令和 3 年 8 月 18 日 (水) ゼミ資料　　M1 楠本 祐暉

進捗報告

1　今週やったこと

- 以前まで読み込めなかった朝食のポジネガクラスのデータを一部修正して読み込んだ.
- 読み込んだデータで, 同じクラスでポジティブとネガティブのラベルが両方立っているデータについて調査した.

2　データの内容

今回読み込んだデータは少なくとも 1 つのラベルが立っているデータ群である. 表 1 に朝食のポジネガクラスを読み込んだデータにおいて, 同じクラスでポジティブとネガティブのラベルが両方立っているデータ数や, 平均, 語彙数などの値について示す. 語彙数は東北大学の BERT を用いて, データの日本語文章から tokenizer で獲得した.

表 1　データのラベルや語彙数についての各値

全データ数	53192
全データにおける語彙数の平均	30.07
全データにおける語彙数の最大値	337
全データにおける語彙数の最小値	4
両ラベルが立っているデータ数	893
立っているラベル数の平均	1.735

次に, 同じクラスでポジティブネガティブ両方のラベルが立っているデータについていくつか具体的に見た. 表 2 に両方のラベルが立っているデータの具体例について示す.

表 2　両方のラベルが立っているデータの具体例

テキスト	朝食 po	朝食 ne	夕食 po	夕食 ne	風呂 po	風呂 ne	サービス po	サービス ne	立地 po	立地 ne	設備 po	設備 ne	部屋 po	部屋 ne
立地、最上階、部屋からの景色、エアウィーヴ等の良い点と比較しても、次の機会に泊まるかは疑問です.	0	0	0	0	0	0	0	1	1	0	1	1	1	1
外観を見て失敗したと思いましたが、中に入ると別世界でした.	0	0	0	0	0	0	0	0	0	0	1	1	0	0
古いながらも大変メンテナンスされていますので清潔でした.	0	0	0	0	0	0	0	0	0	0	1	1	1	1
食事も夕・朝とも質量ともに問題なかったのですが, 逆に朝は量が多すぎるくらいでした.	1	1	1	0	0	0	0	0	0	0	0	0	0	0