

世界中のアイデアを集めるく ずし字コンペの開催

カラーヌワット・タリン

ROIS-DS 人文学オープンデータ共同利用センター

国立情報学研究所



大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

kaggle コンペティションとは？



世界最大のデータサイエンスコミュニティ。会員登録は世界中から300万人以上。

kaggle コンペティションとは？

11 Active Competitions



2019 Data Science Bowl

Uncover the factors to help measure how young children learn

Featured · Code Competition · 3 months to go · video games, children, learning, education

\$160,000
622 teams



TensorFlow 2.0 Question Answering

Identify the answers to real user questions about Wikipedia page content

Featured · Code Competition · 3 months to go · text mining, text data

\$50,000
341 teams



Peking University/Baidu - Autonomous Driving

Can you predict vehicle angle in different settings?

Featured · 2 months to go · image data, image processing

\$25,000
148 teams



RSNA Intracranial Hemorrhage Detection

Identify acute intracranial hemorrhage and its subtypes

Featured · 7 days to go · image data, health foundations and medical research

\$25,000
139 teams



Lyft 3D Object Detection for Autonomous Vehicles

Can you advance the state of the art in 3D object detection?

Featured · 6 days to go · image data, object detection

\$25,000
512 teams

カグルでは常にコンペが開催されている。コンペの課題は医療、生物学、情報学など。さまざまな分野の問題を機械学習で解決する。

kaggle コンペティションとは？



カグルコンペは参加者の国籍、
年齢、技術レベルを問わず、
世界中から誰でも参加できる。

kaggle コンペティションとは？



kaggle

#	Δpub	Team Name	Score ?	Entries
1	—	tascj	0.950	13
2	—	Konstantin Lopuhin	0.950	60
3	—	Kenji	0.944	161
4	▲1	YoudaoOCR	0.942	49
5	▼1	See--	0.940	42
6	—	abc	0.939	15
7	—	K_mat	0.934	20
8	—	t-hanya	0.920	21
9	—	Ollie, Nanashi, and Tom	0.910	35
10	—	Zenkei_R&D	0.903	144
11	—	masayai	0.903	12
12	▲5	Kirill Brodt (shad nsk)	0.901	4
13	▲1	James Day	0.901	33
14	▼1	NEU	0.900	54
15	▼3	s tatsuya	0.900	29
16	▼1	amirassov	0.897	9
17	▼1	Nicholas Yi	0.896	64
18	—	m454lci	0.894	45
19	—	Yuzu-Project	0.891	26
20	—	E_D_	0.888	70

