Title: Chinese Idioms (Final Project)
Authors: Jin Zhao, Kun Li, Xiaojing Yan, Erik Andersen
Date: 11 May 2019

Final Project: Chinese Idioms

Code Submitted By: Erik Andersen

Team Member Contributions:

Jin: add_feature.py, animal features, sentiment features, pinyin features

I took part in brain storming the idea of the project, mainly worked on the add_feature.py, in which I added the animal zodiac information, character number, sentiment information in the corpus. I segmented the pinyin, I used the annotated dataset of Dalian University of Technology to train Naive Bayes classifier to classify the sentiment of the idioms that don't have the sentiment information. We supported the user to search for idioms that are about zodiac animals. Chinese idioms usually have multiple word to refer to the same kind of animal. For example, the "彘/pig" in "杀彘取子", the "豕/pig" in "豕突狼奔", and the "猪/pig" in "牧猪奴戏" are all referring to pig. In our search system, all those idioms can be retrieved for the user by only checking "pig" in the zodiac animal box. However, the characters that refer to animals also appear in people or place names, such as "司马昭", which contains "马/horse", we don't want to retrieve this kind of idioms but at this time, I wasn't able to find an effective NER tool to recognize proper names in idioms.

Kun: add_feature.py, difficulty features

I worked on building features of the project. In order to figure out the difficulty of each idiom, I need to find a benchmark to compare. There are many methods to determine the difficulty of idioms. For instance, we can tell an idiom is hard based on complexity of the character, or we can find frequency using of the idioms.

Then I find a very useful document HSK. HSK has collected news texts from World Forum website. The corpus contains a total of 27965 news articles with more than 20 million Chinese characters. Those news articles fall into more than 15 categories which includes topics such science, culture, technology. Based on those news articles, HSK has identified the four characters in the news corpus. They also found the frequency of each idiom and posted the result to Chinese text computing website. I use the results as a benchmark to divide our corpus into "Easy", "Medium", and "Hard". To be more specific, I first mapped all the HSK corpus with our corpus and if any idiom which is not showed in HSK corpus, I assigned value 1 to its frequency. After mapping, I iterated our corpse using frequency filed and divided idioms into three parts.

Xiaojing: index.py, query.py, templates

I worked on building the index of elasticsearch and using conjunctive search to query the fields in the corpus. I also worked on html templates to display the query results. When I tried to build the index for elasticsearch, I initially met several issues on how to properly index and query the corpus. One main issue that took me quite a long time was forgetting to import the index in the query.py, and this led to the no connection in the elasticsearch. Because of the unclear error message, I did spend a long time to finally figure out the problem.

I also took part in displaying query results on the html pages. When I tried to display our result on localhost, I integrated Bootstrap in the main page (query_page). With the help of Bootstrap, the html page could be formatted better. But due to the complex interaction, I didn't have enough time to format the other pages with Bootstrap as well.

Erik: chengyu_index.py, translation features (including modifications on template files)

One of the biggest challenges for me was integrating translation into the project. This feature was crucial in order to make the project accessible to English-speaking audiences. For indexing, the difficulties mostly lay first in finding corpora that would be easily processed (text files). This is why I used cedict_1_0_ts_utf-8_mdbg.txt, as it was a dictionary with over 100,000 entries that could easily be turned into a json file. Another difficulty lay in preparing and indexing all the translation aspects to be ready for the elasticsearch index, as the chengyu text file and Chinese-English text file dictionary were formatted slightly differently. After using the Stanford CoreNLP system to segment the Chinese sentences as best we could, I stored the regular features and translation-based keys separately in order to save some space in the json files. On the query end, it was very difficult both to properly display the hover box that contained possible translations, which required a lot of editing for each HTML template file. Furthermore, it was difficult to link the translations with their Chinese counterpart, i.e. highlight the corresponding Chinese word when an English search is performed. However, I managed to employ a system that made use of this feature, which will give English speakers more access to our project. Given more time, I would like to add more to the translation features, perhaps even creating a more robust and accurate translation system. There are also some small formatting issues I found with some individual article pages that I would like to examine more closely, and ultimately fix (see 显而易见), as I was unable to find out why these individual pages were displayed as such.

