專題名稱:區塊鏈之旅館推薦系統

指導老師:謝孟諺

專題學生:

資工四A 410516497王培威

**前言**

推薦引擎是常見機器學習的應用，如淘寶、亞馬遜等網站都在使用個性化商品推薦系統。可以說現在我們經常用的各種在線服務都或多或少的使用了推薦技術來提高他們的服務。

專題裡將以旅館為推薦項目，建立推薦系統。建置以自架的Spark可擴展性叢集平台，並用Spark MLlib提供的ALS推薦演算法套件建立推薦系統，也會將推薦系統，以網頁的方式提供使用者推薦服務。

由於資料多樣性和加上與網際網路上的資料可能做連結。處理巨量使用者間接資料必須採用大數據技術進行，收集、過濾、分析、整合，和以至即時推薦運算處理，透過分散式儲存(Hadoop HDFS)、MapReduce 處理運算、和叢集記憶體運算(Resilient Distributed Dataset, RDD)，亦能夠加速巨量資料處理

專題裡也會提供一個旅館訂房的交易平台，應用Ganache區塊鏈環境，並使用Vyper撰寫區塊鏈的智能合約，使得訂房交易具備透明且可信賴性。

1. **系統功能**
2. **註冊和登入**: 系統透過網頁進行註冊跟登入功能，讓使用者創建帳號，才能對其收集推薦相關資訊，以便進行個人化推薦運算。
3. **推薦旅館**: 系統會對用戶進行個人推薦運算，並提供十個推薦的旅館。
4. **上傳旅館**:提供用戶上傳旅館，產生智能合約，透過此合約與其他用戶進行交易。
5. **訂房交易**:提供用戶訂房功能，並產生區塊鏈交易紀錄，可以透過交易歷史紀錄查看。
6. **系統特色**

* 協同過濾推薦機制

1. 協同過濾法(Collaborative Filtering，簡稱CF)，其核心思想為推薦使用者，與具有相似愛好或興趣的其他使用者喜歡的項目。

優點:

* 可以達到個人化推薦
* 不需要內容分析
* 可以發現使用者新的興趣點
* 自動化程度高

缺點:

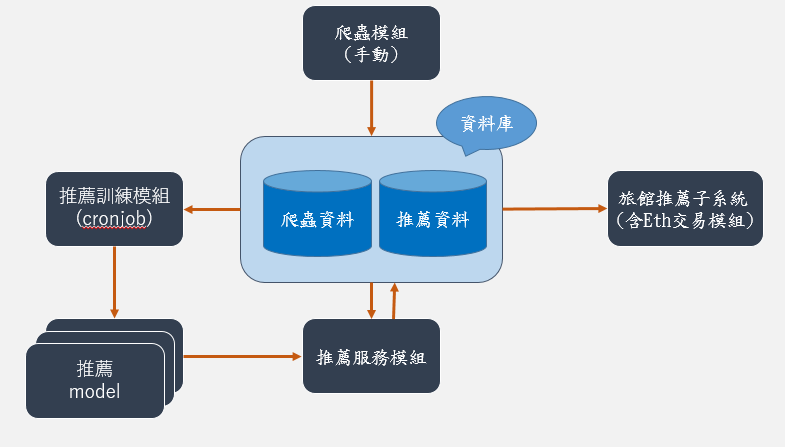
* 冷啟動問題(Cold-start):如果沒有歷史資料就沒辦法分析
* 稀疏問題(Sparsity):大部分的使用者會主動給予評分的非常少，所以造成矩陣的稀疏性。
* 系統延伸性問題(Scalability):當資料量和使用者數量多時計算量和時間會很大，要提供即時反饋有點難度。
* 在Python Spark on Hadoop Yarn叢集架構建立推薦系統

1. MLlib是Spark上分散式機器學習框架。採用Spark Mllib中實現的基於ALS矩陣分解協同過濾推薦算法
2. Spark分散式記憶體式的架構比Hadoop磁碟式的[Apache Mahout](https://zh.wikipedia.org/w/index.php?title=Apache_Mahout&action=edit&redlink=1" \o "Apache Mahout（頁面不存在）)快上10倍，加速推薦運算的進行。
3. Hadoop Distributed File System如同Google File System，提供大量儲存空間
4. Yarn為 Yet Another Resource Negotiator 的縮寫，是一個資源管理系統，用來管理各種分散式運算應用程式所使用的資源
5. Spark是美國柏克萊大學AMPlab所開發的一個開源叢集運算框架，與Apache Hadoop相容，並支援了Hadoop的儲存系統
6. Spark的模組化架構可以支援多種叢集管理器，因此Spark應用程式可以部署到Yarn或獨立叢集上，專題也會運用Yarn部署Spark應用程式到Hadoop叢集上。

