



**МІНІСТЕРСТВО ОСВІТИ, НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ**

**Лабораторна робота №8
Аналіз даних**

Підготував:

студент 4 курсу

групи ФІ-84

Коломієць Андрій Юрійович

E-mail: andrew.kolomiets.work@gmail.com

Київ – 2021

Лабораторна робота №8

Аналіз даних

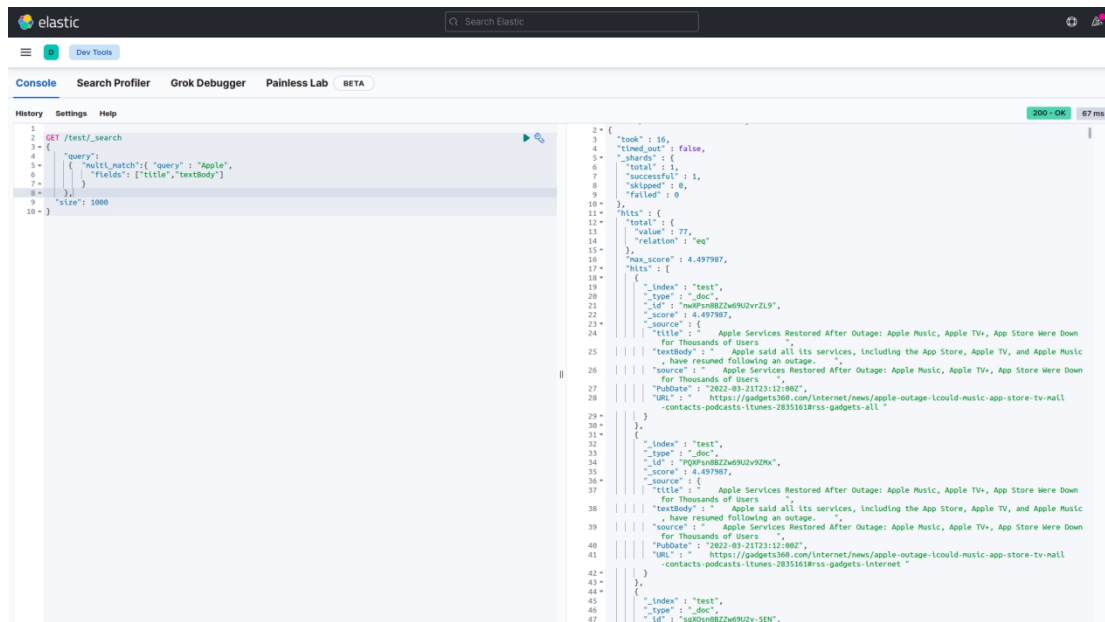
Завдання на самостійну роботу

1. Детально ознайомитись із засобами обробки тестових рядків в мові **Python**.
2. Мовою **Python** самостійно розробити програмний модуль підрахунку ваги слів.
3. Ознайомитись з іншими методами розрахунку ваги слів, зокрема методом **TF-IDF**, дисперсійним методом і методом горизонтальної видимості (**HVG**).

Виконання завдання

Отримання масиву документів

Для подальшої обробки текстових документів, зокрема екстрагування найбільш вагомих, ключових слів, необхідно отримати тестові документи із бази даних **Elasticsearch**.



```
1 GET /test/_search
2 {
3   "query": {
4     "multi_match": { "query": "Apple",
5                      "fields": ["title", "textBody"] }
6   }
7 }
8
9 "size": 1000
10 }
```

```
2 {
3   "took": 16,
4   "timed_out": false,
5   "shards": {
6     "total": 1,
7     "successful": 1,
8     "skipped": 0,
9     "failed": 0
10  },
11  "hits": {
12    "total": {
13      "value": 77,
14      "relation": "eq"
15    },
16    "max_score": 4.497987,
17    "hits": [
18      {
19        "_index": "test",
20        "_type": "_doc",
21        "_id": "mnpFsn8ZzW9UzvZL9",
22        "_score": 4.497987,
23        "_source": {
24          "title": "Apple Services Restored After Outage: Apple Music, Apple TV, App Store Were Down for Thousands of Users",
25          "textBody": "Apple said all its services, including the App Store, Apple TV, and Apple Music, have resumed following an outage.",
26          "source": "Apple Services Restored After Outage: Apple Music, Apple TV, App Store Were Down for Thousands of Users",
27          "pubDate": "2022-03-21T23:12:00Z",
28          "url": "https://gadgets360.com/internet/news/apple-outage-icloud-music-app-store-tv-mail-contacts-podcasts-itunes-2035181rss-gadgets-all"
29        },
30      },
31      {
32        "_index": "test",
33        "_type": "_doc",
34        "_id": "PQdFsn8ZzW9UzvZL9",
35        "_score": 4.497987,
36        "_source": {
37          "title": "Apple Services Restored After Outage: Apple Music, Apple TV, App Store Were Down for Thousands of Users",
38          "textBody": "Apple said all its services, including the App Store, Apple TV, and Apple Music, have resumed following an outage.",
39          "source": "Apple Services Restored After Outage: Apple Music, Apple TV, App Store Were Down for Thousands of Users",
40          "pubDate": "2022-03-21T23:12:00Z",
41          "url": "https://gadgets360.com/internet/news/apple-outage-icloud-music-app-store-tv-mail-contacts-podcasts-itunes-2035181rss-gadgets-all"
42        },
43      },
44      {
45        "_index": "test",
46        "_type": "_doc",
47        "_id": "sgUzn8ZzW9UzvZL9",
```

