

B3 実験 11 月 8 日分課題

1 課題 1

1.1 実験内容

gensim の Word2Vec ライブラリを使い, 与えられたテキストを学習した.

1.2 前処理 と モデルのパラメータ

与えられたテキストにおいて `!?'` のみ残し, その他の半角記号を消去した. その後 `!?'` の 3 つの記号に関しては前後に空白を挟むように調整し, 文章中の全単語を小文字にした. そして, Natural Language Toolkit (nltk) を用いて分かち書きした. その後 `!?'` の記号を文末と定義して 1 文ごとに分けて Word2Vec のモデルの学習を行えるようにした. また, 表 1 に今回作成した Word2Vec モデルのパラメータを示す

表 1: Word2Vec におけるパラメータ

パラメータ 値	分散表現の次元	学習時に利用される文脈の広さ	分散表現を獲得する単語の最小頻度
	500	5	1

1.3 実験結果

適当な名詞, 動詞, 形容詞を 1 つ選択しその単語と近い類似度の単語を 5 個調べた. 今回は名詞として 'alice', 動詞として 'think', 形容詞として 'good' を選択した. 表 2 にそれぞれの結果を示す.

表 2: 'alice' に最も類似度が高い上位 5 単語

単語	'trouble'	'wish'	'what'	'little'	'then'
類似度	0.130898	0.124550	0.104635	0.094204	0.092406

2 ???

3 ???

参考文献