**МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ**

Національний університет “Львівська політехніка”

**ВИВЧЕННЯ ОСНОВ КОРПУСНОЇ ЛІНГВІСТИКИ.**

**КОДУВАННЯ ТЕКСТІВ ПРИ ЇХ ВКЛЮЧЕННІ ДО КОРПУСУ ТЕКСТІВ**

**МЕТОДИЧНІ ВКАЗІВКИ**

до розрахунково графічної роботи

з курсу “Комп’ютера лінгвістика ” для студентів базового напрямку 7.08.04 “Прикладна лінгвістика”

(частина 2)

ЛЬВІВ 2013

**ВИВЧЕННЯ ОСНОВ КОРПУСНОЇ ЛІНГВІСТИКИ. КОДУВАННЯ ТЕКСТІВ ПРИ ЇХ ВКЛЮЧЕННІ ДО КОРПУСУ ТЕКСТІВ.** Методичні вказівки до розрахунково графічної роботи з курсу “Комп’ютерна лінгвістика ” для студентів базового напрямку 7.08.04 “Прикладна лінгвістика” /Укл. А.Б.Романюк - Львів:НУЛП, 2013 - 15с.

Укладачі: А.Б.Романюк к.т.н., доцент

О.І. Дмитраш, аспірант

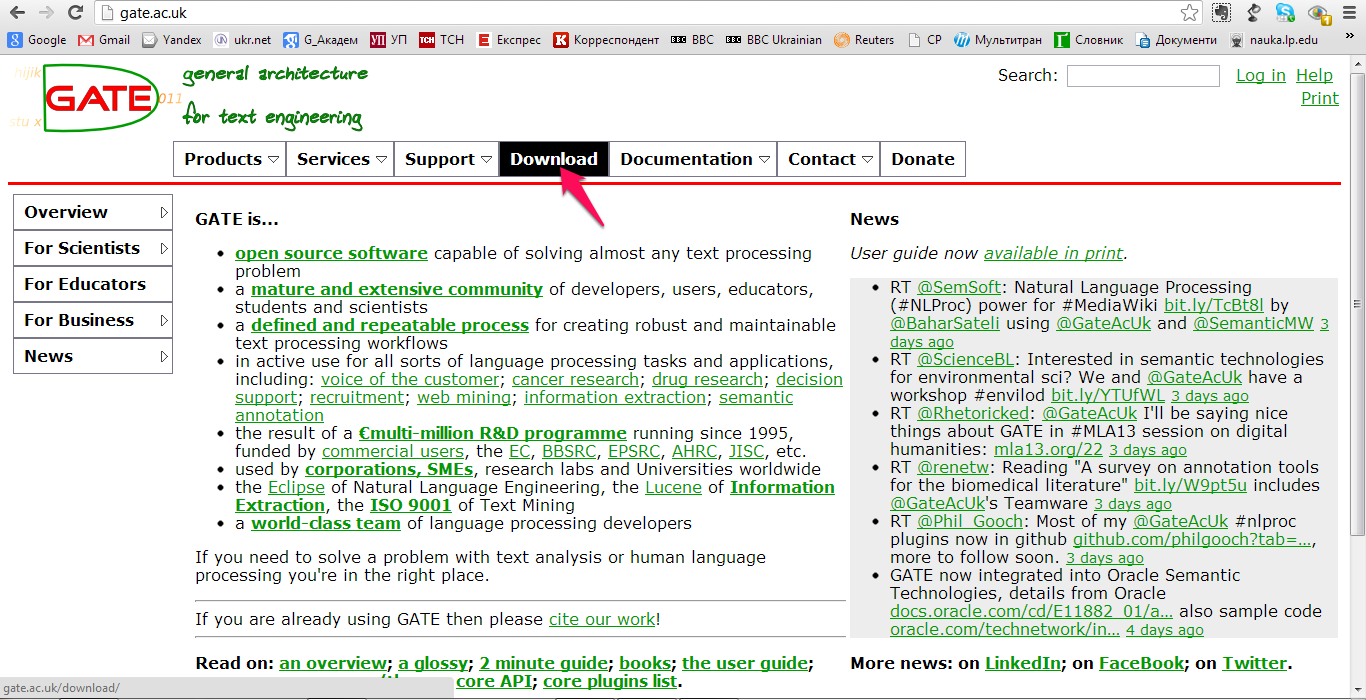
**МЕТА РОБОТИ**

* Вивчення основних понять корпусної лінгвістики.
* Вивчення основ кодування текстів при їх включенні в корпус текстів.

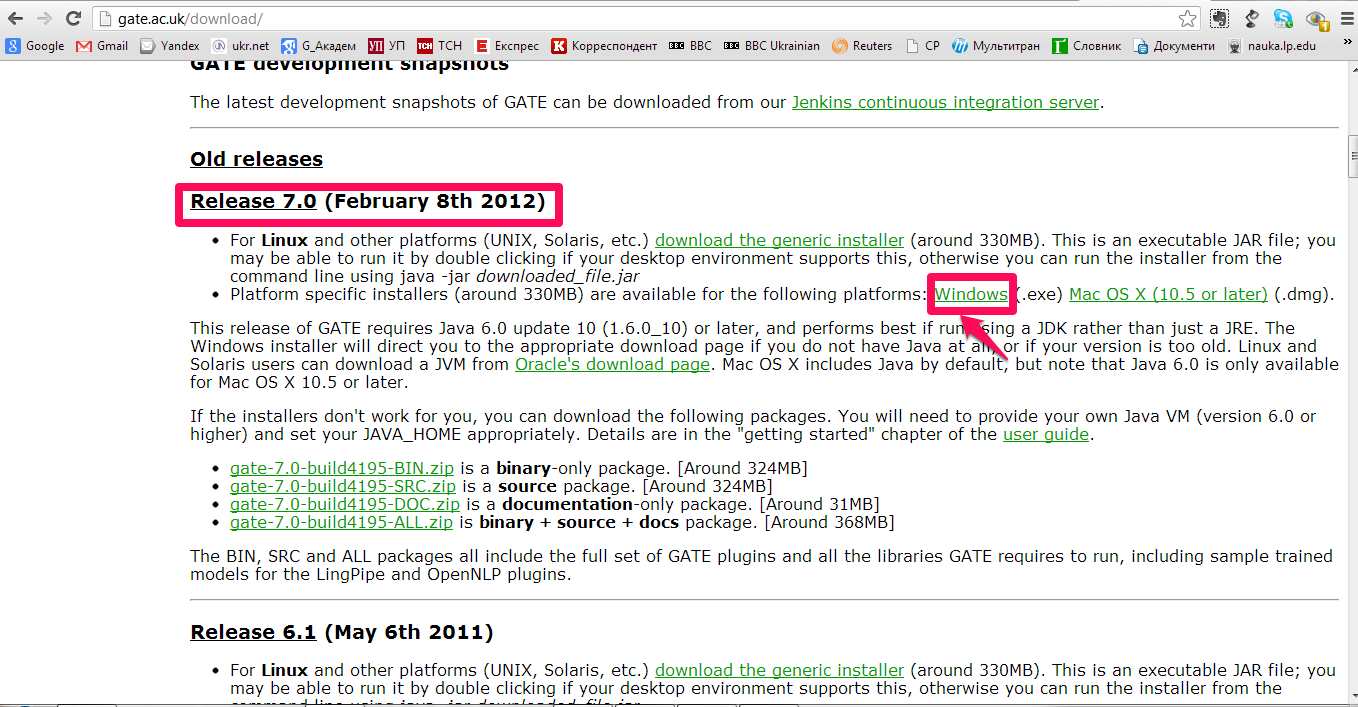
***Хід виконання роботи***

**0. Gate (*Встановити програмне забезпечення: Gate7.0)***

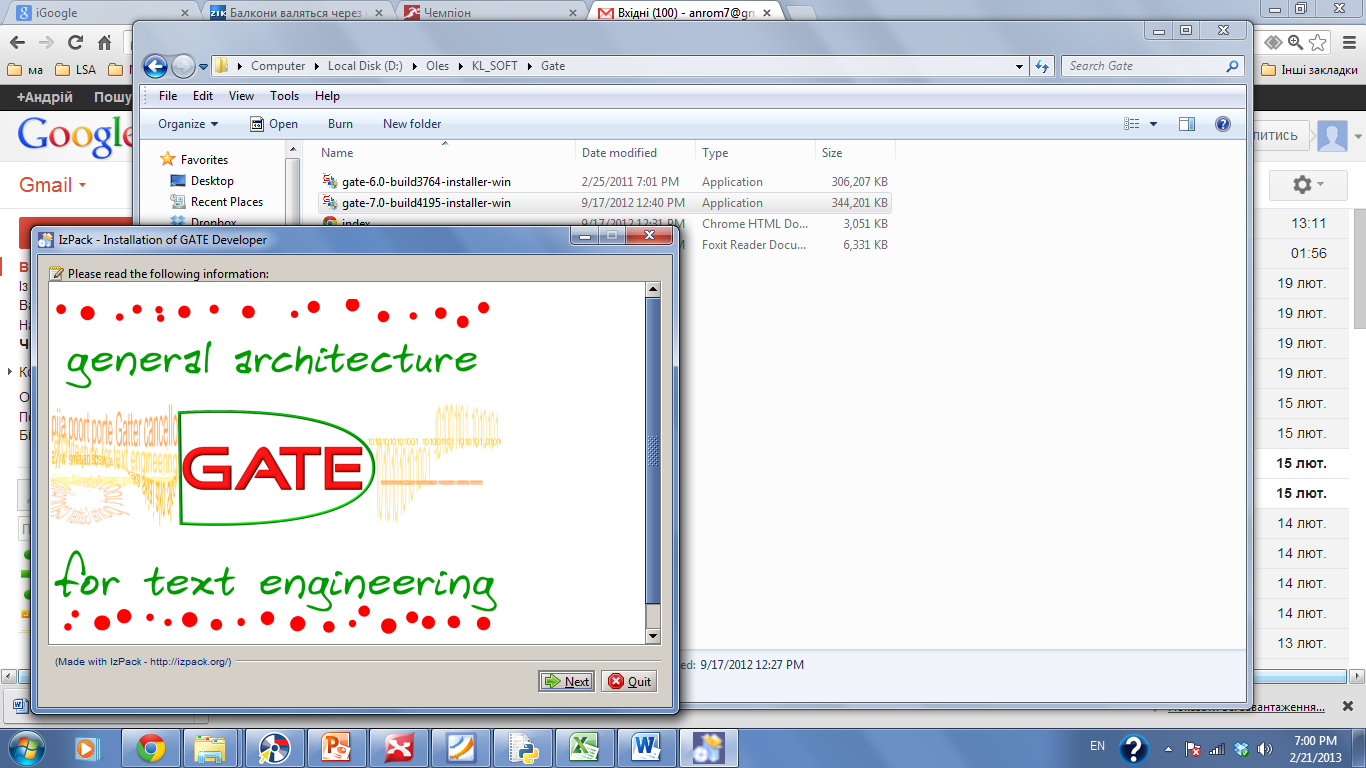
1. Зайти на сайт [*www.gate.ac.uk*](http://www.gate.ac.uk)та перейти на закладку *Download.*



2. Серед запропонованих обрати версію Gate 7.0 та завантажити її відповідно до операційної системи Вашого комп’ютера (у даному випадку Windows).

