Переводчик прайс-листов на китайский #1

Edit New issue

Closed fortuneteller opened this issue on 11 Sep 2019 · 2 comments

fortuneteller commented on 11 Sep 2019 • edited

Нам надо сделать возможность формирования EXCEL файлов оптовых прайс-листов на китайском. Задача разбивается на две подзадачи:

1. Микросервис-переводчик для перевода а) произвольных русских строк и б) Конкатенации родного названия товара (английский или французский) + бренда (английский или французский) на китайский. Для перевода с русского (вариант а)) нужен будет простой UI - таблица с двумя колонками - слева русская строка, справа китайский перевод. Этим будет заниматься человек-переводчик. Для перевода названий бренда + товара (вариант б)) будет использоваться парсер китайского сайта. Но также должна быть возможность ручной правки результатов перевода через отдельный UI (таблица с двумя колонками). Это всё будем делать в данном репозитории. Ограничений по конкретному PHP фреймворку и базам данных нет.
2. Если всё будет ок с пунктом 1, то я дам доступ в наш основной репозиторий CRM. Нужно будет сделать внесение доработок в CRM в модуль генерации оптовых прайсов. Он написан на yii1. Сейчас есть перевод с русского на английский, нужно будет сделать перевод с русского на китайский. Детали уже будут отдельной задачей, но следует иметь её в виду, делая текущую задачу. Это если парсер из пункта 1 будет сделан хорошо.

Везде ниже речь будет идти только про пункт 1

Что такое строка прайс-листа

Пример

Бренд	Товар	Тип	Объём	Пол	Цена
12 Parfumeurs Francais	Bagatelle	Парфюмерная вода	100	Унисекс	11 100 Р

Что мы собираемся переводить на китайский

1. Должна быть сущность **Перевод товара** с двумя ключевыми атрибутами [название, пол]. Пол может быть женский, мужской, унисекс. В рамках решения задачи не нужно знать откуда берётся название, но для понимания поясню, что это

Assignees
No one—assign yourself

Labels
None yet

Projects
None yet

Milestone
No milestone

Linked pull requests
Successfully merging a pull request may close this issue.
None yet

Notifications Customize
Unsubscribe
You're receiving notifications because you're watching this repository.

1 participant

Lock conversation

конкатенация (Бренд + товар) строки прайс-листа. Т.е. для примера выше название будет равно `12 Parfumeurs Francais Bagatelle`. Не должно быть двух сущностей с одинаковыми значениями пар атрибутов название + пол.

Бренд	Товар	Тип	Объём	Пол	Цена
12 Parfumeurs Francais	Bagatelle	Парфюмерная вода	100	Унисекс	11 100 Р

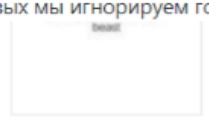
Сущность **Перевод товара** должна также иметь атрибут `статус автоматического перевода` : `запланирован` , `не найдено на сайте` , `успешно переведено`
Перевод товара должен также иметь бинарный флаг `имеет дубли` . Если есть хотя бы два **Перевода товара** с одним и тем же значением по-китайски, обоим **Переводам товара** проставляем флаг = true
Перевод товара должен также иметь бинарный флаг `откорректирован вручную` . Если перевод был редактирован через UI проставляем флаг = true

Этот перевод мы будем делать путём парсинга сайта на китайском (https://www.nosetime.com). При генерации прайс-листа модуль из crm сначала затребует все готовые переводы через некоторый АПИ. Готовые, значит `статус автоматического перевода` = `успешно переведено` или `откорректирован вручную` = `true` и при этом `имеет дубли` = `false` . Мы должны выдать полный список переведённых пар. Если CRM видит, что в прайс-листе есть непереведённые товары - опять же через апи добавляем в очередь на парсинг с сайта набор **Переводов товара** .

Должна быть реализована очередь задач, куда будут попадать все новые **Переводы товара** . Эти пары парсер должен вводить в окно поиска по сайту https://www.nosetime.com и запускать поиск.

Обработка результатов будет идти следующим образом:

Во-первых мы игнорируем года



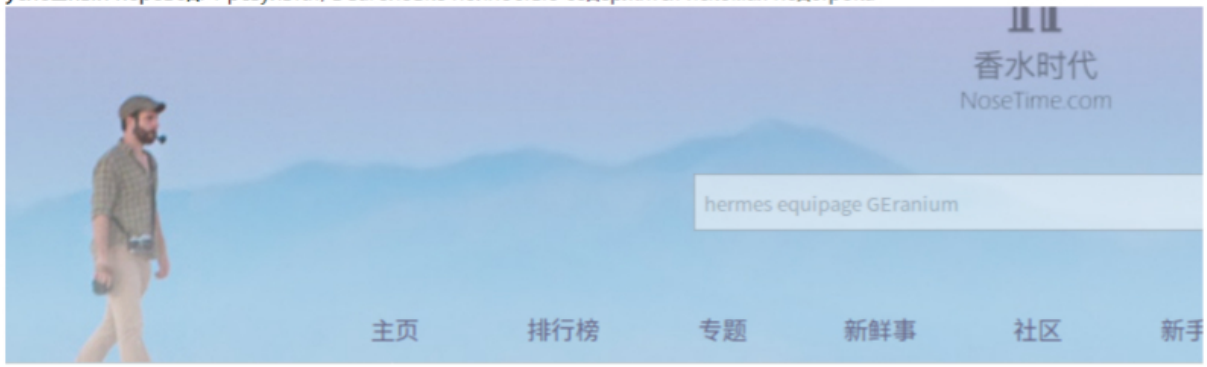
香水:

爱马仕 船员 Hermes Equipage, 19~~78~~
★★★★½ 8.7 分 114 评价
前调：醛,橙子,快乐鼠尾草,肉豆蔻,香柠檬,巴西红木
中调：康乃馨,肉桂,茉莉,铃兰,松树
后调：零陵香豆,广藿香,麝香,橡木苔,香草,香根草

爱马仕 天竺葵船员 Hermes Equipage Geranium, 20~~15~~
★★★★½ 8.9 分 21 评价
气味：老鹳草,辛香料,檀香木

Pin issue
Transfer issue

- **успешный перевод**. 1 результат, в заголовке полностью содержится искомая подстрока



"hermes equipage GEranium" 的搜索结果

帖子:

爱马仕船员equipage

香水:

爱马仕 天竺葵船员 Hermes Equipage Geranium, 2015

★★★★½ 8.9 分 21 评价

气味：老鹳草,辛香料,檀香木

. Прописываем перевод для пары и меняем `статус автоматического перевода` на `успешно переведено`

香水:

爱马仕 天竺葵船员 Hermes Equipage Geranium, 2015

★★★★½ 8.9 分 21 评价

气味：老鹳草,辛香料,檀香木

- результата нет

"montale no such product" 的搜索结果

введён заведомо несуществующий товар

帖子:

montale和旗下mancera算不算性价比很高的沙龙香

蒙塔莱montale有什么适合男生用的香吗

результатов нет

显示更多帖子

香水:

维多利亚的秘密 眉来眼去 Victoria's Secret Such a Flirt Fragrance Mist, 2012

☆☆☆☆☆ 4 人评价

气味：杨桃,芦荟,兰花,洋甘菊

Это что-то вроде "мне повезёт" т.е. рандом

维多利亚的秘密 甜言蜜语 Victoria's Secret Such A Flirt

☆☆☆☆☆ 3 人评价

气味：杨桃,兰花

金光闪烁 Fleurage Such a Boy

☆☆☆☆☆ 0 人评价

Меняем `статус автоматического перевода` на `не найдено на сайте`

- **результаты есть и более одного**

Ищем есть ли среди результатов точное равенство строки запросу. Если есть, то это наш результат

По идее он всегда будет первым, но не стоит на этом основывать логику.

Частный случай. Результатов более чем 1 и строки содержат признаки пола pour femme, pour homme, for women, for men.

香水:

爱慕 黄金女士 Amouage Gold pour Femme, 1983
★★★★½ 9分 197 评价
前调：焚香,铃兰,玫瑰
中调：鸢尾根,茉莉,没药
后调：琥珀,檀香木,麝香,麝猫香,雪松

爱慕 黄金男士 Amouage Gold pour Homme, 1998
★★★★☆ 8.3分 138 评价
前调：野玫瑰果,焚香,铃兰
中调：鸢尾根,茉莉,没药
后调：琥珀,檀香木,广藿香,麝香,麝猫香,橡木苔,雪松

Тогда нужно сопоставлять эти признаки с атрибутом пол сущности **Перевод товара**.
Если пол = женский, то строка содержащия pour femme или for women и в остальном полностью совпадающая с запросом, является нашим результатом. Аналогично если мы ищем перевод мужского товара. Для унисекса это неприменимо.

Для всех переводов, где результаты есть, но ни один не соответствует запросу, нужно вести лог, с возможностью вывода на отдельной странице, чтобы искать какие-то ситуации, не обработанные описанными выше правилами.

2. должна быть сущность **Перевод строки** - атрибут-ключ "значение". Любая подстрока на русском языке (заголовок столбца прайса, тип товара, меры измерения).

