Network science lab #1: intro to Projects

Lejla Džanko
lejla.dzanko@studenti.unipd.it

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

About me

Lejla Džanko

Bachelor degree: Computer Science and Informatics,

Faculty of Electrical Engineering, University of Sarajevo

Master degree: ICT for Internet and Multimedia, UniPD

PhD Student at SWPS University in Warsaw, Poland

Contact me at (after reading the next slide):

lejla.dzanko@studenti.unipd.it

Rules for communication

- Any question about the class material or your own work should be asked on the students' forum! (**check for duplicates**)
- I strongly encourage you to answer each others' questions
- if no one answers in a reasonable amount of time - I will
- sometimes the answer is in the materials or can be found with a quick Google search (documentation & Stack Overflow are your friends)
- if you leave your work for September I'll try to help you but ca⁀̈ promise I'll be available

Outline

1. About projects
2. Learn how to access Twitter
3. Twitter API:
   a. overview of usage (i.e what can we achieve with it)
   b. tiers: possibilities, limitations, rate limits
   c. tips & tricks
4. Lab assignment

About the projects

- you get to: apply network science knowledge + learn some new coding skills + try an interdisciplinary approach to research + make friends = projects are FUN!

- what students think they need for a good project: excellent grasp of all the network science concepts, mad Python skills, tons of data, ...

- what you really need: a cool project idea *(please don't do the movies database),* creativity, innovation, readiness to learn, intuition to interpret the results – A.K.A everything you already have as Master students of ICT/Data Science/Computer Science...

Step #1: idea

- some real life scenario that might inspire you: student life, living abroad, different kinds of communities you are a part of (gaming, forums, etc.)
- current events & causes you care about (protests, human rights violations, celebrities, viral trends)
- what do other people do? IP projects, last year projects, papers published on Google Scholar…
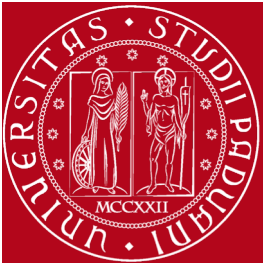
Step #1 and 1/2

- it is not enough to think of a possible network
- you also need to ground yourself in some theory or intuition: **what do I want to test? What do I expect to find in data? How do I go about finding it?**
- again, look at IP presentations – learn from your social science colleagues

- example #1: I want to test if people talk about abortion rights on Twitter differently based on whether they are prolife or prochoice. I expect to find different topics in the two datasets. I will create a network, find communities that correlate to the topics, display the highest ranked words/hashtags and compare them.

- example #2: can I figure out the alliances between characters in a book based on their co-occurrences in a sentence? I will connect the characters if they appear in the same sentence. I will then create communities in the network and compare them to the real communities in the book.

- example #3: I think people usually have more than one shared interest. I will collect data from multiple subforums, connect the people who posted on a same subforum and create a network. I will cluster the network to identify subgroups of subforums. I will identify the bridges in this network…

Step #2: data

- we primarily use social media because it's easily accessible and offers lots of insights but …
- there are datasets from other peoples' projects/papers that are free to download/ you can request them & you can use in a different ways (check Open Science Foundation, Google Scholar)
- there is tons of free datasets out there: medical, biological…
- news articles, books, political speeches, video transcripts, movie scenarios,...

Why Twitter?

- Twitter is a platform for expressing one's opinion
- used by politicians, celebrities, philosophers, bloggers... and "regular" people!
- large and well varied dataset: users from all over the world, of different age, gender, political leanings, etc.
- its use of hashtags allow for collecting data on various (trending) topics
- hashtags are a way to connect Tweets (i.e link between nodes)
- other ways to connect Tweets through retweeting, mentions, likes, replies... different networks can be created
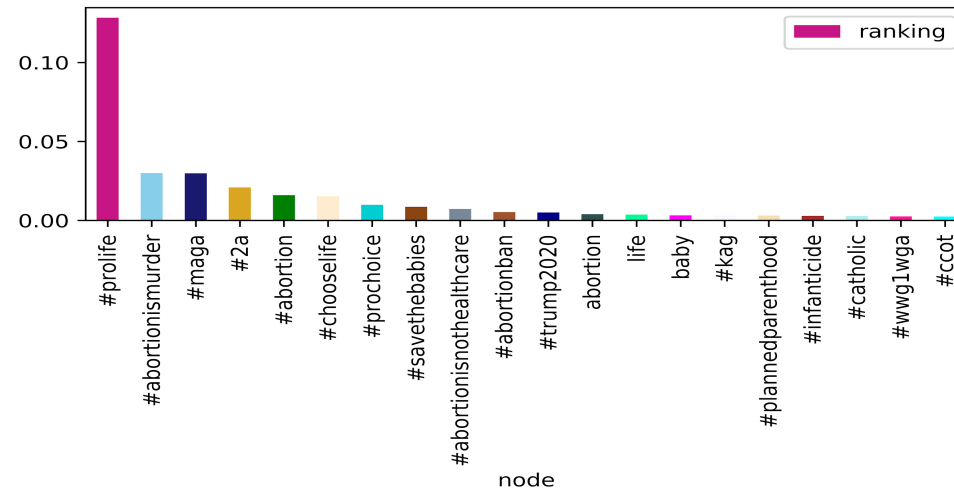- extensive and well documented API with wrappers in most commonly used programming languages

Why Twitter?

