Thu thập dữ liệu song ngữ chữ Nôm - chữ Quốc Ngữ

Giảng viên hướng dẫn:
PGS. TS. Đinh Điền
ThS. Lê Thị Thúy Hằng
ThS. Nguyễn Hồng Bửu Long

Nhóm thực hiện:
Nguyễn Thị Thu Duyên - 22C11005
Đặng Hoàng Minh Triết - 22C11048

Nội dung trình bày

1. Giới thiệu chung
2. Phương pháp triển khai
3. Kết quả thực nghiệm
4. Kết luận và hướng phát triển

Giới thiệu chung

Bối cảnh:

- Các tài liệu bằng chữ Nôm chứa đựng nguồn thông tin và tri thức của dân tộc ta nhiều đáng kể
- Dữ liệu song ngữ chữ Nôm - chữ Quốc Ngữ còn hạn chế

Giới thiệu chung

Phát biểu bài toán:

- **Đầu vào:** Một trang web chứa nội dung chữ Nôm và chữ Quốc Ngữ. Trong khuôn khổ đồ án này chúng em sử dụng trang web Thơ Văn Hán - Nôm | Scribd

- **Đầu ra:** Một tệp excel trong đó có một cột là chữ Nôm, một cột là chữ Quốc Ngữ và một cột là số Nôm tự được trích xuất từ trang web đầu vào

Phương pháp triển khai

1. Thu thập các đường dẫn liên quan từ URL
2. Tải nội dung về các tệp HTML
3. Thu thập nội dung trong các tag cần thiết
4. Thu thập hình ảnh

Phương pháp triển khai

5. Xử lý nội dung ngôn ngữ
   a. Gọi Nom Transliteration HCMUS và HVDict Translate API dịch làm chuẩn
   b. Mapping bản dịch và dữ liệu vừa crawl
   c. Phân loại dữ liệu dựa trên độ tương đồng so với bản dịch
6. Lưu trữ nội dung ngôn ngữ đã xử lý dưới dạng Excel

Kết quả thực nghiệm

1. Thu thập các đường dẫn liên quan từ URL

```
if __name__ == '__main__':
    main("https://www.scribd.com/lists/21643875/Th%C6%A1-V%C4%83n-Han-Nom","/content/drive/MyDrive/XLNNTN-MSc-S2/images")
```

https://www.scribd.com/document/354356904/Thập-Giới-Co-Hồn-Quốc-Ngữ-Văn-Le-Thanh-Tong
https://www.scribd.com/document/351003106/Thien-Nam-Ngữ-Lục-Thơ-Nom
https://www.scribd.com/document/354358928/Cung-Oan-Ngam-Khuc-Nguyễn-Gia-Thiều-Le-Văn-Đặng-Diễn-Nom
https://www.scribd.com/document/354358921/Chinh-Phụ-Ngam-Diễn-Ca-Nguyễn-Văn-Xuan
https://www.scribd.com/document/354359286/Chinh-Phụ-Ngam-Han-Việt
https://www.scribd.com/document/351002527/Ma-Thanh-Tan-Truyện-Lý-Hồng-Phượng
https://www.scribd.com/document/354356908/Bạch-Van-Am-Quốc-Ngữ-Nguyễn-Bỉnh-Khiem
https://www.scribd.com/document/354356906/Bai-Tựa-Đại-Nam-Quốc-Ngữ
https://www.scribd.com/document/354361314/Minh-Tam-Bửu-Giam-Diễn-Nom
https://www.scribd.com/document/354361116/Nữ-Tử-Tu-Tri-Diễn-Nom
https://www.scribd.com/document/354361115/Mộng-Lien-Đinh-Thi-Thảo-Nguyễn-Gia-Tuyển
https://www.scribd.com/document/354359991/Giao-Tử-Phu-Mạc-Đĩnh-Chi
https://www.scribd.com/document/354359995/Hồ-Xuan-Hương-Thi-Tập-Le-Văn-Đặng-Diễn-Nom
https://www.scribd.com/document/354358926/Cac-Bai-Tuyệt-Cu-Trong-Quốc-Am-Thi-Tập-Nguyễn-Đinh-Hoa-Le-Văn-Đặng

