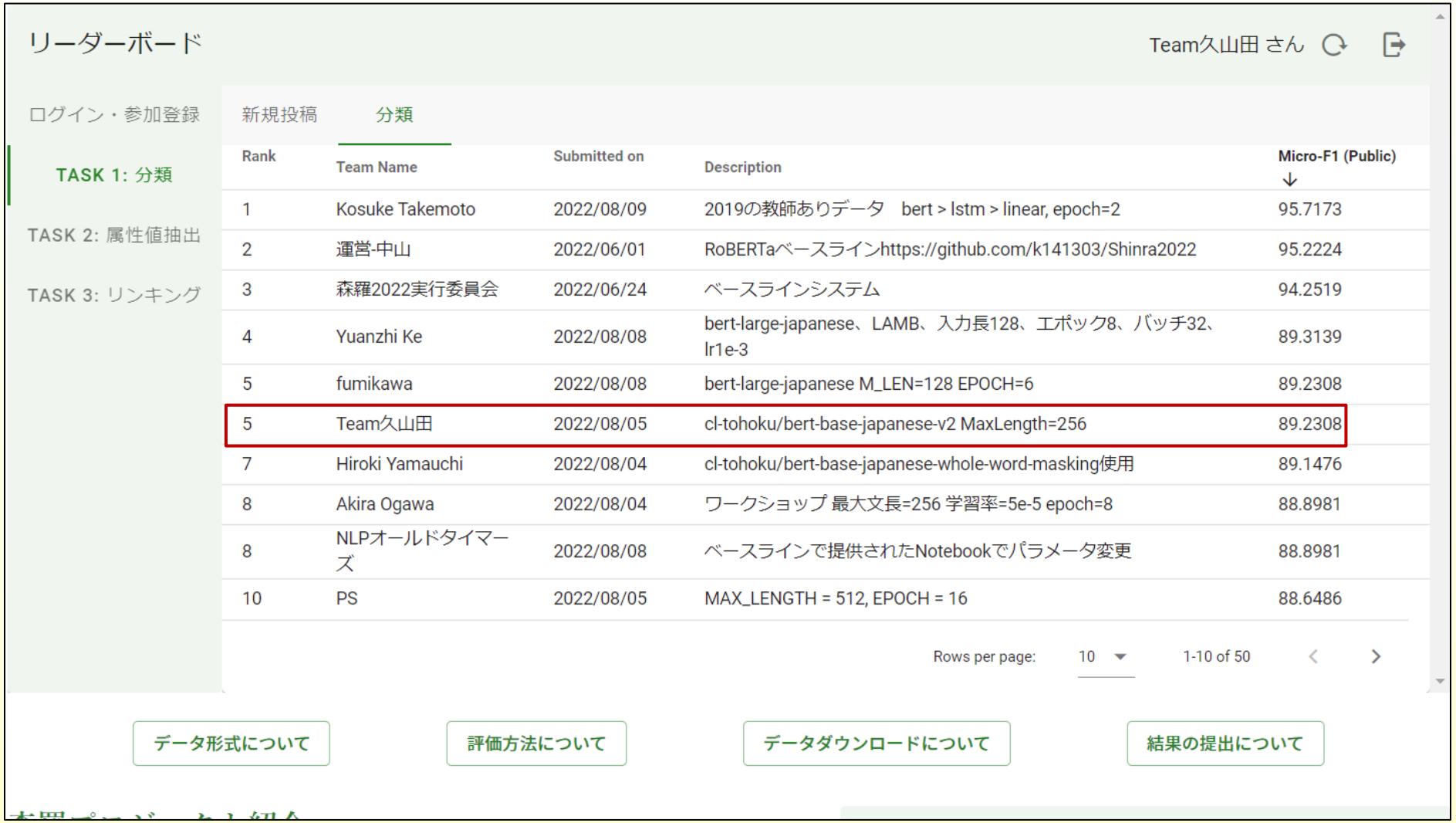
# Wikipedia構造化プロジェクト 森羅2022
## 学生・若手研究者のためのBERTワークショップ
# 招待体験報告 私の工夫