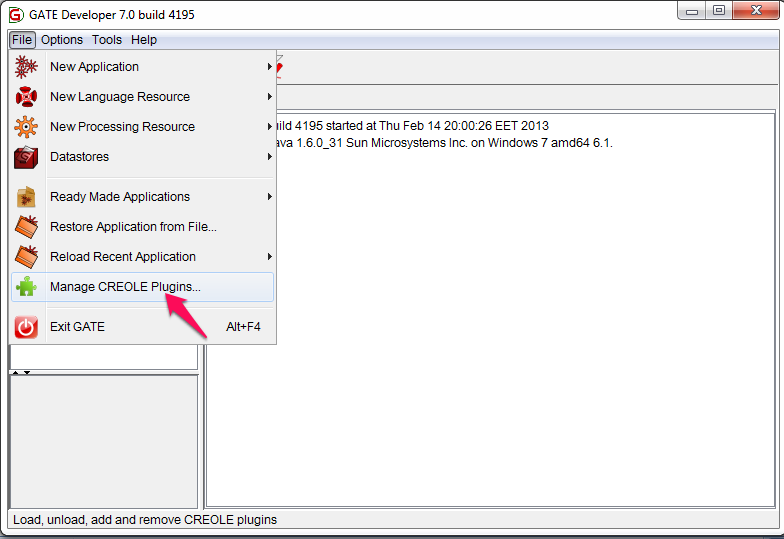


3. Встановити Gate 7.0

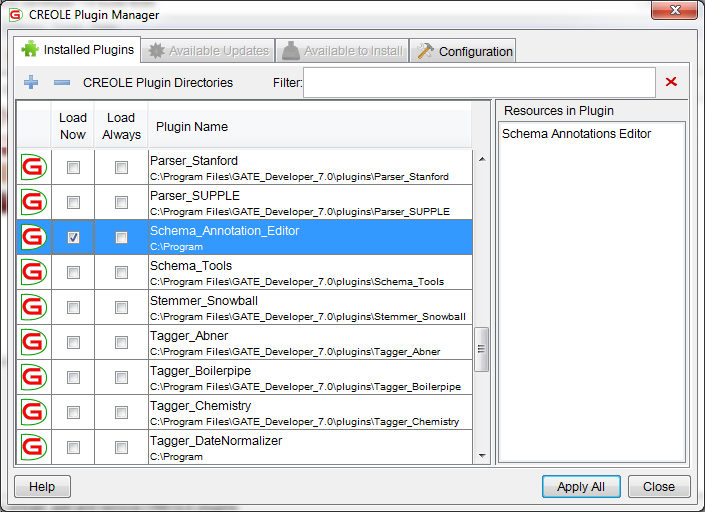


**I. Gate**

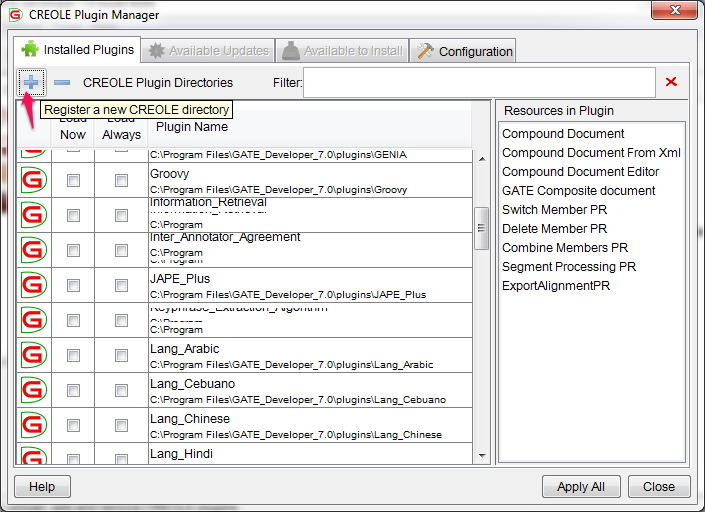
1. Увімкнути *Gate*.
2. *File* ->*Manage CREOLE Plugins*



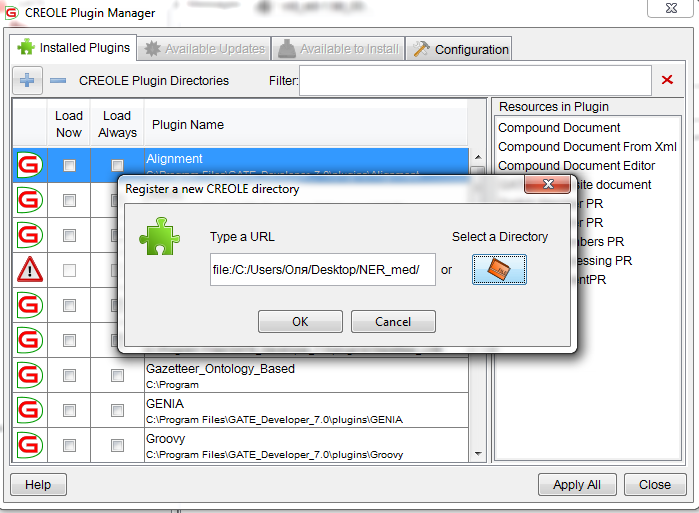
1. Поставити галочку *Load now* для *Schema\_Annotation\_Editor*



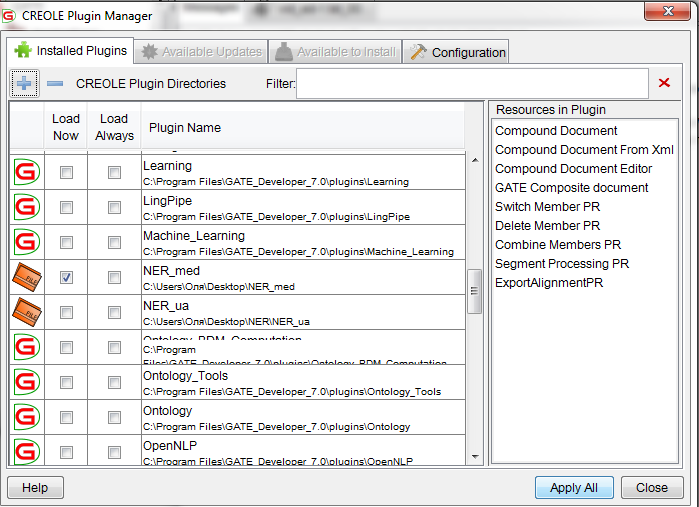
1. Клацнути *+ (Register a new CREOLE directory)*



1. У новому діалоговому вікнінатиснути  *Select directory* та вибрати *NER\_med*



1. Поставити галочку *Load now* для *NER\_med* та натиснути *Apply All*.



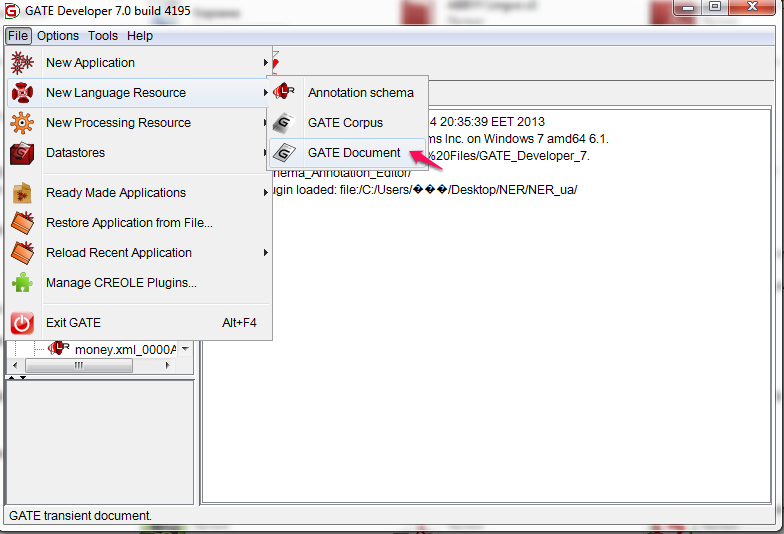
1. **УВАГА!!! Цього разу не потрібно з переліку обирати ANNIE, лише *Schema\_Annotation\_Editor* та *NER\_med***

**II. Вхідні дані**

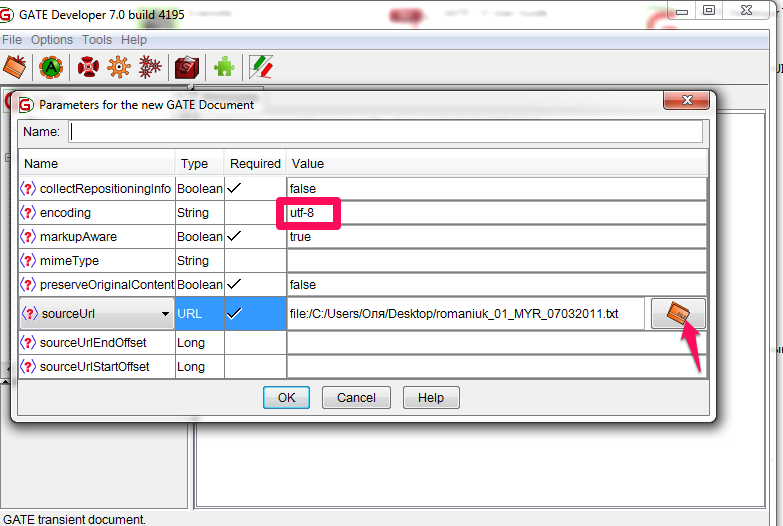
1. Підготувати вхідні дані:
   1. зберегти кожен відгук (файл з відгуками Forum\_files.pdf, відгуки вибирати згідно вказаних сторінок у таблиці із завданням ) в окремому *\*.txt* файлі під назвою *прізвище\_студента\_номер\_відгука\_автор\_дата(день,місяць,рік)*.*txt Кодування файла UFF-8*
   2. зберегти фрагмент статті (файли статей у pdf форматі згідно таблиці із завданням, фрагмент статті, який потрібно опрацювати виділено у файлі Article\_files.rtf) в окремому *\*.txt* файлі під назвою *прізвище\_студента*.*txt Кодування файла UFF-8*

