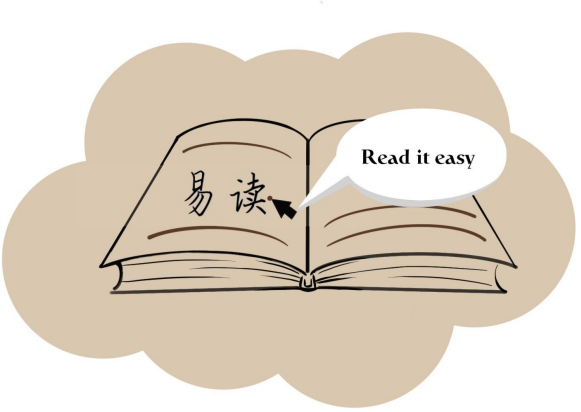
Documentation de ReadItEasy

Bienvenue ! Ceci est la documentation de ReadItEasy, interface graphique web destiné à faciliter la lecture en vue d'aider les utilisateurs à apprendre une langue étrangère.



Pour l'instant, ce projet se concentre sur la langue chinoise vers l'anglais. Plus de langues seront prises en compte au fur et à mesure de son développement.

- Site disponible sur : readiteasy.com
- Github disponible sur : https://github.com/kirianguiller/ReadItEasy

Objectifs

La lecture étant un moyen indispensable d'apprendre une langue, nous nous employons à créer un outil avec interface graphique web qui analyse un livre en montrant ses informations telles que les fréquences des mots et qui affiche spontanément la traduction quand on passe le curseur au-dessus d'un mot. De plus, une fonction de recherche permet à l'utilisateur de trouver les occurrences en contexte d'une chaîne de texte dans le livre en cours.

Le but à long terme de ReadItEasy est d'élaborer des outils d'analyses de difficulté lexicale et syntaxique d'un texte afin de pouvoir proposer à l'utilisateur les livres les plus adaptés à son niveau actuel.

Un autre objectif est d'étoffer les langues cibles et sources disponibles. La langue cible est la langue que veut parler l'utilisateur et la langue source est une langue que l'utilisateur maîtrise.

Structure (Django)

Avant de continuer plus loin dans l'explication du projet, il est important de comprendre que l'application est une application web gérée par Django. Cela nous impose donc d'utiliser certaines conventions/fonctionnalités de ce framework.

Django fonctionne avec un système de MVC (model/view/controller) ce qui fait que les données (model) sont séparées du traitement (view) ainsi que de la vue (controller ou templates).

Nous avons décidé de ne pas utiliser les model pour stocker les données pour le moment bien que cela constitue un axe d'amélioration important pour augmenter la robustesse de code ainsi que la réutilisabilité des données.

Voici une explication succinctes des éléments clés du framework django illustré avec la structure de notre projet. Veuillez vous référer vers la documentation de Django si nécessaire.

/

Voici la racine de notre projet (le dossier racine s'appelle 'ReadItEasy'. Attention à ne pas le confondre avec un de ses dossiers enfant `ReadItEasy/ReadItEasy` qui possède les informations générales sur le projet). La racine va donc avoir comme enfant le projet principal 'ReadItEasy' ainsi que les différentes applications du site (dans notre cas 'users', 'books', 'dictionary'). Django permet de factoriser un projet en différentes applications pour permettre plus de réutilisabilité et de clarté.

Notre dossier racine possède aussi d'autres dossiers qui sont propres à notre projet comme '`data`' (qui contient les livres, dictionnaires, etc), '`utils`' (qui contient les scripts que nous avons fait qui répondent à nos besoin), et '`venv`' (qui contient l'environnement virtuel)

/ReadItEasy

Voici donc l'enfant 'ReadItEasy' qui va posséder les informations du projet ainsi que les templates et static qui seront communs à toute les applications. Les templates sont les modèles html de django qui peuvent contenir des variables pour pouvoir réaliser des pages personnalisées à la requête http de l'utilisateur

**/urls.py

Les scripts python `urls.py` sont présent dans le projet ainsi que dans chaque application. Ce sont ces scripts qui vont rediriger les requêtes http de l'utilisateur vers les différentes vues/views.

