SHAP を用いたトピックモデルの解釈性向上

須永 歩夢

要旨

ニュース記事や口コミなどのテキストデータの分析にはトピックモデルが広く用いられており，特に LDA（Latent Dirichlet Allocation）を活用したレビュー文や SNS 投稿の分類に関する研究が多く行われてきた．しかし，LDA で得られるトピックの解釈性の解釈性には課題があり，各トピックの意味をより明確に把握する手法が求められている．本研究では，トピックモデルの解釈性の向上を目的に，SHAP（SHapley Additive exPlanations）を用いたカテゴリ予測を併用することを提案する．具体的には，LDA で抽出される各トピックの単語分布と，それらの単語に対する SHAP 値を考慮することで，トピックがどのように分類に寄与しているかを分析する．すなわち，トピック内の各単語が持つカテゴリ予測に与える影響の大きさと方向を定量的に評価し，トピックの意味をより直感的に理解することを支援する．提案の有効性を確認するために，ニュース記事を対象とした予備的な実験を行った．その結果，いくつかのケースにおいて，トピックの解釈性が向上することが確認された．