that wrap the HTTP communication
   - **tweepy** for Python and the Twitter API
   - **facebook-sdk** for Python and the Facebook Graph API
3. Tools that perform specific tasks
   - **TweetCat** for gathering tweets of low-frequency languages or published on specific locations

Authentication

- Each API requires you to authenticate with a series of tokens
- Those tokens can be obtained from the social media (https://apps.twitter.com)

TweetCat

Modular command-line tool / set of scripts written (mostly) in Python

- Harvesting Twitter data either via seed terms or from geographical perimeters
  - Python
  - output arrays of JSON objects
- Extracting variables from the harvested data (text, metadata, variables from text)
  - Python
  - output CSV file
- Analysis, inference, visualisation, performed in R (or other tool of choice)

https://github.com/clarinsi/tweetcat

Data harvesting

Data harvesting

Two basic modes for data harvesting

1. LANG mode
    a. Want to collect data published in a specific language
    b. User input
        i. Seed words (used for querying the Search API)
        ii. Languages of interest (langid.py dependency)
2. GEO mode
    a. Want to collect geo-encoded data published in a geographical perimeter
    b. User input
        i. Geographical perimeter (used for listening on the Streaming API)
        ii. Languages of interest for potential filtering

Hands on…

Prerequisites

- Python2.7
- tweepy module
- langid module
- access tokens

Data sharing

Defined? by the Developer Agreement / Terms of Service.

Twitter

You can share user and tweet IDs that can be used to recollect data from the API.

You can publish up to 50k public tweets directly.

What to do when the data is linguistically annotated? https://github.com/clarinsi/tweetpub

Facebook?

As long as you do not sell the data and the data is public, you can share the harvested data.

Variable extraction

Variable extraction

Four variable extraction levels

1. Extraction from the Status object (metadata)
2. Extraction from original text
3. Extraction from lowercased text
4. Extraction from normalised text

Two text extraction principles

1. Lexicon-based (list of words mapped to variable values)
2. Regex-based (regular expressions mapped to variable values)

Hands on...

Data analysis

R

Language variation analysis in BCMS

Bosnian, Croatian, Montenegrin, Serbian

Analyse 16 linguistic variables known to vary between variants

Focus on the linguistic strength of administrative borders - has the continuum been interrupted?

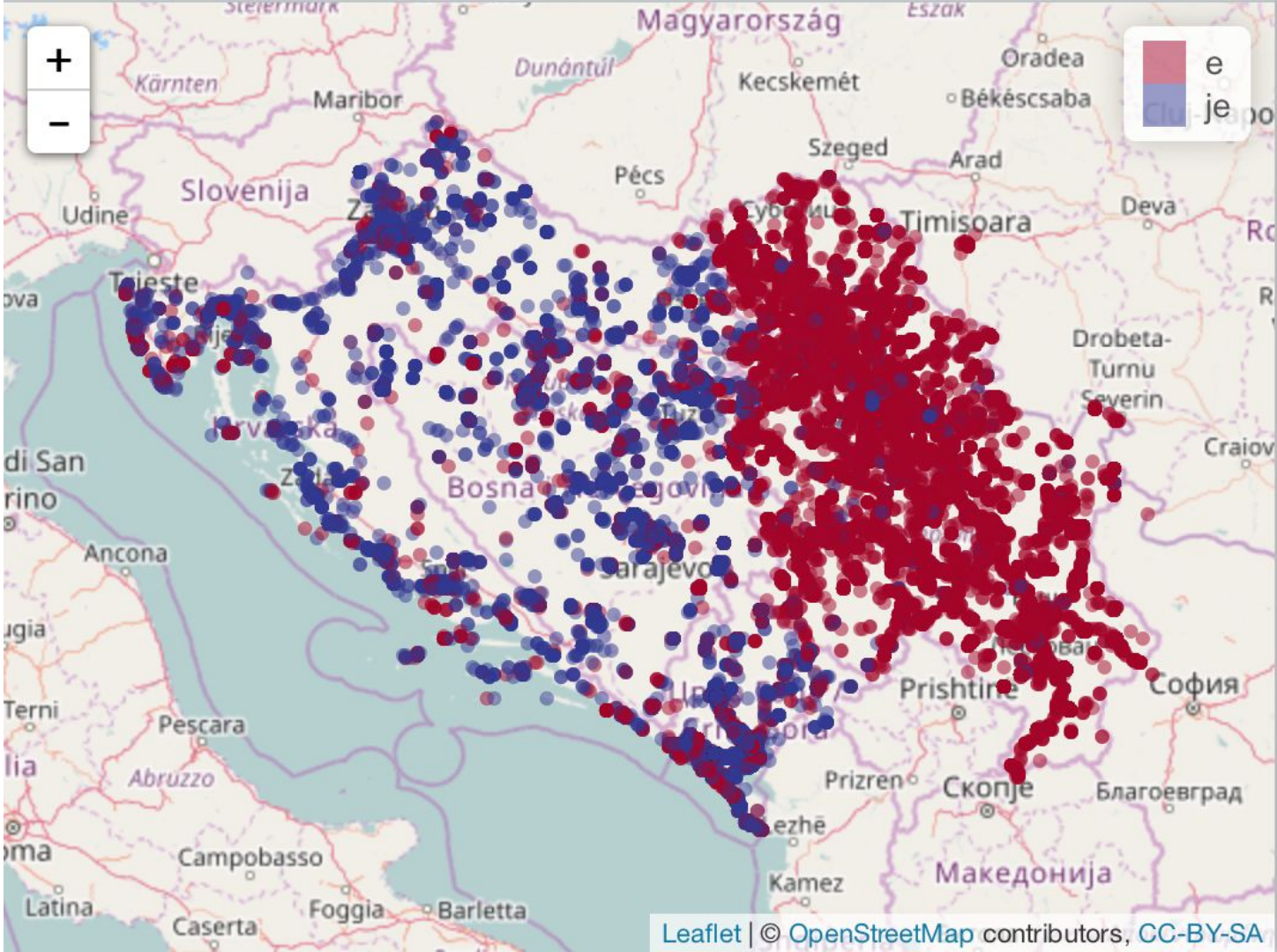
Tweets collected since 2013 (200 million tweets, 10 million geo-encoded, 1 million with relevant linguistic variables)