**/views.py

Les scipts python `views.py` vont servir au traitement de l'informations (et non pas à la vue des html !) en fonction de la requête de l'utilisateur. Une fonction view devra toujours retourner soit une redirection vers une autre vue ou url, soit vers un template html.

**/templates/*.html

Les templates `*.html`, qui sont situés dans les dossiers templates (présents dans le projet ainsi que dans chacunes des applications) sont des modèles html qui permettent d'être personnalisable. On peut y utiliser des boucles, des conditions d'état, des calculs, etc.

**/static/

Les dossiers static (présent dans le projet ainsi que dans chacunes des applications) contiennent les `.css` et images nécessaires pour l'apparence du site.

/ReadItEasy/settings.py

Fichier settings qui contient les configurations du projet. Ce fichier est très important et beaucoup d'erreurs viennent du fait qu'il est mal configuré. C'est dans ce fichier que sont situées les informations sur les différentes applications, sur les chemins des dossiers templates et static, etc...

Nous espérons que ces explications aideront à la compréhension de la structure du projet ainsi que les interactions MVC de django.

Données

Nous disposons de 1000 livres chinois dont seulement une petite partie (3 livres) sont sur le github (car nous sommes sur que ces trois livres sont libres de droits). Nous avons aussi un dictionnaire chinois -> anglais (open source).

Nous proposons deux axes d'accès aux données :

- Pour les 3 livres libres de droits, nous proposons à l'utilisateur d'en choisir un pour l'aide à la lecture. Cette aide tokenize le texte, ajoute des informations sur le mot et les statistiques du texte et donne la possibilité à l'utilisateur de faire des recherches de texte dans le livre.
- Lorsque l'utilisateur lit un texte, nous proposons en plus à l'utilisateur d'obtenir plus d'informations sur un mot en particulier tels que la fréquence et les plus proches voisins word2vec dans notre corpus de 1000 livres.

Ces données sont dans différents fichiers textes et nous y accédons de manière classique pythonic (with open(...) as f). Cependant, il faudra par le futur utiliser les Model Django qui permettront de répertorier ces informations dans une base de données SQL.

Méthodologie

Nous avons dans un premier temps récupéré des livres en chinois.
Nous avons ensuite calculé sur ce corpus les plongements vectoriels (word embeddings) ainsi que les statistiques de fréquences des mots (après tokenisation).
Nous avons ensuite récupéré un dictionnaire open-source (CEDICT) ainsi qu'un repertoire des niveaux HSK d'un mot.
Par la suite, nous nous sommes familiarisé avec l'outil django pour réaliser l'outil qui convient à nos besoin d'aide à la lecture.

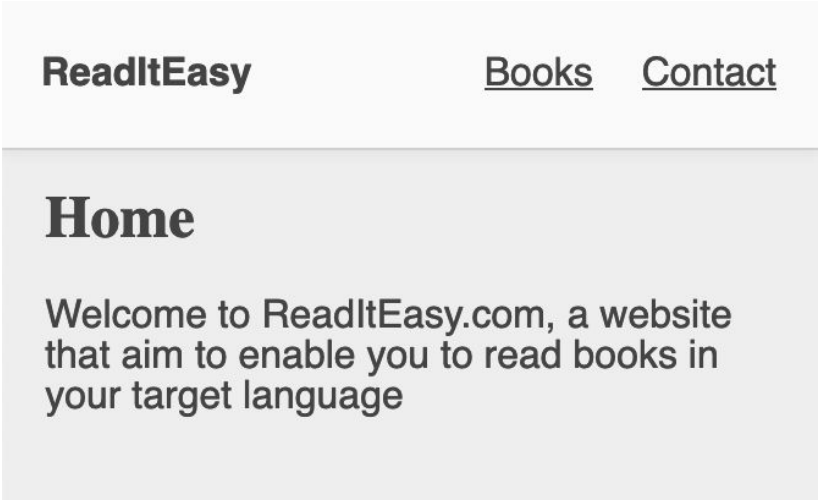
Shuai s'est occupé de trouver les ressources (dictionnaires, livres, etc) et de faire certaines fonctions de traitements tels que les scripts dans le dossier `/script`.
Ces fonctions de traitements sont :

- la fonction qui obtient les fréquences des mots dans un corpus de textes en mandarins.
- la fonction qui trouve dans un texte les chaînes de caractères qui correspondent à la recherche.
- les fonctions qui recherche dans les dictionnaires numériques les informations des mots chinois

