r/marvelstudios:

Community and Content Analysis after the release of Avengers: Endgame

Social Media Analytics project

Roberto Ferrari 852220, Davide Prati 845926, Marco Sallustio 906149

In the Google Drive folder are present the following files:

- **Report.pdf** : the LaTex-written report describing the work done, i.e., the project, the implemented solutions, and the evaluations.
- **Presentation.pdf**: the presentation of the project results, conducted using Colab

- The following Google Colab ore Jupyter notebooks, which constitute the source code:
  - **data_extraction.ipynb**: it contains all the necessary code to interact with the Reddit API using the Python library PRAW, enabling the retrieval of the dataset for analysis. The outcome of this notebook is the raw dataset, which will then undergo further processing.
  - **preprocessing.ipynb**: notebook that encompasses all text preprocessing steps, resulting in the preprocessed dataset in CSV format. Specifically, it includes code for language detection and filtering, lemmatization, and normalization.
  - **exploratory_data_analysis.ipynb:** notebook that contains a dataset description and dataset exploration operations with plots.
  - **sentiment_analysis.ipnyb**: notebook in which sentiment analysis was performed on the dataset after the pre-processing phase.
  - **topic_modeling.ipynb**: it includes a section for text representation to prepare the data for LDA, a section where LDA is performed and one in which we show the graphical results
  - **WordCloud_for_topic_modeling.ipnyb**: it contains just the WordClouds corresponding to the 7 topics obtained before

- json and csv files created and used during the course of the project:
  - **avengers.csv**: just the raw dataset extracted from Reddit
  - **avengers_pre.csv**: dataset resulting from preprocessing.ipynb
  - **avengers_sentiment.csv**: dataset containing the results from the sentiment analysis.
  - **topic_terms.json**: dataframe obtained from topic_modeling.ipnyb and necessary to compute the WordClouds

All packages and libraries are loaded at the beginning of the code files. In case they are not already installed, they can be installed using the standard procedure.