Comp150-NLP Final Project

by Alena Borisenko

°.✧ ╮(⁰▽⁰)╭ ✧.°

Idea

- Look at popular Japanese Twitter accounts
- See what words and characters are used
- Compare to popular Kanji lists and/or literature
- (Maybe) Generate a sentence
- Recommend ways to learn

Why Japanese?

Currently learning the language myself

もし日本語が読める人がいたら、教えて下さい！

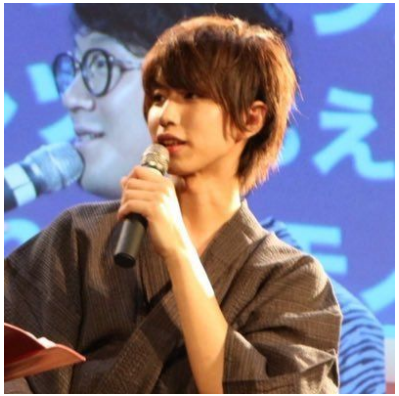
Curious to see how Japanese complicates things

- ➔ Regular expressions? (surprisingly easy)
- ➔ Tokenizer? (unsurprisingly hard)
- ➔ Looking cool? (easier than ever)

Data

Data: Twitter Accounts

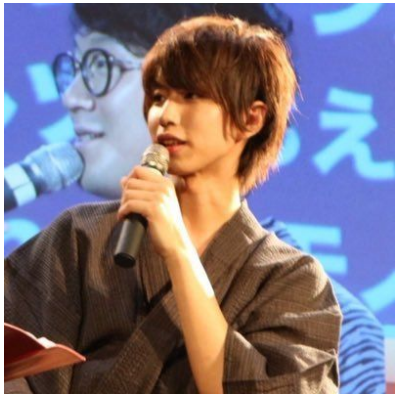
Data: Twitter Accounts



@hajimesyacho

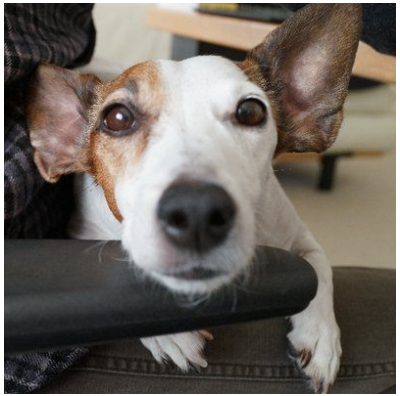
[YouTuber]
[24]

Data: Twitter Accounts



@hajimesyacho

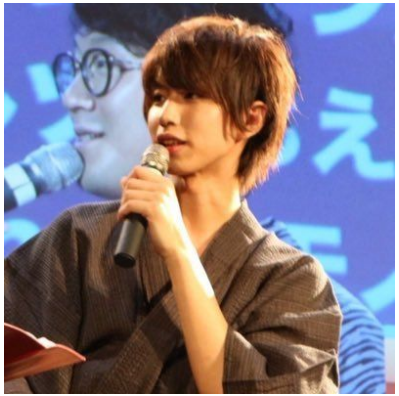
[YouTuber]
[24]



@itoi_shigesato

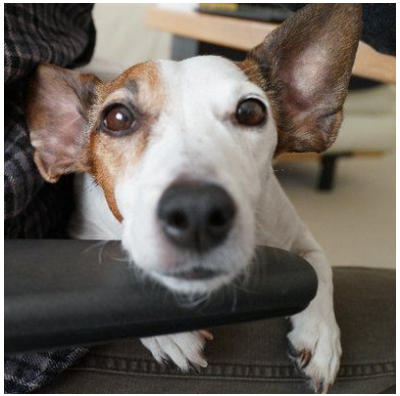
[Writer, Lyricist]
[69]