Team久山田　有吉(尾道市立大)

R04.8.10

# 自己紹介



リーダーボード

Team久山田 さん

ログイン・参加登録

TASK 1: 分類

TASK 2: 属性値抽出

TASK 3: リンキング

新規投稿　分類

| Rank | Team Name | Submitted on | Description | Micro-F1 (Public) ↓ |
|---|---|---|---|---|
| 1 | Kosuke Takemoto | 2022/08/09 | 2019の教師ありデータ　bert > lstm > linear, epoch=2 | 95.7173 |
| 2 | 運営-中山 | 2022/06/01 | RoBERTaベースラインhttps://github.com/k141303/Shinra2022 | 95.2224 |
| 3 | 森羅2022実行委員会 | 2022/06/24 | ベースラインシステム | 94.2519 |
| 4 | Yuanzhi Ke | 2022/08/08 | bert-large-japanese、LAMB、入力長128、エポック8、バッチ32、lr1e-3 | 89.3139 |
| 5 | fumikawa | 2022/08/08 | bert-large-japanese M_LEN=128 EPOCH=6 | 89.2308 |
| 5 | Team久山田 | 2022/08/05 | cl-tohoku/bert-base-japanese-v2 MaxLength=256 | 89.2308 |
| 7 | Hiroki Yamauchi | 2022/08/04 | cl-tohoku/bert-base-japanese-whole-word-masking使用 | 89.1476 |
| 8 | Akira Ogawa | 2022/08/04 | ワークショップ 最大文長=256 学習率=5e-5 epoch=8 | 88.8981 |
| 8 | NLPオールドタイマーズ | 2022/08/08 | ベースラインで提供されたNotebookでパラメータ変更 | 88.8981 |
| 10 | PS | 2022/08/05 | MAX_LENGTH = 512, EPOCH = 16 | 88.6486 |

Rows per page: 10　1-10 of 50

データ形式について　評価方法について　データダウンロードについて　結果の提出について

# 自己紹介

- 氏名 有吉勇介
- 所属・役職等
  - 尾道市立大学 経済情報学部 経済情報学科 教授
- 担当授業
  - 1年 プログラミングI、2年 情報基礎理論、3年 情報システム設計
- 深層学習は、昨年の春休みから
  - 最近のお気に入りは、AIcia Solid
    - https://www.youtube.com/channel/UC2lJYodMaAfFeFQrGUwhlaQ/featured
- 昨日まで、前期試験でした…

# ワークショップ提供のコードからの変更

- cl-tohoku/bert-base-japanese-v2
- !pip install unidic-lite
- SEED=1234
- MAX_LENGTH=256
- EPOCH=12
- Google Colab Pro

# cl-tohoku/bert-base-japanese-v2

- ちょうど読んでいたWeb記事が使っていたから
  - 「JGLUE/MARC-jaをGoogle Colabで評価してみる」
    - https://qiita.com/hideki/items/07f901c3d94bb6204674

- v1との違い
  - 日本語版Wikipedia
    - 2019/9/1 ⇒ 2020/8/31
    - サイズ:2.6GB 、約 17M文 ⇒ 4.0GB、約30M文
  - MeCab使用辞書
    - ipadic ⇒ Unidic
  - 語彙サイズ
    - 32000 ⇒ 32768

- くわしくは
  - 「Huggingface Transformers 入門 (34)
    -東北大学の乾研究室の日本語BERTモデルのv1とv2の比較」
    - https://note.com/npaka/n/nbbf6b38f4b46

# ipadic と Unidic

- ipadic: IPA辞書
  - IPA(情報処理技術者機構)が作っている辞書
  - 2007年から更新されていない
- Unidic
  - 国立国語研究所が作っている辞書
  - 最新版v3.1.0は2021/4/1公開
- くわしくは
  - 「2022年最新版 Python + mecab の周辺事情」
    - https://techtekt.persol-career.co.jp/entry/tech/220614_01