Datamining report3 205713B 朝比奈太郎 2022/06/09

Level 1: コーパスを用意し、pd.DataFrameとして読み込もう。

**コーパスの説明**

・コーパスの概要

Livedoorニュースコーパス

展開するとlivedoorニュースの項目ごとに文書(text)がまとめられており、その中でもdokujo-tsushinというカテゴリを採択。そのままでは文書数852, 総文字数1368059とデータが多かった為、処理しやすいように上から20文書、総文字数(text)29772とした。

・コーパスの取得手順

出典: <https://www.rondhuit.com/download.html#news%20corpus>

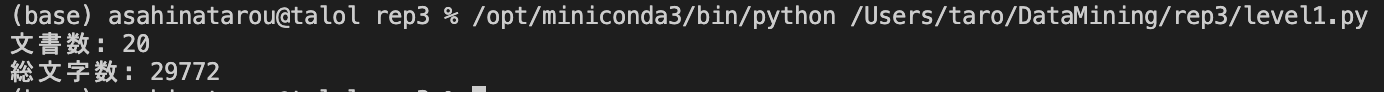
[ldcc-20140209.tar.gz](https://www.rondhuit.com/download/ldcc-20140209.tar.gz)をダウンロード。

・pd.DataFrame.head()の出力結果。

グラフィカル ユーザー インターフェイス, テキスト

自動的に生成された説明

文書数、総文字数



工夫した点。

まず、テキストファイルがデータフレームに適した状態ではなかった(画像１)ため、URL, date、head, textに分け、dfという名のDataFrameとした。その際にURL, date, headは一行ずつ格納することができたが、textは非常に文章量が多いので別途dataという変数を用いて処理した。また、読者に読みやすくするためか改行や空白が所々あったのでそれらの処理も行った。

感想

前処理にかなりの時間が掛かってしまったが、文書内容が非常に面白いのでok.

画像１

テキスト

自動的に生成された説明