Word2Vec

1 英語

テキストファイルからの読み込みのコーディングに苦戦しているため, まだ結果が出ておりません.

2 日本語の Word2Vec

英語など, 多くの言語は, 単語が分かち書きされているため, 単語の切り出しが容易である. 一方, 日本語は, ふつう単語を分かち書きせずに文章を書くため, 単語の切り出しに手間がかかる. 例えば, 「はにわ」という言葉の「は」や「に」が, 助詞として使われているのか, 単語の一部であるのか, 文章中で判断できなければ, 単語の切り分けができないのである. また, 語形変化も豊富であり, 動詞の後ろにたくさんの助動詞が付くことがあり, これも単語の切り出しを難しくしている. 結果として, 形態素解析が必要になる.

ChiVe など, Word2Vec には日本語のモデルも存在するので, 学習済みのモデルで単語ベクトルを取得したり, 単語の類似度を調べることは可能である. また, 任意のウェブページや, テキストファイルを使って学習することも可能である. しかし, 後述の「単語の情報」でも分かる通り, 必ずしも単語で分割されているとは言い難い.

なお，Chive は，Sudachi の開発元である Works applications が公開しているモデルであり, 国立国語研究所の日本語ウェブコーパスで学習されたモデルである．分かち書きには Sudachi を使用している．

3 単語の情報

「研究」「学ぶ」「高い」を選択した. 類似度上位 10 個を取得すると,

「研究」との類似度
(’研究者’, 0.7765358686447144)
(’共同研究’, 0.7702159881591797)
(’究’, 0.7638627290725708)
(’研究開発’, 0.7089934945106506)
(’学術’, 0.6682144403457642)
(’実験’, 0.662083625793457)
(’研究所’, 0.6566915512084961)
(’研究員’, 0.6386178731918335)
(’知見’, 0.6332124471664429)
(’生命科学’, 0.6104219555854797)

研究が含まれる語が多く入っている. また, 学術や実験, 知見といった, 研究と関連しそうだと直感的にも思える語が入っている. また, すべて名詞である.

「学ぶ」との類似度
(’学び取る’, 0.6955525279045105)
(’実践的’, 0.6908577680587769)
(’勉強’, 0.667337954044342)
(’習得’, 0.6619090437889099)
(’教わる’, 0.6601988673210144)
(’講座’, 0.6532350778579712)
(’習う’, 0.6521884799003601)
(’基礎的’, 0.6431999802589417)
(’学び直す’, 0.6262302994728088)
(’独学’, 0.6253233551979065)

学ぶが含まれる, 学び取る, 学び直すといった語が入っており, 学ぶに類似しているが, 教える側の存在が示唆される, 教わる, や習う, という語が入っている. また, 動詞の類義語ではあるが, 実践的や基礎的といった, 学ぶと共に使われる頻度が高いと思われる形容詞や, 学ぶの言い換えとして使用可能な場面が多い, 勉強, 習得といった名詞, 学ぶことの下位カテゴリーと言って良い独学, 学びの場である講座が入っている.

「高い」との類似度
(’低い’, 0.8304073214530945)
(’高め’, 0.7010296583175659)
(’高燥’, 0.6296851634979248)
(’低め’, 0.6122028827667236)
(’上がる’, 0.6088359355926514)
(’下がる’, 0.5925613641738892)
(’高’, 0.5886228084564209)
(’高める’, 0.5559080839157104)

('下げる', 0.5526874661445618)
('非常', 0.5497260689735413)

類似度での検索であるが, 対義語である低いが一番上位に来ており, 2 位との差も約 0.13 と相当なものである. 高い, 低いが含まれた語が多くランクインしており, 高燥, はかなり使用頻度が低そうに思われるが, 高の字が含まれることや, 同じ文脈で使われることからランクインしたのだろう. また, 上がるや下がるなどは, 結果として高さが変わる行為であり, 確かに類似度が高いように思われる. 非常がランクインしているのは, 非常に高い, というセットで使われる頻度が高いことが理由だと思われる.

4 参考文献

https://qiita.com/makaishi2/items/63b7986f6da93dc55edd
https://blog.hoxo-m.com/entry/2020/02/20/090000