* 使用Ganache區塊鏈環境模擬進行交易

1. Ganache是一個為開發者提供的私有Ethereum區塊鏈客戶端，可以用於本地佈署、開發、測試應用程式。專題會使用Vyper以太坊開發語言撰寫智能合約，並配置到Ganache區塊鏈上。
2. **系統相關架構與流程**

* **系統架構圖**



**圖一、系統架構圖**

1. 推薦訓練模組:

(1)主要透過RecommendTrain.py產生推薦model。。

1. 推薦服務模組:

(1)主要透過Recommend.py程式使用model進行個人推薦。

1. 爬蟲模組:
   1. 使用爬蟲程式爬取資料。
   2. 程式使用Crawling\_user\_hotel\_rating.py、Crawling\_trustrelate.py。
2. 旅館推薦子系統:
   * 1. 用flask框架建立推薦網頁，並結合Ganache測試鏈進行交易。
     2. 主要程式: Hotel\_recommend.py、base.html、index.html、login.html、registed.html。
3. 資料庫:

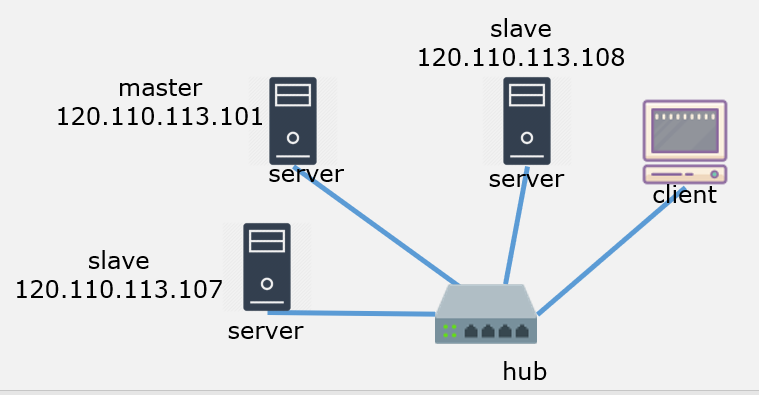
(1)爬蟲資料:包括userid、hotelid、rating、旅館資訊等等。

(2)推薦資料:儲存推薦服務模組產生的個人推薦旅館資料，並提供給推薦子系統

(3)交易紀錄:儲存訂房後產生的交易紀錄。