Data: Twitter Accounts



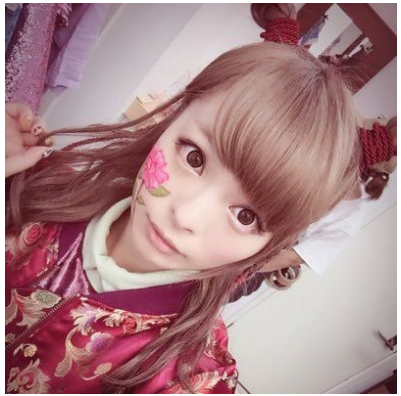
@hajimesyacho

[YouTuber]
[24]



@itoi_shigesato

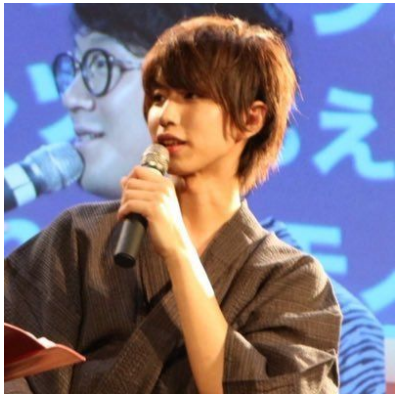
[Writer, Lyricist]
[69]



@pamyurin

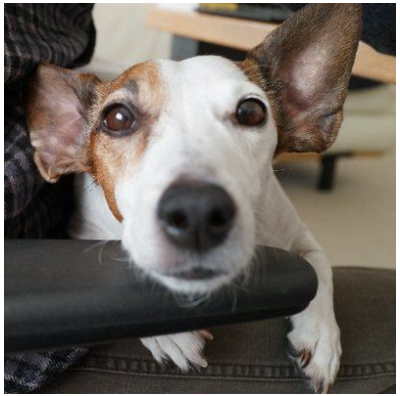
[J-Pop Singer]
[24]

Data: Twitter Accounts



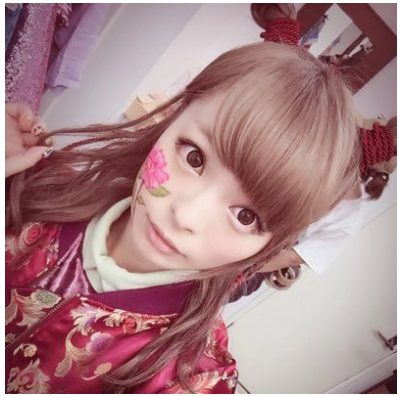
@hajimesyacho

[YouTuber]
[24]



@itoi_shigesato

[Writer, Lyricist]
[69]



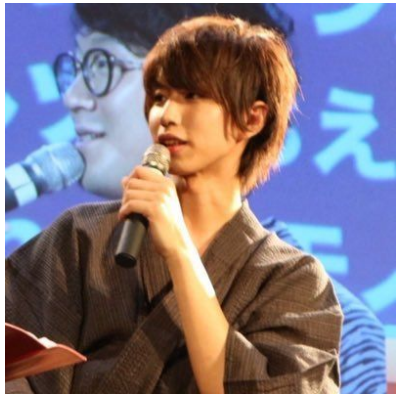
@pamyurin

[J-Pop Singer]
[24]

@nhk_news

[News Org]
[N/A]

Data: Twitter Accounts

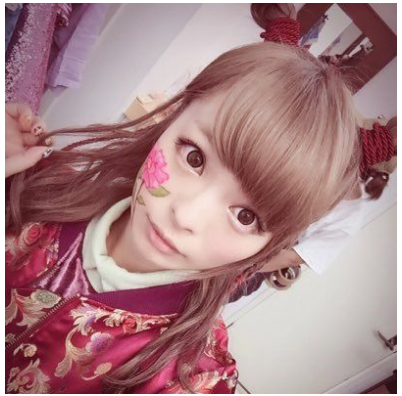


@hajimesyacho

[YouTuber]
[24]

@itoi_shigesato

[Writer, Lyricist]
[69]



@pamyurin

[J-Pop Singer]
[24]