Kết quả thực nghiệm

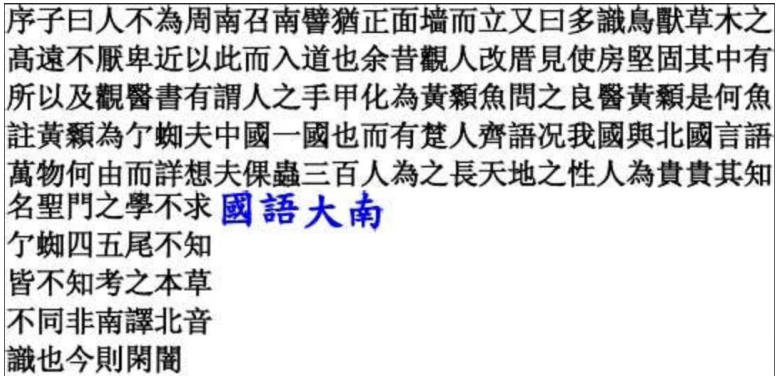
2. Tải nội dung các tệp HTML


```
<!DOCTYPE HTML><html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en" xmlns:og="http://opengraphprotocol.org/schema/"
xmlns:fb="http://www.facebook.com/2008/fbml"><head prefix="og: http://ogp.me/ns# scribd-com: http://ogp.me/ns/apps/scribd-com#">
<link rel="preconnect" href="https://s-f.scribdassets.com/"><link rel="preconnect" href="https://html.scribdassets.com"><link
rel="preload" crossorigin="anonymous" href="https://s-
f.scribdassets.com/webpack/assets/fonts/source_sans_pro/regular/source_sans_pro_regular.latin.e8ecbdac.woff2" as="font"
type="font/woff2"><link rel="preload" crossorigin="anonymous" href="https://s-
f.scribdassets.com/webpack/assets/fonts/source_sans_pro/semibold/source_sans_pro_600.latin.76017e81.woff2" as="font"
type="font/woff2"><link rel="preload" href="https://cmp.osano.com/AzZdHGSGtpxCq1Cpt/4e10b135-d113-4574-a477-
270ace40bba7/osano.js?language=en" as="script"><script>window.Scribd = window.Scribd || {}; window.Scribd.config =
{"js_logging_enabled":false,"raise_js_callback_errors":false,"session_domain":".scribd.com","csrf_token_url":"https://www.scribd.
com/csrf_token","facebook":{"app_id":136494494209,"permissions":"email,public_profile","namespace":"scribd-
com","link_url":"/facebook_link"}}; window.Scribd.webpack_public_path = "https://s-f.scribdassets.com/webpack/monolith/";
window.Scribd.AssetPath = "https://s-f.scribdassets.com/";</script>
<style>.outer_page,.outer_page_container{position:relative}.outer_page{display:block;font-size:16px;margin:10px auto
20px;overflow:hidden;contain:strict}.outer_page.book_view{display:inline-block}.outer_page.blurred_page{-webkit-user-drag:none;-
webkit-user-select:none;-moz-user-select:none;-ms-user-select:none;user-select:none}.newpage{white-
space:nowrap;position:relative;top:0;left:0;text-
rendering:auto;color:#000}.image_layer{width:0;height:0;position:absolute;top:0;left:0}.image_layer
```

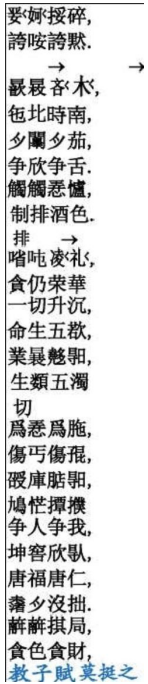

1898-Quảng-Tập-Viem-Văn-Edmond-Nordemann.html

Kết quả thực nghiệm

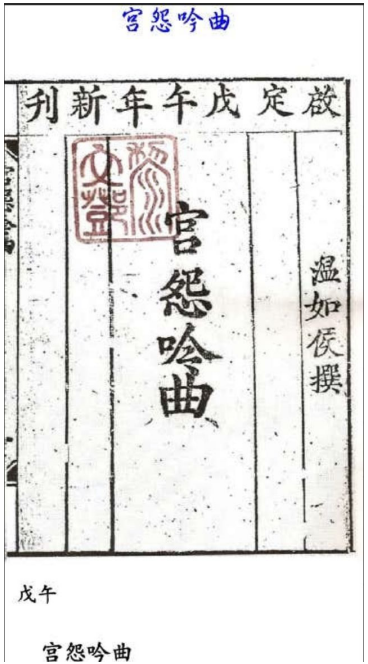
4. Thu thập hình ảnh


序子曰人不為周南召南譬猶正面墻而立又曰多識鳥獸草木之
高遠不厭卑近以此而入道也余昔觀人改厝見使房堅固其中有
所以及觀醫書有謂人之手甲化為黃顙魚問之良醫黃顙是何魚
註黃顙為𠯿蜘夫中國一國也而有楚人齊語況我國與北國言語
萬物何由而詳想夫倮蟲三百人為之長天地之性人為貴貴其知
名聖門之學不求 國語大南
𠯿蜘四五尾不知
皆不知考之本草
不同非南譯北音
識也今則閑閣