Результати виконаної операції збережемо в файл **Apple.txt**.

```
1 {
2   "took": 16,
3   "timed_out": false,
4   "shards": {
5     "total": 1,
6     "successful": 1,
7     "skipped": 0,
8     "failed": 0
9   },
10  "hits": {
11    "total": {
12      "value": 77,
13      "relation": "eq"
14    },
15    "max_score": 4.497987,
16    "hits": [
17      {
18        "_index": "test",
19        "_type": "_doc",
20        "_id": "mdPsn88Zw69Uzv9ZL9",
21        "_score": 4.497987,
22        "_source": {
23          "title": "Apple Services Restored After Outage: Apple Music, Apple TV, App Store Were Down For Thousands of Users",
24          "textBody": "Apple said all its services, including the App Store, Apple TV, and Apple Music, have resumed following an outage.",
25          "source": "Apple Services Restored After Outage: Apple Music, Apple TV, App Store Were Down For Thousands of Users",
26          "pubdate": "2022-03-21T23:12:00Z",
27          "url": "https://gadgets360.com/Internet/news/apple-outage-icould-music-app-store-tv-mall-contacts-podcasts-itunes-2835161/rss-gadgets-all"
28        }
29      },
30      {
31        "_index": "test",
32        "_type": "_doc",
33        "_id": "PQXpsn88Zw69Uzv9ZmX",
34        "_score": 4.497987,
35        "_source": {
36          "title": "Apple Services Restored After Outage: Apple Music, Apple TV, App Store Were Down For Thousands of Users",
37          "textBody": "Apple said all its services, including the App Store, Apple TV, and Apple Music, have resumed following an outage.",
38          "source": "Apple Services Restored After Outage: Apple Music, Apple TV, App Store Were Down For Thousands of Users",
39          "pubdate": "2022-03-21T23:12:00Z",
40          "url": "https://gadgets360.com/Internet/news/apple-outage-icould-music-app-store-tv-mall-contacts-podcasts-itunes-2835161/rss-gadgets-Internet"
41        }
42      },
43      {
44        "_index": "test",
45        "_type": "_doc",
46        "_id": "sg0Xpsn88Zw69Uzv-5EN",
47        "_score": 3.955235,
48        "_source": {
49          "title": "Apple Event (March 8) Highlights: All the Announcements From Peek Performance Event",
50          "textBody": "Apple event on March 8 unveiled the new 5G iPhone SE, apart from a new iPad Air, and the Mac Studio machines based on brand new Apple silicon. This was our live blog coverage of
```

Формування словника та знаходження найчастіших слів

Для формування словника слів із отриманого масиву документів **D**, що розміщуються в файлі **Apple.txt**, у середовищі **Python** об'єднаємо вміст всіх полів **"title"** і **"textBody"**, замінимо всі розподільні символи пропусками, і за допомогою регулярних виразів визначимо всі слова. Після сортування словника слів визначимо лише унікальні слова і абсолютну частоту їх появ у масиві **D**. Для кожного слова **t**, це значення **tf(t)**. Фрагмент вихідного коду мовою **Python** наведено нижче:

Dictionary

```
#!/bin/python3
```

```
import re
import string
```

```
#-----
# For regular expression use : https://regex101.com/
#-----
# Dictionary of JSON file print
#-----
```

```

f = open("./Apple.json", "r")
t = f.read()
f.close()

json = t.split('\n')
t = ""

for i in range(len(json)):
    t = t+" "+json[i]

# print(t)

title = re.findall('"title" : "(.+?)"source"', t)
t = ""

for i in range(len(title)):
    t = t+" "+title[i]

t = re.sub('"textBody" :', '', t)

t = re.sub('[ "" [\]\/?0-9\",()$+»«\-.:;._\'`\'\'--]', ' ', t)

t = t.upper()
t = re.sub('\s\w\s', ' ', t)
t = re.sub('\s\w\w\s', ' ', t)
t = re.sub('\s\s+', ' ', t)
word = t.split(' ')
word.sort()

# Dictionary building
d = {}
old = ""
n = 0

for i in range(len(word)):
    if (word[i] == old):
        n = n+1
    else:
        # print(old,n)
        d[old] = n
        old = word[i]
        n = 1

d[old] = n

```

```

# print(d)
sorted_dict = {}
sorted_keys = sorted(d, key=d.get, reverse=True)

# [1, 3, 2]
for w in sorted_keys:
    sorted_dict[w] = d[w]

print(sorted_dict)

#-----
# Stop word : https://code.google.com/archive/p/stop-words/
#-----

# Print frequency of word

#-----

f = open("./stop.txt","r")
t = f.read()
f.close()

t=t.upper()
stop =t.split('\n')
M=50
j=1

sorted_dict = {}
sorted_keys = sorted(d, key=d.get, reverse=True) # [1, 3, 2]

for w in sorted_keys:
    sorted_dict[w] = d[w]
    pr=0
    for i in range(len(stop)):
        if (stop[i] == w):
            pr=1
    if (pr == 0):
        print(j,w,sorted_dict[w])
        j=j+1
    if (j > M):
        break

