На Урок
освітній проект

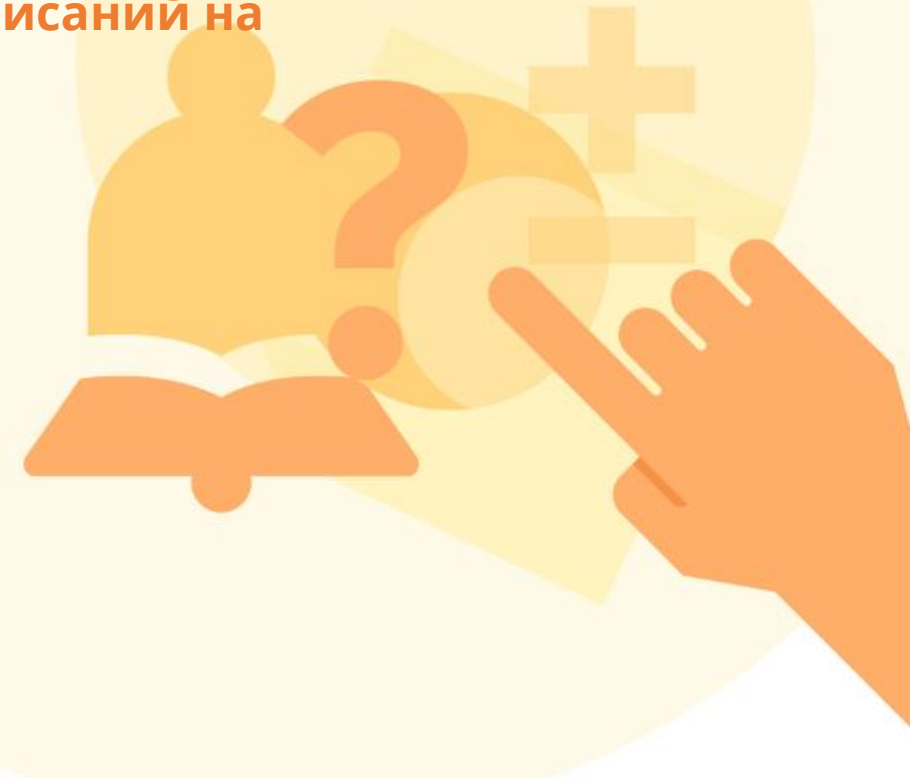
Комп’ютерна лінгвістика: чому гуманітаріям важливо вивчати програмування

На Урок
освітній проект

Якщо ви готові до трансляції, напишіть відповідь на запитання:

ПЕРЕВІРКА ЗВ’ЯЗКУ

Як ви вважаєте, чи може комп’ютерна програма аналізувати текст, який написаний на природній мові?



МІСЦЕ ДЛЯ ТРАНСЛЯЦІЇ

ПРО ЛЕКТОРА

СЕРГІЙ ПЕТРОВИЧ

- Учитель фізики та інформатики м. Вінниця.
- **Кандидат педагогічних наук зі спеціальності «Теорія і методика професійної освіти».**
- **Фіналіст премії Global Teacher Prize Ukraine-2019.**
- **Учитель Всеукраїнської школи онлайн.**

МІСЦЕ ДЛЯ ТРАНСЛЯЦІЇ

На Урок
освітній проект

ПРО НАС

- Освітній журнал
- Бібліотека розробок
- Онлайн-тести
- Інтернет-конференції
- Курси
- Вебінари
- Інтенсиви
- Лабораторні роботи
- Олімпіади та конкурси
- Проєкти

МІСЦЕ
ДЛЯ
ТРАНСЛЯЦІЇ

info@naurok.com.ua

https://www.facebook.com/naurok.com.ua

На Урок
освітній проект

ПЛАН ВЕБІНАРУ

1. Чим цікава комп’ютерна лінгвістика.
2. Великі дані, машинне навчання і природні мови.
3. Приклад обробки тексту художнього твору.

МІСЦЕ ДЛЯ ТРАНСЛЯЦІЇ

Чому комп’ютерна лінгвістика може бути цікава вашим учням

Комп'ютерний лінгвіст – це відносно нова професія, що знаходиться на стику математики й лінгвістики. Підійде для юнаків та дівчат, які віддають перевагу точним наукам і схильні до самонавчання. Пропоную познайомитися із цією хоч і новою, але дуже перспективною професією.

Фахівці з комп'ютерної лінгвістики розробляють алгоритми розпізнавання людської мови, систем машинного перекладу тощо. Професія підходить для тих учнів, хто зі шкільних предметів віддає перевагу іноземній та українській мові, літературі, математиці й інформатиці.

Комп'ютерна лінгвістика – це наука, яка має безліч напрямків і дозволяє вирішити величезну кількість важливих завдань. Фахівці цієї професії беруть участь у створенні алгоритмів і програм, що використовуються для отримання даних, розробки онлайн-словників, онлайн-перекладачів, аналізу текстів художніх творів тощо. Наприклад, алгоритми розпізнавання усного мовлення використовуються в системах розумних будинків, сучасних гаджетах. Такі технології полегшують життя звичайних користувачів і людей із обмеженими можливостями.

На Урок
освітній проект

МІСЦЕ
ДЛЯ
ТРАНСЛЯЦІЇ

Особливості професії комп’ютерного лінгвіста

Як прикладна наука комп'ютерна лінгвістика зародилася в США в другій половині XX століття. Сьогодні ця сфера активно розвивається, адже величезна кількість користувачів із різних точок світу використовують інтернет, комп'ютери, аксесуари для пошуку й обробки інформації, аналітики, навчання – вирішення будь-яких завдань. Комп'ютерні лінгвісти виконують великий обсяг робіт, спрямованих на створення:

- QA-систем (система запитань-відповідей);
- алгоритмів машинного перекладу;
- генераторів тексту;
- електронних словників і баз даних;
- систем вилучення та пошуку інформації, розпізнавання мови та інших продуктів, алгоритмів.

