План розробки застосунку з аналізу якості повітря

Прелюдія

- [x] Для проведення аналізу обираємо Твіттер. Попри нижче охоплення в Україні у порівнянні з ФБ та Інстаграм, Твіттер значно менше обмежує користувачів-скреперів.
- [x] Можна працювати через API або напряму знімати інформацію з веб-сторінки. Другий варіант безкоштовний, перший варіант дозволяє безкоштовно зняти 500 тис. твітів за місяць (також є безкоштовний тариф у 10 млн твітів на місяць для академічних досліджень). Крім того перший варіант надає можливість доступатись до ширшого переліку полів.
  - [x] З огляду на попередній пункт потрібно отримати обліковку на Порталі Розробника Твіттер. Це займе певний час через веріфікацію службою Твіттер наданої вами інформації. Можуть, як у моєму випадку поставити додаткові питання. Тобто потрібна чіткість формулювань і певна терплячість.
  - [ ] Портал Розробника Твіттер дозволяє командну роботу над проєктом шляхом підключення інших розробників до проєкту.

Зчитування твітів

- [x] Спочатку відлагодимо просту процедуру зчитування твітів та їнього збереження у БД (для початку CSV-файлу може вистачити).
  - [x] Запустимо тестовий приклад на базі статті Collecting tweets from Twitter API v2 using Python 3 | Towards Data Science
    - [x] Це результат тестового запуску Python читає Твіттер. Збираємо твіти з допомогою Twitter API... | by Olegh Bondarenko | Sep, 2021 | Medium
  - [ ] Є аналогічний, але простіший приклад Searching for Tweets with Python. | by Martin Šiklar | Towards Data Science

Вибір твітів

Вибір твітів здійснюється за трьома складовими: часовий інтервал, просторовий регіон, контекст.

- [x] **Час:** Відлагодити вибірку (selection) твітів для визначеного періоду (інтервалу) часу - поля `start_time` і `end_time` в параметрах запиту `query_params`

- [ ] **Простір:** Розібратись і відлагодити вибірку (selection) твітів за географічною ознакою.
  - [ ] Приналежність регіону, місту тощо - Filtering Tweets by location. Твіттер може надавати гео прив'язку за місцем реєстрації обліковки та/або за місцем публікації твіта для *geo-tagged* твітів. Є можливість шукати за координатами і радіусом навколо. Можна фільтрувати за містом (та навіть районом міста), але у Преміум підписці.
  - [ ] Безпосереднє отримання географічних координат твітів.
- [ ] **Контекст:** Розібратись і відлагодити вибірку (selection) твітів, що стосуються якості повітря, включно із запахами.
  - [ ] Розібратись, як працюють складні запити у Твіттері.
  - [ ] Розробити запит стосовно якості повітря для укр, рос та англ мов.
    - [ ] Чи визначає сам Твіттер мову твіту `lang:uk`? Це було б ідеальне рішення
    - [ ] Визначення мови з аналізу тексту твіту, як варіант
    - [ ] Врахування словоформ і відповідників трьома мовами: укр, рос та англ
    - [ ] Залучення синонімів

Аналіз контексту

- [ ] **Визначення класу факторів** з контексту (наприклад, запиленість, загазованість, конкретний запах, свіжість, вологість)
- [ ] **Sentiment Analysis:** Аналіз сприйняття контексту (фактора)
  - [ ] Розібратись з визначенням ступеню і знаку емоційного сприйняття (polarity)

Подальший розвиток

- [ ] Збереження результатів аналізу у БД
- [ ] Представлення результатів аналізу: графіки, карти
- [ ] Співставлення показників інструментальних систем моніторингу якості повітря з Твіттер-моніторингом. Тут партнером може виступити, зокрема, Airly.
- [ ] Бізнес-модель
- [ ] Інші комерційні застосування, у першу чергу, маркетингові дослідження