Statistical Learning Project

1st Milestone

Group 01 – Stefano Rando 1745758, Negin Amininodoushan 1915791, Marco Muscas 1883544, Leonardo Placidi 1761588, Davide Zingaro 1873954

Research Title

The Pandemic behind the Pandemic: Italian Media response to COVID-19.

Abstract

As quarantined citizens of 2020, in the middle of the Covid-19 outbreak, we feel like to be affected by one more spreading desease:the COVID-19 media response! At the start we will focus on Italian data(and if we will get meaningful results we will extend it to larger pool), since is anyway one of the most affected countries and was the first big outbreak out of China. We will analyze reflected effects of COVID-19 in not only news and journals but also in social media, so we will have an understanding of how different categories of media sources (scientific ones, social media, news websites) approached the COVID-19 outbreak day by day, reacted and evolved. We expect to see a development of news similar to the spread of a virus, reaching possible insights on how digital information can mimic a biological pandemic.

Main research aim & framework

We felt so overwhelmed by the informations about COVID-19, like we can't go online without this name popping out, so we started to think of this info as a desease itself! Our main goal in fact will be to compare the growth and behaviour of COVID-19 and Media, and how news starts to change and oppress internet more and more, becoming a digital pandemic. We will also studies anomalies in Media behaviour, such as the presence of COVID-19 where he does not belong, such as gaming communities, and we will thinking about this isolated media as 'quarantined' digital individuals, we will see the fact that as in the digital world, also in the real world Quarantene has limits. Also we feel confident in this double comparing, because we know about an ongoing research by David L. Buckeridge and Amine Kamen(https://reporter.mcgill.ca/mcgill-professor-receives-500k-to-assess-online-community-and-public-health-responses-to-covid-19/), that have been chosen to participate in the Government of Canada's global effort to address the public health challenges of the COVID-19 outbreak, in particular using online news media to assess community and public health responses to COVID-19.

Coronavirus news day by day: https://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1.

Data source(s)

We will use python and its libraries to parse the web, while we will keep updating the virus outbreak data for Italy from the website of the Protezione Civile(https://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1) that updates every day Github repositories with the data, so once again we will use python to extract the csv and build our model.\ A first list of data sources to get an idea:

1. Regional and National Newspapers like https://www.quotidiani.net/giornali_regionali.htm or https://www.repubblica.it/.
2. Scientific Websites like https://www.lescienze.it/.
3. Social networks like https://www.reddit.com/.
4. Youtube for comments https://www.youtube.com/?gl=IT&hl=it.
5. Specialized websites(not directly affected by Covid growth) like https://multiplayer.it/.

Data collection

We will start using a website specifically designed to index articles from more than 3000 news outlets across all Italy and using basic parsing techniques we will be able to get information about title, content and authoring outlet. In the meantime we will start collecting data from selected online data outlets (selected by topic such as newspapers, online communities, social networks), parsing wherever the virus is quoted. There is not a fixed number of data, but we expect to gather data as much as possible to build a trustwhorty model to study its evolution and behaviour. The difficulties could be generated by websites being heterogeneous and therefore requiring dedicated "miners" but we are strong. A small sample of data, deriving only from news outlets, got us more than 14'000 samples: we expect much more once we access public reactions. Mainly, we will use Python and some of its native libraries(such as BeautifulSoup, Pandas) to collect data.