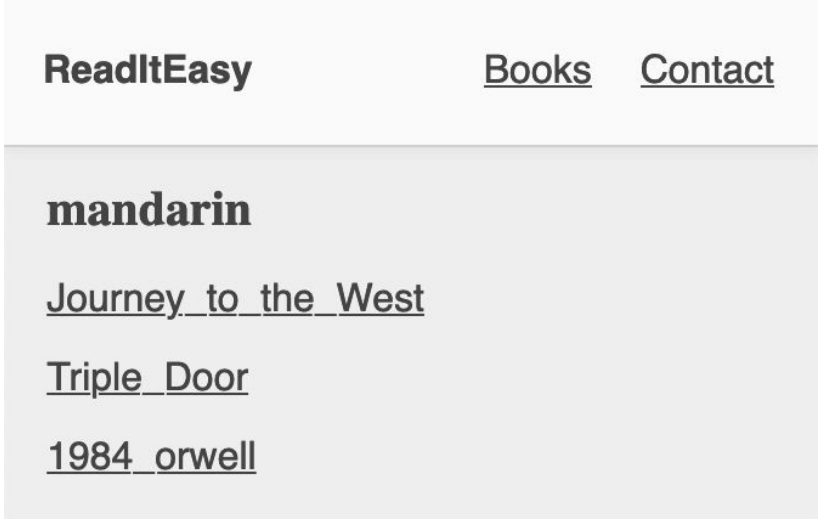
Kirian s'est occupé de faire la partie graphique et serveur du site ainsi que de la fonction qui calcul les words embeddings d'un corpus pour ensuite avoir les mots les plus similaires d'autres mots.

Résultats

Ce projet est maintenant disponible sur readiteasy.com et sera maintenu à long terme.



Voici ci-dessus l'index de ReadItEasy. Accédons à la bibliothèque en cliquant sur « Books ».



Nous disposons pour le moment de trois livres en mandarin libres de droit comme exemples. Testons les fonctionnalités de ReadItEasy avec « Journey to the West ». Il affiche à gauche le corps du livre et un bouton qui permet de tourner la page ; à droite les fréquences des mots.

ReadItEasy Books Log Out Profile Contact

Journey_to_the_West

next chapter

第一回

灵根育孕源流出心性修持大道生

诗曰：

混沌未分天地乱，茫茫渺渺无人见。

自从盘古破鸿蒙，开辟从兹清浊辨。

覆载群生仰至仁，发明万物皆成善。

欲知造化会元功，须看西游释厄传。

盖闻天地之数，有十二万九千六百岁为一元。将一元分为十二会，乃子、丑、寅、卯、辰、巳、午、未、申、酉、戌、亥之十二支也。每会该一万八百岁。且就一日而论：子时得阳气，而丑则鸡鸣；寅不通光，而卯则日出；辰时食后，而巳则挨排；日午天中，而未则西蹉；申时晡而日落酉，戌黄昏而人定亥。譬于大数，若到戌会之终，则天地昏缯而万物否矣。

再去五千四百岁，交亥会之初，则当黑暗，而两间人物俱无矣，故曰混沌。又五千四百岁，亥会将终，贞下起元，近子之会，而复逐渐开明。邵康节曰

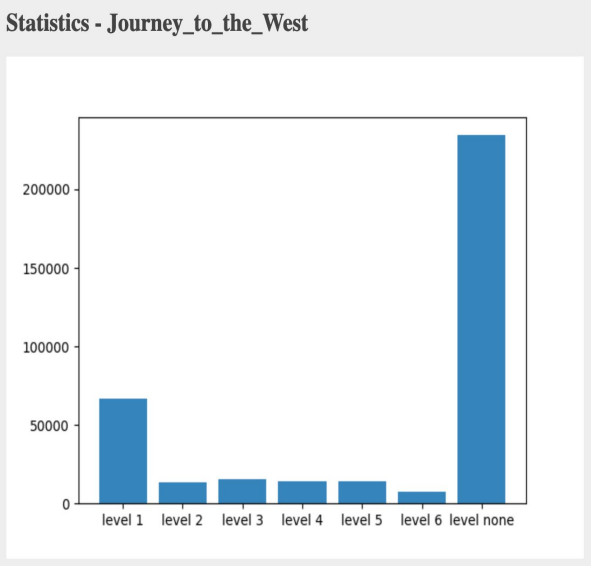
Char	Book Rank	Book Freq	Corpus Rank	Corpus Freq
道	0	23128	60	1480
了	1	19866	2	28410
我	2	15849	3	16642
的	3	14355	1	57319
他	4	14292	4	14620
你	5	14128	7	10749
那	6	11386	32	3045
行者	7	11135	28132	2
是	8	11116	5	14451
也	9	7620	10	7953
在	10	7592	6	12504
有	11	6664	14	5389
又	12	6628	21	3949
去	13	6549	24	3571
与	14	5921	63	1439
来	15	5743	33	2854
这	16	5519	18	4583
八戒	17	4588	74655	0
就	18	4574	11	7596
却	19	4394	64	1414