**III. Завантажити вхідні дані**

1. Клацнути *File* ->*New Language Resources* ->*Gate Document*



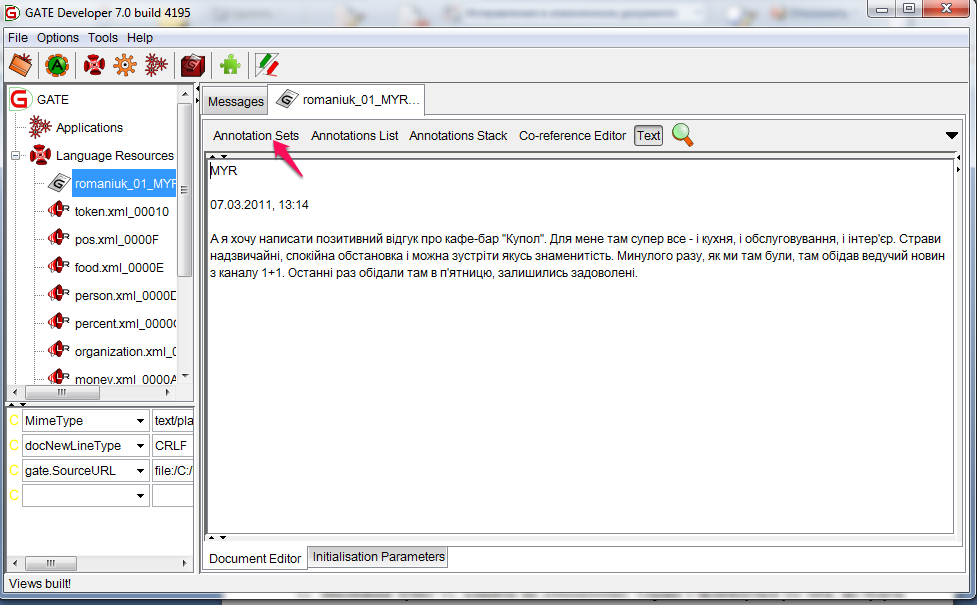
1. Обирати необхідний файл та вписати ***utf-8*** в поле *encoding* та натиснути *OK.*



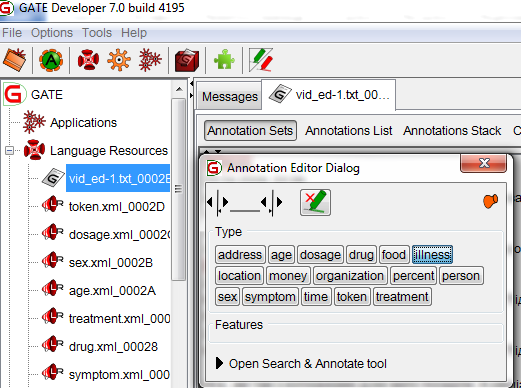
1. Цей текстовий файлик з’явився у структурі зліва. Щоб відобразити вміст файлу, на нього потрібно клацнути двічі.
2. Додати всі текстові файли до *Language Resources*
3. **УВАГА!!! Цього разу ми не додаємо усі файлики до корпусу і не застосовуємо ANNIE**

**IV. Розмітка**

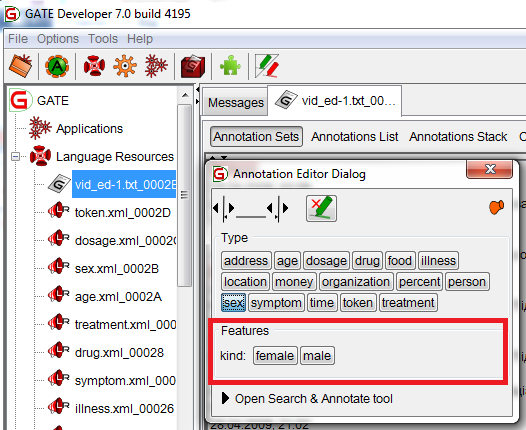
1. Відкрити файл та клацнути на *AnnotationSets*. Справа з’являтимуться усі теги, які будуть використовуватись для розмітки



1. Щоб присвоїти певному текстовому фрагменту потрібний тег, потрібно виділити цей фрагмент тексту (NE – Named Entity) і в вікні, що з’явилось, вибрати потрібний тег



1. Переважна більшість тегів міститиме в собі підтеги, так звані *Features*, серед яких також потрібно обрати відповідний для конкретного фрагменту тексту



**УВАГА!!! Якщо Вам потрібно закрити віконечко з тегами *Annotation Editor Dialog,* потрібно натиснути Esc.**

1. У кожному відгуці потрібно виділити усі ті фрагменти тексту, які можна віднести до однієї з цих категорій:

*Табл. 1.*

**Типи Named Entities**

|  |  |  |
| --- | --- | --- |
| **TYPE** | **FEATURES** | **EXAMPLES** |
| **address** | street | вул. Личаківська,  пл. Ринок 5,  вул. Ст.Бандери 28 |
| postcode | 79000, K1A 0A6, 10003 |
| phone | (0322)725569, +38(044)358053, 0936788743 |
| email | [sekine@cs.nyu.edu](mailto:sekine@cs.nyu.edu),  [rector@lp.edu.ua](mailto:rector%40lp.edu.ua) |
| url | <http://lp.edu.ua>, www.bbc.com |
| **time** | date | 1.02.2012, 14 квітня, восени, День Незалежності, 2013р., липень, понеділок, вихідні дні, 4 роки тому |
| time | 12:10, 6 год., північ |
| period | 10 діб, 30 хвилин, 2 години, півдня, один тиждень, 3 роки |
| **location** | сontinental\_region | Європа, Азія, Північна Америка |
| country | Україна, Японія, США, Канада |
| region | Західна Україна, Київська обл., |
| city | Лондон, Харків, Торонто |
| **money** |  | 50 грн., $40 |
| **organization** |  | Microsoft, Світоч, Мапа |
| **percent** |  | 100%, 22,5% |
| **person** | name | Андрій Худо, А.В.Іванченко, Зірко В.К. |
| position | **пацієнт, хворий**, директор, офіціант, професор, лікар |
| nickname | Nezabudka, ola-lyola |
| **food** |  | яблучний пиріг, сік, суші, картопля, м’ясні страви, пиво, чай, солодощі |
| **token** | pos | N (Noun - Іменник) |
| V (Verb - Дієслово) |
| A (Adjective - Прикметник) |
| P (Pronoun - Займенник) |
| R (Adverb - Прислівник) |
| S (Adposition - Прийменник) |
| C (Conjunction - Сполучник) |
| M (Numeral - Числівник) |
| Q (Particle - Частка) |
| I (Interjection - Вигук) |
| Y (Abbreviation - Скорочення) |
| X (Residual - Залишок) |
| lemma |  |
| morphology | Тег згідно http://nl.ijs.si/ME/V4/msd/html/msd-uk.html |
|  |  |  |
| **illness** |  | Інфаркт, грип, злоякісна пухлина, ішемічна хвороба серця, харчове отруєння, остеопороз, цукровий діабет, автоімунний тиреоїдит, фіброзно –кістозна остеодистонія, хронічний панкреатит, рак нирок, вертебро-базилярний інсульт, вузловий зоб, колоїдний зоб, пневмонія, і т.д. |
| **symptom** |  | Ламкість кісток, слабкість, свербіж, подразнення, біль у грудях, одинофальгія, втома, спазм, відчуття печіння, невралгія, зниження апетиту, кардіалгія, печія, кольки, задишка, гіркота у роті, потемніння в очах, кашель, озноб, і т.д. |
| **drug** |  | Протиконвульсні препарати, розчин Рінгера, таміфлю, розчин амінокапронової кислоти, цитофлавін, норадреналін, іт.д. |
| **treatment** |  | пункційна біопсія, через шкірні пункцій ні втручання, лапароскопічна холицестомія, назобіліарне дренування, через печінкове дренування жовчного міхура, ультразвуковий контроль, паліативне лікування, і т.д. |
| **age** |  | 60-74 рр., 41- 50 років, підлітки, дорослі, похилого віку, старші 50 років, 45 років, **Звертайте увагу на контекст, щоб відрізнити від *time->period*** |
| **sex** | female | жінки |
|  | male | чоловіки |
| **dosage** |  | (дозування лікарських засобів) 0,75 мг; 0, 003 мг |

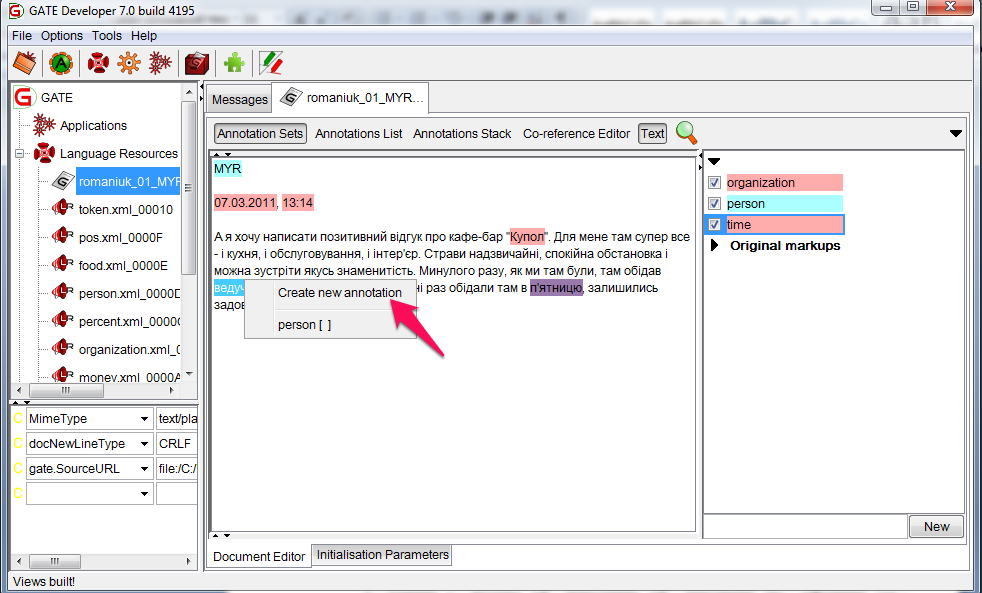
**Перечитайте спочатку повністю статтю, а тоді починайте тегування!**