Model & Methods

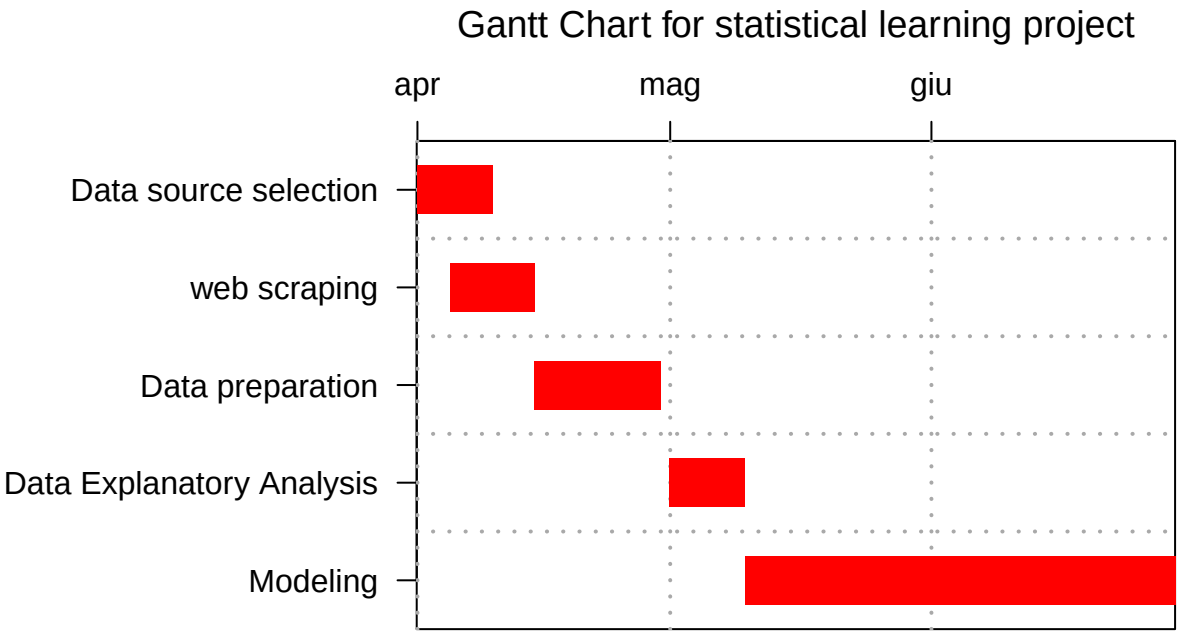
Our plan is to develop a model for the spread of media responses to COVID-19. In particular we would like to be able to do some kinds of preditcions like "How many articles can we expect on a given website given the pandemic situation?" or "What to we expect is the emotional response to a pandemic situation?". The methods we'll try to use will be decided based on how the data will behave. Anyway we can already say that for analysing specific articles we will need some Text Analytics algorithms like Naive-Bayes, Decision trees or SVM. But again we need to perform a bunch of tasks and see what methods fits best for our purposes. Moreover we need to develop an interpolation / extrapolation model for predicting the spread of COVID-19 cases as well as of news talking about COVID-19.

Software/Hardware Toolkit

We will use Python and some of its native libraries(such as BeautifulSoup, Pandas) to collect data and libraries (such as Tensorflow or keras) and specific websites APIs when available. Our laptops will be enough

and we will not develop or use a dedicated hardware, but if we will feel the need we will use dedicated cloud services like AWS.

Project Timeline



- Task 1 : selecting a pool of journals, news webpage, social media (1 April- 10 April)
- Task 2 : web scraping (5 April - 15 April) end date for data collection is still to be defined (as it is an ongoing phenomenon)
- Task 3 : data preparation (15 April- 30 April)
- Task 4 : data explanatory analysis (1 May - 10 May)
- Task 5 : start modeling (10 May- God knows when)

References

- Mining of massive datasets by Anand Rajaraman and Jeffrey Ullman, http://infolab.stanford.edu/~ullman/mmds/book.pdf.
- An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani http://faculty.marshall.usc.edu/gareth-james/ISL/.
- The Elements of Statistical Learning Jerome by H. Friedman, Robert Tibshirani, Trevor Hastie https://www.springer.com/gp/book/9780387848570.
- ResearchGate https://www.researchgate.net/publication/317701706_Deep_Learning_approach_for_sentiment_analysis_of_short_texts.
- Youtube https://www.youtube.com/watch?v=Cr6VqTRO1v0.
- Our Darth Master https://www.dss.uniroma1.it/it/dipartimento/persone/brutti-pierpaolo.