```

Result-1

{'THE': 166, 'APPLE': 152, 'AND': 127, 'IPHONE': 101, 'NEW': 59, 'WITH': 55, 'EVENT': 40, 'PRO': 40, 'AIR': 36, 'MAC': 29, 'MACBOOK': 29, 'FOR': 28, 'HAS': 28, 'IPAD': 27, 'HAVE': 26, 'ITS': 25, 'ARE': 23, 'SAID': 22, 'THAT': 22, 'ALSO': 21, 'FROM': 21, 'MACOS': 21, 'STUDIO': 21, 'MARCH': 19, 'RUSSIA': 18, 'WHILE': 18, 'WILL': 18, 'BE': 17, 'LAUNCH': 17, 'GOOGLE': 16, 'MINI': 16, 'TECH': 15, 'BEEN': 14, 'DISPLAY': 14, 'IOS': 14, 'MICROSOFT': 14, 'USERS': 14, 'COMPANY': 13, 'MAX': 13, 'APPLE&': 12, 'GIANTS': 12, 'IPADOS': 12, 'LAUNCHED': 12, 'NEXT': 12, 'PEEK': 12, 'STEVE': 12, 'YEAR': 12, 'MONTEREY': 11, 'OVER': 11, 'REPORT': 11, 'REPORTEDLY': 11, 'SERIES': 11, 'SHIPMENTS': 11, 'TIPPED': 11, 'CHIP': 10, 'EXPECTED': 10, 'META': 10, 'MODELS': 10, 'PERCENT': 10, 'RUSSIAN': 10, 'TECHNOLOGY': 10, 'THIS': 10, 'WAS': 10, 'WATCH': 10, 'WEEK': 10, 'WERE': 10, 'BASED': 9, 'DEBUT': 9, 'DELL': 9, 'INCLUDING': 9, 'INDIA': 9, 'MAY': 9, 'MOST': 9, 'SILICON': 9, 'APPS': 8, 'FACE': 8, 'FOUNDER': 8, 'IMAC': 8, 'INCH': 8, 'MASK': 8, 'NOW': 8, 'PERFORMANCE': 8, 'PRODUCTS': 8, 'SERVICES': 8, 'UKRAINE': 8, 'WOZNIAK': 8, 'ACCORDING': 7, 'ALL': 7, 'BIONIC': 7, 'BUT': 7, 'CHIPS': 7, 'CONTROL': 7, 'HAD': 7, 'OPERATIONS': 7, 'OTHER': 7, 'SOME': 7, 'UNLOCK': 7, 'WEARING': 7, 'ALONGSIDE': 6, 'APP': 6, 'BIG': 6, 'DARK': 6, 'FEATURE': 6, 'GREEN': 6, 'INTEL': 6, 'LENOVO': 6, 'MODE': 6, 'PATENT': 6, 'SCREEN': 6, 'SOON': 6, 'SPRING': 6, 'THEIR': 6, 'UNIVERSAL': 6, 'WATCHOS': 6, 'WEBSITES': 6, 'AMAZON': 5, 'AMONGST': 5, 'AN': 5, 'BRAND': 5, 'BRING': 5, 'COMPANIES': 5, 'COVID': 5, 'DELAYED': 5, 'DESIGN': 5, 'EMPLOYEE': 5, 'FOLLOWING': 5, 'GLOBAL': 5, 'GROWTH': 5, 'LIKELY': 5, 'MAKER': 5, 'MILLION': 5, 'RELEASED': 5, 'TOP': 5, 'TUESDAY': 5, 'UNTIL': 5, 'UNVEILED': 5, '&': 4, 'ABOUT': 4, 'AFTER': 4, 'AGREEMENT': 4, 'AMD': 4, 'AMID': 4, 'ANNOUNCED': 4, 'BELARUS': 4, 'BEST': 4, 'CHARGED': 4, 'COULD': 4, 'COUNTERPOINT': 4, 'CRIMEA': 4, 'CUPERTINO': 4, 'CUTTING': 4, 'DANGERS': 4, 'DEFRAUDING': 4, 'DEPARTURE': 4, 'DEVICES': 4, 'DISPUTE': 4, 'DIVERGENT': 4, 'EMPLOYEES': 4, 'EXTERNAL': 4, 'FEDERATION': 4, 'FIRMWARE': 4, 'FLEXIBLE': 4, 'FOLLOWED': 4, 'FOXCONN': 4, 'GETS': 4, 'HIS': 4, 'HOW': 4, 'HP': 4, 'IMMERSED': 4, 'INTERESTS': 4, 'INVASION': 4, 'JOBS': 4, 'LATE': 4, 'LATEST': 4, 'LAWMAKERS': 4, 'LEADING': 4, 'LICENCE': 4, 'LIMITING': 4, 'LIST': 4, 'LOOKING': 4, 'MAPS': 4, 'MARK': 4, 'MODEL': 4, 'MUSIC': 4, 'OLD': 4, 'OLED': 4, 'OUTAGE': 4, 'PART': 4, 'PAST': 4, 'PERFORMANCE&': 4, 'POWERED': 4, 'PURSUING': 4, 'RELEASE': 4, 'REMAINED': 4, 'RESTORE': 4, 'RESTRICTED': 4, 'REVOLUTIONARY': 4, 'RUMOURED': 4, 'SAFARI': 4, 'SALES': 4, 