Description: For this project, our goal was to create an effective information retrieval system for Chinese idiomatic expressions. Chinese idioms are incredibly diverse and numerous, and the term *idiomatic expression* has many possible translations in Chinese, including 成语 [*chéngyǔ*] (a set phrase usually containing 4 or 5 characters) and 俗语 [súyǔ] (a vernacular idiom). Chinese also makes use of other types of phrases, such as 歇后语 [xiēhòuyǔ] (two-part allegorical sayings). Our starting index contains 13,279 such expressions, which contains name, pinyin (with diacritical marks flattened), and description as structured data. For unstructured data, we used a regular expression to grab important elements found in the description, such as usage and source. To Chinese people, selecting a correct idiom to use is very important, so our search mechanism contains features to filter based on aspects important to Chinese culture

(such as a specific animal, or positive/negative connotation). As an added bonus, we also included translation features to aid Chinese learners.

Dependencies:
Python 3.6
Flask 1.0.2
Elasticsearch 6.7
Elasticsearch-dsl
zhon==0.2.0
thulac==0.2.0
nltk==3.3
pandas==0.24.1

Build Instructions: Before running, the module flask needs to be installed. The process can be done as follows.

1. pip3 install Flask

Some sources also suggest installing the virtual environment, which can be done as follows.

1. pip3 install virtualenv

Also, please make sure that nltk is installed. nltk can be installed as follows.
sudo pip3 install -U nltkthe

Next, the CoreNLP and ElasticSearch servers both need to be started. First download coreNLP version 3.9.2 from the Stanford website: https://stanfordnlp.github.io/CoreNLP/

Then, run the coreNLP server with the following command (make sure to be in the directory corresponding to the coreNLP package you downloaded):
java -mx3g -cp "*" edu.stanford.nlp.pipeline.StanfordCoreNLP -props StanfordCoreNLP-chinese.properties -file chinese.txt -outputFormat text

java -Xmx4g -cp "*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer \
> -serverProperties StanfordCoreNLP-chinese.properties \
> -preload tokenize,ssplit,pos,lemma,ner,parse \
> -status_port 9001 -port 9001 -timeout 15000

Then, run the elasticsearch server, making sure to be in the directory where you downloaded elasticsearch to:
./bin/elasticsearch

Files to be Submitted: (excluding this file)
Python Files: chengyu_index.py, add_feature.py, index.py, query.py
HTML Files: query_page.html, page_SERP.html, page_targetArticle.html

Txt Files: 13279_chengyu.txt, cedict_1_0_ts_utf-8_mdbg.txt, HSK.txt
Xlsx Files: sentiment_vocab.xlsx
Json files: test.json

Run Instructions:

Build the corpus index:

python chengyu_index.py

Add features:

python add_features.py

Build the elasticsearch index

python index.py

Run the elasticsearch query

python query.py

Index Build Time: (if applicable) - [approx. 3 minutes for chengyu_index.py, 18 seconds for add_feature.py, around 7-10 seconds for index.py]

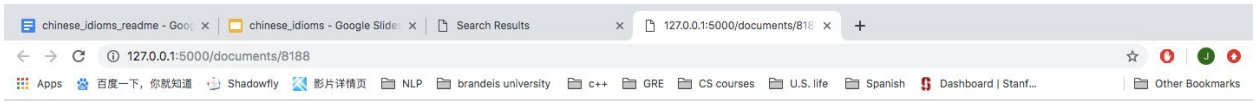
Modules: chengyu_build_index.py, add_features.py, index.py, query.py

chengyu_build_index.py: Builds the shelf files according to what we needed
add_features.py: Processes each Chengyu in the corpus and adds the zodiac animal information, the sentiment information, the difficulty level of the Chengyu and segments the pinyin.
index.py: Passes the necessary information to the elasticsearch index
query.py: Handles the elasticsearch queries

Testing:

We handpicked 20 documents from our corpus, the list below contains the names of the idioms.
test_list = ["杀鼋教子", "六畜兴旺", "韬光养晦","踏破铁鞋无觅处，得来全不费工夫", "恶性循环", "以其人之道，还治其人之身","鼠窜狼奔", "犬兔俱毙", "爱礼存羊", "替罪羊", "鸡烂嘴巴硬","狂吠猞犴", "阿猫阿狗", "知小谋大", "徒读父书", "前言不搭后语", "有其父必有其子", "疑人勿用，用人勿疑", "司马昭之心，路人皆知", "公说公有理，婆说婆有理"]

