



～ちはやふる～
産後カルタ大会

2019/7/25(Thu.)



5/30 待望の書籍発売！

全国書店にて
5/30
— 発売 —

産後カルタ

あるある！これがリアルなママライフ

著者：ママスタセレクト

発行：ぴあ株式会社 定価：1,100円（税別）

＼Instagramで大人気！／

多数テレビでも取り上げられた「#産後カルタ」がついに書籍化！現役ママたちが描く、愛と笑いと涙の出産&育児のあるあるエピソードに共感すること必至！！子育てで疲れたママに笑顔をもたらす1冊です！





読み札と取り札が一体化！

・ あいうえお順に
ならんでいない！

・ 旦那カルタなど
種類がいろいろある！





こんなカルタもある！

- ・ カタカナが混ざってる！
- ・ 数字も混ざってる！





カルタは全部でなんまいあるの？

- カルタ画像一覧はインスタグラムにありました！
 - https://www.instagram.com/mamastar_manga/
- Selenium(セレンウム)を使ってスクレイピング！
 - JavaScriptを使ったWebページの為
- 全部で900枚ほどありました(Wow!)
 - 2019年7月時点



あいうえお順にならべたいな！（その①）

• Google Cloud VisionAPI で文字を認識

- 意外と正しく取れない！（認識率50%）
- アニメ画像中に埋め込まれた文字の認識は得意でないらしい！



<読み札>

もうやめて！ じいじばあばが大暴走
家のキャパ越え買いすぎよ



<VisionAPI認識結果>

うやめて！ | しいじばあばが |
家のキャバ越え | 買いすぎよ



あいうえお順にならべたいな！(その①)

• Google Cloud VisionAPI で文字を認識

- 意外と正しく取れない...(認識率50%)
- アニメ画像中に埋め込まれた文字の認識は得意でないらしい...



<読み札>

でんちが全く追いつかない。
電池替えて～ ママ、コシ音なれへん...



<VisionAPI認識結果>

で、|んちが 全く| 追いつかない。|
電池替えて| ママ、コシ| 容れへん...
| 鳴らない楽器 | かない動物。



あいうえお順にならべたいな！(その②)

・ディープラーニングで文字認識モデルを自作

- {ひらがな(73)+カタカナ(73)+数字(0~9)}の計156クラス分類
- 教師データ:「ひらがな」はNDLラボから取得。その他は自前で収集
 - ・NDLラボ: 国立国会図書館(National Diet Library)の実験的なサービスを提供するサイト

The screenshot shows the NDL Lab website interface. On the left is a navigation menu with links like Home, Experimental Services, and Experimental Data. The main content area is titled 'Text Image Dataset (Hiragana 73 Characters Version) Test Release'. It includes a date '2016/09/27' and a paragraph explaining that the dataset is derived from the National Diet Library's digital collection. Below the text, there is a grid of 15 sample images showing the hiragana characters 'あ', 'い', and 'う' in various styles and backgrounds, each with a corresponding filename.

NDL Lab

ホーム » 実験データ » 文字画像データセット » 文字画像データセット(平仮名73文字版)を試験公開しました

文字画像データセット(平仮名73文字版)を試験公開しました

2016/09/27

国立国会図書館デジタルコレクションのインターネット公開資料から、平仮名の文字画像を切り出したデータセットを試験公開しました。画像形式はPNGで、計8万画像あります。