'SECURITY': 4, 'SET': 4, 'SHOWING': 4, 'SINCE': 4, 'SIZES': 4, 'SPECIFICATIONS': 4, 'SPOTTED': 4, 'STARTED': 4, 'STILL': 4, 'STORE': 4, 'TAKING': 4, 'THEM': 4, 'THING': 4, 'TIES': 4, 'TIKTOK': 4, 'TIPSTER': 4, 'TODAY': 4, 'ULTRA': 4, 'UPCOMING': 4, 'VALUABLE': 4, 'WEATHER': 4, 'WHAT': 4, 'WHEN': 4, 'WILAN': 4, 'WIRELESS': 4, 'WITHDRAWN': 4, 'WORKING': 4, 'ABILITY': 3, 'AGAINST': 3, 'AHEAD': 3, 'ANTITRUST': 3, 'APPLES': 3, 'ARRIVE': 3, 'BRANDS': 3, 'BUILD': 3, 'CHI': 3, 'COMPANY'S': 3, 'DURING': 3, 'EARPHONES': 3, 'ENABLES': 3, 'EXCITING': 3, 'FIRST': 3, 'FIXES': 3, 'GET': 3, 'HEADPHONES': 3, 'HIT': 3, 'IDEA': 3, 'INTERNAL': 3, 'INTRODUCE': 3, 'KUO': 3, 'MACHINES': 3, 'MADE': 3, 'MANUFACTURER': 3, 'NEWS': 3, 'OUR': 3, 'OUT': 3, 'PER': 3, 'PRODUCTION': 3, 'QUALITY': 3, 'REDESIGNED': 3, 'SMARTPHONES': 3, 'SOCS': 3, 'SPORT': 3, 'SUPPLIER': 3, 'SUPPORT': 3, 'SUSPENDED': 3, 'TABLET': 3, 'TO': 3, 'UNDER': 3, 'UPDATE': 3, 'UPDATED': 3, 'UPDATES': 3, 'USING': 3, 'VIRTUAL': 3, 'WEBKIT': 3, ':': 2, 'ABLE': 2, 'ACTIVE': 2, 'ADDING': 2, 'AKA': 2, 'ALLOWED': 2, 'ALLOWS': 2, 'ALPHABET': 2, 'ALPINE': 2, 'ANALYST': 2, 'ANDROID': 2, 'ANNOUNCEMENT': 2, 'ANNOUNCEMENTS': 2, 'ANYTIME': 2, 'APART': 2, 'APPEARS': 2, 'APPLE'S': 2, 'ASIAN': 2, 'ASSERTIONS': 2, 'AUDIT': 2, 'BETA': 2, 'BIKES': 2, 'BILL': 2, 'BLOG': 2, 'BLOOMBERG'S': 2, 'BOOSTER': 2, 'BROWSER': 2, 'BUG': 2, 'BUSINESS': 2, 'CANALYS': 2, 'CANCELLING': 2, 'CAR': 2, 'CARRIER': 2, 'CATALINA': 2, 'CEO': 2, 'CHANNEL': 2, 'CHIEF': 2, 'CITING': 2, 'CODE': 2, 'COLOUR': 2, 'COLOURWAY': 2, 'COLOURWAYS': 2, 'COME': 2, 'COMING': 2, 'COMPLY': 2, 'CONCERNS': 2, 'CONFIRMED': 2, 'CONNECTIVITY': 2, 'CONSUMERS': 2, 'CONTINUED': 2, 'COST': 2, 'COUNTRY': 2, 'COVERAGE': 2, 'CRISIS': 2, 'CRORE': 2, 'DAVIDSON': 2, 'DEDICATED': 2, 'DEVELOPMENT': 2, 'DIRECTLY': 2, 'DISCONTINUED': 2, 'DISCUSSED': 2, 'DISPUTES': 2, 'DOMINATES': 2, 'DOWN': 2, 'DROPPED': 2, 'EEC': 2, 'ENABLED': 2, 'END': 2, 'ENGINEER': 2, 'ENTRY': 2, 'EQUIPMENT': 2, 'ESSENTIALLY': 2, 'EVERYTHING': 2, 'EXPECT': 2, 'EXPLOITED': 2, 'FACEBOOK': 2, 'FALL': 2, 'FEDERAL': 2, 'FEMALE': 2, 'FIRM&': 2, 'FLAW': 2, 'FORD': 2, 'FOUND': 2, 'GADGETS': 2, 'GENERATION': 2, 'GIANT': 2, 'GLOBALLY': 2, 'GOES': 2, 'GOING': 2, 'GURMAN': 2, 'HALF': 2, 'HARLEY': 2, 'HEAVYWEIGHTS': 2, 'HERE'S': 2, 'HIGH': 2, 'HIGHLIGHTS': 2, 'HITTING': 2, 'IMPORTANT': 2, 'IMPOSE': 2, 'IMPRESSIONS': 2, 'INCLUDE': 2, 'INDIVIDUAL': 2, 'INFLUENTIAL': 2, 'INTO': 2, 'INVITE': 2, 'IS': 2, 'ISSUES': 2, 'JOINED': 2, 'KEY': 2, 'KICKBACKS': 2, 'LATER': 2, 'LAUNDERING': 2, 'LEAKED': 2, 'LED': 2, 'LEVEL': 2, 'LITIGATIONS': 2, 'LIVE': 2, 'LIVESTREAM': 2, 'LIVESTREAMED': 2, 'LOW':