Test examples:



杀一儆百

score: 68.439476

name: 杀一儆百

english: execute one as a warning to a hundred

pinyin: shā yī jǐng bǎi

zodiac:

difficulty: Hard

sentiment: Positive

afterword:

riddle:

source: 《汉书·尹翁归传》：“其有所取也，以一警百，吏民皆服，恐惧改行自新。”

story:

synonym: 杀鸡儆猴、惩前毖后

antonym: 既往不咎、宽大为怀

description: :

儆： 警告。 处死一个人， 借以警戒许多人。

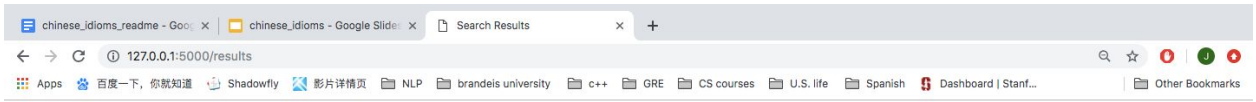
1. to warn
2. to admonish

》： “其有所取也， 以一警百， 吏民皆服， 恐惧改行自新。”

usage: :

连动式； 作谓语、 宾语； 含褒义

char_num: 4



Chengyu (Chinese Idiom) Search

General Query Search:

Monkey

Search Pinyin:
Select Zodiac Sign: /
Select Sentiment: /
Select Difficulty: /
Select Character Number:
Search