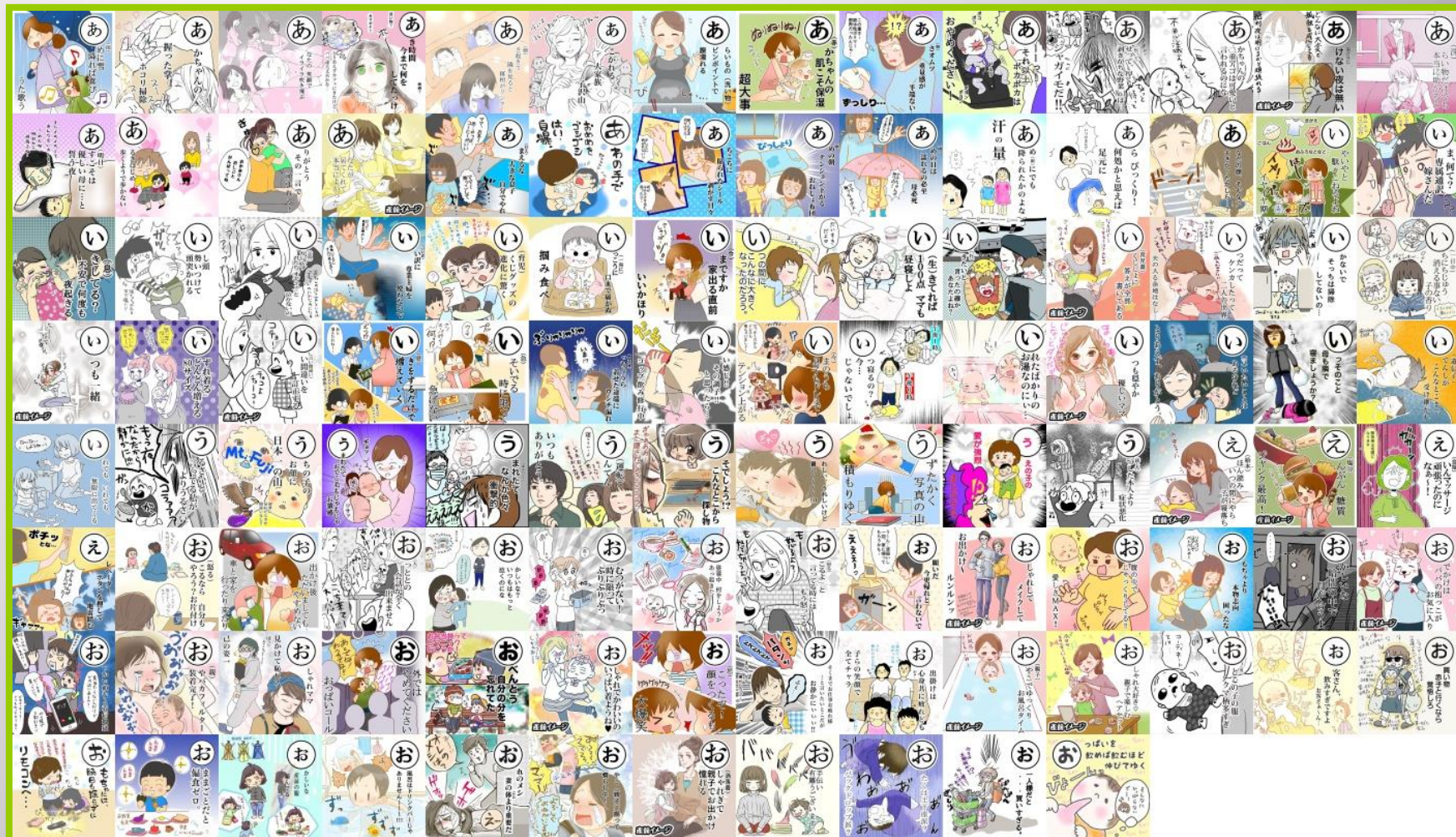
文字画像の例:

あ	あ	あ	あ	あ
1900_754949_0023.png	1910_808569_0198.png	1920_958609_0028.png	1930_1873255_0111.png	1940_1239059_0066.png
い	い	い	い	い
1900_808821_0042.png	1910_839986_0036.png	1920_906916_0039.png	1930_1175565_0024.png	1940_1873816_0084.png
う	う	う	う	う
1900_890806_0066.png	1910_813236_0091.png	1920_961711_0105.png	1930_1176246_0041.png	1940_1026632_0029.png

<https://lab.ndl.go.jp/cms/hiragana73>



結果①(あいうえお)





結果②(ちはやぶる)





結果③(いろはにほへと)





結果④(さんごかるた)