Діяльність комп'ютерних лінгвістів має важливе соціальне значення, її результати застосовуються в сфері машинного навчання, Big Data. Фахівці вільно працюють із SQL, технологіями обробки мови, різноманітними бібліотеками, програмуванням, створюють персональних помічників (Siri).

Grammarly – українська онлайн-платформа на основі штучного інтелекту для допомоги у спілкуванні англійською мовою, запущена у 2009 році. Grammarly підвищує якість письмового спілкування, пропонуючи рекомендації щодо правильності (граматика та механіки письма), чіткості (стислість та зрозумілість), захопливості (словниковий запас та розмаїття) та тону повідомлення (формальність, ввічливість і впевненість). Має представництва у Києві, Сан-Франциско, Нью-Йорку та Ванкувері.
Продукт Grammarly доступний для декількох інтерфейсів і пристроїв: як вебредактор, настільні додатки для Windows і Mac, браузерне розширення (для Google Chrome, Safari, Mozilla Firefox, Microsoft Edge), додаток для iPad, мобільні клавіатури (iOS, Android) та надбудова для Microsoft Office.
https://www.grammarly.com/

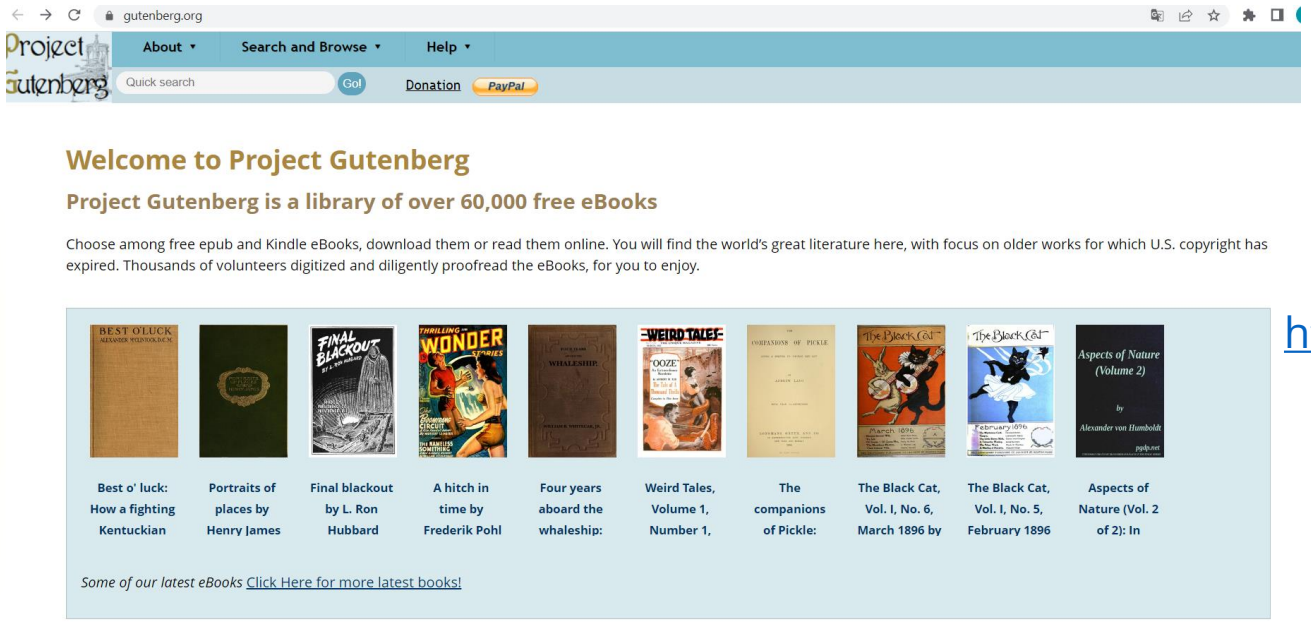
Звідки беруться комп’ютерні лінгвісти?

- Лінгвісти, що навчилися програмувати
- Програмісти, що навчилися лінгвістичної теорії
- Власне комп’ютерні лінгвісти за освітою (рідкісний вид)

На Урок
освітній проект

МІСЦЕ
ДЛЯ
ТРАНСЛЯЦІЇ

https://eadh.org/



https://www.gutenberg.org/

Об’єм даних, які ми створюємо, продовжує дедалі зростати; наприкінці 2021 року об’єм доступної цифрової інформації збільшиться з 5 зетабайт сьогодні до 50 зетабайт у 2021 (1 зетабайт = 1 мільйон мільйонів терабайт). Уявіть, що з одного жорсткого диску на 1 зетабайт можна дивитись кіно у форматі 4K впродовж 63 мільйонів років. Це багато, надзвичайно багато даних =)

На сьогоднішній день майже кожна з наших дій в онлайні залишає цифровий слід. Ми генеруємо дані щоразу, як підключаємось до мережі, використовуємо смартфони з GPS, спілкуємось із друзями через соціальні мережі та месенджери, або коли робимо покупки. Таким чином, можна впевнено сказати, що ми залишаємо цифрові сліди разом із усім, що включає в себе цифрову дію. Окрім цього, кількість машиногенерованих даних також стрімко зростає. Дані генеруються та обмінюються, коли наші «розумні» домашні пристрої спілкуються один із одним, або з їх домашніми серверами. Промислові машини на заводах і фабриках по всьому світі дедалі більше оснащуються датчиками, які також збирають і передають дані.

