自然言語処理
―準備、データ作成―

https://github.com/satoyoshiharu/nlp

データ作成の位置づけ

- ニューラルネットのアプリを組む時、コードはパッケージのクラスを利用すれば簡潔に書けます。一方、実は、トレーニングしたりテストするデータを準備する法に、多くの工数がとられます。
- 自然言語処理で、Wikipediaを利用する場合、重宝するツールに触れます。

WikiExtractor

- WikipediaのテキストDumpを取り出すツールとして、wikiextractorがあります。
  - https://github.com/attardi/wikiextractor
- これを使って、以下のサイトからデータを入手できます。
  - Wikipediaのテキストダンプサイト：https://dumps.wikimedia.org/jawiki/latest/
  - 全記事の圧縮ファイル：jawiki-latest-pages-articles.xml.bz2

BeatifulSoup

- WikipediaのXML/HTMLからテキストを抜き出すツールとして、BeautifulSoupが利用されています。
  - https://www.crummy.com/software/BeautifulSoup/bs4/doc/
  - 日本語訳->http://kondou.com/BS4/

課題

- 「データ作成.ipynb」というノートをコピーし、冒頭の準備をやった後、指定の課題をやってください。

確認クイズ

- ありません。