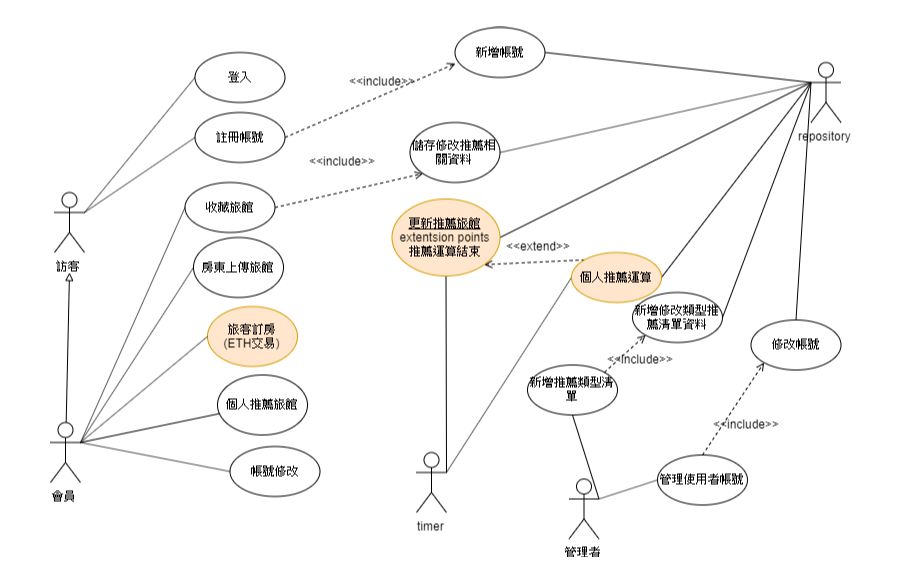
* **網路架構圖**

使用spark建立一台master，兩台slave的分散式叢集。

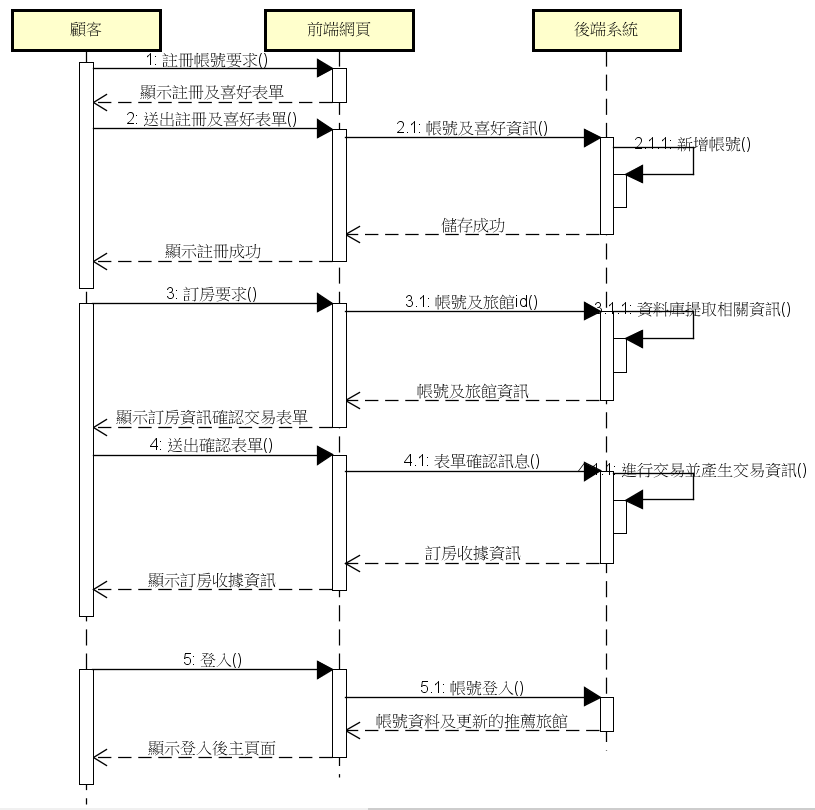


**圖二、網路架構圖**

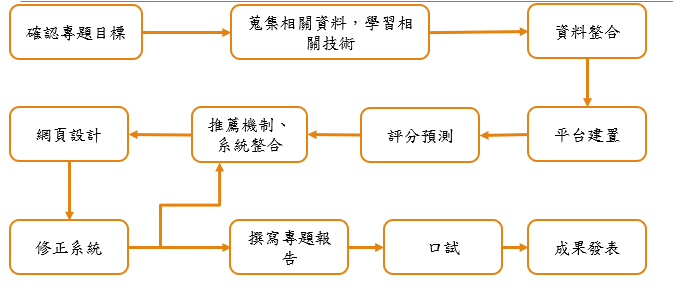
* **UseCase**



* **循序圖**



* **開發流程圖**



1. **使用對象**

提供所有房東上傳旅館的網頁，並提供旅客推薦服務以及訂房交易平台。

1. **使用環境**

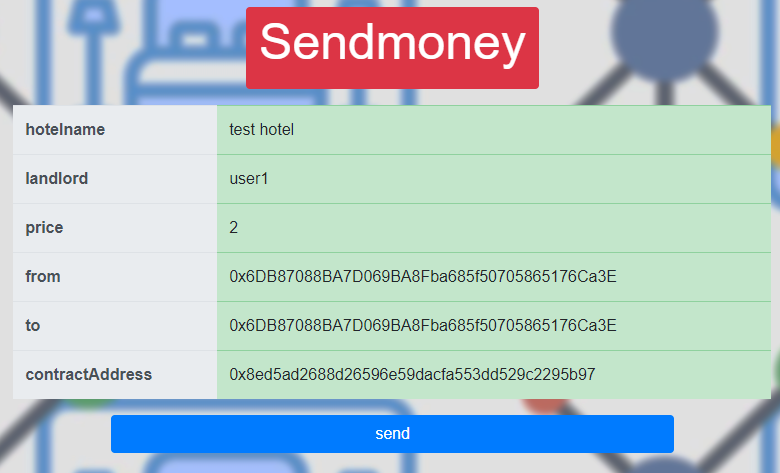
|  |  |  |  |
| --- | --- | --- | --- |
|  | 節點1 | 節點2 | 節點3 |
| 作業系統 | Ubuntu14.04 | Ubuntu14.04 | Ubuntu14.04 |
| 處理器 | Intel Core i7-4790 3.60GHz | Intel Core i7-4790 3.60GHz | Intel Xeon CPU E5-2630 2.3 |
| 記憶體 | 4.8GB | 7.7GB | 3.9GB |
| 虛擬機器 | 是 | 是 | 是 |
| Hadoop版本 | 2.8.5 | 2.8.5 | 2.8.5 |
| Spark版本 | 2.4.0 | 2.4.0 | 2.4.0 |