2, 'LTPO': 2, 'MAINLY': 2, 'MAJOR': 2, 'MAKE': 2, 'MAKERS': 2, 'MANUFACTURING': 2, 'MARKETS': 2, 'MEDIA': 2, 'MICROSITE': 2, 'MING': 2, 'MINORITY': 2, 'MIX': 2, 'MONEY': 2, 'NATIONS': 2, 'NEARLY': 2, 'NEITHER': 2, 'NOISE': 2, 'NOR': 2, 'OF': 2, 'OPEN': 2, 'ORBITAL': 2, 'OTHERS': 2, 'OUTSIDE': 2, 'OVERBLOWN': 2, 'PANEL': 2, 'PARTNER': 2, 'PENDING': 2, 'PHONES': 2, 'PLACE': 2, 'PODCAST': 2, 'POPULAR': 2, 'PORSCHÉ': 2, 'POWER': 2, 'PREFER': 2, 'PREFERENCE': 2, 'PRESSES': 2, 'PRESSURE': 2, 'PREVIOUS': 2, 'PRICE': 2, 'PROCESSOR': 2, 'PRODUCT': 2, 'PROJECTS': 2, 'PROOF': 2, 'PROSECUTORS': 2, 'PUBLISHERS': 2, 'RECEIVED': 2, 'RECOVER': 2, 'REFERENCES': 2, 'REPORTED': 2, 'REQUIRE': 2, 'RESPECTIVELY': 2, 'RESPONSE': 2, 'RESTORED': 2, 'RESUMED': 2, 'ROUGHLY': 2, 'RUNNING': 2, 'SAMSUNG': 2, 'SANCTIONING': 2, 'SANCTIONS': 2, 'SAYS': 2, 'SCHEDULED': 2, 'SCREENS': 2, 'SECOND': 2, 'SECURITIES': 2, 'SELLING': 2, 'SENT': 2, 'SETTLED': 2, 'SETTLING': 2, 'SEVERAL': 2, 'SHADES': 2, 'SHENZHEN': 2, 'SHIPPERS': 2, 'SHORTAGE': 2, 'SIDELOADING': 2, 'SIGN': 2, 'SIGNED': 2, 'SIGNIFICANT': 2, 'SMALLER': 2, 'SMARTPHONE': 2, 'SMARTPHONES': 2, 'SOC': 2, 'SONY': 2, 'SOURCE': 2, 'SPECULATED': 2, 'START': 2, 'STATE': 2, 'STEALING': 2, 'STEPS': 2, 'STOPPED': 2, 'STOPS': 2, 'STORAGE': 2, 'SUGGESTED': 2, 'SUR': 2, 'SUSPENDING': 2, 'TAKEN': 2, 'TALK': 2, 'TALLER': 2, 'TESTING': 2, 'THAT'S': 2, 'THESE': 2, 'THEY'RE': 2, 'THOUSANDS': 2, 'THREE': 2, 'THROUGH': 2, 'TOGGLE': 2, 'TOLD': 2, 'TREATMENT': 2, 'TWO': 2, 'UNDERGO': 2, 'UNVEILING': 2, 'USE': 2, 'USER': 2, 'VEDANTA': 2, 'VERSION': 2, 'VIEWING': 2, 'VOLUNTARY': 2, 'WEE': 2, 'WEEKS': 2, 'WEEK'S': 2, 'WELL': 2, 'WHAT'S': 2, 'WON': 2, 'WORK': 2, 'WRITTEN': 2, 'YET': 2, 'YOUTUBE': 2, '‘SIDELOADING’': 2, 'ABOVE': 1, 'ACCOMMODATE': 1, 'ACER': 1, 'ACROSS': 1, 'ACTIVELY': 1, 'ACTUALLY': 1, 'ADJUSTS': 1, 'AFFECTING': 1, 'AGREES': 1, 'ALLOWING': 1, 'ALMOST': 1, 'ANC': 1, 'APPROVED': 1, 'AROUND': 1, 'ASSEMBLER': 1, 'ASSESSING': 1, 'AT': 1, 'AUTHORITY': 1, 'AVAILABLE': 1, 'BACK': 1, 'BACKLIGHTING': 1, 'BARRING': 1, 'BATTERY': 1, 'BECOME': 1, 'BEGINNING': 1, 'BEGUN': 1, 'BLASS': 1, 'BLOOMBERG': 1, 'BLUETOOTH': 1, 'BLUME': 1, 'BOTH': 1, 'BRINGING': 1, 'BUDGETS': 1, 'BUDS': 1, 'CAMERA': 1, 'CHAIN': 1, 'CHART': 1, 'CHASSIS': 1, 'CHINA': 1, 'CHINESE': 1, 'CHIPSET': 1, 'CITED': 1, 'CITY': 1, 'CIVIL': 1, 'CLAIM': 1, 'CLAIMED': 1, 'COMES': 1, 'COMMISSION': 1, 'COMMON': 1, 'COMMUNITY': 1, 'COMPANY&': 1, 'COMPARED': 1, 'COMPETITION': 1, 'COMPLAIN': 1, 'COMPLAINTS': 1, 'COMPLIANCES': 1, 'COMPONENT': 1, 'CONGLOMERATE': 1, 'CONTENDER': 1, 'COPUTER': 1, 'CORPORATE': 1, 'COVERS': 1, 'CURRENT': 1, 'CURRENTLY': 1, 'CUTOUT': 1, 'DATA': 1, 'DATE': 1, 'DEBUTED': 1, 'DECISION': 1, 'DELIVERING': 1, 