Термін **Big Data** означає збір усіх цих даних і нашу здатність використовувати їх на користь людства у широкому спектрі галузей, у тому числі у бізнесі.

ЗАСТОСУВАННЯ BIG DATA

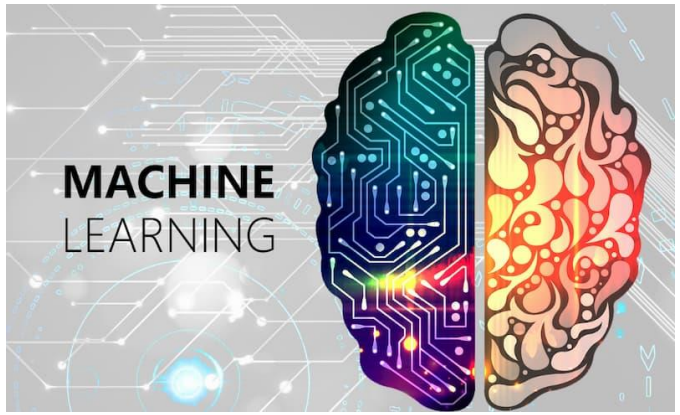
Охорона здоров'я. Зібрана інформація про перебіг хвороб, лікування, ліки, походження захворювання допомагає врятувати людей, яких ще пів століття тому вважали невиліковними.

Правоохоронні органи. Big data можна використати для прогнозування сплеску криміналу. Наступний крок – попередження та стримувальні заходи.

Попередження природних та техногенних катастроф. Прогноз може врятувати багато життів. Збирають дані, обробляють показники датчиків, а вже на їх основі визначають дату та місце катаклізму.

Запобігання шахрайству. Вже є успішні кейси про те, як великі дані допомогли простежити та попередити шахрайські операції в банках.

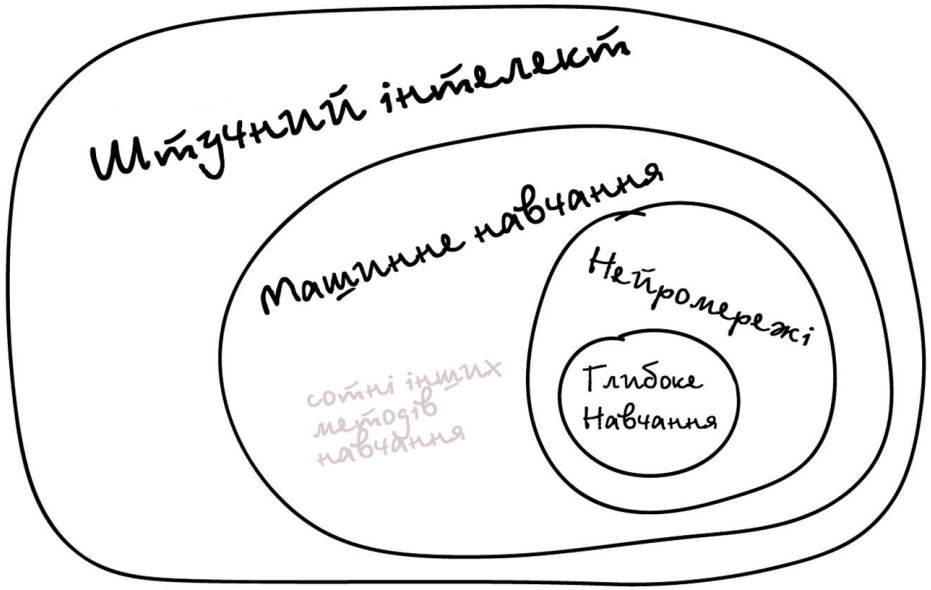
Бізнес. Великі дані використовують для створення проєктів: вивчають і аналізують вимоги та відгуки клієнтів, залучають та утримують цільову аудиторію, прогнозують популярність продуктів тощо.



МІСЦЕ
ДЛЯ
ТРАНСЛЯЦІЇ

Машинне навчання (англ. machine learning) – це підгалузь штучного інтелекту в галузі інформатики, яка часто застосовує статистичні прийоми для надання комп'ютерам здатності «навчатися» (тобто, поступово покращувати продуктивність у певній задачі) з даних, без того, щоби бути програмованими явно.

Назву «машинне навчання» (англ. machine learning) було започатковано у 1959 році Артуром Семюелем. Еволюціонувавши з досліджень розпізнавання образів і теорії обчислювального навчання в галузі штучного інтелекту, машинне навчання досліджує вивчення та побудову алгоритмів, які можуть навчатися й робити передбачення з даних, – такі алгоритми долають слідування строго статичним програмним інструкціям, здійснюючи керованими даними прогнози або ухвалювання рішень: шляхом побудови моделі з вибіркових входів.
Машинне навчання застосовують в ряді обчислювальних задач, у яких розробка та програмування явних алгоритмів з доброю продуктивністю є складною або нездійсненною; до прикладів застосувань належать фільтрування електронної пошти.



Штучний інтелект – назва всієї області, як біологія або хімія.

Машинне навчання – це розділ штучного інтелекту. Важливий, але не єдиний.