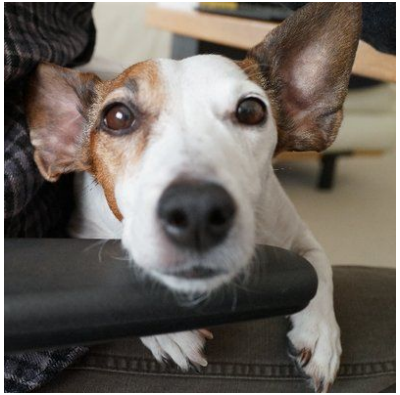
@nhk_news

[News Org]
[N/A]

Data: Twitter Accounts

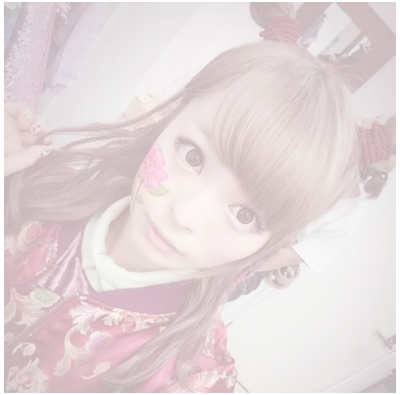
@hajimesyacho

[YouTuber]
[24]



@itoi_shigesato

[Writer, Lyricist]
[69]



@pamyurin

[J-Pop Singer]
[24]

@nhk_news

[News Org]
[N/A]

Data: Kanji

---

tangorin.com

Lists of common kanji

JLPT (N5 to N1)

- ➔ Japanese Language Proficiency-Test
- ➔ For non-native speakers (like TOEFL)

Jōyō (G1 to G7)

- ➔ Taught in primary and secondary school
- ➔ Issued by the Japanese Ministry of Education

Tangent: JLPT Learning Curve is Insane

JLPT N5

日 一 国 人 年 大 十 二 本 中 長 出 三 時 行 見 月 後 前 生 五 間 上 東 四 今 金 九 入 学 高 円 子 外 八 六 下 来 気 小 七 山 話 女 北 午 百 書 先 名 川 千 水 半 男 西 電 校 語 土
木 聞 食 車 何 南 万 毎 白 天 母 火 右 読 友 左 休 父 雨

JLPT N4

会 同 事 自 社 発 者 地 業 方 新 場 員 立 開 手 力 問 代 明 動 京 目 通 言 理 体 田 主 題 意 不 作 用 度 強 公 持 野 以 思 家 世 多 正 安 院 心 界 教 文 元 重 近 考 画 海 売 知 道
集 別 物 使 品 計 死 特 私 始 朝 運 終 台 広 住 真 有 口 少 町 料 工 建 空 急 止 送 切 転 研 足 究 楽 起 着 店 病 質 待 試 族 銀 早 映 親 験 英 医 仕 去 味 写 字 答 夜 音 注 帰 古
歌 買 悪 図 週 室 歩 風 紙 黒 花 春 赤 青 館 屋 色 走 秋 夏 習 駅 洋 旅 服 夕 借 曜 飲 肉 貸 堂 鳥 飯 勉 冬 昼 茶 弟 牛 魚 兄 犬 妹 姉 漢

