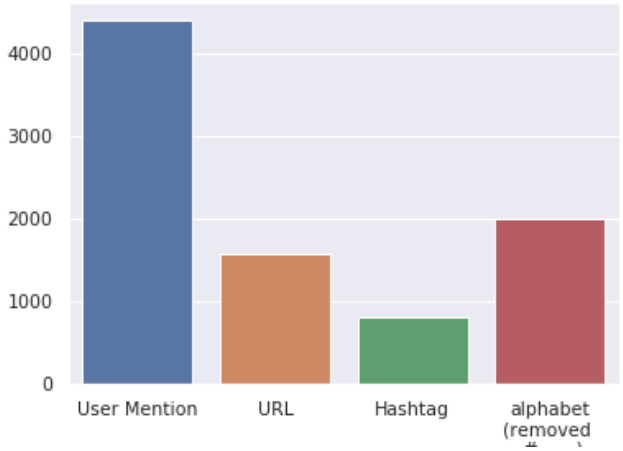
進捗報告

181-T3201　高木　裕仁

○　日本語と英語の混じったツイート件数分析

□ツイートの前処理で頻繁に処理されるもの
（→）URL・ユーザーメンション・ハッシュタグ
　これらに比べてどうなのかを Twitter API を申請して１万件のデータで調べた

ユーザーメンション：4401
URL：1564
ハッシュタグ：799
英語（アルファベット）
を含むもの：1991



□　アルファベットを含むツイートは多い
　　※　ユーザーメンション・URL のものは含まない
　　※　＃happy などのハッシュタグの後ろのものは含まない

□　コードを書いていて感じたこと
　　（→）日本人のツイートには w（笑いを示す）が案外多い
　　　これを前処理することはセンチメント分析には有用？

ハッシュタグの
うしろのアルファベット
も含む：2063
英語（アルファベット）
を含むもの：1991
w を含むもの：516
２文字以上：1549