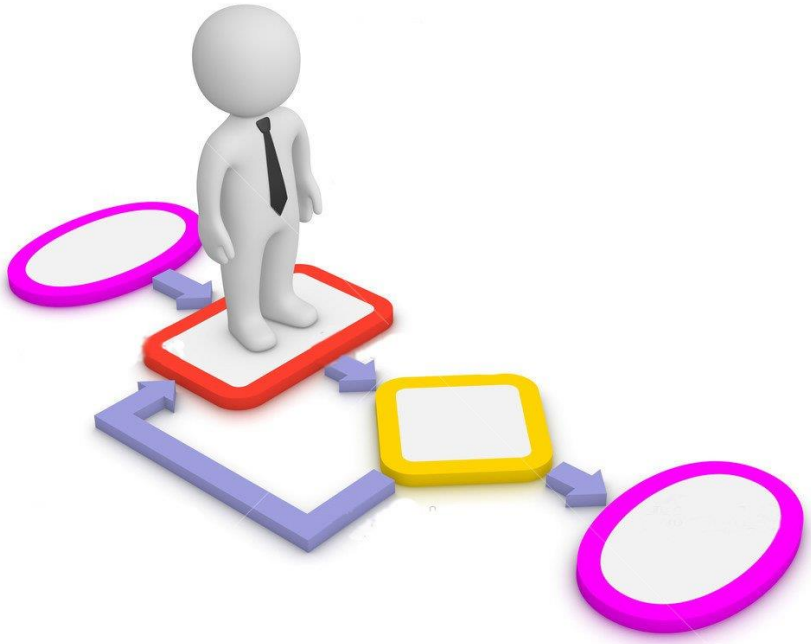
Нейромережі – один із типів машинного навчання. Популярний, але є й інші, не гірші.

Глибоке навчання – архітектура нейромереж, один із підходів до їх побудови та навчання. На практиці мало хто відрізняє, де глибокі нейромережі, а де не дуже. Кажуть назву конкретної мережі і все.

На Урок
освітній проект

МІСЦЕ
ДЛЯ
ТРАНСЛЯЦІЇ

Сьогодні в машинному навчанні є чотири основні напрями



Перші алгоритми прийшли до нас ще в 1950-х роках із чистої статистики. Вони розв'язували формальні задачі – шукали закономірності в числах, оцінювали близькість точок у просторі та вираховували напрямки.

Сьогодні на класичних алгоритмах тримається добра половина інтернету. Коли ви зустрічаєте в браузері інформацію «Рекомендовані покупки» на сайті.

Класичне навчання люблять ділити на дві категорії – із учителем і без. Часто можна зустріти їх англійські назви – **Supervised і Unsupervised Learning.**

У першому випадку у машини є якийсь учитель, який говорить їй як правильно. Розповідає, що на цій картинці кішка, а на цій собака. Тобто вчитель вже заздалегідь розділив усі дані на кішок і собак, а машина навчається на конкретних прикладах.

У навчанні без учителя, машині просто надають купу фотографій тварин на стіл і кажуть «розберися, хто тут на кого схожий». Дані не розмічені, у машини немає вчителя, і вона намагається сама знайти будь-які закономірності.

Очевидно, що з учителем машина навчиться швидше і точніше, тому в бойових задачах його використовують набагато частіше. Ці задачі діляться на два типи: класифікація – передбачення категорії об'єкта, і регресія – передбачення місця на числовій прямій.

Класифікація

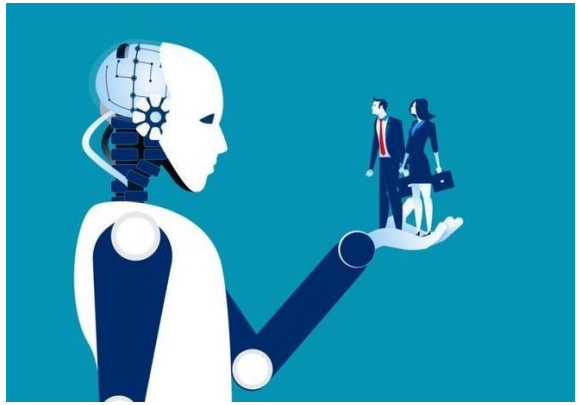
«Поділяє об'єкти за заздалегідь відомою ознакою. Шкарпетки за кольорами, документи за мовами, музику за жанрами»

Сьогодні використовують для:

- Спам-фільтрів
- Визначення мови
- Пошуку схожих документів
- Аналізу тональності
- Розпізнавання рукописних букв і цифр
- Визначення підозрілих транзакцій

Класифікація речей – найпопулярніша задача у всьому машинному навчанні. Машина в ній як дитина, яка навчається розкладати іграшки: роботів в один ящик, танки в інший.

Для класифікації завжди потрібен учитель – розмічені дані з ознаками і категоріями, які машина буде вчитися визначати за цими ознаками. Далі класифікувати можна що завгодно: користувачів за інтересами – так роблять алгоритмічні стрічки, статті за мовами та тематиками – важливо для пошукових систем, музику за жанрами – згадайте плейлисти, навіть листи у вашій поштовій скриньці.



Обробка природної мови (Natural Language Processing – NLP) – це міждисциплінарна галузь, яка стоїть на перетині комп'ютерних наук, штучного інтелекту та обчислювальної лінгвістики, основним проблемним полем якої є забезпечення взаємодії між комп'ютерами та людськими (природними) мовами.

Інтелектуальний аналіз тексту (Text mining) – напрям інтелектуального аналізу даних і штучного інтелекту, метою якого є отримання високоякісної інформації з колекцій текстових документів за допомогою застосування методів машинного навчання та обробки природної мови. Основна задача Text mining полягає в тому, щоб виявити інформацію, яка, можливо, невідома та прихована в контексті іншої інформації. Це досягається за допомогою різних методологій аналізу; обробка природної мови – одна з них, вона виконує лінгвістичний аналіз, що допомагає машині «читати» текст.