1. **開發工具**
2. Scala-2.11.6
3. Java-8
4. Python-3.7.0
5. Eclipse
6. Virtualbox
7. Jupyter Notebook
8. MongoDB
9. Ganache
10. Viper 0.1.0b13+commit.375fb27
11. **系統畫面**

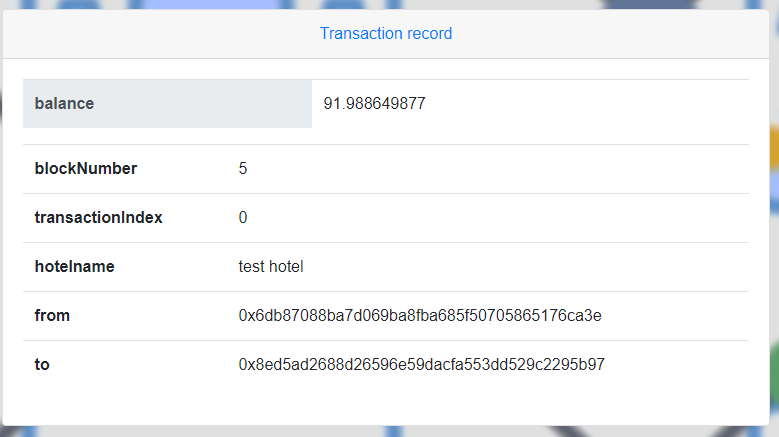
* **網頁畫面**
  1. 首頁



* 1. 交易清單

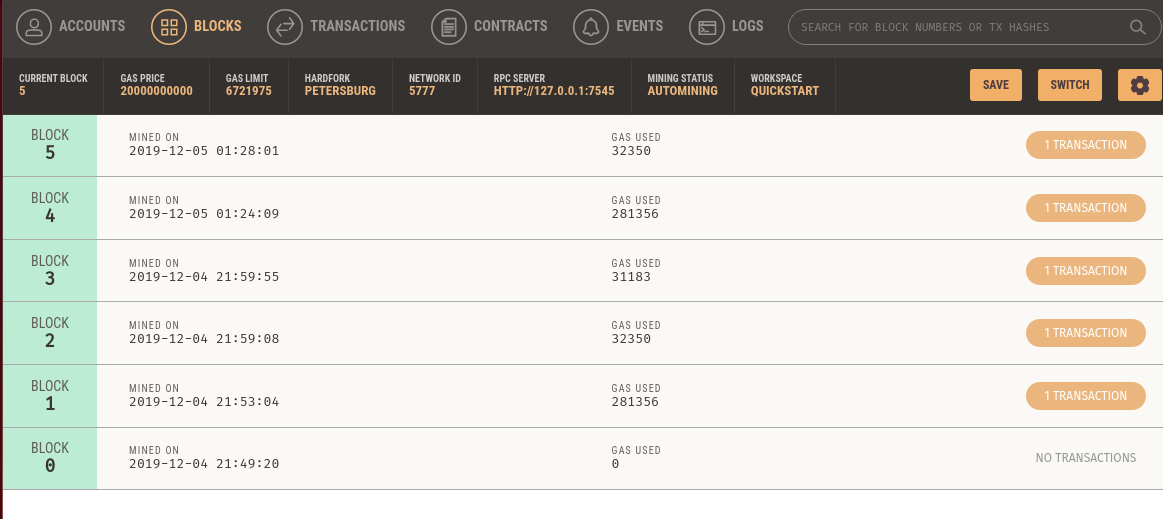


* 1. 旅館網頁上交易後產生的交易收據紀錄畫面

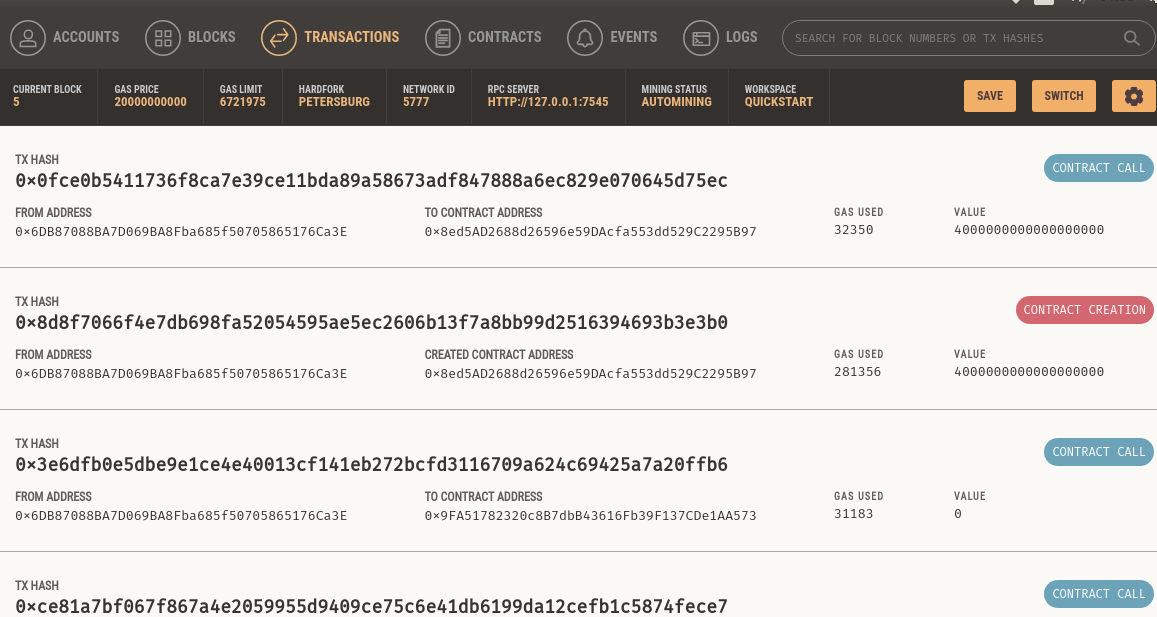


* **Ganache區塊鏈畫面**

1. 區塊鏈產生Block畫面



1. 區塊鏈產生Transaction畫面

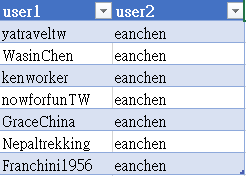


* **爬蟲資料相關畫面**

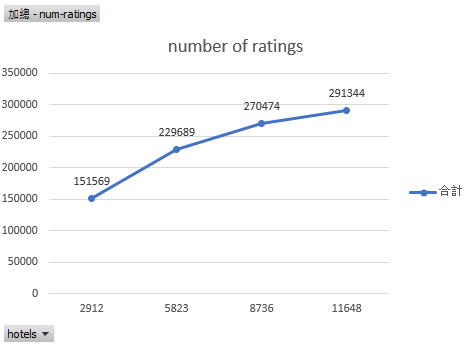
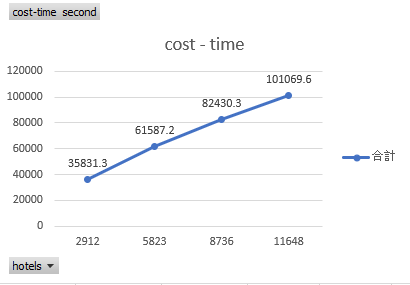
1. 爬取的評分資料



1. 爬取信賴關係資料



1. 爬蟲模組爬取效能



**(a) (b)**

1. **成本分析**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 項 目 名 稱 | 說 明 | 單位 | 數量 | 單 價 | 小 計 | 備 註 |
| 臺幣(元) | 臺幣(元) |
| 個人電腦 | 專案之進行 | 部 | 3 | 26000 | 78000 | 由系上實驗室提供 |
| 消耗性器材 | 印表機消耗材料、紙張等 | 批 | 1 | 1000 | 1000 | 由系上實驗室提供 |
| 雜支費 | 印刷費、文具等 | 批 | 1 |  | 500 | 自行負擔 |
| 共 計 | | | | | 101500 |  |

1. **結論**

本專題的實驗系統應用在旅館推薦服務上，主要有flask網頁框架建立的網頁平台，應用Spark叢集平台為後端分散式推薦運算，使用其ALS推薦演算法來建立推薦系統，也應用區塊鏈的智能合約在旅館訂房交易上。

本專題有參加Tanet投稿，並獲得優秀論文獎。

1. **附錄**

Tanet優秀論文獎