lep
lijep

mleko
mlijeko

smeh
smijeh

devojka
djevojka



lep
lijep

mleko
mlijeko

smeh
smijeh

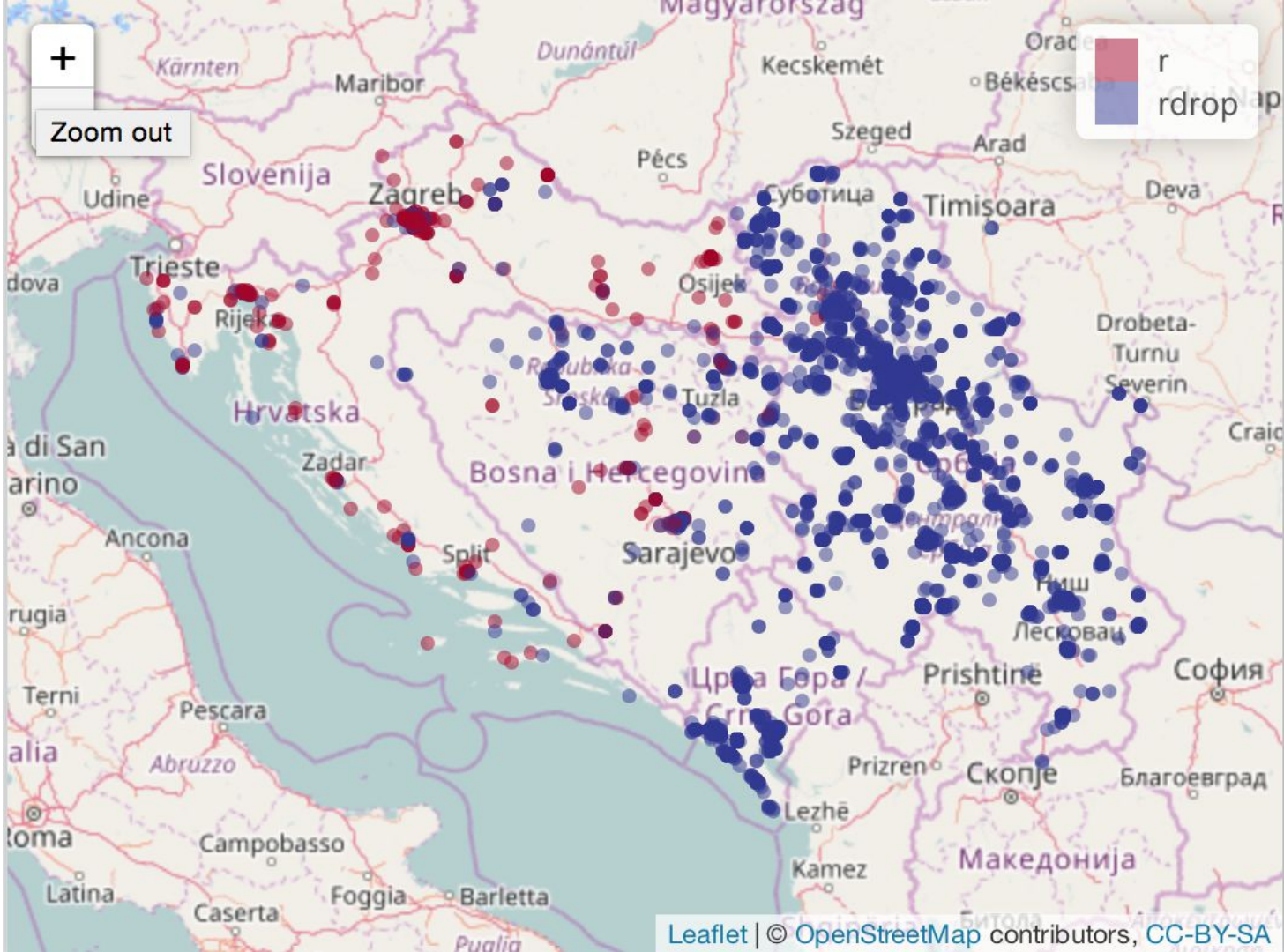
devojka
djevojka

jučer
juče

večer
veče

također
također

navečer
naveče

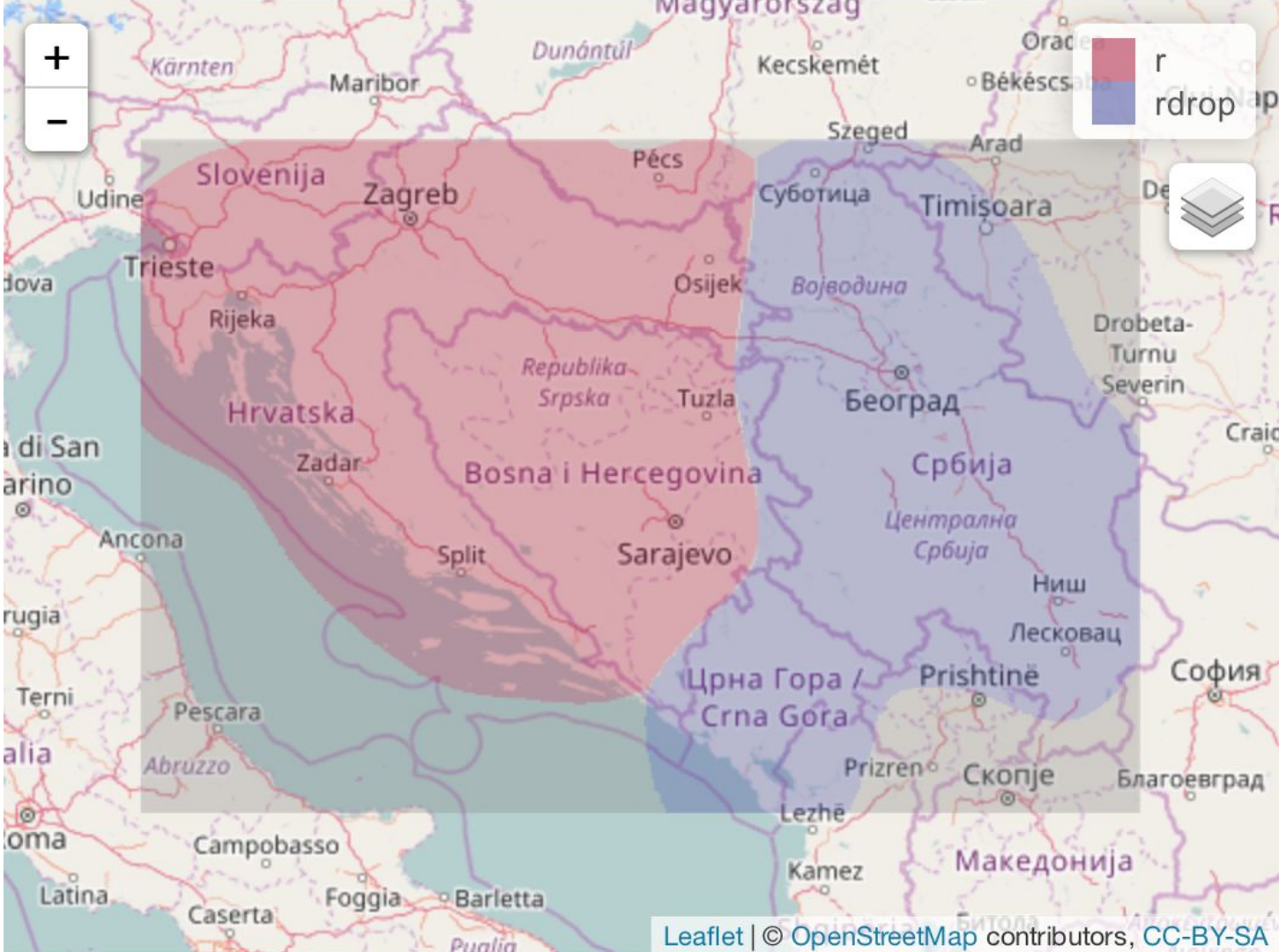


jučer
juče

večer
veče

također
također

navečer
naveče



Harvesting, Processing and Visualising Geo-Encoded Data from Social Media

Nikola Ljubešić
Department of Knowledge Technologies
“Jožef Stefan” Institute, Ljubljana