実は、文字認識モデルを作いたかったのです

<現状>

- 予測結果の答え合わせを行い、修正しているため正しい順番で出力されている
- 自作した文字認識モデルの認識精度は80%程度

<今後>

- 文字認識モデルの精度向上: 目標90%以上
- リアルタイム予測に基づき、その場で精確に並べ替えができれば理想！(cf.エッジコンピューティング)



おまけ①: こんなカルタもありますよ

・ フォントかるた

- フォントの種類で札を見分ける
- 計48フォント

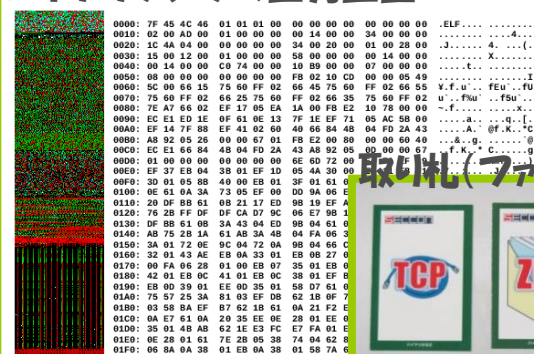
・ バイナリかるた

- バイナリダンプの出力画面からファイル形式を見分ける
- セキュリティコンテスト「SECCON」でバイナリ解析力を競う競技



<https://www.fontkaruta.com/>

バイナリダンプの出力画面

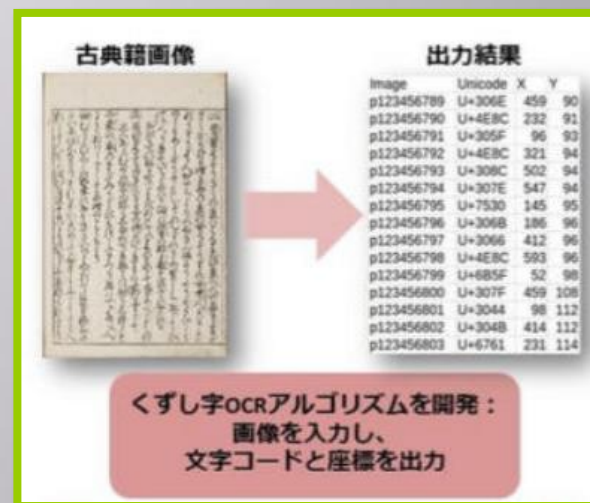


<http://kozoes.jp/binary-karuta/>



おまけ②: Kaggleでくずし字認識コンペ!

- 機械学習部も日本の文字文化への扉を叩きます!
- コンペ名:
 - 「くずし字認識: 千年に及ぶ日本の文字文化への扉を開く」
 - 「Kuzushiji Character Recognition:
~Opening the Door to A Thousand Years of Japanese Literate Culture~」
- 主催機関:
 - 人文学オープンデータ共同利用センター [CODH]
 - 国立情報学研究所 [NII]
 - 国文学研究資料館 [国文研]
- コンペ期間:
 - 2019年7月18日~10月14日(3ヶ月ちょい)
- 賞金総額: **\$15,000** (1~5位で一律\$3,000)
 - 2019年11月11日: 表彰式「日本文化とAI」





おまけ②: Kaggleでくずし字認識コンペ!

くずし字の特徴をいくつか知っていますか?

