## Wikipedia National Geography

**Пропозиція про проведенню дослідження**

**Вимога на систему**

|  |  |  |
| --- | --- | --- |
| 1. | Спонсорування проекту | Ініціатор проекту: Андрій Дмитрук  Проект не потребує жодних додаткових ресурсів чи спонсорування – дані, які будуть оброблятись належать до загального доступу і не потребують значних ресурсів для обробки. |
| 2. | Потреба | Потреба в даному проекті існує, адже в інтернеті не знайдено жодного іншого сайту, який би ідентично повторював можливості даного проекту.  Не зважаючи на існування багатьох проектів з подібним функціоналом (на приклад, [WikiGalxy](http://wiki.polyfra.me/), [TraceMedia](http://tracemedia.co.uk/portfolio/mapping-wikipedia/)), вони не надають деякі з зазначених тут функцій, або не видають достатню кількість інформації для української мови.  Водночас, даний проект може слугувати як джерелом інформації для зацікавлених в статистиці користувачів Вікіпедії, так і містити цінні статистичні дані.  (Детальніше про можливості і використання можна дізнатись на сторінці проблеми і завдань проекту) |
| 3. | Вимоги | Зберігати значні кількості даних, отриманих з Wikipedia API, обробка даних перед збереженням для заощадження місця і збільшення подальшої швидкодії програми.  Створювати карту за наявними даними (або отримувати і зберігати нові дані для карти залежно від ситуації) з візуальним виводом інформації на карті.  Створювати різну карту залежно від вказаних параметрів (мови, або кількох мов, території, на яку виводити, що саме виводити).  (Детальну інформацію про очікувані можливості проекту можна прочитати в документі проблеми і завдань проекту)  Дуже важливою умовою успішного виконання програмою вимог є можливість відфільтровувати написані (перекладені) ботами сторінки енциклопедії і зображати їх окремо від загальних. (Способами реалізації цього є фільтрування імені і user\_agent користувачів на наявність інформації про автоматизацію виконання певних функцій користувачем – ботом; або отримання завчасно створених списків зареєстрованих ботів у Вікіпедії і порівняння користувачів до них)  (Детальніше про вимоги до проекту і можливість отримання такої інформації на основі вікіпедії описано нижче в документі) |
| 4. | Вигоди | Створені даною програмою дані зможуть дати детальну інформацію, для відображення активності користувачів Wikipedia в тих чи в інших місцях, активності написання сторінок про різні регіони і країни і показати реальну картину залежності цих двох факторів для версій енциклопедії різними мовами. |
| 5. | Питання та обмеження | Граничні терміни здачі етапів розробки проекту:   1. Визначення теми дослідження – 25 лютого 2. Робота з типами даних – 11 березня 3. Розробка функцій і структур даних проекту – 25 березня 4. Проведення дослідження – 6 травня 5. Здача проекту – 25 травня   Для збільшення швидкості обробки даних, можна здійснити спробу отримання більших прав на запити Wikipedia API. |

**Основні завдання дослідження. Можливості отримання необхідних даних для дослідження з Wikipedia API.**

Дані, які можна зібрати з Wikipedia API та які задовольняють вимоги до завдання даного проекту і з яких можна отримати необхідну інформацію при обробці:

1. Отримання списку усіх сторінок (тільки офіційних сторінок, а не додаткових сторінок, redirect сторінок, сторінок з профілями користувачів) за допомогою query:

Параметри запиту:

* list = allpages
* apfilterredir = noredirects (всі сторінки, окрім сторінок, які здійснюють перевід на інше посилання)

Результат запиту: список сторінок, з зазначеними id та іменами сторінок.

1. Отримання внутрішнього вмісту сторінки по її імені або id.

Параметри запиту:

* prop = revisions
* rvprop = ids | timestamp | content | comment | tags | user
* titles = <page names> АБО ids = <page ids>

Отриманий результат може містити внутрішній текст редакції і інформацію про редакції сторінки залежно від значення параметру rvparam. Для отримання вмісту, необхідно задати rvparam = content. Для отримання інформації про редакцію, необхідно задати rvparam = user (користувач, який здійснив редакцію), rvparam = timestamp (час коли було здійснено редакцію). Можна водночас отримувати інформацію про декілька з значень цього параметру, задавши їх водночас (rvparam = user|timestamp)

При простому запиті з цими параметрами, ми отримаємо інформацію про одну редакцію сторінки (це буде остання версія сторінки). Цей функціонал і буде використано для збирання геоданих про координати місця на сторінці. Необхідно буде переглянути текст сторінки і знайти теги вікіпедії, які відповідають за координати - {{ Coord | parameters }}.

Для отримання інформації про користувачів і дату, корисним буде надсилати запит на більш ніж одну редакцію даної сторінки або отримувати інформацію про першу редакцію. Для цього можна використати наступні параметри:

* rvlimit = max АБО <number> - задати кількість редакцій, на які ми здійснюємо запит
* rvdir = newer АБО older – задати напрям, в якому буде посортовано список редакцій

(від старішої до новішої АБО від новішої до старішої)

Не варто об’єднювати запит на вміст сторінки і інформацію про редакції (авторів і дату), адже текст займає набагато більше місця і буде значно заповільнювати запит і його обробку

Даний запит дасть змогу отримувати координати сторінок, дату створення сторінки, інформацію про користувачів, які створили і редагували сторінку (зокрема в декого – їх ip)

1. Отримання посилань на версії даної сторінки інших мов.

Параметри запиту:

* param = langlinks
* lllimit = max
* titles = <page names> АБО ids = <page ids>

Результат запиту: список усіх версій даної сторінки в інших мовах. Список містить код мови і назву сторінки на даній мові. Цієї інформації достатньо для надсилання запиту на отримання даної сторінки в Вікіпедії.

Ці 3 з багатьох можливих поєднань параметрів query (action = query) запиту на Wikipedia API, дають можливість зібрати усю необхідну для реалізації даного проекту інформацію. Крім того, API дає змогу отримувати до 50 (100 для користувачів) сторінок за один запит або 500 (1000 для користувачів), якщо ідеться про списки, або інші запити, які не містять багато тексту чи інформації.

Інформації, зібраної з Вікіпедії буде достатньо для проведення статистичної оцінки активності, адже за інформацією, розміщеною на сторінці [WikiProject Geographical coordinates](https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates), кількість сторінок з географічними мітками в версії енциклопедії українською мовою сягає 31%, англійської – 20%. Це значні цифри, враховуючи, що більшість статей Вікіпедії (на приклад, про природничі науки) не розповідають про певне місце на планеті і відповідно не можуть мати координат.

Крім того, за інформацією, наданою Вікіпедією, в залежності від мови кількість анонімних редакцій в залежності від мови, може сягати більш ніж 50% від всіх редакцій. З таких редакцій можна легко отримати ip автора, а отже і приблизні координати здійснення редагування або створення. Цієї інформації без необхідності отримання жодних додаткових прав, буде достатньо для проведення успішного дослідження і аналізу. На жаль, неможливо отримати координати зареєстрованих в Вікіпедії користувачів, адже це суперечить [політиці Wikimedia Foundation](https://wikimediafoundation.org/wiki/Privacy_policy).