1-51059560b7.jpg


娶婀挼碎,
誇唆誇默.
→ →
聂裹吝木,
包北時南,
夕闌夕茄,
争欣争舌.
觸觸悉慮,
制排酒色.
排 →
啃吨凌礼,
食仍荣華
一切升沉,
命生五欸,
業裹觳鞊,
生類五濁
切
爲悉爲施,
傷丐傷飛,
瑕庫賠鞊,
鳩恺撢攫
争人争我,
坤窖欣馭,
唐福唐仁,
橐夕沒拙.
薜薜棋局,
貪色貪財,
教子賦莫挺之

1-31837694ee.jpg



1-a58d5af9a6.jpg

Kết quả thực nghiệm

3. Thu thập nội dung trong các tag cần thiết

```
1647;1954;798;亂時1. Loạn Thờ i
1820;862;1662.75;天地風塵
1954;862;1820;紅顏多屯
2088;862;1954;悠悠彼蒼兮誰造因
2223;862;2088;鼓鼙聲動長城月
2388;663;2223;5 烽火影照甘泉雲
2491;862;2388;九重按劍起當席
2625;862;2491;半夜飛檄 傳將軍
2759;862;2625;清平三百年天下
2893;862;2759;從 此 戎衣鷹武臣
3058;663;2893;10 使星天門催曉發
3161;862;3058;行人重法輕離別
3295;862;3161;弓箭兮在腰
3430;862;3295;妻孥兮別袂
3564;862;3430;獵獵旌旗兮出塞愁
3729;663;3564;15 喧喧簫鼓兮辭家怨
3832;862;3729;有怨兮分攜
3966;862;3832;有愁兮契闊
4100;862;3966;良人二十吳門豪
4234;862;4100;投筆硯兮事弓刀
4399;663;4234;20 直把連城 獻明聖
4502;862;4399;願將尺劍斬天驕
4637;862;4502;丈夫千里志馬革
4771;862;4637;泰山一擲輕鴻毛
4905;862;4771;便辭閨閫從 征 戰
5070;663;4911.0;25 西風鳴鞭出渭橋
1856;2321;5070;Thiên địa phong tr ần
1989;2321;1856;Hồng nhan đa truân
2121;2321;1989;Du du bi thươ ng hề thuỳ tạo nhân
2254;2321;2121;Cổ bề thanh động Tr ườ ng Thành nguyệt
2387;2321;2254;Phong hỏa ảnh chiếu Cam Tuyền vân
```

Kết quả thực nghiệm

5. Xử lý nội dung ngôn ngữ

```python
import json
import time

def hcmus_translate(text):
    url = 'https://api.clc.hcmus.edu.vn/sentencepairs/90/1'
    response = requests.request('POST', url, data={'nom_text': text})
    time.sleep(0.1)

    try:
        result = json.loads(response.text)['sentences']
        result = result[0][0]['pair']['modern_text']
        return result
    except:
        #print(f'[ERR] "{text}": {response.text}')
        return 'Cannot translate this text.'
```

```python
def hvdic_translate(text):
  count_Nom_Character = 0
  output = ""
  meaning = []
  if(text is not None):
    url = 'https://hvdic.thivien.net/transcript-query.json.php'
    headers = { 'Content-Type': 'application/x-www-form-urlencoded; charset=UTF-8' }

    # Request phonetics for Hán Việt (lang=1) first, if the response result is not
    # Hán Việt (contains blank lists) => Request phonetics for Nôm (lang=3)
    for lang in [1, 3]:
        payload = f'mode=trans&lang={lang}&input={text}'
        response = requests.request('POST', url, headers=headers, data=payload.encode())
        time.sleep(0.1)
        try:
            result = json.loads(response.text)['result']
        except:
            print(f'[ERR] {text}: {response.text}')
            result = {}
    for phonetics_dict in result:
      if phonetics_dict['t'] == 3 and len(phonetics_dict['o']) >= 0:
        output += phonetics_dict['i']
        meaning.append(phonetics_dict['o'])
        count_Nom_Character += 1
  return output, meaning, count_Nom_Character
```