'DESIGNED': 1, 'DESKTOP': 1, 'DESKTOPS': 1, 'DESPITE': 1, 'DEVELOPER': 1, 'DISPLAYS': 1, 'DOMINATED': 1, 'DRAIN': 1, 'DROP': 1, 'DUE': 1, 'EAR': 1, 'EARLY': 1, 'EASE': 1, 'EASED': 1, 'ECONOMIC': 1, 'ECONOMICAL': 1, 'EMAIL': 1, 'ENJOYED': 1, 'EURASIAN': 1, 'EUROPEAN': 1, 'EVAN': 1, 'EVEN': 1, 'EXAMPLE': 1, 'EXPERIMENTING': 1, 'FACTORS': 1, 'FASHIONED': 1, 'FAVOUR': 1, 'FEATURES': 1, 'FEW': 1, 'FIGHT': 1, 'FILIPPO': 1, 'FINANCE': 1, 'FINE': 1, 'FINES': 1, 'FIRMS': 1, 'FIX': 1, 'FORCING': 1, 'FORM': 1, 'FORUMS': 1, 'FRENCH': 1, 'FUTURE': 1, 'GEN': 1, 'GLOBE': 1, 'GOOD': 1, 'GOVERNMENT&': 1, 'GREW': 1, 'HALTS': 1, 'HAPPEN': 1, 'HERE': 1, 'HERES': 1, 'HIRED': 1, 'HIRES': 1, 'HOLE': 1, 'HOLIDAYS': 1, 'HOME': 1, 'HOOD': 1, 'HYBRID': 1, 'IMMINENT': 1, 'IMPROVE': 1, 'IN': 1, 'INDEPENDENT': 1, 'INDIAN': 1, 'INTRODUCED': 1, 'INVENTORY': 1, 'INVESTORS': 1, 'JANUARY': 1, 'KNOW': 1, 'KNOWLEDGE': 1, 'LAPTOP': 1, 'LAPTOPS': 1, 'LAST': 1, 'LEADS': 1, 'LIFE': 1, 'LOCAL': 1, 'LUNAR': 1, 'M': 1, 'MANAGEMENT': 1, 'MANUFACTURERS': 1, 'MARGRETHE': 1, 'MARKET': 1, 'MATTER': 1, 'MESSAGING': 1, 'MIGHT': 1, 'MIKE': 1, 'MINIMAL': 1, 'MONDAY': 1, 'MONITOR': 1, 'MULTIPLE': 1, 'NATURE': 1, 'NEED': 1, 'NETHERLANDS&': 1, 'NOT': 1, 'OFFICE': 1, 'OLDEST': 1, 'OLIVER': 1, 'ONE': 1, 'ONLINE': 1, 'ONLY': 1, 'OPTION': 1, 'ORDER': 1, 'OWN': 1, 'OWNERS': 1, 'PAR': 1, 'PARTNERED': 1, 'PARTNERS': 1, 'PAY': 1, 'PCS': 1, 'PEOPLE': 1, 'PILL': 1, 'POLICIES': 1, 'POOR': 1, 'PORTS': 1, 'POSITIONED': 1, 'POSSIBLE': 1, 'POWERFUL': 1, 'PREDECESSOR': 1, 'PRICED': 1, 'PRICES': 1, 'PROPOSAL': 1, 'PROVIDE': 1, 'PUBLICATION': 1, 'PUNCH': 1, 'QUARTER': 1, 'RAISED': 1, 'RANGE': 1, 'RARE': 1, 'RATHER': 1, 'REALME': 1, 'REBUKE': 1, 'REDDIT': 1, 'REFRESH': 1, 'REGISTERING': 1, 'REGULAR': 1, 'RELEASES': 1, 'REMAINS': 1, 'RESEARCH': 1, 'RESOLVES': 1, 'RESTRICTIONS': 1, 'RETAIL': 1, 'RETAINS': 1, 'RIGHTS': 1, 'ROLLED': 1, 'ROLLING': 1, 'RULES': 1, 'RUNS': 1, 'SAME': 1, 'SATELLITE': 1, 'SAW': 1, 'SEMICONDUCTORS': 1, 'SERVER': 1, 'SERVERS': 1, 'SHAPED': 1, 'SHARE': 1, 'SHAREHOLDERS': 1, 'SHIPMENT': 1, 'SHIPPED': 1, 'SHORTAGES': 1, 'SHOT': 1, 'SHOULD': 1, 'SHOWN': 1, 'SHOWS': 1, 'SMARTWATCH': 1, 'SOFTWARE': 1, 'SPORTS': 1, 'STOCKPILE': 1, 'SUCH': 1, 'SUMMER': 1, 'SUPER': 1, 'SUPPLIES': 1, 'SUPPLY': 1, 'TABLETS': 1, 'TEAM': 1, 'THAN': 1, 'THERE'S': 1, 'THROUGHOUT': 1, 'TIDE': 1, 'TITLE': 1, 'TITLED': 1, 'TOPPED': 1, 'TOW': 1, 'TRUE': 1, 'UNION&': 1, 'UNITS': 1, 'UNLOCKING': 1, 'URGE': 1, 'URGING': 1, 'US': 1, 'VARIOUS': 1, 'VERGE': 1, 'VERSIONS': 1, 'VESTAGER': 1, 'VS': 1, 'WAIT': 1, 'WEBSITE': 1, 'WHAES': 1, 'WHICH': 1, 'WHOLE': 1, 'WIDE': 1, 'WITNESS': 1, 'WORKPLACES': 1, 'WORKSTATION': 1, 'WORLD&': 1, 'WORLD'S': 1, 'XIAOMI': 1, 'YOU': 1, '‘SUMMER’': 1}