JLPT N3

政 議 民 連 対 部 合 市 内 相 定 回 選 米 実 関 決 全 表 戦 経 最 現 調 化 当 約 首 法 性 要 制 治 務 成 期 取 都 和 機 平 加 受 続 進 数 記 初 指 権 支 産 点 報 済 活 原 共 得 解
交 資 予 向 付 際 由 面 告 反 判 認 参 利 組 信 消 在 件 番 側 任 引 求 害 次 昨 論 官 増 係 乗 残 感 情 投 声 念 変 助 打 労 存 幸 吹 娘 誤 洗 慣 礼 窓 昔 貧 怒 祖 泳 杯 疲 皆 鳴 腹 煙 眠 怖 耳 頂 箱 晩 寒 髪 忙
役 構 割 費 横 深 説 難 優 夫 収 断 石 違 良 候 神 番 規 術 備 宅 害 育 警 席 太 訪 乗 客 否 想 師 声 登 易 速 絶 雪 忘 歯 酒 悲 晴 掛 到 寝 暗 盗 吸 陽 御 婚 喜 浮 忘 雪 吹 娘 誤 洗 慣 礼 窓 昔 貧 怒 祖 泳 杯 疲 皆 鳴 腹 煙 眠 怖 耳 頂 箱 晩 寒 髪 忙
非 観 察 段 込 顔 申 様 余 王 港 妻 呼 達 熟 夢 君 閉 険 緒 折 草 暮 酒 悲 晴 掛 到 寝 暗 盗 吸 陽 御 婚 喜 浮 忘 雪 吹 娘 誤 洗 慣 礼 窓 昔 貧 怒 祖 泳 杯 疲 皆 鳴 腹 煙 眠 怖 耳 頂 箱 晩 寒 髪 忙
冷 適 婦 寄 込 顔 申 類 余 王 返 妻 背 熟 夢 君 閉 険 緒 折 草 暮 酒 悲 晴 掛 到 寝 暗 盗 吸 陽 御 婚 喜 浮 忘 雪 吹 娘 誤 洗 慣 礼 窓 昔 貧 怒 祖 泳 杯 疲 皆 鳴 腹 煙 眠 怖 耳 頂 箱 晩 寒 髪 忙
越 欲 痛 笑 互 束 似 列 探 逃 遊 迷 夢 君 閉 険 緒 折 草 暮 酒 悲 晴 掛 到 寝 暗 盗 吸 陽 御 婚 喜 浮 忘 雪 吹 娘 誤 洗 慣 礼 窓 昔 貧 怒 祖 泳 杯 疲 皆 鳴 腹 煙 眠 怖 耳 頂 箱 晩 寒 髪 忙
才 靴 恥 偶 偉 猫 幾

JLPT N2

党 協 総 区 領 県 設 改 府 査 委 軍 団 各 島 革 村 勢 減 再 税 営 比 防 補 境 導 副 算 輸 述 線 農 州 武 象 域 額 欧 担 準 賞 辺 造 被 技 低 復 移 個 門 課 脳 極 含 蔵 量 型 況 針
専 谷 史 階 管 兵 接 停 爆 希 封 毒 叫 拝 氷 乾 棒 祈 拾 粉 糸 綿 汗 銅 湿 瓶 咲 召 缶 隻 脂 蒸 肌 耕 鈍 泥 隅 灯 辛 磨 麦 姓 筒 鼻 粒 詞 胃 畳 虜 机 濯 塔 灰 沸 菓 帽 枯 涼 舟 貝 符 憎 皿
均 億 圧 芸 署 伸 停 爆 陸 玉 波 籍 帯 汚 賢 腕 兆 床 郵 固 橋 岸 則 仲 毛 緑 尊 祝 板 測 戸 央 豊 骨 柔 殿 届 濃 液 衣 肩 零 幼 荷 泊 黄 甘 臣 浅 掃 雲 掘 捨 軟 沈 凍 乳 恋 紅 郊 腰 炭 踊 冊 勇 械 菜 珍 卵 湖 皮 泉 星 永 著 漁 喫 干 虫 庫 貯 刊 硬 埋 湯 溶 柱
坂 底 布 寺 宇 巨 震 希 触 胸 砂 塩 賢 枚 複 郵 仲 栄 札 板 祝 柔 骨 傾 届 巻 燃 跡 包 駐 弱 紹 雇 替 預 焼 簡 章 臓 律 贈 照 薄 群 秒 奥 詰 双 刺 純 翌 快 片 敬 悩 温 季 仏 績 築 貨 混 昇 池 血 温 敬 悩 泉 皮 湖 喫 涼 舟 貝 符 憎 皿
祭 袋 筆 訓 浴 童 宝 封 胸 毒 叫 拝 氷 乾 棒 祈 拾 粉 糸 綿 汗 銅 湿 瓶 咲 召 缶 隻 脂 蒸 肌 耕 鈍 泥 隅 灯 辛 磨 麦 姓 筒 鼻 粒 詞 胃 畳 虜 机 濯 塔 灰 沸 菓 帽 枯 涼 舟 貝 符 憎 皿
鉱 涙 匹 孫 鋭 枝 塗 軒 毒 叫 拝 氷 乾 棒 祈 拾 粉 糸 綿 汗 銅 湿 瓶 咲 召 缶 隻 脂 蒸 肌 耕 鈍 泥 隅 灯 辛 磨 麦 姓 筒 鼻 粒 詞 胃 畳 虜 机 濯 塔 灰 沸 菓 帽 枯 涼 舟 貝 符 憎 皿
肯 燥 畜 挟 曇 滴 伺

