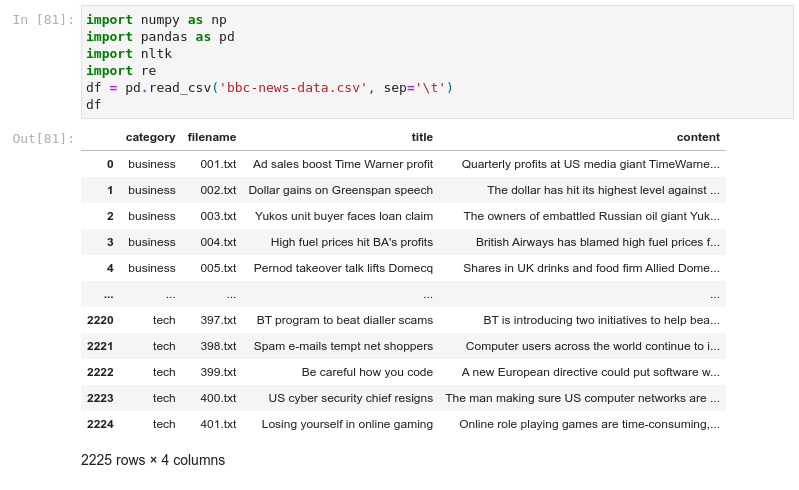
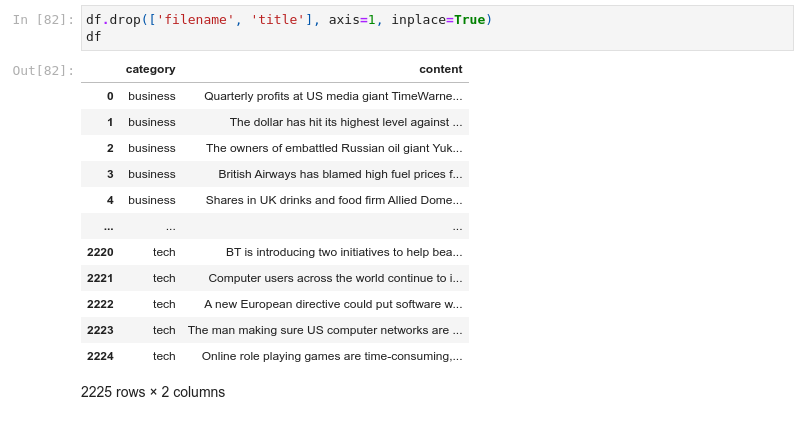
# Виконання

## Створення моделі TD-IDF.

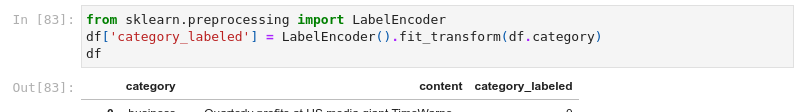
Для початку імпортуємо модулі та зчитаємо файл.

  
  
Рисунок 3.1 - Зчитування файлу

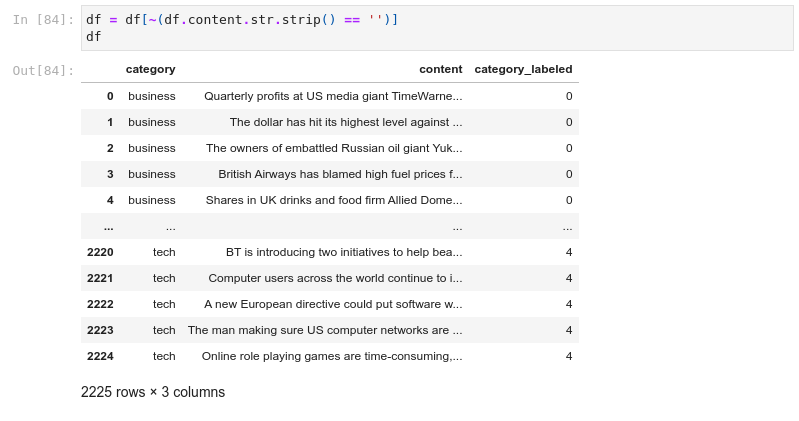
Видалимо колонки 'filename' та 'title', щоб класифікувати модуль лише за категоріями.

  
  
Рисунок 3.2 - Видалення колонок 'filename' та 'title'

Перетворимо категоріальні змінні у числові за допомогою класу LabelEncoder з модуля sklearn.preprocessing.

  
  
Рисунок 3.3 - Перетворення категоріальних змінних в числові

Видалимо порожні документи, якщо вони є.