Kết quả thực nghiệm

5. Xử lý nội dung ngôn ngữ

```python
import Levenshtein

def calculate_similarity(string1, string2):
    distance = Levenshtein.distance(string1, string2)
    max_length = max(len(string1), len(string2))
    similarity = (max_length - distance) / max_length * 100
    return similarity

string1 = "cho con ả thị đào"
string2 = "c h o c o n ả t h ị đ à o "

similarity_percentage = calculate_similarity(string1, string2)
print(f"Similarity: {similarity_percentage}%")

```

```
Similarity: 65.38461538461539%
```

```python
string = remove_unprintable_characters("诺□伴貝诺□□ Nớ c non bạn với nớ c non nhà.")

Nom_String, Meaning, Rate = hvdic_translate(string)
VN_String = parallel_remove_duplicates(Nom_String, string)

print(Rate)
print(check_String_Match(VN_String, Nom_String, Meaning))
print(Nom_String)
print(VN_String)

```

```
0.4444444444444444
('Reliable', ' nớ nớ bạn với')
诺伴貝诺
 Nớ c non bạn với nớ c non nhà.
```

Kết quả thực nghiệm

6. Lưu trữ nội dung ngôn ngữ đã xử lý dưới dạng tệp Excel

Top	Left	Start_top	Text	Translated by Document	Translated by Dictionary
547	663	0	TRÍCH Thập Giới Cô Hồn Quốc Ngữ Văn Lê Thánh Tông Soạn		
1428	663	1308	, 蔑瓶蔑鉢蔑袈裟		
1549	641	1428	Náu ⍰	hi'n giG /Gm	>a nhG?耨於廬伽茹
1669	663	1549	@inh "= ngI	/8u hư5ng /In t!iAn, 經芏楼香論篆	
1790	663	1669	定退禪院俸磋花	thi'n oay hoa? ---in---[Original] *3nh /ui thi'n iAnJ EKng oay hoa	định lui thiền viện bụng xoay hoa
1910	663	1790	Th8n t8m !>a +ạ	h quê hG hMu, ⍰	
795	663	673	S⍰ th3nh	On t!uy'n N3nh ThP	h
918	663	795	\⍰t m⍰		hVng quên E'
1038	663	918	L⍰		ao +ao ^hko /r ngư⍰i taX⍰⍰2
1188	None	1038	"5 (Thiên ⍰⍰n "3a /4&/ @A than !Bng		
1429	663	1309	*iểm huyAt tdm /ong ^h⍰p mIi nhG?黠⍰:œžŸ=	NgI	T⍰n g / u 8 n t h i ê n " a / ⍰ n / o n g + a , š luận) , › œ 7
1550	663	1429	Tạo "Oi phư5ng,	h8n "ạp tuy#t,C⍰ng	T⍰n g / u 8 n t h i ê n " a / ⍰ n / o n g + a , š luận) , › œ 7
1670	663	1550	T!ông tFm ^hP, mbt ⍰[y hoa?⍰:¡⍰O=	T⍰ng /u8n thiên "A /⍰n /ong +a,š論),›œ7	T⍰n g / u 8 n t h i ê n " a / ⍰ n / o n g + a , š luận) , › œ 7
1796	663	1675.5	Long EGn h<	ứ ⍰⍰m nhi'u th#,œ¢£¤⍰⍰	T⍰n g / u 8 n t h i ê n " a / ⍰ n / o n g + a , š luận) , › œ 7
1916	663	1796	sận th3nh th⍰i +uy "I	mIi	a?¥⍰¦§¨Ÿ⍰
3154	663	3034	Qao th⍰p ai ai "'u gi⍰p "ưW	,高⍰⍰0	
3275	663	3154	⍰i⍰p ngư⍰i +ao	hVng gi⍰p mFnh taX⍰⍰12	
553	663	5151	Qh`	mOng nhMng mŸi +⍰ giao	a?óô(õ⍰ö⍰
673	663	553	Ti#	⍰u8n ^hôn ti#	, ti#
795	641	673	晌	tháng ---in---[Original] NgGy tháng ai hdu ^_ "Wi taXTúUû2	tháng
947	None	795	"1 (Thư5ng	<& / @A than !Bng	
2806	663	2685	sui /Gm "Gn, ti#ng ⍰ư7ng, ti#ng	a?.弹嗜唱嗜⍰	
2926	663	2806	Qhkp miAng	⍰ng nhau !Bng /3	h +⍰,割/+⍰

