ゆる自然言語処理 -word2vec編-

自然言語とは

- 日本語や英語など，人がコミュニケーションで利用する言語
  e.g. テキスト, 会話音声
- プログラミング言語と対になる
- 翻訳や要約，生成をするのが自然言語処理

今回やること

- 単語のベクトル化
- ベクトル化した単語のコサイン類似度により，単語間の類似度を求める
- 単語同士の加減算

手法

- ベクトル化したモデルの用意
  - ベクトル化されたモデル(学習済みモデル)をダウンロード
  - コーパス(大量のテキストを集めたもの)の用意&モデルの学習(単語のベクトル化)
- やる

環境

- Google colab
- 各種ライブラリ
  e.g. gensim, MeCab

モデルの作成

- 日本語Wikipediaの記事をダウンロード
- MeCabを使って文章を分かち書き(私は昨日です．　-> 私 は 昨日 です ．)
- gensimを使ってモデルの学習(単語のベクトル化)
    - 数行書けば自動でやってくれる
    - このページの各工程はどれもそこそこ時間がかかる

デモンストレーション

- みせます

つまったところ

- とくになし
- 待ち時間が長いくらい

結論

- 単語同士の加減算ができるのはなんとなく面白いですよね
- ライブラリのおかげでそんなに難しくないので，みんなもやってみてね