* INTRODUCTION

The goal of this project is to try and reproduce for Italy
the results reported in this article for the USA:

[1] D.J. McIver & J. S. Brownstein (2014), "Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time", PLoS Comput Biol 10(4): e1003581
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003581

Data:

1) OFFICIAL DATA ON INFLUENZA IN ITALY. The Italian health protection agency runs a flu surveillance program called "Influnet" that uses sentinel doctors.
The project is described here:
http://www.iss.it/iflu/index.php?lang=1&anno=2016&tipo=4

The official data are reported in PDF files (see the "PDF" folders for examples)

These data will be the ground truth.

2) WIKIPEDIA PAGE VIEW DATA. Wikimedia Foundation makes available several
datasets, tools and APIs to work with page view data. A summary can be found here:
https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics

* PART 1

1.1 - Process the Wikipedia pageview data for the "Influenza" page of the Italian Wikipedia (https://it.wikipedia.org/wiki/Influenza), aggregate the pageviews on a weekly time scale, and plot the resulting time series of page views for the current year and - ideally - also for previous years.

1.2 - Compare the time series from the official Influnet surveillance system with the time series of pageviews obtained in 1.1.
Compute some measure of correlation between the two time series.

* PART 2

2.1 - Try to find other Wikipedia pages related to flu whose pageview time series are correlated with the Influnet signal. Compute their weekly pageview time series for the last year and - if possible - for the previous years,and plot them together with the Influnet signal as in 1.1.

2.2 - For each of the selected Wikipedia pages, compute the same correlation with the Influnet time series that you computed in 1.2. Which of these correlations are strongest ?

* PART 3

3.1 - Build a regression model that predicts the Influnet incidence for a given week based on the Wikipedia pageview data for the same week. Features are the Wikipedia pageview counts for the "Influenza" page, for all the pages selected in Part 2.Evaluate the performance of your model via cross-validation.

3.2 - Add these features to the model:

- the Influnet incidence for the week preceding the target week
- the pageview counts for all the pages you selected for the week preceding the target week

Re-train your model and evaluate its performance via cross-validation.