JLPT N1

氏 統 保 第 結 派 案 策 基 価 提 挙 応 企 検 藤 沢 裁 証 援 施 井 護 展 態 鮮 視 条 幹 独 宮 率 衛 張 監 環 審 義 訴 株 姿 閣 衆 評 影 松 撃 佐 核 整 融 製 票 渉 響 推 請 器 士 討
攻 崎 督 授 催 及 憲 離 激 摘 系 批 郎 健 盟 従 修 隊 織 拡 故 振 弁 就 異 献 厳 維 浜 遺 塁 邦 素 遣 抗 模 雄 益 緊 標 宣 昭 廃 伊 江 僚 吉 盛 皇 臨 踏 壊 債 興 源 儀 創 障 継 筋
闘 葬 避 司 康 善 逮 迫 惑 崩 紀 聴 脱 級 博 締 救 執 房 撤 削 密 措 志 載 陣 我 為 抑 幕 染 奈 傷 択 秀 徴 弾 償 功 拠 秘 拒 刑 塚 致 繰 尾 描 鈴 盤 項 喪 伴 養 懸 街 契 掲 躍 棄
邸 縮 還 属 慮 枠 恵 露 沖 緩 節 需 射 購 揮 充 貢 鹿 却 端 賃 獲 郡 併 徹 貴 衝 焦 奪 災 浦 析 譲 称 納 樹 挑 誘 紛 至 宗 促 慎 控 智 握 宙 俊 銭 渋 銃 操 携 診 託 撮 誕 侵 括 謝
駆 透 津 壁 稲 仮 裂 敏 是 排 裕 堅 訳 芝 綱 典 賀 扱 顧 弘 看 訟 戒 祉 誉 歓 奏 勧 騒 閥 甲 縄 郷 揺 免 既 薦 隣 華 範 隠 徳 哲 杉 釈 己 妥 威 豪 熊 滞 微 隆 症 暫 忠 倉 彦 肝 喚
沿 妙 唱 阿 索 誠 襲 懇 俳 柄 驚 麻 李 浩 剤 瀬 趣 陥 斎 貫 仙 慰 序 旬 兼 聖 旨 即 柳 舎 偽 較 覇 詳 抵 脅 茂 犠 旗 距 雅 飾 網 竜 詩 繁 翼 潟 敵 魅 嫌 斉 敷 擁 圏 酸 滅 罰 礎 腐
脚 潮 梅 尽 僕 桜 滑 孤 炎 賠 句 鋼 頑 鎖 彩 摩 励 縦 輝 蓄 軸 巡 稼 瞬 砲 噴 誇 祥 牲 秩 帝 宏 唆 阻 泰 賄 撲 堀 菊 絞 縁 唯 膨 矢 耐 塾 漏 慶 猛 芳 懲 剣 彰 棋 丁 恒 揚 冒 之 倫
陳 憶 潜 梨 仁 克 岳 概 拘 墓 黙 須 偏 雰 遇 諮 狭 卓 亀 糧 簿 炉 牧 殊 殖 艦 輩 穴 奇 慢 鶴 謀 暖 昌 拍 朗 寛 覆 胞 泣 隔 浄 没 暇 肺 貞 靖 鑑 飼 陰 銘 随 烈 尋 稿 丹 啓 也 丘 棟
壌 漫 玄 粘 悟 舗 妊 熟 旭 恩 騰 往 豆 遂 狂 岐 陛 緯 培 衰 艇 屈 径 淡 抽 披 廷 錦 准 暑 磯 奨 浸 剰 胆 繊 駒 虚 霊 帳 悔 諭 惨 虐 翻 墜 沼 据 肥 徐 糖 搭 盾 脈 滝 軌 俵 妨 擦 鯨
荘 諾 雷 漂 懐 勘 栽 拐 駄 添 冠 斜 鏡 聡 浪 亜 覧 詐 壇 勲 魔 酬 紫 曙 紋 卸 奮 欄 逸 涯 拓 眼 獄 尚 彫 穏 顕 巧 矛 垣 欺 釣 萩 粛 栗 愚 嘉 遭 架 鬼 庶 稚 滋 幻 煮 姫 誓 把 践 呈