При потребі користуйтеся доданими медичними словниками та

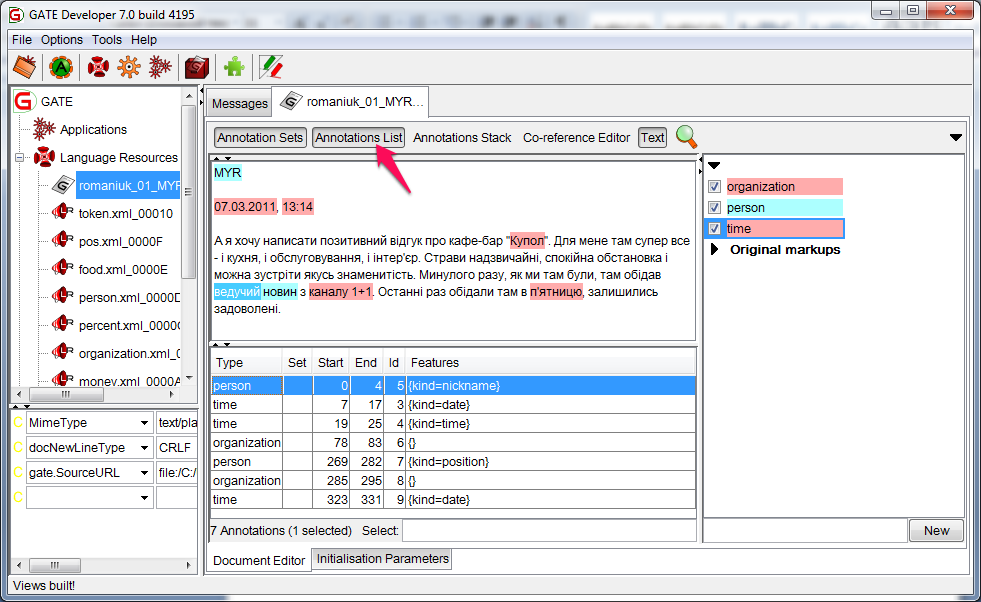
<http://slovnik.com.ua/?grupa=6>

<http://www.lingvo.ua/uk/Search/uk-uk?searchMode=Interpretations&fromSlovnyk=1>

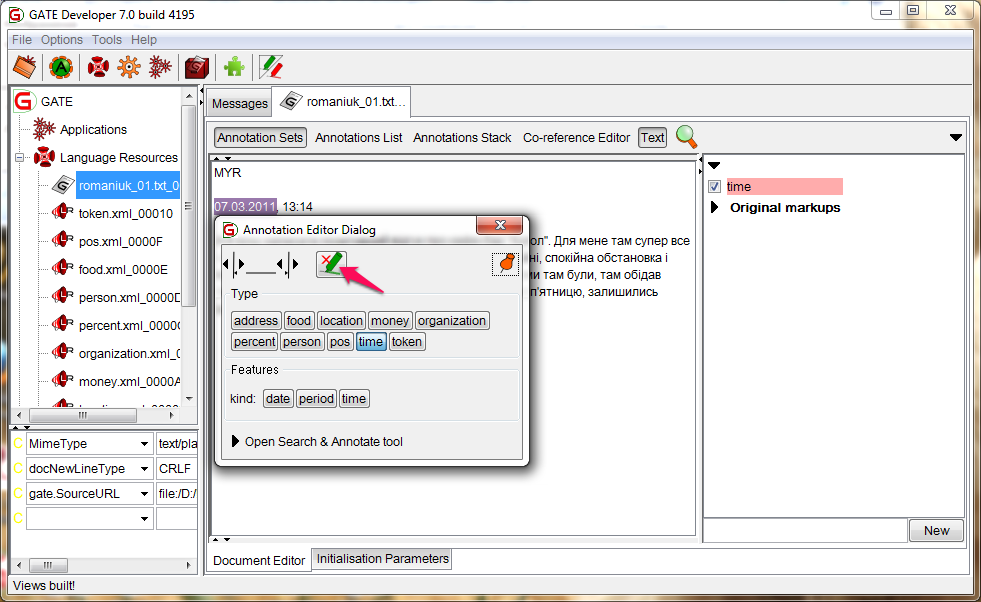
1. Якщо два фрагменти з різними тегами накладаються, потрібно виділити фрагмент і вибрати *Create new annotation* у випадаючому вікні



1. **УВАГА!!!! Як і в першій частині розрахункової роботи кожне слово у виділеному фрагменті тексту (Named Entity) потрібно виділити тегом *token*. АЛЕ лише тегом *token*, не вказуючи features (pos, lemma, morphology).**
2. **УВАГА!!! Тепер потрібно виділяти тегом *token* 2 слова ДО і 2 слова ПІСЛЯ кожного виділеного фрагмента тексту (Named Entity) У МЕЖАХ РЕЧЕНЯ!!!** Тобто, якщо після Named Entity стоїть крапка, то 2 слова після уже тегувати не потрібно. Аналогічно: якщо Named Entity - перше слово/словосполучення у реченні, то 2 слова перед ним тегувати не потрібно. **Для цих 2 слів до і 2 слів після Named Entity потрібно виділяти усі features (pos, lemma, morphology).**
3. Клацнувши на *AnnotationsList* біля *AnnotationSets*, можна побачити усі проставлені мітки та їх характеристики. Клацнувши на якусь із них, відповідний фрагмент тексту почне мигати і можна буде змінити розмітку при потребі.

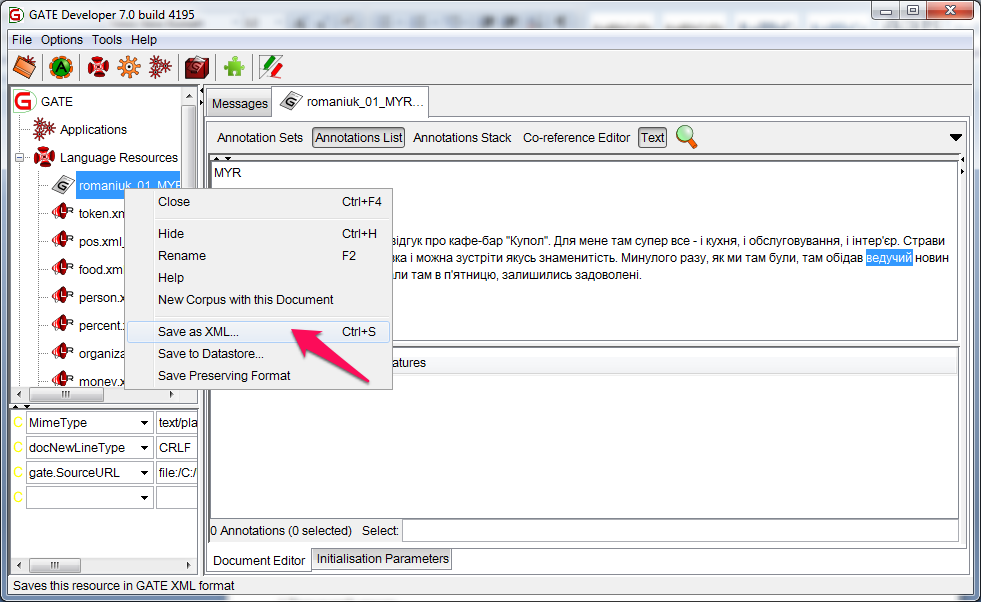


1. Для того, щоб видалити певну мітку, потрібно виділити ще раз фрагмент тексту, якому присвоєна мітка і у новому вікні натиснути на іконку видалення мітки:



**V. Збереження результатів**

1. Клацнути правою кнопкою на файлі у структурі справа і вибрати *Save as xml*



1. Збережіть результати роботи під назвою

Відгуки - файлі під назвою *прізвище\_студента\_номер\_відгука\_автор\_дата(день,місяць,рік)*.*xml*

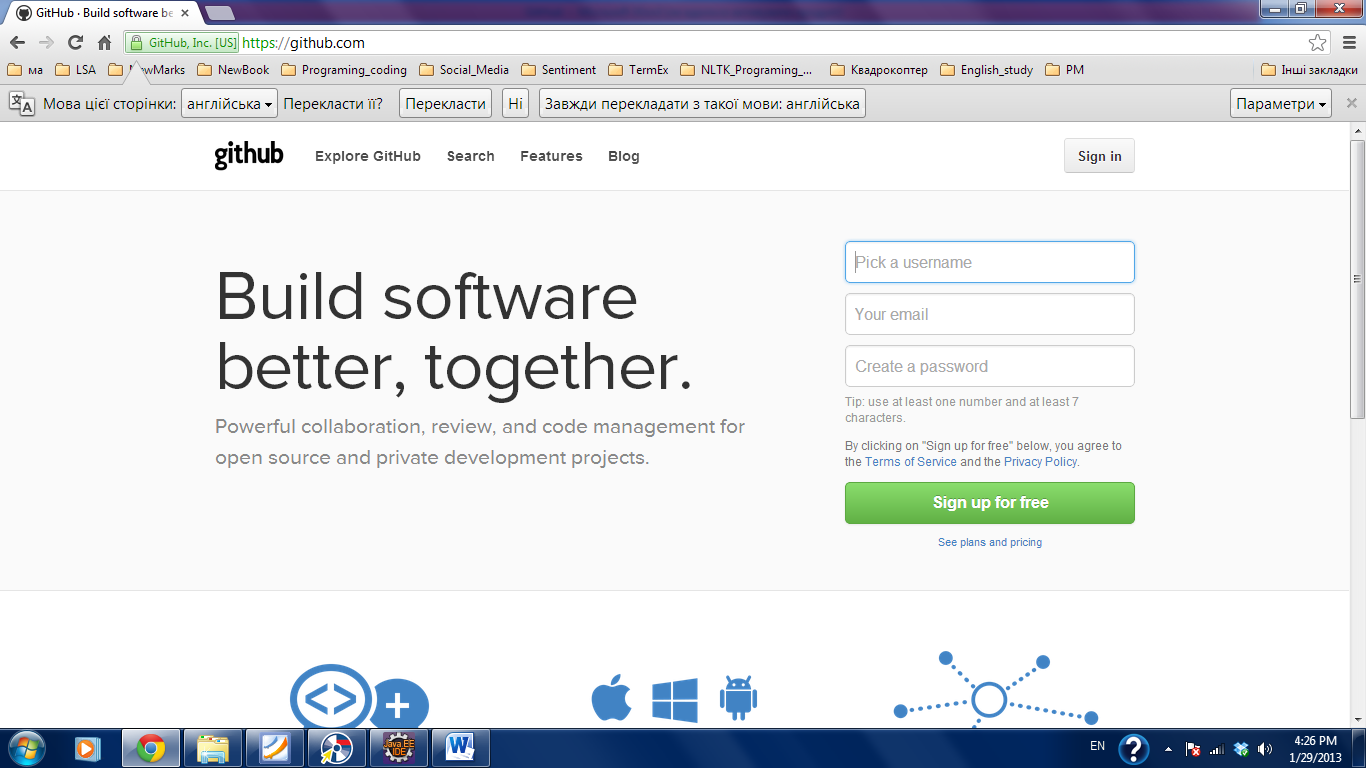
Фрагмент статті *-* файлі під назвою *прізвище\_студента*.*xml*

**VI. Надсилання результатів**

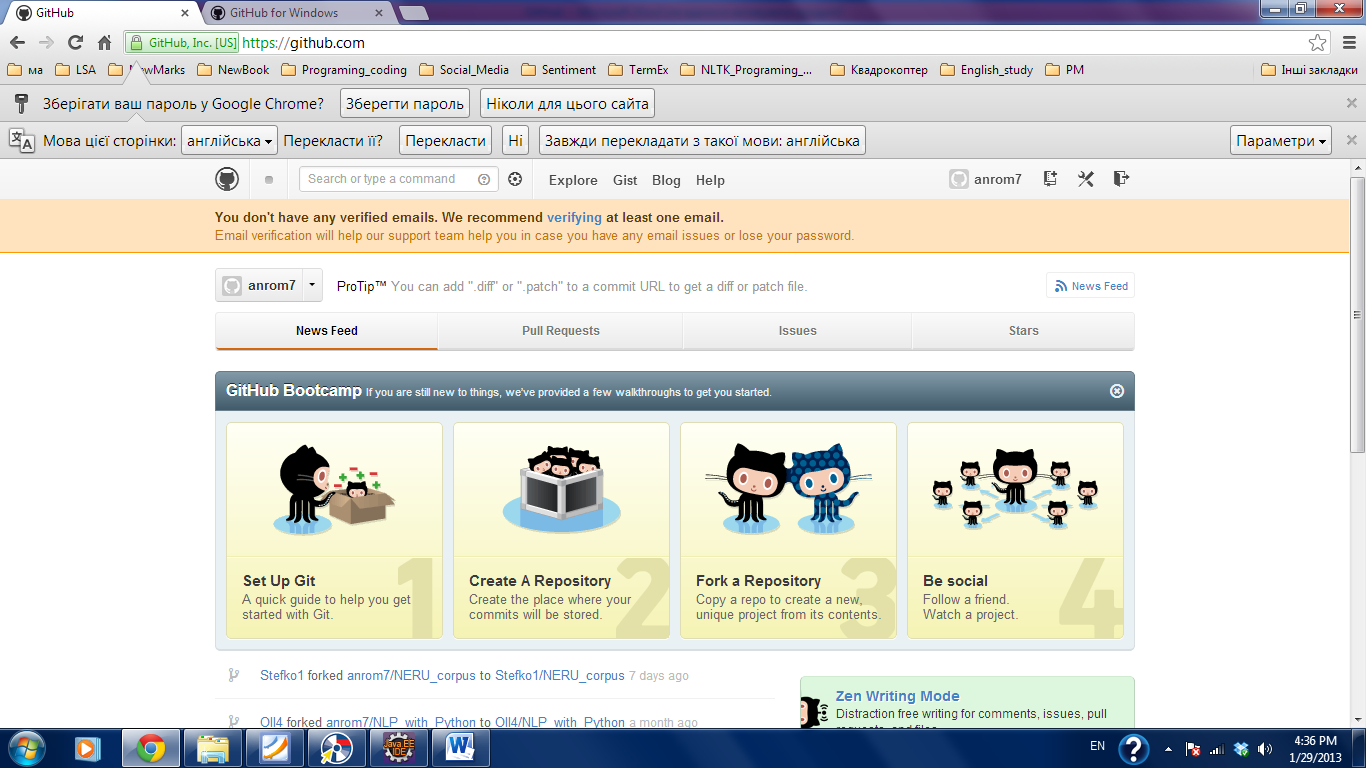
1. Надсилання отриманих розмічених текстів та колективна робота над ними здійснюватиметься за допомогою одного з найбільших веб-серверів для спільної розробки програмного забезпечення GitHub. Опис налаштування GitHub та [GitHub for Windows](http://windows.github.com/) подано нижче.