За допомогою NLP можна розробляти програмне забезпечення, яке може витягувати інформацію з вихідного тексту, знаходити зв'язок між словами, робити синтаксичний розбір, визначати емоційне ставлення автора, досліджувати тексти на схожість, а також розпізнавати людську мову.

Що таке аналіз тексту?

Аналіз тексту – це процес вивчення неструктурованих даних, які представлені у формі тексту. Її завдання – отримати уявлення про патерни та теми, що цікавлять.

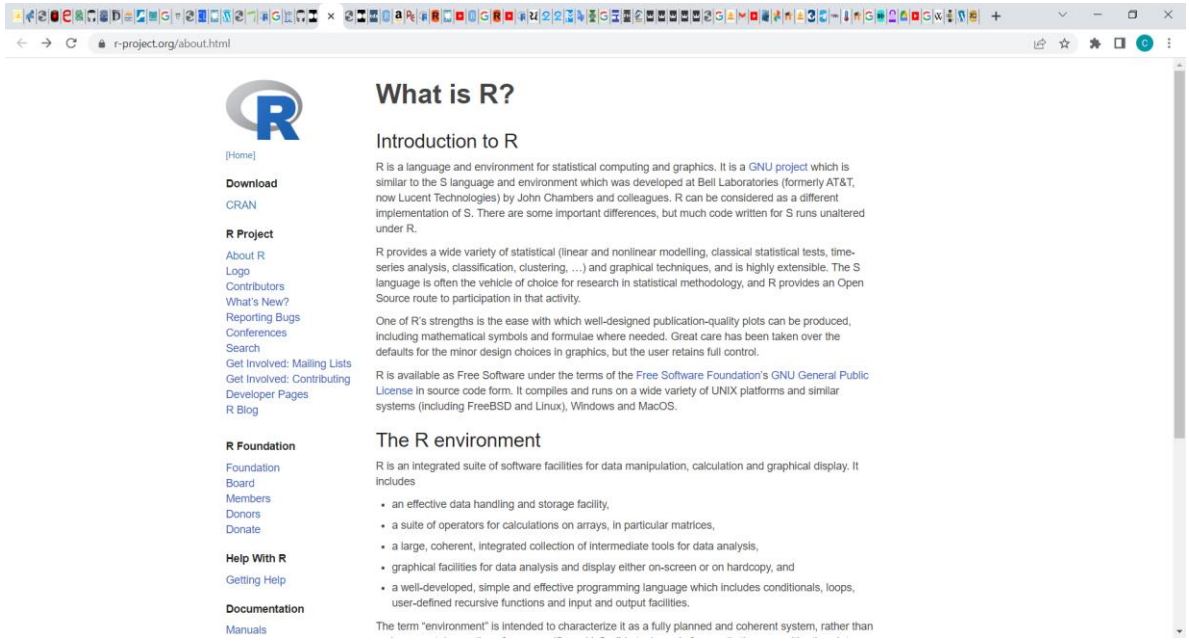
Чому аналіз тексту важливий?

Є багато причин щодо текстового аналізу. Найголовніша з них – розібратися в настроях та емоціях, які використовуються у додатках та сервісах, які ми відвідуємо щодня. Завдяки текстовій аналітиці ми можемо отримувати важливу інформацію з твітів, електронних листів, текстових повідомлень, реклами, карт тощо.

На Урок
освітній проект

Будемо використовувати базові можливості мови R, які знадобляться для текстової аналітики.

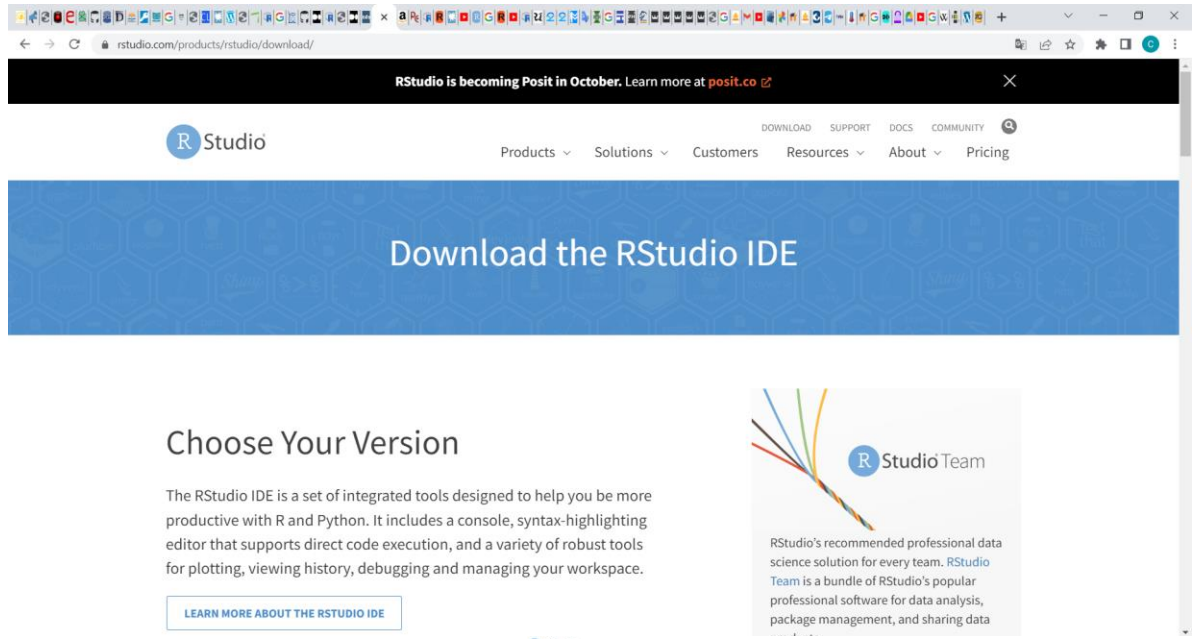
R – це мова та середовище для статистичних обчислень і графіки.
R надає широкий спектр статистичних (лінійне та нелінійне моделювання, класичні статистичні тести, аналіз часових рядів, класифікація, кластеризація, ...) і графічних методів.