疎 仰 剛 疾 征 砕 謡 嫁 謙 后 嘆 菌 鎌 巣 頻 琴 班 棚 潔 酷 宰 廊 寂 辰 霞 伏 碁 俗 漠 邪 晶 墨 鎮 洞 履 劣 那 殴 娠 奉 憂 朴 亭 淳 怪 鳩 酔 惜 穫 佳 潤 悼 乏 該 赴 桑 桂 髄 虎 盆
晋 穂 壮 堤 飢 傍 疫 累 痴 搬 晃 癒 桐 寸 郭 尿 凶 吐 宴 鷹 賓 虜 陶 鐘 憾 猪 紘 磁 弥 昆 粗 訂 芽 庄 傘 敦 騎 寧 循 忍 怠 如 寮 祐 鵬 鉛 珠 凝 苗 獣 哀 跳 匠 垂 蛇 澄 縫 僧 眺 亘
呉 凡 憩 媛 溝 恭 刈 睡 錯 伯 笹 穀 陵 霧 魂 弊 妃 舶 餓 窮 掌 麗 綾 臭 悦 刃 縛 暦 宜 盲 粋 辱 毅 轄 猿 弦 稔 窒 炊 洪 摂 飽 冗 桃 狩 朱 渦 紳 枢 碑 鍛 刀 鼓 裸 猶 塊 旋 弓 幣 膜
扇 腸 槽 慈 楊 伐 駿 漬 糾 亮 墳 坪 紺 娯 椿 舌 羅 峡 俸 厘 峰 圭 醸 蓮 弔 乙 汁 尼 遍 衡 薫 猟 羊 款 閲 偵 喝 敢 胎 酵 憤 豚 遮 扉 硫 赦 窃 泡 瑞 又 慨 紡 恨 肪 扶 戯 伍 忌 濁 奔
斗 蘭 迅 肖 鉢 朽 殻 享 秦 茅 藩 沙 輔 媒 鶏 禅 嘱 胴 迭 挿 嵐 椎 絹 陪 剖 譜 郁 悠 淑 帆 暁 傑 楠 笛 玲 奴 錠 拳 翔 遷 拙 侍 尺 峠 篤 肇 渇 叔 雌 亨 堪 叙 酢 吟 逓 嶺 甚 喬 崇 漆
岬 癖 愉 寅 礁 乃 洲 屯 樺 槙 姻 巌 擬 塀 唇 睦 閑 胡 幽 峻 曹 詠 卑 侮 鋳 抹 尉 槻 隷 禍 蝶 酪 茎 帥 逝 汽 琢 匿 襟 蛍 蕉 寡 琉 痢 庸 朋 坑 藍 賊 搾 畔 遼 唄 孔 橘 漱 呂 拷 嬢 苑
巽 杜 渓 翁 廉 謹 瞳 湧 欣 窯 褒 醜 升 殉 煩 巴 禎 劾 堕 租 稜 桟 倭 婿 慕 斐 罷 矯 某 囚 魁 虹 鴻 泌 於 赳 漸 蚊 葵 厄 藻 禄 孟 嫡 尭 嚇 巳 凸 暢 韻 霜 硝 勅 芹 杏 棺 儒 鳳 馨 慧
愁 楼 彬 匡 眉 欽 薪 褐 賜 嵯 綜 繕 栓 翠 鮎 榛 凹 艶 惣 蔦 錬 隼 渚 衷 逐 斥 稀 芙 詔 皐 雛 惟 佑 耀 黛 渥 憧 宵 妄 惇 脩 甫 酌 蚕 嬉 蒼 暉 頒 只 肢 檀 凱 彗 謄 梓 丑 嗣 叶 汐 絢
朔 伽 畝 抄 爽 黎 惰 蛮 冴 萌 旺 壱 偲 瑠 允 侯 蒔 鯉 弧 遥 舜 瑛 附 彪 卯 但 綺 芋 茜 凌 皓 洸 毬 婆 緋 鯛 怜 邑 倣 碧 啄 穣 酉 悌 倹 柚 繭 亦 詢 采 紗 賦 眸 玖 弐 錘 諄 倖 痘 笙
侃 裟 洵 爾 耗 昴 銑 莞 伶 碩 宥 滉 晏 伎 朕 迪 綸 且 竣 晨 吏 燦 麿 頌 箇 楓 琳 梧 哉 澪 匁 晟 衿 凪 梢 丙 颯 茄 勺 恕 蕗 瑚 遵 瞭 燎 虞 柊 侑 謁 斤 嵩 捺 蓉 茉 袈 燿 誼 冶 栞 墾
勁 菖 旦 椋 叡 紬 胤 凜 亥 爵 脹 麟 莉 汰 瑶 瑳 耶 椰 絃 丞 璃 奎 塑 昂 柾 熙 菫 諒 鞠 崚 濫 捷