**Якщо Ви уже створили свою гілку для першої частини розрахункової роботи, то одразу перейдіть до пункту 34.**

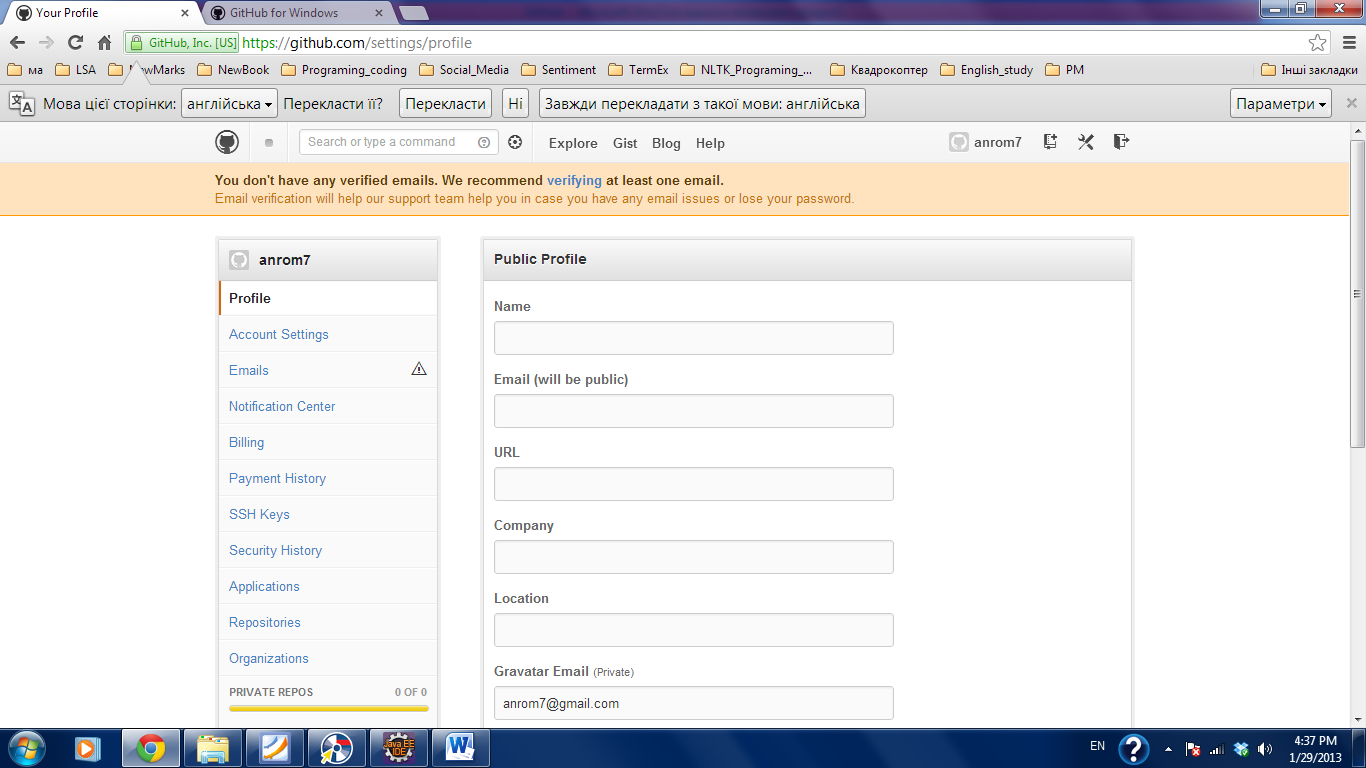
1. Зайти на сайт https://github.com/



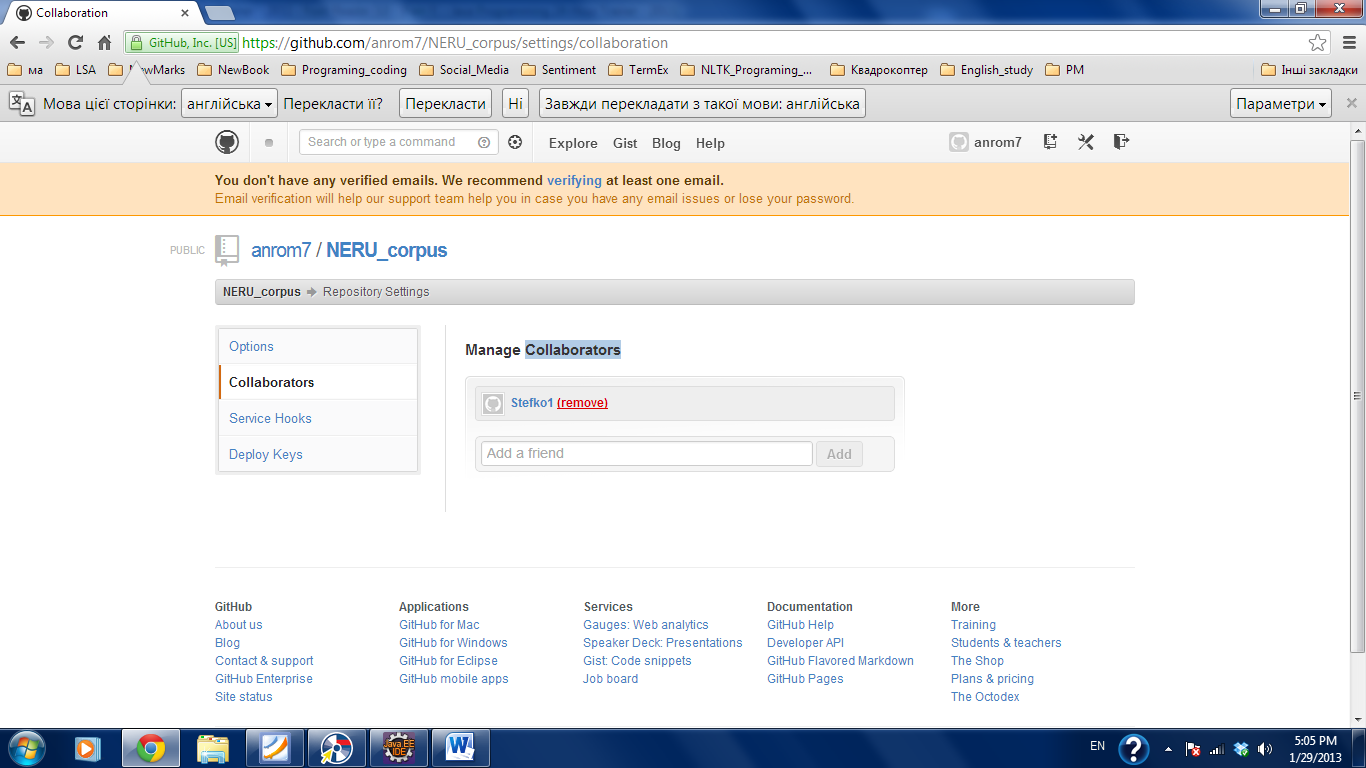
1. Створити обліковий запис (User account)



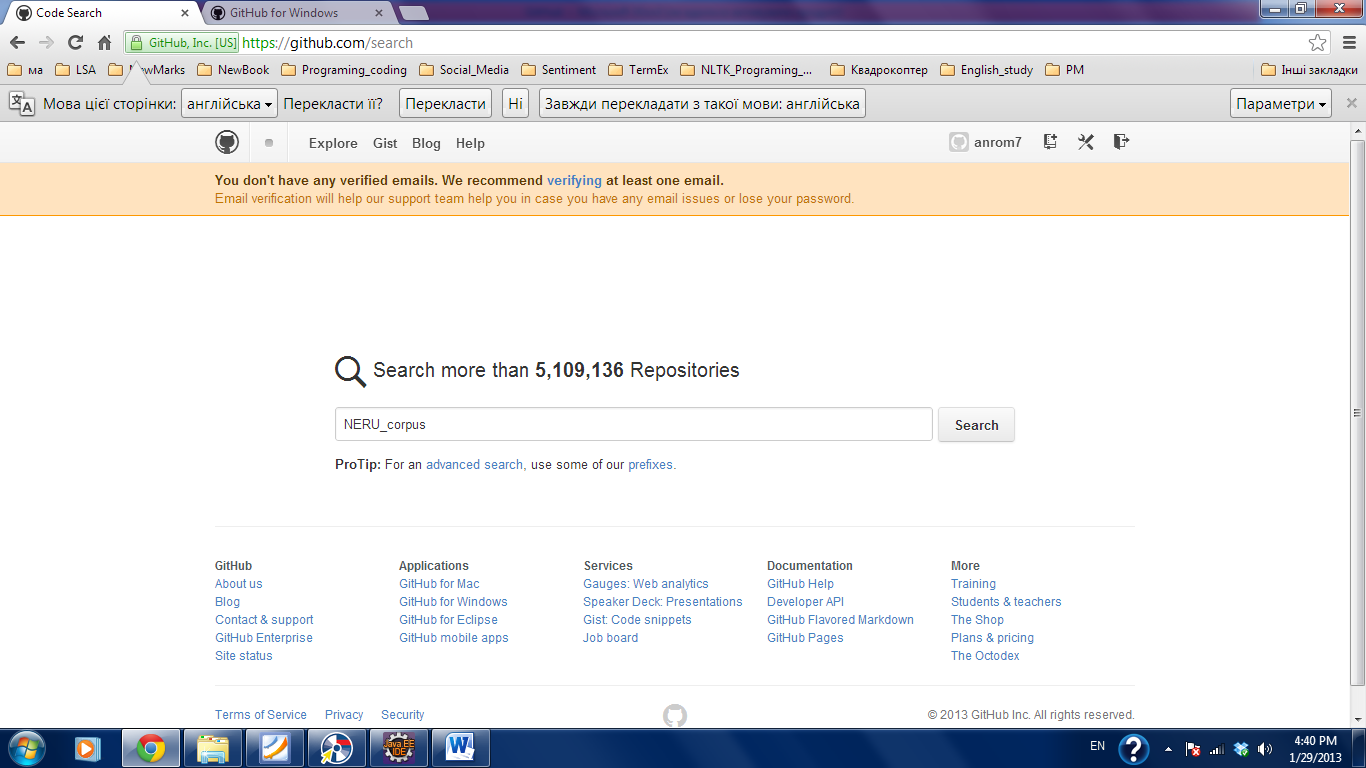
1. Заповнити налаштування облікового запису