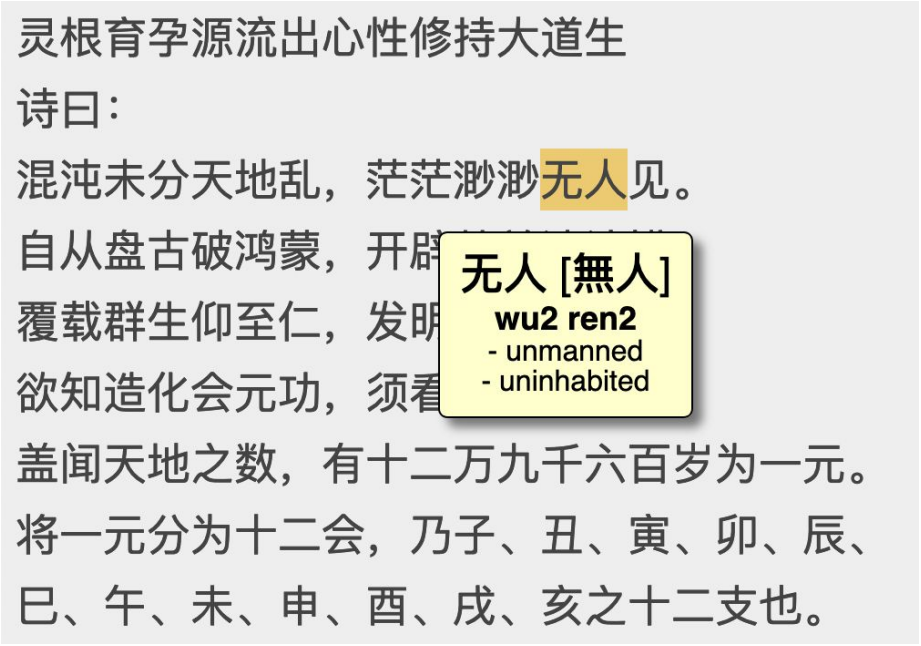
-> See book statistics <-

Enter text to search

En cliquant sur « See book statistics », on voit la distribution des mots selon leur niveau de difficulté défini par HSK. Cependant, il arrive souvent que beaucoup de mots n'appartiennent à aucune de ces six classes.

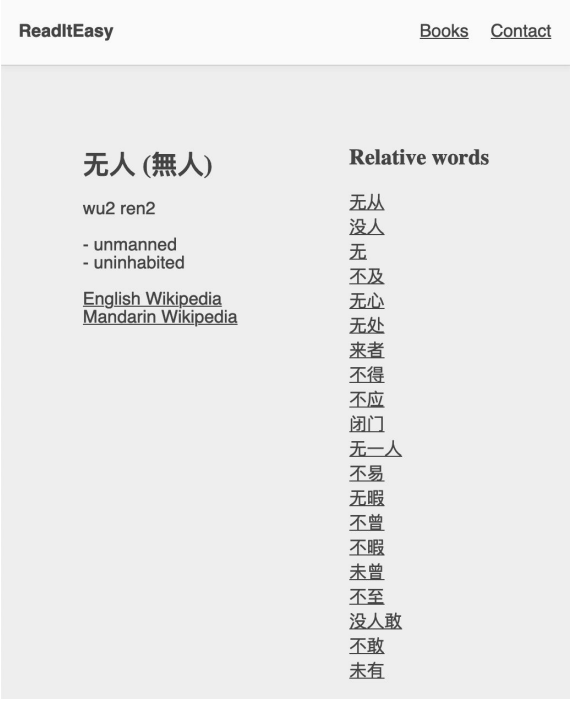
Quand on passe le curseur au-dessus du mot « 无人 » (non-habité ; ne personne), on voit immédiatement son écriture en chinois traditionnel, son pinyin ainsi que sa traduction en anglais.