Бренд	Товар	Тип	Объём	Пол	Цена
12 Parfumeurs Francais	Bagatelle	Парфюмерная вода	100	Унисекс	11 100 Р

Тут всё аналогично пункту 1 но проще - должен быть метод api для добавления сущности **Перевод строки** в UI для перевода. После добавления строка должна появиться в интерфейсе для ручного перевода. Переводчик вносит перевод. Должен быть метод для получения полного словаря переведённых строк. Crm запрашивет полный словарь переводов и всё, для чего не найдено переводов - добавляем снова в список, если ещё ранее не было добавлено.

fortuneteller commented on 13 Sep 2019 • edited ▾ Author

Дополнение по сущности **Перевод товара**. Если возможно, примите, пожалуйста, эту правку. Я согласен на некоторое увеличение числа часов.

Убираем из задания

Отменяется атрибут `имеет дубли`. Его не нужно реализовывать.

Добавляем в задание

`Название бренда` и `название товара` в сущности **Перевод товара** должны быть отдельными полями. Да, поиск будет идти по их конкатенации (Бренд + товар), но храниться они должны отдельно. Кроме того, сам перевод нужно разбивать на части и хранить отдельно перевод товара, отдельно перевод бренда. Перевод нужно разбивать по пробелу. В китайском переводе даже если название товара или бренда состит из нескольких слов, то иероглифы идут без пробелов между словами. Но между названием товара и названием бреда есть пробел. Т.е. будет всего две непрерывные последовательности иероглифов - первую прописываем в перевод бренда, вторую в перевод товара.

вот пример

> 解放橘郡 激情喷射 Etat Libre d'Orange Secretions Magnifiques, 2006
> ★★½☆☆ 5.1 分 158 评价
> 气味：鸢尾花,檀香木,红没药,椰子,牛奶,海藻

> 解放橘郡 腐尸 Etat Libre d'Orange Charogne, 2008
> ★★★½☆ 7.4 分 91 评价
> 气味：皮革,生姜,百合,粉红胡椒,香草,茉莉,焚香,依兰,香柠檬,小豆蔻,桂花,安息香,劳丹脂,黄葵,麝香

Если последовательность иероглифов без пробела только одна - значит это либо название бренда, либо название товара, тогда мы НЕ сохраняем ни в перевод бренда, ни в перевод товара ничего, перевод товара получает статус `не найдено на сайте`. Т.е. обрабатываем так же, как если бы вообще не была найдена строка.

Вот пример

蝴蝶工匠 Papillon Artisan Perfumes Salome, 2015

Итого

Сущность **Перевод товара** будет иметь следующие атрибуты

`название бренда`
`название товара`
`перевод бренда`
`перевод товара`
`статус автоматического перевода` : (`запланирован` , `не найдено на сайте` , `успешно переведено`)
`откорректирован вручную`

Если пользователь редактирует перевод и оба перевода (перевод бренда и перевод товара) оказываются непустыми, то данный **Перевод товара** должен получить статус `успешно переведено`

Детализация по методам АПИ

Предлагаю сделать так:

Метод апи для принятия новых переводов товара должен принимать json структуру вида

```
[
...,
{"brand": "brand name", "product": "product name"},
{"brand": "brand name1", "product": "product name1"},
...
]
```

Методов апи для возвращения словаря будет два. Один для брендов, один для товаров. Будем считать, что разных переводов брендов не бывает. Так мы сэкономим на том, что перевод брендов не будет передаваться много раз.

```
{
...
   "brand name": "тут перевод",
   "brand name1": "тут перевод1",
...
}
```

Аналогично для товаров

```
{
...
   "product name": "тут перевод",
   "product name1": "тут перевод1",
...
}
```

fortuneteller commented on 30 Sep 2019 Author

Как я уже говорил, мне нужна будет помощь с traefik, так как я никогда с ним не работал. Не получается его завести

```
andrey@andrey:~/Projects/translator-parser/traefik$ ./up.sh
Error response from daemon: network with name web already exists
Recreating traefik ...
Recreating traefik ... done
andrey@andrey:~/Projects/translator-parser/traefik$ docker-compose up traefik
traefik is up-to-date
Attaching to traefik
traefik    | 2019/09/30 13:32:10 command traefik error: field not found, node: tls
traefik    | 2019/09/30 13:32:11 command traefik error: field not found, node: tls
traefik    | 2019/09/30 13:32:12 command traefik error: field not found, node: tls
traefik    | 2019/09/30 13:32:13 command traefik error: field not found, node: tls
traefik    | 2019/09/30 13:32:14 command traefik error: field not found, node: tls
traefik exited with code 1
```


fortuneteller closed this on 3 Oct 2019

Write Preview

Leave a comment

Attach files by dragging & dropping, selecting or pasting them.

Reopen issue Comment

Remember, contributions to this repository should follow our GitHub Community Guidelines.