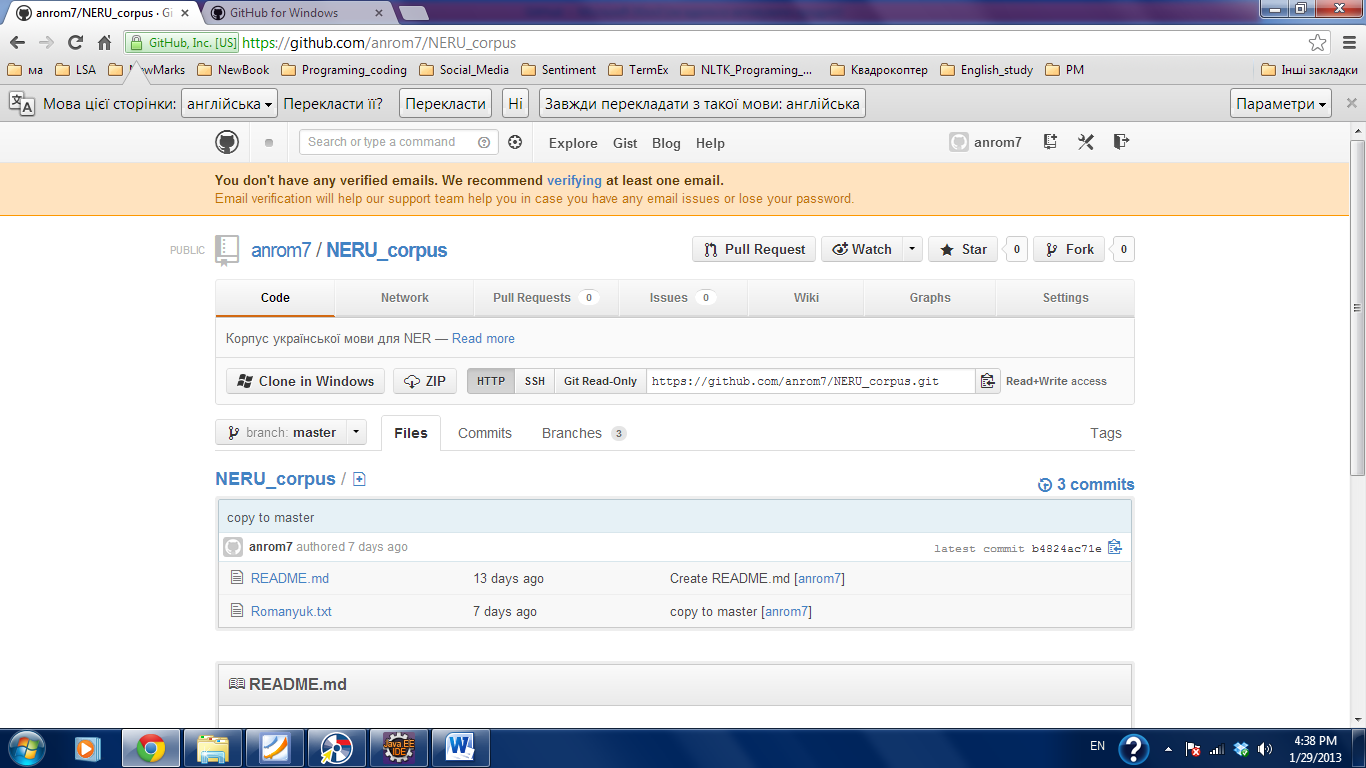


1. Дані облікового запису (ім’я користувача) надіслати керівнику ([anrom7@gmail.com](mailto:anrom7@gmail.com)) та зачекати поки вас додадуть у Collaborators