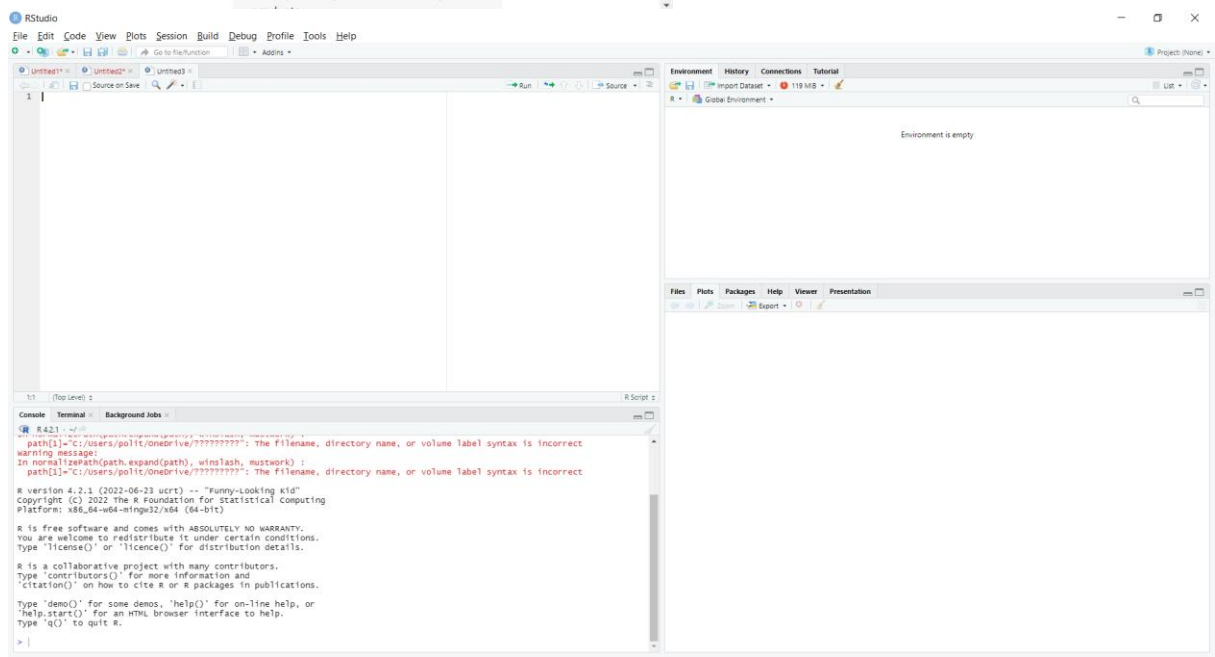


https://www.r-project.org/about.html

На Урок
освітній проект

МІСЦЕ
ДЛЯ
ТРАНСЛЯЦІЇ



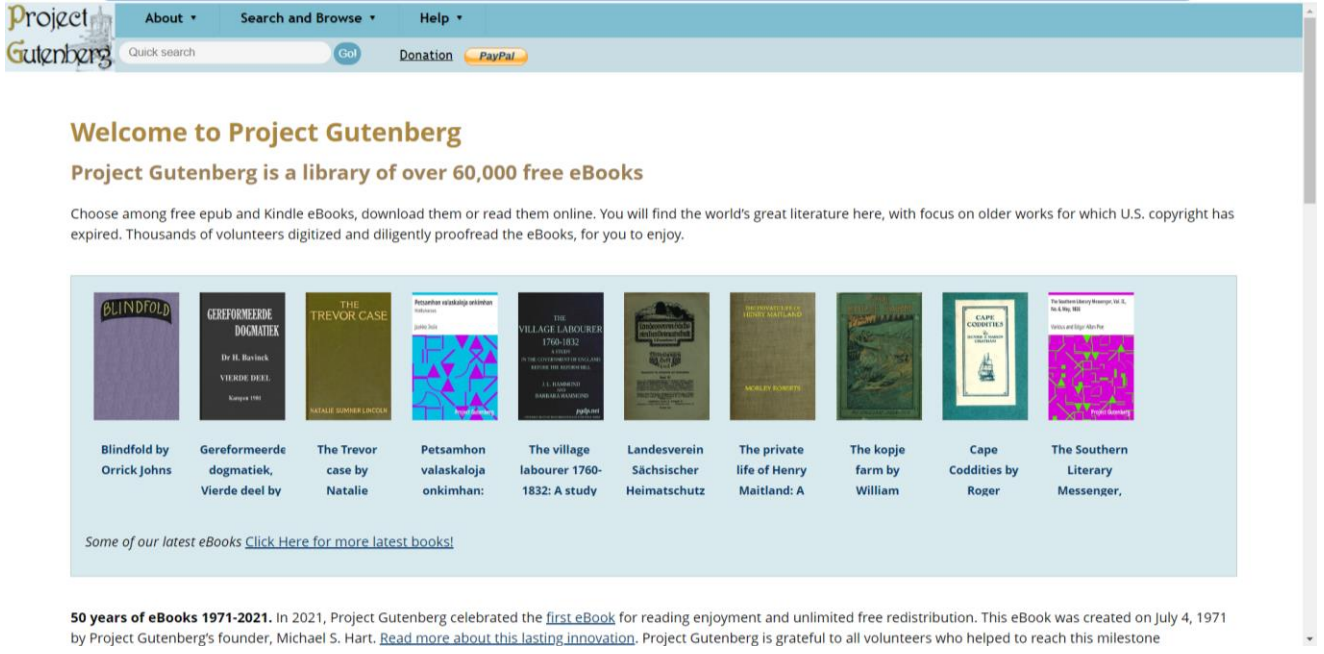


Приклад аналізу тексту художнього твору Марка Твена «The Adventures of Tom Sawyer». Завдання: знайти найпопулярніші слова та їх кількість у тексті твору.

Спершу встановимо пакет gutenbergr, щоб отримати доступ до бібліотеки загальнодоступних книг і публікацій.

Пакет gutenbergr допоможе завантажити та обробити відкриті роботи з колекції Project Gutenberg. У нього входять також інструменти для завантаження книг (і аналізу заголовка/інформації в нижньому колонтитулі), а також повний датасет метаданих Project Gutenberg, який можна використовувати для пошуку слів.

https://www.gutenberg.org/



Для встановлення та завантаження бібліотеки в R Studio потрібно ввести команду:

```
install.packages("gutenbergr")library(gutenbergr)
```

Книги Марка Твена
Зараз ми збираємось витягти кілька книг авторства Марка Твена з бібліотеки gutenbergr.

У бібліотеці Gutenbergr кожна книга маркована ідентифікаційним номером (ID). Нам він знадобиться для встановлення їхнього місцезнаходження.

Пригоди Геккельберрі Фінна – Gutenbergr ID: 76
Пригоди Тома Сойєра – gutenbergr ID: 74
Простаки за кордоном – Gutenbergr ID: 3176
Життя на Міссісіпі – Gutenbergr ID: 245

```
mark_twain <- gutenberg_download(c( 76 , 74 , 3176 , 245 ))
```

За допомогою функції gutenberg_download беремо книги та зберігаємо їх у датафреймі mark_twain.

Скріншот датафрейму mark_twain

	gutenberg_id	text
1	74	THE ADVENTURES OF TOM SAWYER
2	74	
3	74	By Mark Twain
4	74	
5	74	(Samuel Langhorne Clemens)
6	74	
7	74	
8	74	
9	74	
10	74	CONTENTS
11	74	
12	74	CHAPTER I. Y-o-u-u Tom-Aunt Polly Decides Upon her Duty...
13	74	Music--The Challenge--A Private Entrance
14	74	
15	74	CHAPTER II. Strong Temptations--Strategic Movements--Th...
16	74	Beguiled
17	74	
18	74	CHAPTER III. Tom as a General--Triumph and Reward--Dismal
19	74	Felicity--Commission and Omission

