Pipeline:

1. Cominciare da facta.news **OK**
2. Usare scraper (tool) per scaricare tutte le pagine di facta [usare HTTrack] [<https://facta.news>] **SCARICARE AL DIAG**
3. Python: estrazione fact-checking json [quante pagine lo hanno]
   1. “Generated by Full Fact Claim Review Schema”
   2. "claimReviewed"
   3. "itemReviewed"
   4. "reviewRating"

**OK**

1. Stopwords removal (https://stackoverflow.com/questions/54573853/nltk-available-languages-for-stopwords)
2. JSON to DataFrame **OK**
   1. Flatten JSON/dict: <https://www.freecodecamp.org/news/how-to-flatten-a-dictionary-in-python-in-4-different-ways/>
   2. dict to DataFrame https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.from\_dict.html
3. Dati utili:
   1. URL del sito di fact-checking ("url")
   2. Titolo ("claimReviewed")
   3. Label ("reviewRating")
   4. ###Abstract / short-text (probabilmente non c’è)
   5. URL originale (dipende dal sito di fact-checking; su facta è il primo link nel testo) [salvare tutti i link con relativo testo del href; li trovi su <div class="edgtf-post-text-main">]
      1. Tirare fuori link originale (se link a web.archive)
      2. Se non trovato, cercare su web.archive / cache
4. Identificazione lingua titolo [<https://pypi.org/project/langdetect/>] **TODO**
5. Analisi dati **TODO**
   1. Sui titoli che stanno nel JSON
   2. Barplot/pie-chart delle lingue
   3. keywords → word cloud
      1. remove stopwords **OK**
   4. Selecting the optimal number of topics
      1. See NTB
   5. topic detection
      1. word-cloud per topic
      2. Interpretazione del word-cloud
   6. Statistiche utili?
      1. Articoli per lingua
      2. Differenza article-datePublished — review-datePublished
      3. Articoli per fact-checking website

* Aggiungere altri news outlets
* Recuperare pagina notizia originale → download html

Google Fact Check Tools: <https://toolbox.google.com/factcheck/explorer/search/covid;hl=it>

Lista siti fact-checking:

<https://www.agcom.it/siti-di-fact-checking>

Pipeline:

1. Cominciare da facta.news **OK**
2. Usare scraper (tool) per scaricare tutte le pagine di facta [usare HTTrack] [<https://facta.news>] **SCARICARE AL DIAG**
3. Python: estrazione fact-checking json [quante pagine lo hanno]
   1. “Generated by Full Fact Claim Review Schema”
   2. "claimReviewed"
   3. "itemReviewed"
   4. "reviewRating"

**OK**

1. Stopwords removal (<https://stackoverflow.com/questions/54573853/nltk-available-languages-for-stopwords>) **OK**
2. JSON to DataFrame **OK**
   1. Flatten JSON/dict: <https://www.freecodecamp.org/news/how-to-flatten-a-dictionary-in-python-in-4-different-ways/>
   2. dict to DataFrame https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.from\_dict.html
      1. Se non trovato, cercare su web.archive / cache
3. Identificazione lingua titolo [<https://pypi.org/project/langdetect/>] **TODO**
4. Analisi dati **TODO**
   1. Sui titoli che stanno nel JSON
   2. Barplot/pie-chart delle lingue **OK**
   3. keywords → word cloud **OK**
      1. remove stopwords **OK**
   4. Selecting the optimal number of topics
      1. See NTB
   5. topic detection
      1. word-cloud per topic
      2. Interpretazione del word-cloud
   6. Statistiche utili?
      1. Articoli per lingua
      2. Differenza article-datePublished — review-datePublished
      3. Articoli per fact-checking website