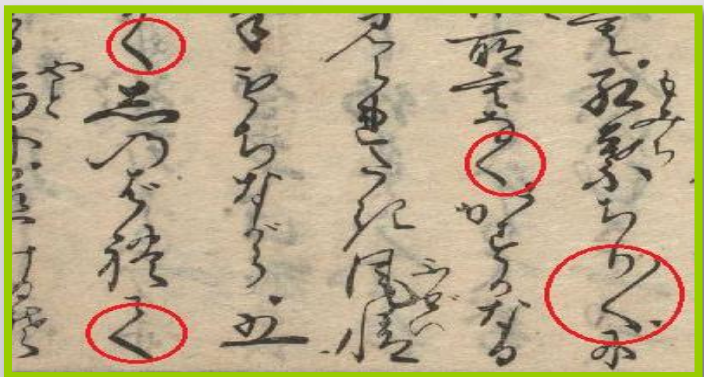
単一文字のバリエーション



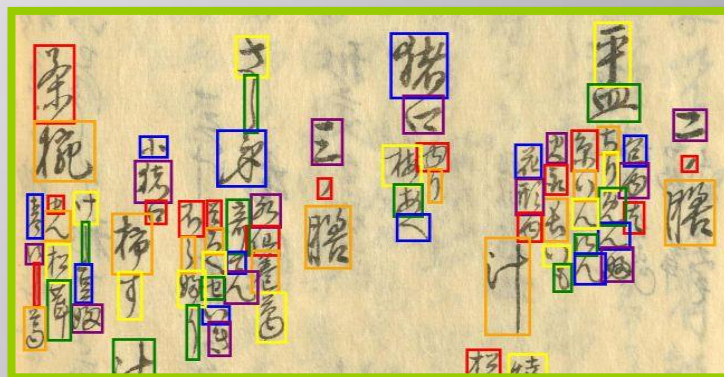
連結・重なりが多い



異なる文字の類似性



様々なレイアウトが可能





おまけ②: Kaggleでくずし字認識コンペ!

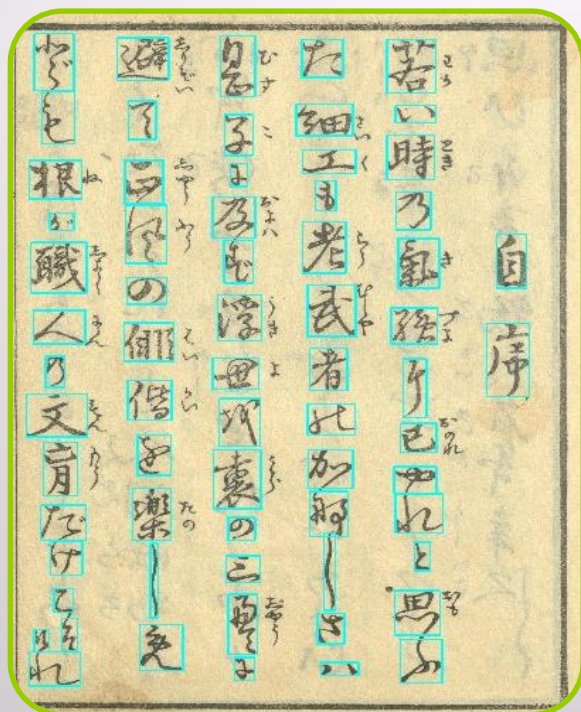
くずし字画像から、文字と書かれた位置を予測!

学習用データ:

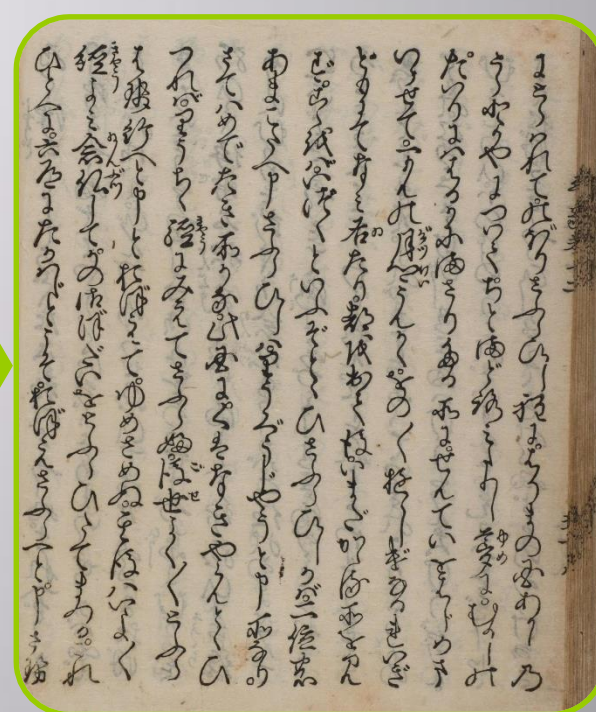
画像 + 文字と書かれた位置

予測用データ:

画像のみ



くずし字認識モデル





Fin.