Back-end

Tweepy

Simplifies use of Twitter API

Handles authentication

Search limitations:

- ➔ 200 tweets per search
- ➔ 7 day window
- ➔ ~3200 recent tweets for a single user
- ➔ Limited number of requests

Solution: get_tweets.py

Set up tweepy

Get recent tweets from api

Write an .xml file

tweets.xml

ー ー ー

```
<tweets>
    <tweet>
        <num>1</num>
        <id_str>938251171617128449</id_str>
        <user>pamyurin</user>
        <lang>ja</lang>
        <text>
            鍵が壊れてトイレに閉じ込められて大変だった、、ふぅ（ ◠‿◠ ）
        </text>
    </tweet>

    <tweet>
        <num>2</num>
        <id_str>937951688958263296</id_str>
        <user>pamyurin</user>
        <lang>ja</lang>
        <text>
            ○✧ https://t.co/hihYrB1m7L
        </text>
    </tweet>
</tweets>
```

get_kanji.py

Read in the .csv kanji guides from tangorin

Get the kanji

Write to separate file for later use

project.py

Set up Flask

Use a regular expression to isolate kanji

Use a tokenizer to isolate words (WIP)

Compare results to JLPT/Jōyō kanji lists

Generate a sentence (WIP)

Front-end

index.html

Simple HTML form

POSTs submitted username to script via Flask

result.html

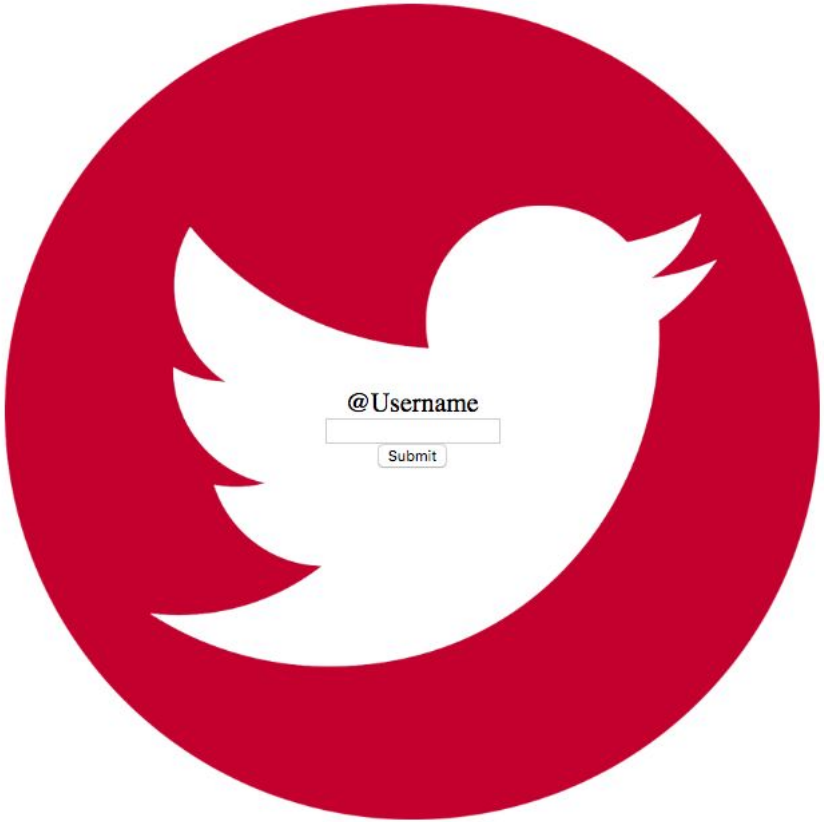
Another simple HTML template

Rendered from project.py with resulting data

Currently mostly wrestling with Flask

Demo

[Warning: It’s not pretty]


