
進捗報告

1 今週やったこと

- 先週のゼミで提案いただいたネガポジ判定を行いました。

2 内容について

- 実装環境は googlecolaboratry です。
- 使用したのは東北大学の BERT で、データセットは岡田先生の github に挙げられている評判分析チェック用データを用いました。
- データセットを取り込み、ラベルデータと文章データに分けてから文章データを分かち書きし、tokenizer を用いてベクトルにしました。それから tensorflow による多層パーセプトロンで訓練データを学習し、accuracy を用いてテストデータと予測結果との正答率を求めました。(0.845 でした)
- 訓練データ数は 4000、テストデータ数は 2000 でした。

3 疑問点など

- pytorch に関する知識がなかったので tensorflow を使いました。
- 文章データによってはベクトルに変換した後の要素数に大きな開きがあり、最も少ないもので 23、多いもので 1807 でした。入力に用いる要素数を最大 (1807) にするとエラーが出るため、全要素の平均よりも少し多い程度の値を入力要素数として設けましたが適切であったのかがわかりません。