In personal experience:

- easy to work with
- large amount of data
- collected data can be used to build sentiment analysis models
- offers interesting (and sometimes unexpected insights!)

What exactly is an API?

- a web API allows us to access publicly exposed endpoints of a web application via URI
- we do this using a request - response message system, which is typically expressed in JSON or XML
- we can send a request using a REST client (f.e. Postman) or we can do it through the code

```
curl --request POST \
  --url https://api.twitter.com/1.1/tweets/search/30day/<ENV>.json \
  --header 'authorization: Bearer <BEARER_TOKEN>' \
  --header 'content-type: application/json' \
  --data '{
              "query":"from:TwitterDev lang:en",
              "maxResults": "100",
              "fromDate":"<YYYYMMDDHHmm>",
              "toDate":"<YYYYMMDDHHmm>"
              }'
```

Twitter API: overview

- from official Twitter documentation: "The Twitter API provides the tools you need to contribute to, engage with, and analyze the conversation happening on Twitter"
- basically a tool to collect Twitter data: Tweets, Users, Lists, Trends, ... (we are only interested in Tweets)
- it has two versions: **v1.1** and **v2**
- offers different tiers: standard, basic, elevated, premium, enterprise, academic
- in order to use Twitter API we need a **developer's account** which gives us authentication keys

Twitter API: overview

- we build a query using URL (f.e, https://api.twitter.com/1.1/search/tweets.json or https://api.twitter.com/2/) plus an URL-encoded query
- query consists of query text, additional parameters and operators
- query text is mandatory and the latter two depend on the version and tier
- we send a HTTP GET request to Twitter API and receive a response

Twitter API: tiers

API VERSION v1.1

Standard (free) - good for trying out the API

Premium - offers more options relevant to data collection but requires
a paid subscription
        - however it offers a sandbox (free but somehow limited
version)

Enterprise - custom tailored to the client (not relevant to us)

API VERSION v2

Basic and elevated – something like the standard in v1.1

Academic research – offers more data for free, currently the only free
tier that can get full history using the v2 API
 - you need to be a member of the academia and have a research
project to obtain it

All tiers require a developer account!

Twitter API: tiers

Standard

- behaves similarly to (but not exactly like) the Search UI feature available in Twitter mobile or web clients
- focused on relevance, not completeness (i.e some users/Tweet might be missing)
- returns only results sampled within the last 7 days (!)
- has a rate limit of 450 requests per time window (15 minutes)

A more detailed overview can be found at Official Twitter API documentation page for standard search API.
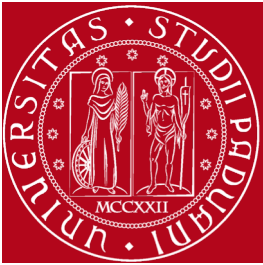
Premium

- it offers a free tier (called *sandbox*) or the true "premium"
- the difference is in the :
- **amount of data that can be fetched within a request** (100 vs 500 Tweets)
- **number of requests made** (50 per month vs 2500)
- **the rate limit on the number of requests per time frame**
- there is also a difference between the operators you can use for the query but we can manage with the ones available for premium level (even the standard ones are fine)

What now?

- I will use Twitter only in labs → you're fine :)

- I want to use Twitter for my project:

  - - … and I want to collect recent (up to last 7 days) data → you're fine, use what we learned today

  - - … and I have a moderate amount of tweets to download → apply for elevated access, then for premium sandbox, use the v1.1 code

  - - … and I need LOTS of data → if the professor gives a green light, we'll share academic access with you and then use the v2 code for full archive

What now?

- In any case you need a good strategy: when to collect, what to collect (tweets? accounts?), how to collect (search by hashtag or keyword? which one?)

- A good project can be done with one or two weeks worth of data, or even data from a strategically picked day…

- … or the data can already be out there and you don't need to collect it via twitter (but it's cool to know how anyway)

- I can help you plan it if you have Steps #1 and #1&1/2 figured out

- don't procrastinate on data collection as it can be pretty time consuming!

References

https://developer.twitter.com/en/docs

https://colab.research.google.com/

Questions