Result-2

1 APPLE 152
2 AND 127
3 IPHONE 101
4 NEW 59
5 EVENT 40
6 PRO 40
7 AIR 36
8 MAC 29
9 MACBOOK 29
10 HAS 28
11 IPAD 27
12 HAVE 26
13 ITS 25
14 SAID 22
15 ALSO 21
16 MACOS 21
17 STUDIO 21
18 MARCH 19
19 RUSSIA 18
20 WHILE 18
21 LAUNCH 17
22 GOOGLE 16
23 MINI 16
24 TECH 15
25 BEEN 14

26 DISPLAY 14
27 IOS 14
28 MICROSOFT 14
29 USERS 14
30 COMPANY 13
31 MAX 13
32 APPLE& 12
33 GIANTS 12
34 IPADOS 12
35 LAUNCHED 12
36 NEXT 12
37 PEEK 12
38 STEVE 12
39 YEAR 12
40 MONTEREY 11
41 OVER 11
42 REPORT 11
43 REPORTEDLY 11
44 SERIES 11
45 SHIPMENTS 11
46 TIPPED 11
47 CHIP 10
48 EXPECTED 10
49 META 10
50 MODELS 10

Питання до практичної роботи

1. Як відбувається вибір найбільш вагомих слів за заданою тематикою?

Будемо застосовувати частотний метод, а саме із сформованого на базі аналізу масиву словника відберемо найбільш частотні слова.

2. Яка подальша мета використання отриманих слів по заданій тематиці?

Подальша мета використання отриманих слів може залежати від сфери їх вживання, наприклад в данній лабораторній роботі вживався детасет, пов'язаний з цифровими пристроями. Згідно цього можна визначати масштаби продажу, використання певної торгової марки мобільних виробів, чи комп'ютерів. Робити певні прогнози, пов'язані з ринком цифрових пристойів, та досліджувати впливи на такі системи. Щодо інших областей застосування, потрібно опиратися на ту сферу робочу, котра досліджується.

3. У чому полягає суть іншого методу розрахунку ваги слів, зокрема методу **TF-IDF**?

***TF-IDF** — статистичний показник, що використовується для оцінки важливості слів у контексті документа, що є частиною колекції документів чи корпусу. Вага (значимість) слова пропорційна кількості вживань цього слова у документі, і обернено пропорційна частоті вживання слова у інших документах колекції. Показник **TF-IDF** використовується в задачах аналізу текстів та інформаційного пошуку. Його можна застосовувати як один з критеріїв релевантності документа до пошукового запиту, а також при розрахунку міри спорідненості документів при кластеризації. Найпростішу функцію ранжування можна визначити як суму **TF-IDF** кожного терміну в запиті. Більшість просунутих функцій ранжування ґрунтуються на цій простій моделі.*

Основна ідея моделі **tf-idf** полягає в наступному: якщо слово **w** часто зустрічається в документі **d** і рідко зустрічається в інших документах, то вважається, що слово **w** має хорошу здатність розрізняти, що підходить для статті **d** та інші статті виділяються.

<https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089>