Found 62 results. Showing 1 - 10

Next

杀鸡骇猴 **score: 23.773533**

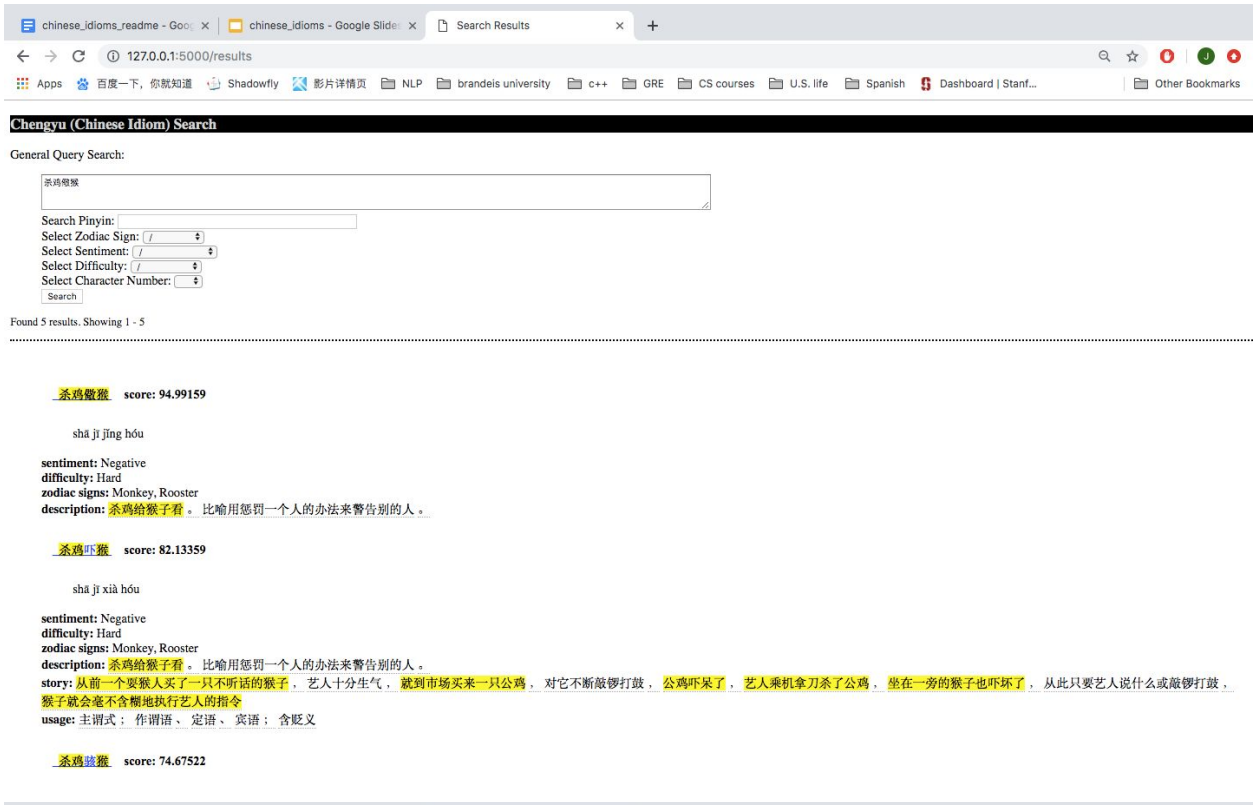
shā jī hài hóu

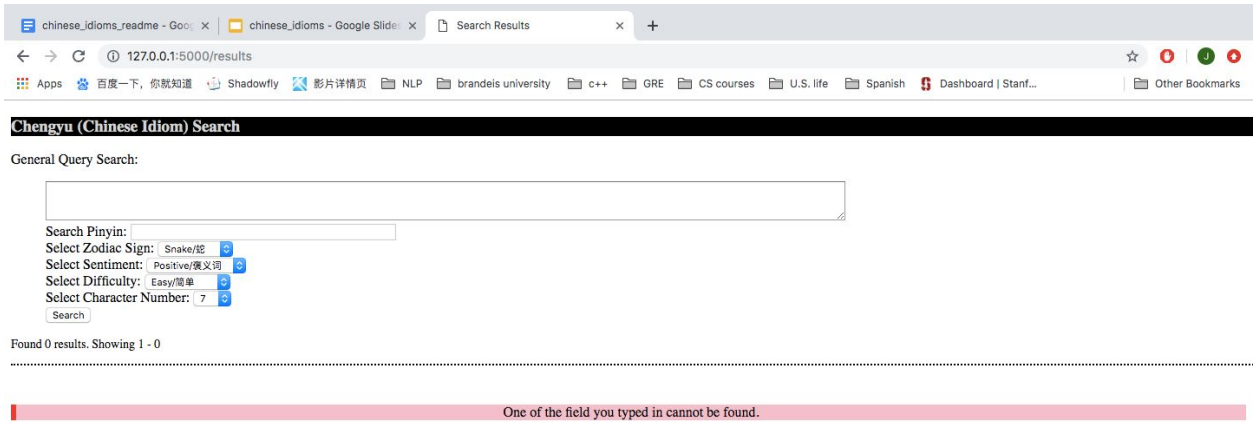
english: kill the chicken to frighten the monkey
sentiment: Negative
difficulty: Hard
zodiac signs: Monkey, Rooster
description: 杀鸡给猴子看。 比喻惩罚一人以恐吓或警戒其他人
source: 清 李宝嘉 《官场现形记》 第53回： “俗话说得好， 叫做‘杀鸡骇猴’， 拿鸡子宰了， 那猴儿自然害怕。”
story: 从前一个耍猴人买了一只不听话的猴子， 艺人十分生气， 就到市场买来一只公鸡， 对它不断敲锣打鼓， 公鸡吓呆了， 艺人乘机拿刀杀了公鸡， 坐在一旁的猴子也吓坏了， 从此只要艺人说什么或敲锣打鼓， 猴子就会毫不含糊地执行艺人的指令
usage: 主谓式； 作谓语、 定语、 宾语； 含贬义

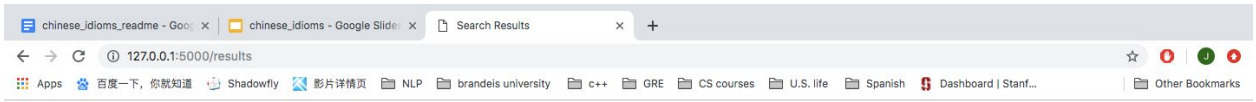
千难万险 **score: 4.86845**

qiān nán wàn xiǎn

sentiment: Negative
difficulty: Hard
description: 形容困难和危险极多
source: 元 杨景贤 《西游记》 第五本第四折： “火焰山千难万险。”
usage: 联合式； 作谓语、 宾语； 形容困难和危险极多