@Username
Submit

Stats for user @hajimesyacho

Percent of JLPT kanji used:

- N5: 87.34%
- N4: 86.75%
- N3: 66.76%
- N2: 34.88%
- N1: 15.58%

Percent of Jōyō kanji used:

- G1: 85.00%
- G2: 75.00%
- G3: 72.50%
- G4: 52.00%
- G5: 43.24%
- G6: 36.46%
- G7: 20.77%

Most common kanji:

[('日', 255), ('今', 214), ('動', 171), ('画', 170), ('寝', 83), ('大', 75), ('気', 73), ('食', 57), ('張', 53), ('人', 48), ('頑', 46), ('出', 42), ('間', 42), ('丈', 40), ('夫', 40), ('岡', 39), ('最', 38), ('様', 38), ('来', 37), ('時', 36)]

Similar writing styles:

WIP...

Stats for user @nhk_news

Percent of JLPT kanji used:

- N5: 100.00%
- N4: 98.80%
- N3: 94.55%
- N2: 86.92%
- N1: 53.49%

Percent of Jōyō kanji used:

- G1: 98.75%
- G2: 95.00%
- G3: 95.50%
- G4: 94.50%
- G5: 92.43%
- G6: 91.16%
- G7: 61.55%

Most common kanji:

[('日', 657), ('大', 592), ('国', 533), ('人', 526), ('会', 496), ('北', 425), ('本', 379), ('朝', 376), ('発', 371), ('鮮', 362), ('年', 341), ('事', 328), ('議', 308), ('中', 295), ('新', 286), ('相', 281), ('見', 262), ('開', 261), ('子', 253), ('場', 245)]

Similar writing styles:

WIP...

Final Notes

Remaining Tasks

- ➔ Stop relying on pre-generated .xml file
- ➔ Figure out how to deal with Auth keys/secrets
- ➔ Keep trying to get the tokenizer working
- ➔ Prettify
- ➔ Add WaniKani statistics if there is time

Questions?

Thank you!

(´ ▽ `)ノ