На Урок
освітній проект

```
# A tibble: 1,149 x 2
   word        lexicon
   <chr>       <chr>
 1 a           SMART
 2 a's         SMART
 3 able        SMART
 4 about       SMART
 5 above       SMART
 6 according   SMART
 7 accordingly SMART
 8 across      SMART
 9 actually    SMART
10 after       SMART
# … with 1,139 more rows
```

МІСЦЕ ДЛЯ ТРАНСЛЯЦІЇ

Ідентифікація стоп-слів

Коли ви проаналізуєте будь-який текст, побачите, що завжди будуть траплятися надмірні слова, які можуть змінити результат залежно від того, які шаблони чи тенденції ви намагаєтеся виявити. Їх називають стоп-словами. Вам самим вирішувати, чи ви хочете видалити їх, а в моєму прикладі ми безумовно їх приберемо.
Для початку нам потрібно завантажити бібліотеку *tidytext:*

```
library(tidytext)
```

Далі ми переглянемо stop_words у всій базі даних R (не в книгах Марка Твена).

```
data(stop_words)
```

На Урок
освітній проект

МІСЦЕ
ДЛЯ
ТРАНСЛЯЦІЇ

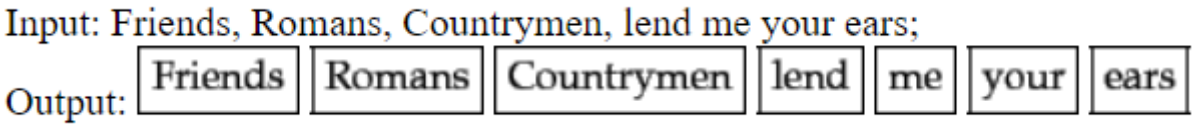
Токенізація та видалення стоп-слів

Для видалення стоп-слів і токенізації нашого тексту використовуємо метод «конвеєра» з бібліотеки dplyr.

Токенізація – це метод розділення тексту на невеликі частини (слова).

Іншими словами, токенізація «розрізає» речення на окремі слова. У процесі аналізу тексту це дає програмі структуру даних, яка необхідна для роботи.

Приклад токенізації



Ми поєднаємо в один ланцюжок кілька кроків, щоб видалити стоп-слова, поки відбуватиметься токенізація датафрейму *mark_twain.*

```
library(dplyr)
tidy_mark_twain <- mark_twain %>%
  unnest_tokens(word, text) %>% # tokenize
   anti_join(stop_words) # remove stop words
print(tidy_mark_twain)
```

Кроки:

Команда unnest tokens за вхідними даними потрібна, щоб визначити, що саме ми хочемо піддати токенізації і яким чином отримати доступ до тексту.
За допомогою anti_join ми значною мірою виключаємо всі слова, які знаходяться у базі даних stop_words.
Зберігаємо всю нашу роботу до нової змінної tidy_mark_twain.