Thập-Giới-Co-Hồn-Quốc-Ngữ-Văn-Le-Thanh-Tong.xlxs

Kết quả thực nghiệm

6. Lưu trữ nội dung ngôn ngữ đã xử lý dưới dạng tệp Excel

Top	Left	Start_top	Text	Translated by Document	Translated by Dictionary
1032	796	907	寒氷喝喝	hàn băng ---in---[Original] Hàn băng hắt hắt,	hàn băng hát hát
1171	796	1032	這	giá ---in---[Original] Giá lạnh căm căm,	giá
1305	796	1171.5	Gieo xuống một khi,		
1438	796	1310.9375	魂漂魄落	. Hồn xiêu phách lạc.	hồn xiêu phách lạc
1563	796	1510	冷漂	xiêu] ---in---[Original] [lãnh=lạnh][phiêu xiêu]	lạnh xiêu
1671	796	1568.25	57-60		
1789	796	1671	狂銅数	chó đồng ---in---[Original] Chó đồng miệng sủa,	chó đồng sổ
1922	899	1795.34375	仍火煙	yên, ---in---[Original] Ra những hỏa yên,	những hoả yên
2057	796	1927.9375	隊番	đòi phen, ---in---[Original] Chạy rảo đòi phen,	đòi phen
2189	809	2062.0	罪人覚落	. Tội nhơn xác lác.	tội nhân giác lạc
2315	796	2261	[giác xác]		
2415	796	2315	61-64		
2533	796	2415	鉄驢鉄馬	. Thiết lư thiết mã,	thiết lư thiết mã
2667	796	2605	鉄獣鉄鷹	thiết thiết thú ưng, ---in---[Original] Thiết thú thiết ưng,	thiết thú sắt ưng
2800	796	2672.9375	哺哏些	cắn ---in---[Original] Mổ cắn người ta,	bộ cắn ta
2936	796	2805.96875	昌排愕哺	xương bày ---in---[Original] Xương bày ngan ngácbộổ]	xương bày ngạc phô

Giao-Tử-Phu-Mạc-Đĩnh-Chi.xlxs

Kết quả thực nghiệm

6. Lưu trữ nội dung ngôn ngữ đã xử lý dưới dạng tệp Excel

	Top	Left	Start_top	Text	Translated by Document	Translated by Dictionary
2						
3	568	1871	0	征婦吟曲	征婦吟曲Chinh Phụ Ngâm Khúc征婦吟曲Chinh Phụ Ngâm Khúc	chinh phụ ngâm khúc
4	798	2565	568	鄧陳琨著		đặng trần côn trứ
5	1647	1954	798	亂時	loạn thờ ---in---[Original] 1. Loạn Thờ i	loạn thời
6	1820	862	1662.75	天地風塵	Thiên địa phong tr ần	thiên địa phong trần
27	4637	862	4502	丈夫千里志馬革	Tr ượ ng phu thiên lý chí mã cách	trượng phu thiên lý chí mã cách
28	4771	862	4637	泰山一擲輕鴻毛	Thái Sơ n nhất tr ịch khinh hồng mao	thái sơn nhất trịch khinh hồng mao
29	4905	862	4771	便辭閨閫從 征 戰	Tiện từ khuê khổn tùng chinh chiến	tiện từ khuê khổn tùng chinh chiến
30	5070	663	4911.0	25 西風鳴鞭出渭	Tây phong minh tiên xuất Vy kiều	25 tây phong minh tiên xuất vị kiều
31	5492	1762	5040	院	viện ---in---[Original] Viện Việt Học - 1 -	viện
32	549	1740	293.25	二出征	xuất chinh ---in---[Original] 2. Xuất Chinh	nhị xuất chinh
33	722	663	562.5	渭橋頭清水滿	Vy kiều đầu thanh thuỷ câu	vị kiều đầu thanh thuỷ câu
34	856	663	722	清水邊青草途	Thanh thuỷ biên thanh thảo đồ	thanh thuỷ biên thanh thảo đồ
35	991	663	856	送君處兮心悠悠	Tống quân xứ hề tâm du du	tống quân xứ hề tâm du du

Chinh-Phụ-Ngam-Han-Việt.xlxs

Kết luận

- Kết quả thực nghiệm cho thấy sự khả quan và đáng mong đợi trong phương pháp triển khai
- Tuy nhiên cần cải tiến và có hướng phát triển thêm cho các trường hợp khuyết chữ, lỗi phông, ký tự lạ…

Hướng phát triển

- Việc xử lý các chữ bị khiếm khuyết, lỗi phông, ký tự lạ cần được cân nhắc và cải tiến thêm
- Đối với dữ liệu ảnh đã thu thập được, nhóm hướng đến việc ứng dụng các tài nguyên có sẵn để trích xuất dữ liệu
- Một số trang chứa dữ liệu dạng pdf cần được thu thập và xử lý riêng

Cám ơn mọi người đã lắng nghe