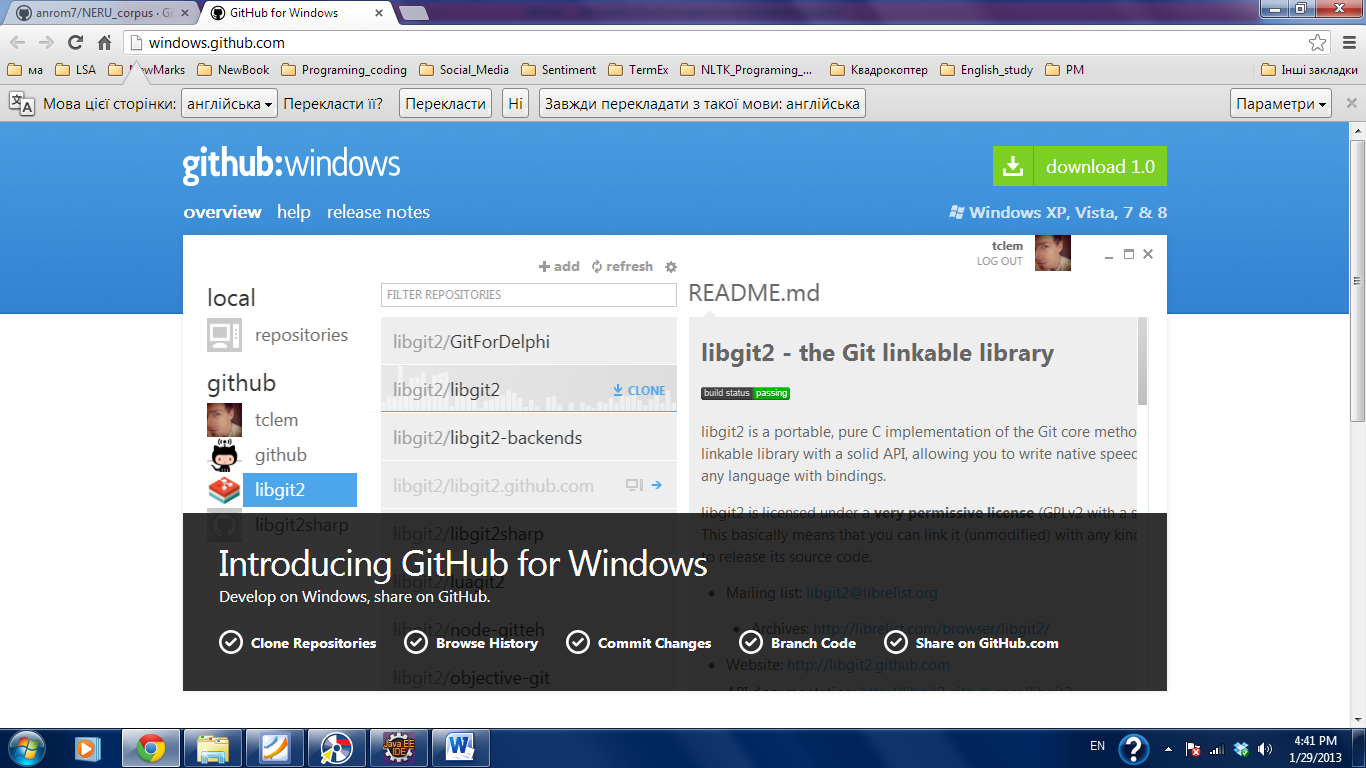


1. Знайти проект NERU\_corpus

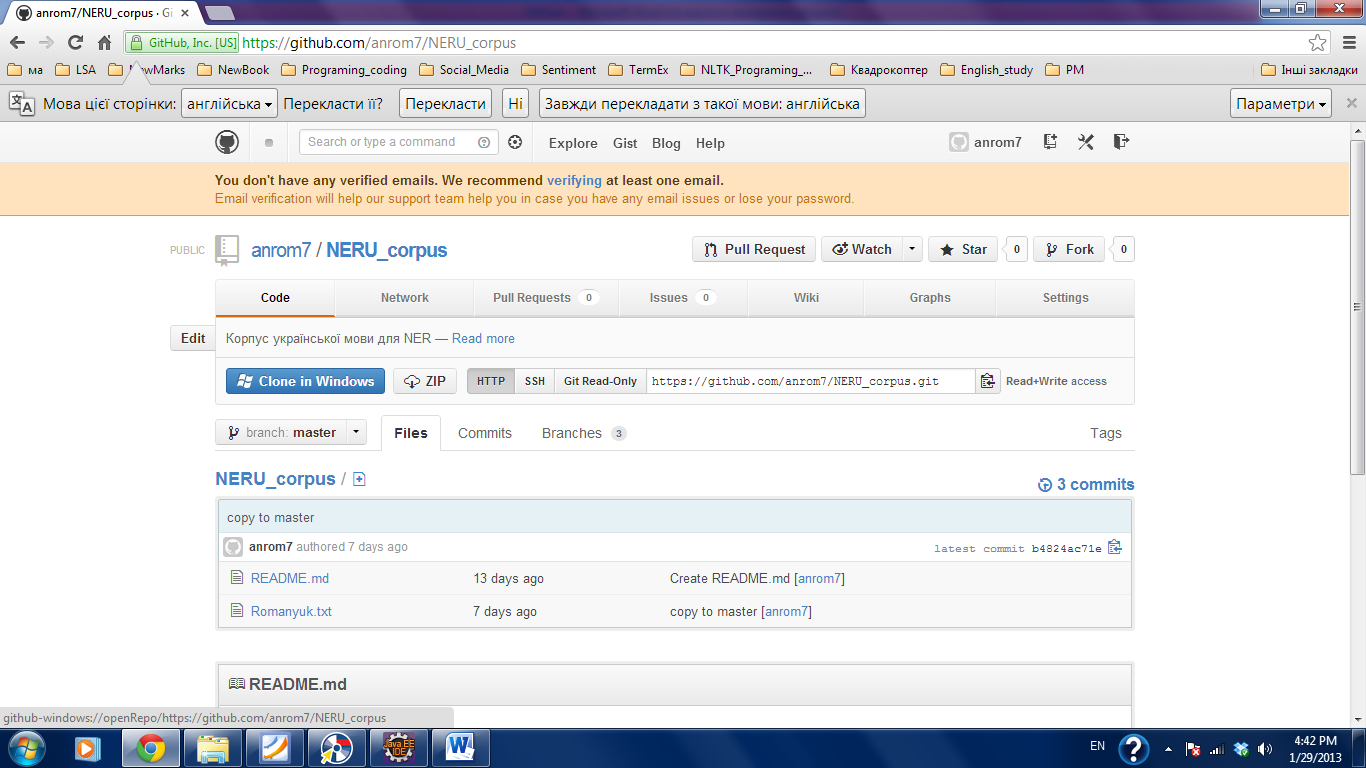




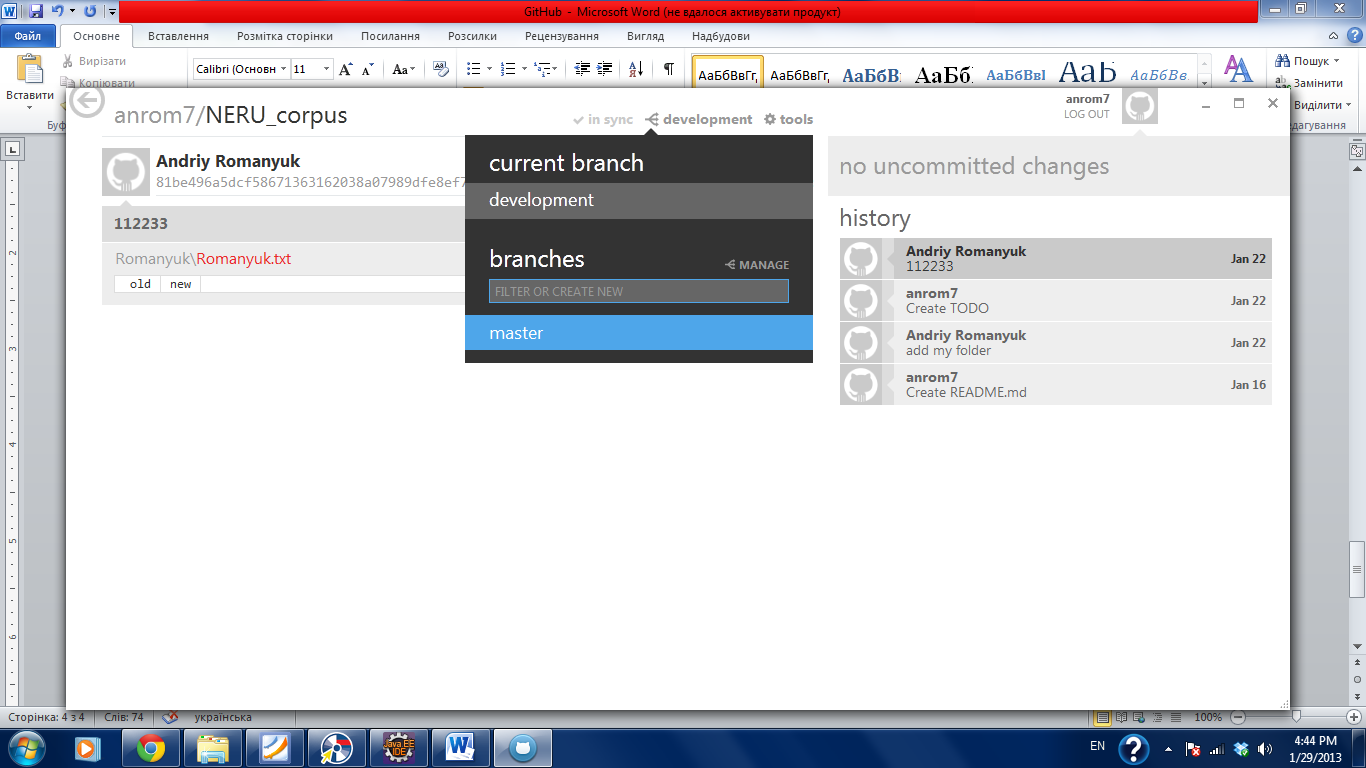
1. Завантажити та встановити [GitHub for Windows](http://windows.github.com/)



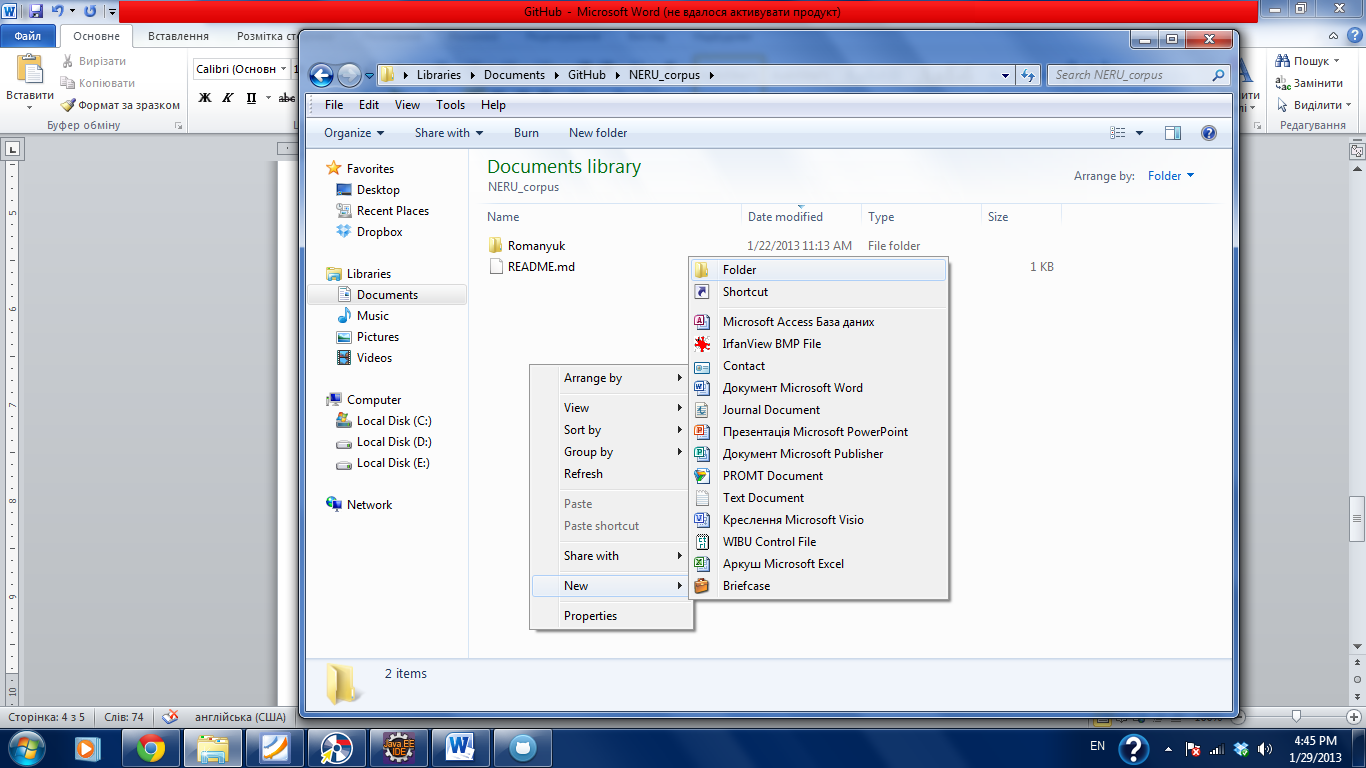
1. Клонувати проект



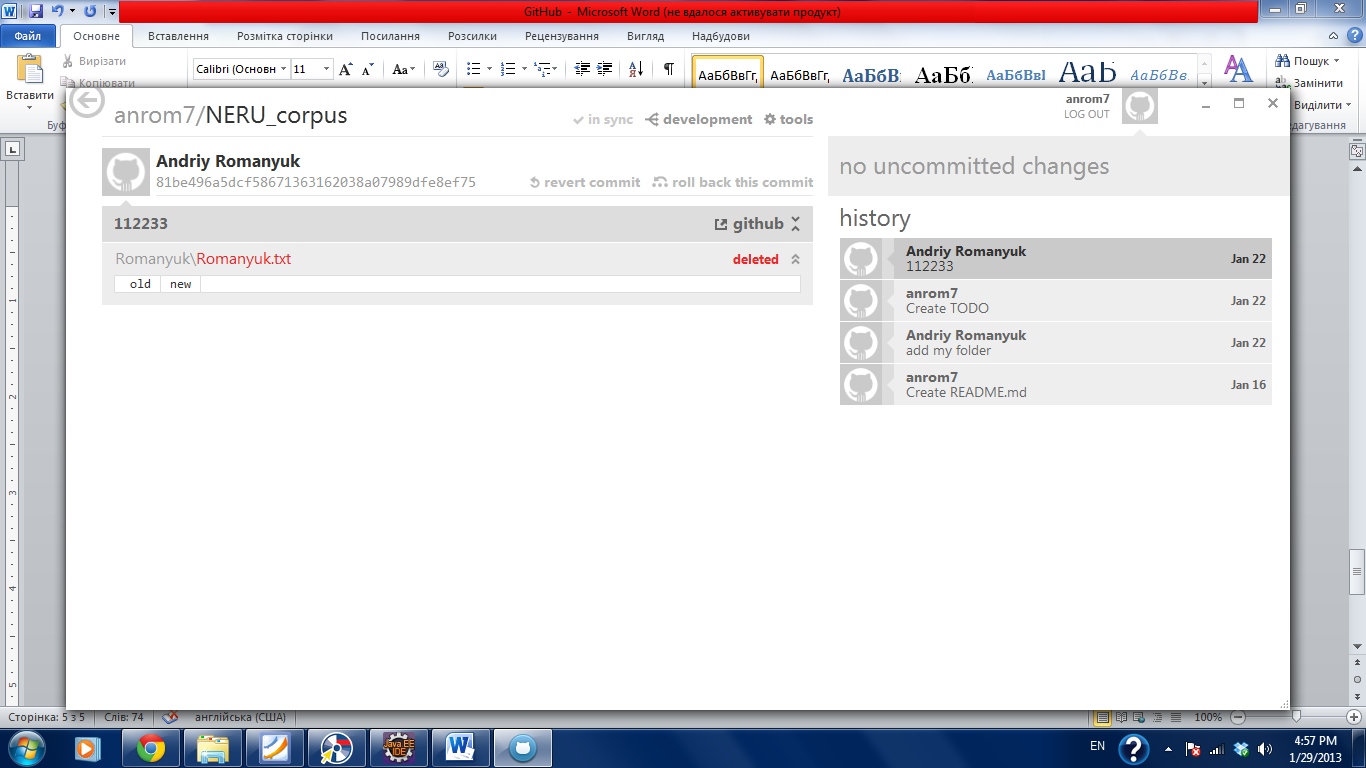
1. Створити нову гілку назвою яка відповідає прізвищу в [GitHub for Windows](http://windows.github.com/) **(ЛИШЕ У ВИПАДКУ, ЯКЩО ВИ НЕ СТВОРЮВАЛИ ГІЛКИ ДЛЯ ПЕРШОЇ ЧАСТИНИ РОЗРАХУНКОВОЇ РОБОТИ!!!)**