  
  
Рисунок 3.4 - Видалення порожніх документів

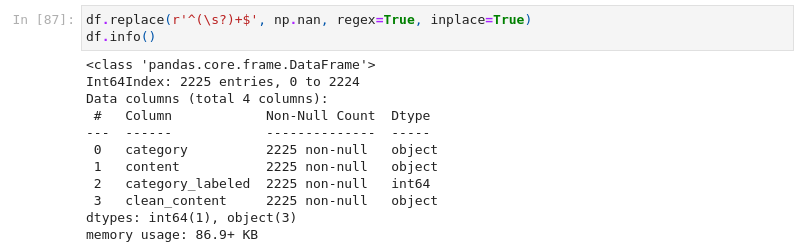
Визначимо стоп-слова англійської мови.

  
  
Рисунок 3.5 - Стоп-слова

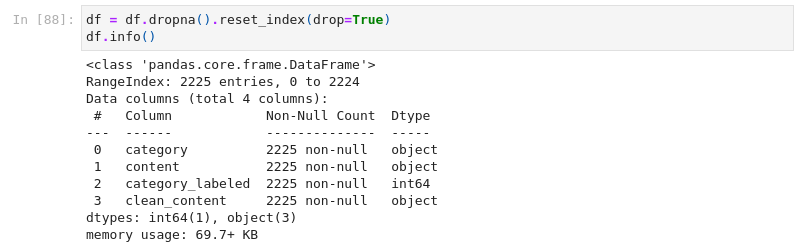
Визначимо функцію, що виконує попередню обробку документу. Застосуємо декоратор np.vectorize для того, щоб функція могла працювати з корпусами.

  
  
Рисунок 3.6 - Обробка документів

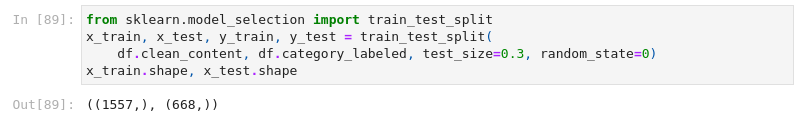
У результаті попередньої обробки деякі документи могли стати порожніми. Тому потрібно це перевірити.

  
  
Рисунок 3.7 - Перевірка на присутність пустих значень

Видалимо пусті значення.

  
  
Рисунок 3.8 - Видалення пустих значень

Розділимо дані на навчальні та тестові, отримаємо чотири масиви: документи для навчання, документи для тестування, мітки для навчання, мітки для тестування.

  
  
Рисунок 3.9 - Розділення даних на навчальні та тестові

Використаємо TF-IDF модель для представлення тексту у векторному вигляді.

  
  
Рисунок 3.10 - TF-IDF модель

## KNN

Оберемо декілька методів класифікації: KNN, Logistic Regression, SVC, Random Forest, Naive Bayes, Gridient Boosting.

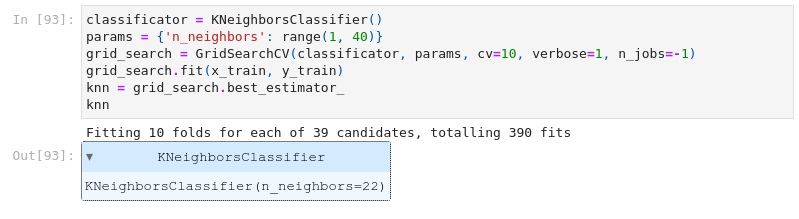
Зберігатимемо результати тестування моделей у списку results.

  
  
Рисунок 3.11 - Список результатів

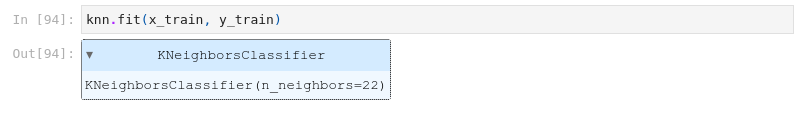
Для виконання роботи методу KNN імпортуємо sklearn.neighbors.KNeighborsClassifier та sklearn.model\_selection.GridSearchCV.

  
  
Рисунок 3.12 - Імпортування модулів

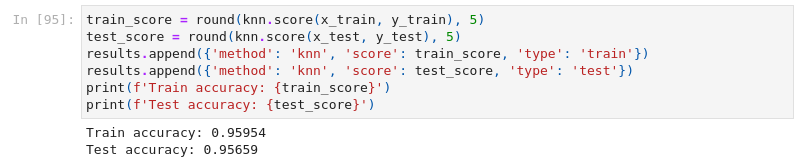
Визначимо, які варіанти параметрів найкраще вирішують дану задачу.