chinese_idioms_readme - Goo
chinese_idioms - Google Slide
Search Results
127.0.0.1:5000/results
Apps
百度一下，你就知道
Shadowfly
影片详情页
NLP
brandeis university
c++
GRE
CS courses
U.S. life
Spanish
Dashboard | Stanf...
Other Bookmarks
Chengyu (Chinese Idiom) Search
General Query Search:
杀鸡儆猴
Search Pinyin:
Select Zodiac Sign:
Select Sentiment:
Select Difficulty:
Select Character Number:
Search
Found 5 results. Showing 1 - 5
杀鸡儆猴 score: 94.99159
shā jī jǐng hóu
sentiment: Negative
difficulty: Hard
zodiac signs: Monkey, Rooster
description: 杀鸡给猴子看。比喻用惩罚一个人的办法来警告别的人。
杀鸡吓猴 score: 82.13359
shā jī xià hóu
sentiment: Negative
difficulty: Hard
zodiac signs: Monkey, Rooster
description: 杀鸡给猴子看。比喻用惩罚一个人的办法来警告别的人。
story: 从前一个耍猴人买了一只不听话的猴子，艺人十分生气，就到市场买来一只公鸡，对它不断敲锣打鼓，公鸡吓呆了，艺人乘机拿刀杀了公鸡，坐在一旁的猴子也吓坏了，从此只要艺人说什么或敲锣打鼓，猴子就会毫不含糊地执行艺人的指令
usage: 主谓式；作谓语、定语、宾语；含贬义
杀鸡骇猴 score: 74.67522


chinese_idioms_readme - Goo
chinese_idioms - Google Slide
Search Results
127.0.0.1:5000/results
Apps
百度一下，你就知道
Shadowfly
影片详情页
NLP
brandeis university
c++
GRE
CS courses
U.S. life
Spanish
Dashboard | Stanf...
Other Bookmarks
Chengyu (Chinese Idiom) Search
General Query Search:
Search Pinyin:
Select Zodiac Sign: Snake/蛇
Select Sentiment: Positive/褒义词
Select Difficulty: Easy/简单
Select Character Number: 7
Search
Found 0 results. Showing 1 - 0
One of the field you typed in cannot be found.



Chengyu (Chinese Idiom) Search

General Query Search:

Search Pinyin: xian
Select Zodiac Sign: /
Select Sentiment: /
Select Difficulty: /
Select Character Number:
Search

Found 217 results. Showing 1 - 10

Next

先知先觉 **score: 5.6242237**

xiān zhī xiān jué

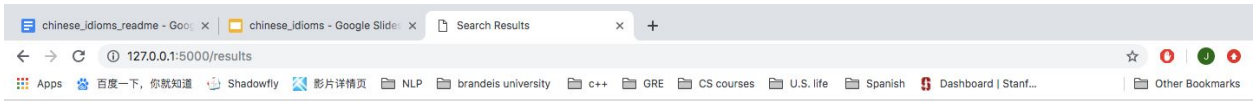
english: a person of foresight
sentiment: Positive
difficulty: Hard
description: 指认识事理较一般人为早的人。
source: 《孟子 万章下》：“使先知觉后知，使先觉觉后觉也。”
usage: 联合式；作主语、宾语、定语；含褒义

1. attributive (modifier)

鲜车怒马 **score: 4.6567945**

xiān chē mǎ

sentiment: Positive
difficulty: Hard
zodiac signs: Horse
description: 崭新的车，肥壮的马。形容服用讲究，生活豪华。
source: 《后汉书 第五伦传》：“蜀地肥饶，掾史家赀多至千万，皆鲜车怒马，以财货自达。”



Chengyu (Chinese Idiom) Search

General Query Search:

Search Pinyin:
Select Zodiac Sign: /
Select Sentiment: Positive/褒义词
Select Difficulty: /
Select Character Number:
Search

Found 5893 results. Showing 1 - 10

Next

Warning: Over 500 search results! We recommend you narrow your search.

资深望重 **score: 0.83627725**

zī shēng wàng zhòng

sentiment: Positive
difficulty: Hard
description: 资格老，声望高。
source: 宋朝苏轼《答试馆职人启》：“非独使之业广而材成，抑将待其资深而望重。”"

众难群移 **score: 0.83627725**

zhòng nán qú yí

sentiment: Positive
difficulty: Hard
description: 众人心中都有疑难。
source: 三国蜀 诸葛亮《后出师表》：“群疑满腹，众难塞胸。”"