Результат роботи на наступному слайді:

```
# A tibble: 182,716 x 2
   gutenberg_id word
          <int> <chr>
 1           74 adventures
 2           74 tom
 3           74 sawyer
 4           74 mark
 5           74 twain
 6           74 samuel
 7           74 langhorne
 8           74 clemens
 9           74 contents
10           74 chapter
# … with 182,706 more rows
```

Зверніть увагу, що відколи кожний об'єкт зазнав токенізації, у нас стало 182,706 рядків! Це тому, що кожне слово тепер перебуває в окремому рядку.

Частотний розподіл слів

Наше завдання – знайти патерни в даних. Частотний розподіл слів – це чудовий спосіб побачити, які слова використовуються найчастіше.

```
tidy_mark_twain %>%
  count(word, sort=TRUE)
```

```
# A tibble: 22,635 x 2
   word        n
   <chr>   <int>
 1 time     1226
 2 tom       966
 3 river     681
 4 day       667
 5 hundred   565
 6 night     562
 7 people    531
 8 water     508
 9 head      468
10 chapter   419
# … with 22,625 more rows
```

Якщо ви читали цю книгу Марка Твена, то не здивуєтесь, що Том це друге за популярністю слово. А ще цікавіше, що найпопулярніше слово «час» зустрічається в тексті 1226 разів!

На Урок
освітній проект

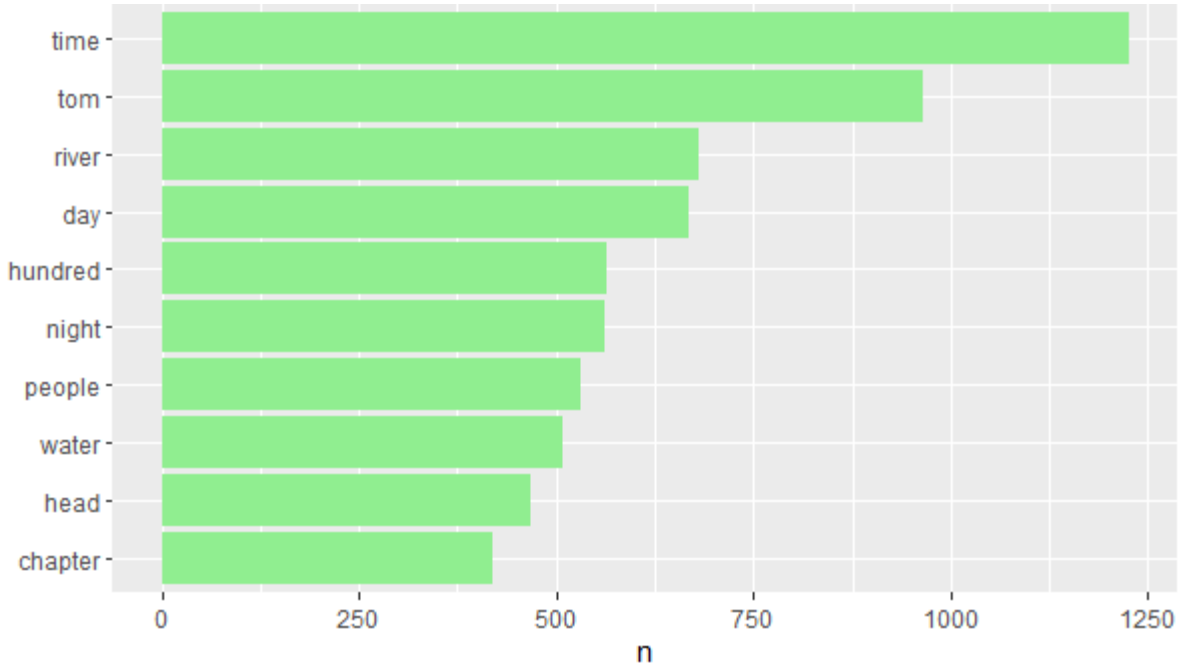
Візуалізація даних

За допомогою бібліотеки ggplot2 ми можемо додати певний візуальний контекст, щоб побачити, які слова частіше вживаються в тексті.

```
library(ggplot2)
freq_hist <-tidy_mark_twain %>%
  count(word, sort=TRUE) %>%
  filter(n > 400) %>%
  mutate(word = reorder(word,n)) %>%
  ggplot(aes(word, n)) +
  geom_col(fill='lightgreen')+
  xlab(NULL)+
  coord_flip()
print(freq_hist)
```

Деякі ключові кроки, які потрібно пройти, щоб отримати точний графік:

filter() потрібен, щоб переконатися, що ми не виведемо кількість кожного слова, яке зустрічається в тексті. Це було б надто багато. Так ми встановлюємо межу слів, які зустрічаються частіше ніж 400 разів.
mutate() потрібен, щоб організувати уявлення слів у кращому вигляді.
coord_flip() потрібен, щоб розгорнути граф і зробити його привабливішим візуально.



На Урок
освітній проект

#МАЮ_СИЛУ_НАВЧАТИ

Команда «На Урок»
продовжує працювати для вчителів України

На Урок
освітній проект

ДЯКУЮ ЗА УВАГУ!

ЧЕКАЮ НА ВАШІ ЗАПИТАННЯ В ЧАТІ

МІСЦЕ
ДЛЯ
ТРАНСЛЯЦІЇ