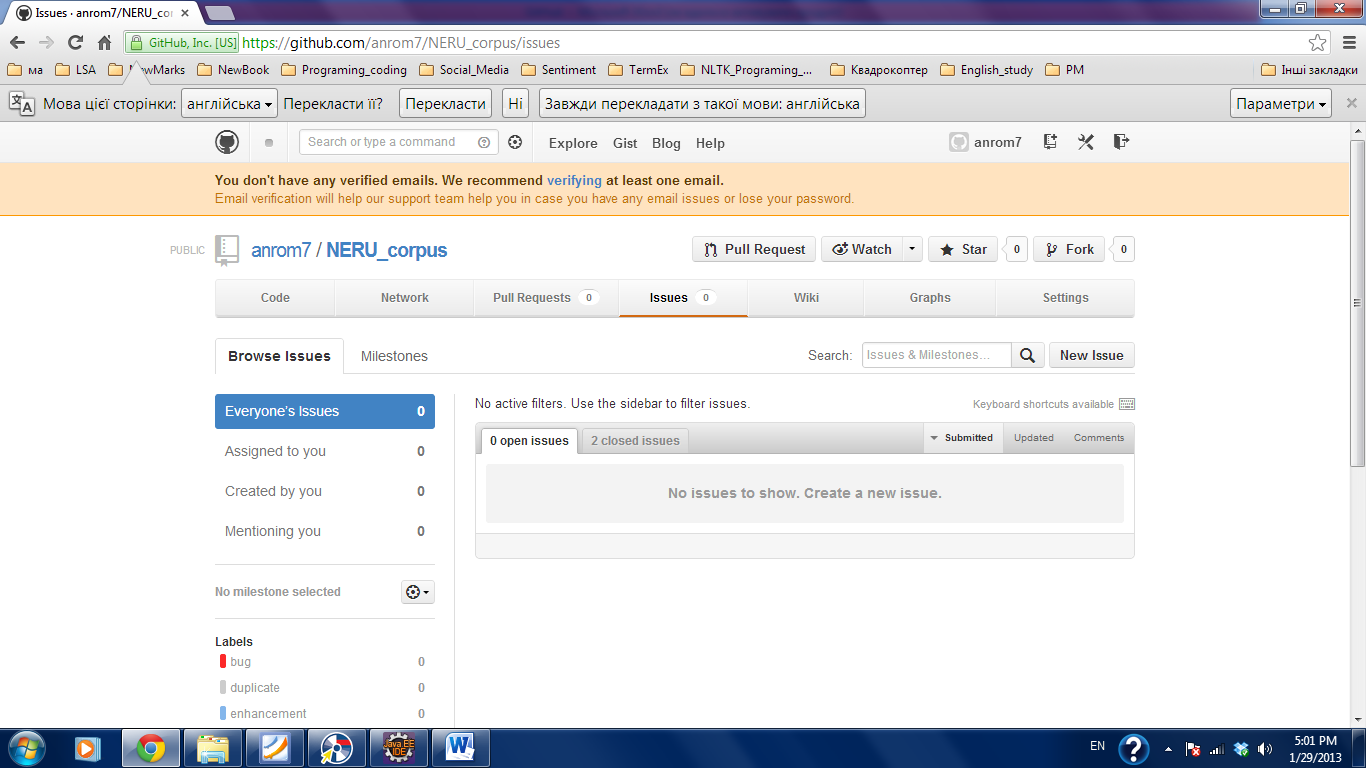
1. Створити папку з назвою ***Пізвище студента\_med (наприклад: Dmytrash\_med)***



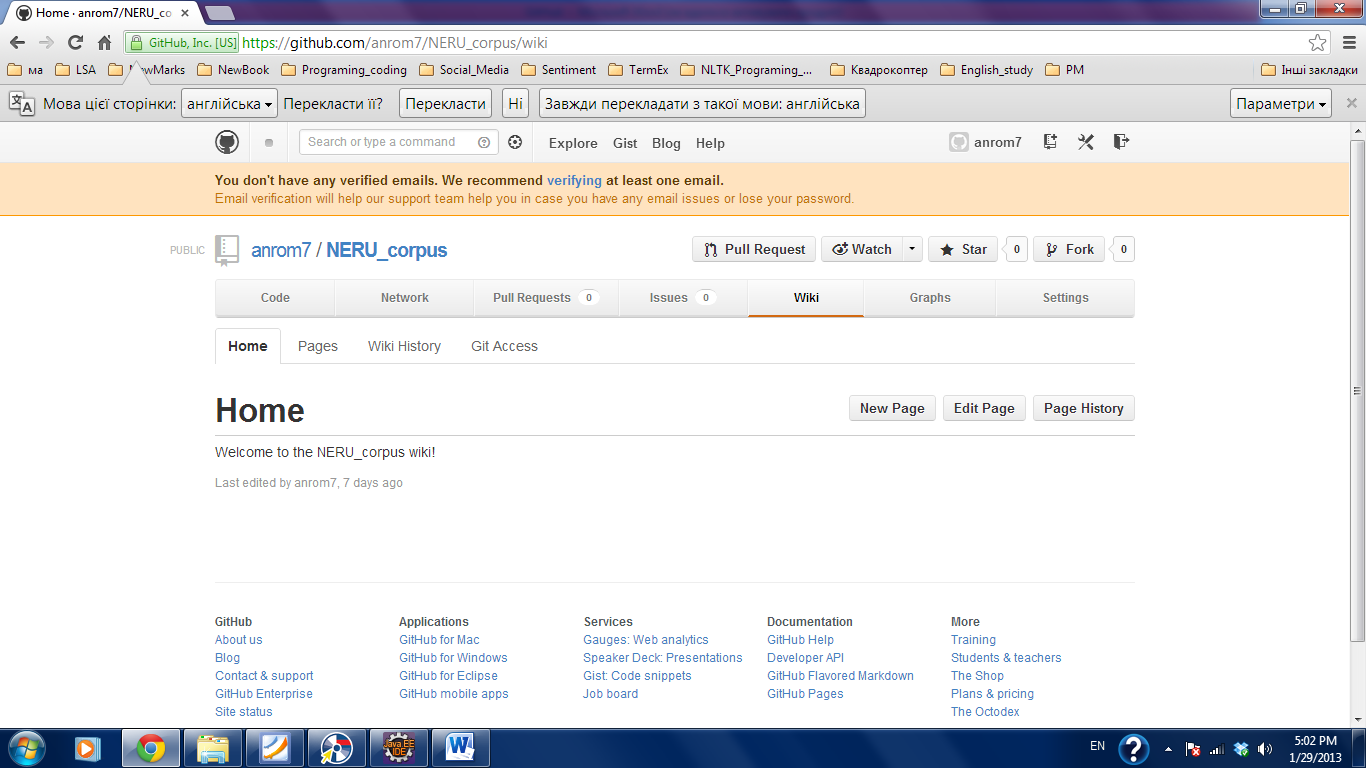
1. Надіслати зміни на сервер



1. Файли завдання будуть знаходитися у вашій гілці
2. При виконанні індивідуального завдання результати (файли) надсилати (синхронізовувати) на сервер
3. Питання додавати у Issues



1. Опис розрахунково графічної роботи переглядати у Wiki



|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Прізвище та ім'я** | **Файл Forum\_files.pdf | номери сторінок** | **Файл Article\_files.docx** |
| 1 | Баген | 1-2 | 6-10.pdf |
| 2 | Білоус | 3-4 | 6-11.pdf |
| 3 | Борис | 5-6 | 7-9.pdf |
| 4 | Булич | 7-8 | 7-10.pdf |
| 5 | Волков | 9-10 | 7-11.pdf |
| 6 | Єнілєєва |  |  |
| 7 | Іванів | 11-12 | 7-12.pdf |
| 8 | Карюк | 13-14 | 7-13.pdf |
| 9 | Кланцак | 15-16 | 7-14.pdf |
| 10 | Комар | 17-18 | 10-14.pdf |
| 11 | Матвіїїв | 19-20 | 11-15(2) .pdf |
| 12 | Обріза | 21-22 | 11-15.pdf |
| 13 | Петруха | 23-24 | 12-16.pdf |
| 14 | Плагова |  |  |
| 15 | Покладок | 25-26 | 12-18.pdf |
| 16 | Прокопяк |  |  |
| 17 | Прокулевич | 27-28 | 13-16.pdf |
| 18 | Рос | 29-30 | 14-16.pdf |
| 19 | Стрельченко | 31-32 | 15-18(1) .pdf |
| 20 | Терех |  |  |
| 21 | Тимечко | 33-34 | 15-18.pdf |
| 22 | Уколов | 35-36 | 16-21(2) .pdf |
| 23 | Фуярчук | 37-38 | 16-21.pdf |
| 24 | Щудло | 39-40 | 17-20.pdf |
| 25 | Антонюк | 41-42 | 17-21.pdf |
| 26 | Бучко |  |  |
| 27 | Везденко | 43-44 | 17-23.pdf |
| 28 | Генсіровська | 45-46 | 19-23(1) .pdf |
| 29 | Голуб | 47-48 | 51-53+2.pdf |
| 30 | Горін | 49-50 | 52-55+3.pdf |
| 31 | Дроздовська | 51-52 | 53-57+3.pdf |
| 32 | Климчак |  |  |
| 33 | Козачук | 53-54 | 54-57+3.pdf |
| 34 | Кузьо |  |  |
| 35 | Марунишин | 55-56 | 55-58+3.pdf |
| 36 | Марцинюк | 57-58 | 58-60+2.pdf |
| 37 | Мельник | 59-60 | 58-64+6.pdf |
| 38 | Михайлюк | 61-62 | 59-62+3.pdf |
| 39 | Морозова | 63-64 | 60-62+2.pdf |
| 40 | Мохонова | 65-66 | 61-64(1)+3.pdf |
| 41 | Оленюк | 67-68 | 61-65+4.pdf |
| 42 | Пегитяк |  |  |
| 43 | Сидоренко | 69-70 | 61-65+4.pdf |
| 44 | Соснова | 71-72 | 63-68+6.pdf |
| 45 | Фурсачик | 73-74 | 65-67+2.pdf |
| 46 | Чубко | 75-76 | 65-68+2.pdf |
| 47 | Яворська |  |  |
| 48 | Якимчук | 77-78 | 65-71+6.pdf |
| 49 | Андрощук | 79-80 | 66-70(2)+4.pdf |
| 50 | Гавриляк | 81-82 | 68-71+3.pdf |
| 51 | Закала | 83-84 | 69-71+2.pdf |
| 52 | Кобрин | 85-86 | 71-76+6.pdf |
| 53 | Кушнірук | 87-88 | 72-75+3.pdf |
| 54 | Пелех | 89-90 | 72-76+4.pdf |
| 55 | Ткачук | 91-92 | 74-78+3.pdf |
| 56 | Труняк | 93-94 | 76-80+4.pdf |
| 57 | Хімка | 95-96 | 77-79+2.pdf |
| 58 | Цимбаліста | 97-98 | 77-81+4.pdf |
| 59 | Брикайло | 99-100 | 79-86+6.pdf |
| 60 | Воляник | 101-102 | 80-83+3.pdf |
| 61 | Дирда | 103-104 | 81-84+3.pdf |
| 62 | Камінська | 105-106 | 94-99+4.pdf |
| 63 | Маковецька | 107-108 | 96-101+4.pdf |
| 64 | Мащакевич | 109-110 | 97-101+4.pdf |
| 65 | Федух | 111-112 | 102-109+6.pdf |
| 66 | Химич | 113-114 | 103-106+4.pdf |
| 67 | Шилімова | 115-116 | 107-113(1)+6.pdf |
| 68 | Штангрет | 117-118 | 114-116+2.pdf |

**ЗМІСТ ЗВІТУ**

1 Титульний аркуш.

2 Мета роботи.

3 Вихідні результати.

В електронному вигляді наступні файли зібрані в папку назва якої відповідає прізвищу студента(Stepanyuk):

Файли завдання в форматі \*.txt з відповідними назвами;

Файли з результатами роботи в форматі \*.xml.

4. Висновок.