  
  
Рисунок 3.13 - Визначення найкращого параметра

Натренуємо модель з найкращим параметром.

  
  
Рисунок 3.14 - Тренування моделі K-Nearest Neighbors

Визначимо точність моделі на тренувальних та тестових даних.

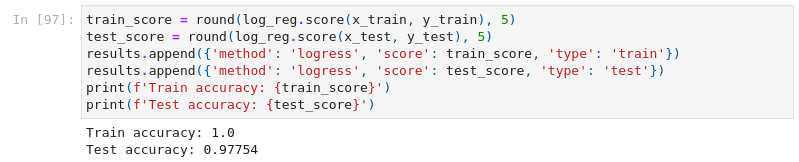
  
  
Рисунок 3.15 - Точність моделі K-Nearest Neighbors

## Logistic Regression

Для виконання роботи методу Logistic Regression імпортуємо sklearn.linear\_model.LogisticRegression. Визначимо найкращі параметри моделі, передавши в неї параметри регуляризації.

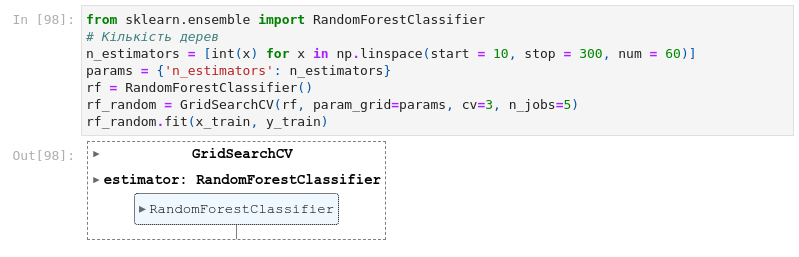
  
  
Рисунок 3.16 - Тренування моделі Logistic Regression

Визначимо точність моделі на тренувальних та тестових даних.

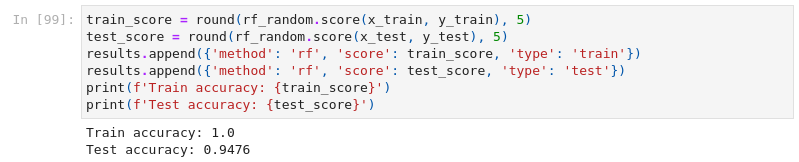
  
  
Рисунок 3.17 - Точність моделі Logistic Regression

## Random Forest

Для виконання роботи методу Random Forest імпортуємо sklearn.ensemble.RandomForestClassifier. Визначимо найкращі параметри для моделі. У випадку Random Forest параметри включають кількість дерев рішень та кількість характеристик, які враховуються кожним деревом під час поділу вузла і використовуються для поділу кожного вузла, отриманого під час навчання. Імпортуємо sklearn.model\_selection.RandomizedSearchCV.

  
  
Рисунок 3.18 - Тренування моделі Random Forest

Визначимо точність моделі на тренувальних та тестових даних.

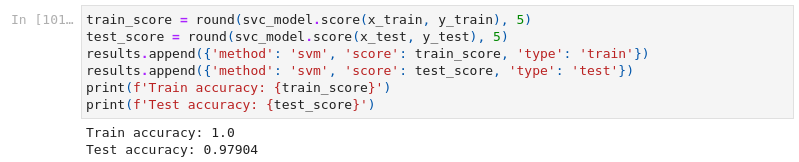
  
  
Рисунок 3.19 - Точність моделі Random Forest

## SVM

Для виконання роботи методу SVM імпортуємо sklearn.svm.SVC.

