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。

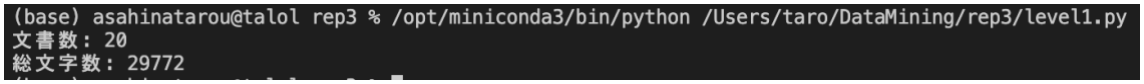
・コーパスの取得手順

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

ldcc-20140209.tar.gz をダウンロード。

・pd.DataFrame.head()の出力結果。

文書数、総文字数



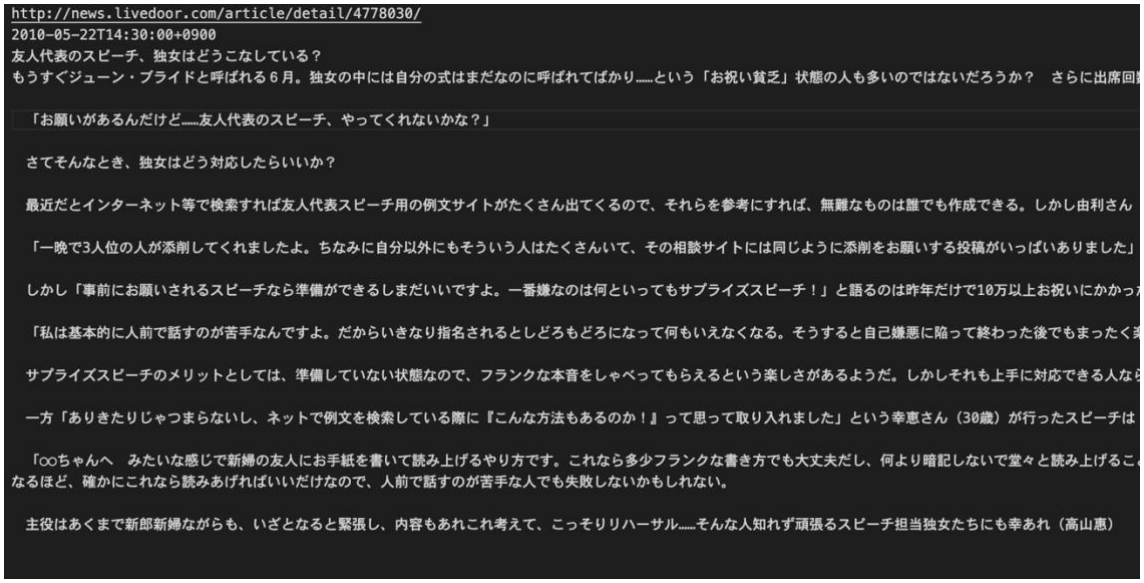
工夫した点。

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


http://news.livedoor.com/article/detail/4778030/
2010-05-22T14:30:00+0900
友人代表のスピーチ、独女はどうこなしている？
もうすぐジューン・ブライドと呼ばれる６月。独女の中には自分の式はまだなのに呼ばれてばかり……という「お祝い貧乏」状態の人も多いのではないだろうか？　さらに出席回
「お願いがあるんだけど……友人代表のスピーチ、やってくれないかな？」
さてそんなとき、独女はどう対応したらいいか？
最近だとインターネット等で検索すれば友人代表スピーチ用の例文サイトがたくさん出てくるので、それらを参考にすれば、無難なものは誰でも作成できる。しかし由利さん
「一晩で3人位の人が添削してくれましたよ。ちなみに自分以外にもそういう人はたくさんいて、その相談サイトには同じように添削をお願いする投稿がいっぱいありました」
しかし「事前にお願いされるスピーチなら準備ができるしまだいいですよ。一番嫌なのは何といってもサプライズスピーチ！」と語るのは昨年だけで10万以上お祝いにかかっ
「私は基本的に人前で話すのが苦手なんですよ。だからいきなり指名されるとしどろもどろになって何もいえなくなる。そうすると自己嫌悪に陥って終わった後でもまったく楽
サプライズスピーチのメリットとしては、準備していない状態なので、フランクな本音をしゃべってもらえるという楽しさがあるようだ。しかしそれも上手に対応できる人なら
一方「ありきたりじゃつまらないし、ネットで例文を検索している際に『こんな方法もあるのか！』って思って取り入れました」という幸恵さん（30歳）が行ったスピーチは
「○○ちゃんへ　みたいな感じで新婦の友人にお手紙を書いて読み上げるやり方です。これなら多少フランクな書き方でも大丈夫だし、何より暗記しないで堂々と読み上げるこ
なるほど、確かにこれなら読みあげればいいだけなので、人前で話すのが苦手な人でも失敗しないかもしれない。
主役はあくまで新郎新婦ながらも、いざとなると緊張し、内容もあれこれ考えて、こっそりリハーサル……そんな人知れず頑張るスピーチ担当独女たちにも幸あれ（高山恵）