En cliquant sur ce mot (en jaune), on voit à droite son classement des fréquences tant dans le livre que dans le chapitre actuel. Ensuite, cliquez sur ce mot (en rouge) dans le classement, on voit d'autres mots liés ainsi que les liens vers Wikipédia.

Char	Book Rank	Book Freq	Corpus Rank	Corpus Freq
几时	787	139	8438	10
闻得	788	139	57570	0
自有	789	139	4484	20
狂风	790	139	10836	7
不用	791	139	553	175
老儿	792	139	38877	1
特	793	139	3771	24
求经	794	139	0	0
路旁	795	139	9591	8
无人	796	136	2446	39
阎王	797	136	12619	6
名	798	136	1861	52
不怕	799	136	1048	93
饭	800	136	1210	81
不须	801	136	23265	3
可以	802	136	71	1335
天尊	803	136	58370	0
旁边	804	136	596	159
观音菩萨	805	136	43623	1
嘴脸	806	136	12390	6



Si vous voulez savoir davantage de ce mot, une case de recherche (sous les statistiques) vous permet de trouver toutes les phrases contenant ce mot, ce qui vous facilite grandement l'apprentissage des usages de ce mot.

-> See book statistics <-

无人

Search

无人 in the book Journey_to_the_West

混沌未分天地乱，茫茫渺渺无人见

又见那洞门紧闭，静悄悄杳无人迹

举世无人肯立志，立志修玄玄自明

风起处，惊散了那傲来国君王，三市六街，都慌得关门闭户，无人敢走

若道半个不字，教你顷刻化为鸏粉！”猴王听说，心中大怒道：“泼毛神，休夸大口，少弄长舌！我本待一棒打死你，恐无人去报信，且留你性命，快早回天，对玉皇说：他甚不用贤！老孙有无穷的本事，为何教我替他养马？你看我这旌旗上字号，若依此字号升官，我就不动刀兵，自然的天地清泰；如若不依，时间就打上灵霄宝殿，教他龙床定坐不成！”这巨灵神闻此言，急睁睛迎风观看，果见门外竖一高竿，竿上有旌旗一面，上写着“齐天大圣”四大字

那里不见老君，四无人迹

迅速严冬如指拈，逍遥四季无人管

因妻李氏缢死，撇下儿女无人看管，小人情愿舍家弃子，捐躯报国，特与我王进贡瓜果，谢众大王厚恩

你怎么不分皂白，一顿打死？全无一点慈悲好善之心！早还是山野中无人查考，若到城市，倘有人一时冲撞了你，你也行凶，执着棍子，乱打伤人，我可做得白客，怎能脱身？”悟空道：“师父，我若不打死他，他却要打死你哩

”行者道：“这样怪物，不打死他，反留他在何处用哩？”菩萨道：“我那落伽山后，无人看管，我要带他去做个守山大神

久无人出，行者性急，跳起身入门里看处，原来有向南的三间大厅，帘栊高控

这如今茶水不得见面，灯火也无人管，虽熬了这一夜，但那匹马明日又要驮人，又要走路

三藏见了，叫：“八戒、沙僧，悟空才说这里旷野无人，你看那里不走出一个人来了？”八戒道：“师父，你与沙僧坐着，等老猪去看看来

Conclusion

Nous sommes enjoué des résultats obtenues pour le moment et avons pour projets de continuer à améliorer le site.
Nous sommes conscient qu'il y a encore beaucoup d'axes d'amélioration pour le site comme par exemple :

- Utiliser une base de donnée SQL pour stocker les informations (au lieu d'ouvrir des dictionnaires python)
- Utiliser des requêtes AJAX et du lazy loading pour ne pas avoir à charger tout un chapitre en mémoire
- Créer des fonctions de difficultés lexicales et syntaxiques
- Ajouter des langues
- Améliorer le styling du site.

Licence