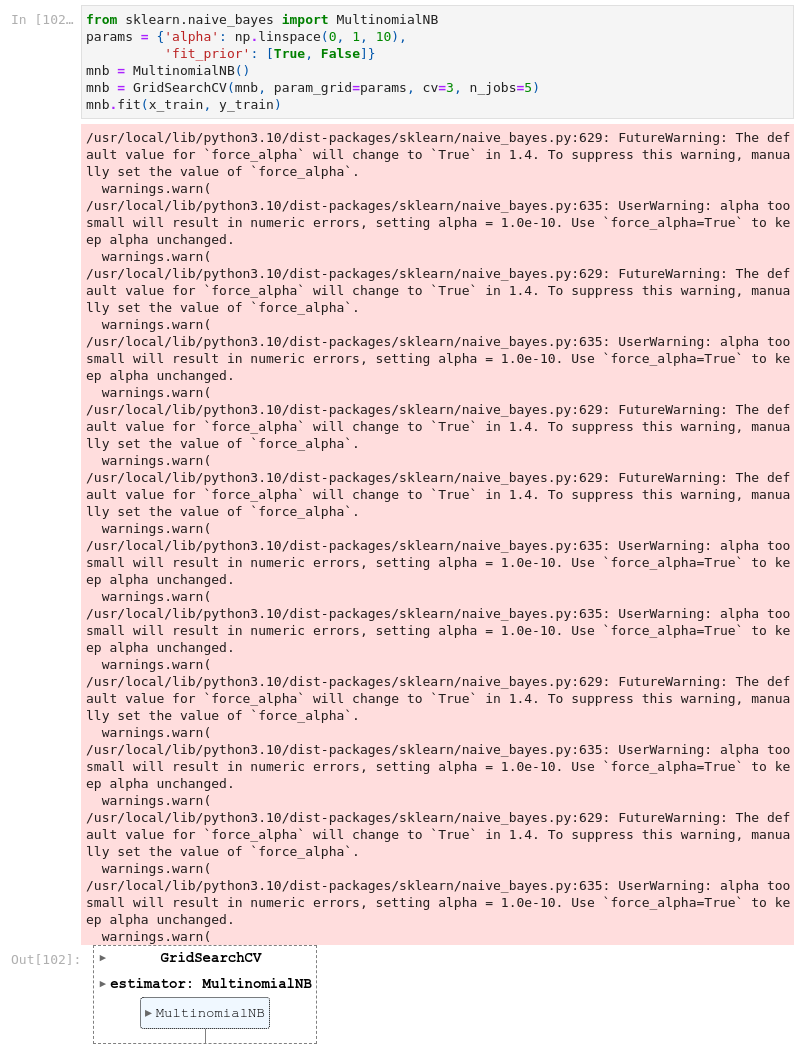
  
  
Рисунок 3.20 - Тренування моделі SVM

Визначимо точність моделі на тренувальних та тестових даних.

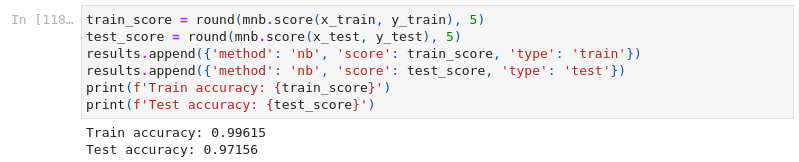
  
  
Рисунок 3.21 - Точність моделі SVM

## Naive Bayes

Для використання алгоритму імпортуємо клас MultinomialNB з модуля sklearn.naive\_bayes.

  
  
Рисунок 3.22 - Тренування моделі Naive Bayes

Визначимо точність моделі на тренувальних та тестових даних.

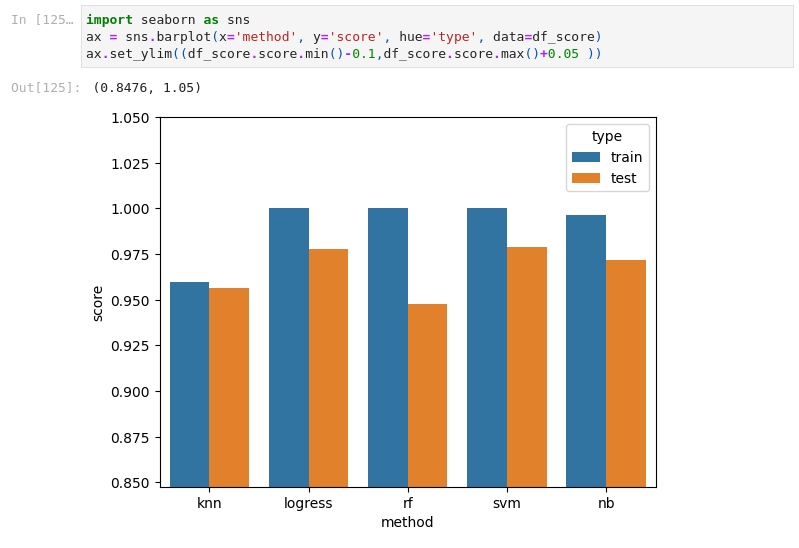
  
  
Рисунок 3.23 - Точність моделі Naive Bayes

## Порівняння результатів

Проаналізувавши окремо кожен із методів, проведемо порівняння даних методів.

  
  
Рисунок 3.24 - Датафрейм результатів

Для наочності побудуємо гістограму.

  
  
Рисунок 3.25 - Результати моделей