参加者が、自ら開発したアルゴリズムの結果をカグルのサイトにアップロードすると、自動的にスコアが計算され、リーダーボードが更新される。

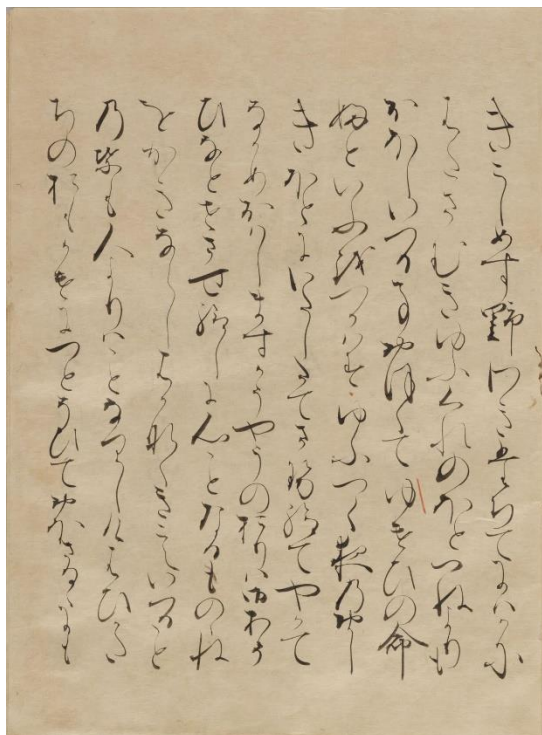
kaggle くずし字認識コンペティション



日本で主催したコンペとしては3例目。カグルで初めての人文系コンペ。

kaggle くずし字認識コンペティション：くずし字データセット

学習データ：画像



学習データ：文字のUnicodeと文字の画像座標

1	image_id	labels							
2	100241706_00004_2	U+306F 1231 3465 133 53 U+304C 275 1652 84 69 U+3044 1495 1218 143							
3	100241706_00005_1	U+306F 1087 2018 103 65 U+304B 1456 1832 40 73 U+304B 2036 1722 65							
4	100241706_00005_2	U+306F 572 1376 125 57 U+306E 1551 2080 69 68 U+3078 891 1731 163 6							
5	100241706_00006_1	U+3082 1455 3009 65 44 U+516B 1654 1528 141 75 U+309E 1714 1106 80							
6	100241706_00007_2	U+309D 1201 2949 27 33 U+309D 1196 1539 27 36 U+309D 729 2209 27 3							
7	100241706_00008_1	U+25B2 1056 929 96 39 U+309D 379 1098 21 43 U+25B2 1302 928 92 43 U							
8	100241706_00008_2	U+25B2 1648 955 95 44 U+25B2 1887 947 96 45 U+25B2 924 940 108 47 U							
9	100241706_00009_1	U+3078 1551 2071 104 41 U+3078 323 1473 135 43 U+3078 1060 1799 120							
10	100241706_00009_2	U+309D 1452 1423 20 37 U+3078 690 2535 121 41 U+309D 237 1709 25 4							
11	100241706_00010_1	U+3078 587 1127 100 40 U+3064 1103 1480 60 45 U+306F 1322 2076 75 4							
12	100241706_00010_2	U+309D 493 1972 20 37 U+25B2 697 935 100 44 U+309E 481 3348 56 48 U							
13	100241706_00011_1	U+309D 1823 1225 21 35 U+25B2 1548 936 91 43 U+25B2 580 920 93 43 U							
14	100241706_00011_2	U+309D 1208 1258 24 45 U+25B2 432 932 111 57 U+3078 1638 1404 115 6							

画像は28冊の古典籍から、全3881枚。文字数は約68万文字。



くずし字認識コンペティション：くずし字データセット

テストデータ：画像のみ。全4551枚



kaggle くずし字認識コンペティション：結果

7月19日～10月14日まで約3ヶ月

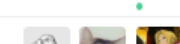
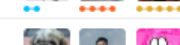
参加チーム： 293 チーム

参加者数： 338 人

結果提出： 2652 回（認識された画像1200万枚）



くずし字認識コンペティション：結果

#	△pub	Team Name	Notebook	Team Members	Score ?	Entries
1	—	tascj			0.950	13
2	—	Konstantin Lopuhin			0.950	60
3	—	Kenji			0.944	161
4	▲1	YoudaoOCR			0.942	49
5	▼1	See--			0.940	42
6	—	abc			0.939	15
7	—	K_mat			0.934	20
8	—	t-hanya			0.920	21
9	—	Ollie, Nanashi, and Tom			0.910	35
10	—	Zenkei_R&D			0.903	144
11	—	masayai			0.903	12
12	▲5	Kirill Brodt (shad nsk)			0.901	4
13	▲1	James Day			0.901	33
14	▼1	NEU			0.900	54
15	▼3	s tatsuya			0.900	29

日本語やくずし字の知識がなくても参加ができたため、上位にはさまざまな国の機械学習エンジニア、研究者が入った。

kaggle くずし字認識コンペティション：意義

- くずし字認識が機械学習の研究対象であることを世界の人々に知ってもらえた。
- カグルで開催したため、数百人の参加者を集めることができた。
- コンペでみんなが競ったため、優れたアルゴリズムが生まれた。
- どんなアルゴリズムがくずし字認識に適しているかを、一気にテストすることができた。
- コンペから得られた知識は、今後のくずし字認識研究に有用である。

kaggle くずし字認識コンペティション：人文学研究へ

- 国文学研究資料館がデータセットを作成したことが、コンペの開催につながった。
- 主催者はくずし字に関する基礎知識を参加者に提供し、日本文化、特に古典文学への関心を高めた。
- 日本文学とAIのように、一見無関係のように思える分野につながりを作った。
- 情報学者や人文学者だけではコンペの開催は不可能だった。人文学と情報学のコラボレーションが重要である。

kaggle くずし字認識コンペティション：くずし字OCR現状とこれから

- コンペのテストデータは比較的きれいに木版印刷された古典籍を対象とするため、全国の300万点以上の古典籍で常に90%の認識精度を達成できるわけではない。
- 手書きの手紙や古文書はテストしていないが、OCRの需要は大きいため、今後は取り組む必要がある。
- コンペでは約4300文字種の認識を競ったが、古語に表れる文字種はもっと多い。異体字への対応も必要である。
- こうした問題を解決するためには、手法の改良よりも、**オープンなくずし字データセット**の構築がより重要である。

Kaggleくずし字認識 : <http://codh.rois.ac.jp/competition/kaggle/>

KuroNetサービス : <http://codh.rois.ac.jp/kuronet/>

KuroNet論文 : <https://arxiv.org/abs/1910.09433>

