Slide 1 – Hi, I'm Gianluca Tasciotti and in this presentation I'm going to show you my thesis, which is named Scientific Research Visualization: A Case Study on Covid-19 Literature

Slide 2 – Here, in the table of contest is resumed the content of the presentation.

---- OR ----

And the table of contest previews those points: in the "Introduction" we are going to introduce the main topic of the thesis and the dataset used, in the "text mining" we will see the techniques applied to the dataset, in the "clustering" point, the algorithms

Slide 3 – So, the aim of this work is to track the evolution of the scientific research along the time using a visualization platform.

Slide 4 – The scientific research in question is the coronavirus, the hottest topic of the moment, and thanks to the dataset CORD-19 and the relative paper, I'm going to explain how is build and composed. In this first picture we can see how the dataset is build, so Semantic Scholar collects all the papers from several resources, like WHO, arxiv, ecc… and it parses those papers in a JSON file, so we have a paper in a semi structured form. The query used to scrape those documents is a simple one, which uses keywords like: "COVID-19", "COVID", "SARS", "MERS", "coronavirus", …. In this second image we can observe how this dataset is composed, so we have a collection of papers that are published until today and the most relevant thing is these papers are incomplete or the text is missing and therefore, we had to do a preprocessing on this dataset.

Slide 5 – The preprocessing did on this dataset is the following: we decided to take into account only papers that are published from 2020 onwards and we tokenized the abstract with respect to the full text for time issue, the number of words was too huge to be processed. We noticed that there were not only English papers, but also German, Spanish, Frances papers, so we removed these kinds of words. These words are not the only one to be removed, but we removed some irrelevant stopwords like "Covid-19", "background", "copyright", "abstract", …. and we apply lemmatization on the final result. Additionally, we took only papers that have at least 10 words at the end of this pipeline. In the end, we had about 36438 papers and we applied TF-IDF, using unigrams and bigrams.

Slide 6 – So once we have cleaned our dataset, we needed to understand how many clusters are present in our dataset and we used the elbow curve and we decided to pick K = 3. In the picture below, we can see one of the first tentative to draw the elbow curve without taking in consideration all the previous steps. On the right there are the wordclouds of the clusters found.

Slide 7 – So, the question was if we can improve this curve and we used SVD with 500 features and the elbow curve obtained is this one, it doesn't change a lot, just a little flection on the 3 and therefore we choose K=3. In the picture below we can see the clusters represented and, on the right, the wordclouds.

Slide 8 – The last approach used is t-SNE and we can immediately see an evident flection near the region 3 and consequently we choose K = 3.

Slide 9 – We have decided to do topic classification because we want to understand how the countries move their footprints against a new phenomenon like COVID-19.

In this visualization tool we can see two sliders on the right, the first represents the number of papers and the second the time. In the middle we have a worldwide map where the countries change colors based on how many papers they have published based on time. On the right we have this barchart where are showed as default the top 5 of the countries which have published papers. Each barchart represent the number of papers according to the topic classification.
The plot below tracks the number of cases of covid-19 week-by-week, along the time.

Slide 10 – So let me describe this little demo, where we observe a spike of cases between February and April and consequently, we make a focus on the same range of time for the papers. We can see the most papers published in this period come from China with respect to the other countries and in particular USA, which we have seen before they were in first position if we consider all the 2020. We want to focus on Europe, selecting the five majority countries, Italy, UK, France, Germany and Spain and what we want to show is how the countries studied the COVID-19. Italy, since we were one of the first countries to be hit by the covid, the first topic studied is the propagation of the virus, this in our opinion to avoid a possible pandemic, while the other states focused on finding a model to track the evolution of the infections. If we compare this situation with the last part of the year, we can see a different scenario, where the USA becomes the first and China second in terms of papers published and looking the situation in Europe, we can see the countries start to analyze the clinical risks placing the evolution of cases in the background.

Slide 11 – The pros and cons are summed up in this slide.

- We can say that we have built a different model of tracking the research which is not the usual citation network.
- After some efforts, we were able to do topic classification which are a good starting point
- and the application showed is independent from the research context, no matter of what we want to speak, if we place another dataset with the same features, the tool is able to draw it.

The difficulties ran into the developing of the application were

- the effort of manage a huge amount of data because this needed a lot of time to be processed,
- the absence of a backend in order to process data and therefore sent it to the frontend
- and we were sure that applying a NN we could find some other topics inside the dataset.

Slide 12 – I hope you enjoyed our work and thanks for the attention.