Zadanie 4 - Levenshtein distance and Longest Common Subsequence

April 27, 2021

1 Levenstein Distance

1.1 Paweł Kruczkiewicz

21.04.2021 r. Zadanie dotyczy wykorzystania odległości edycyjnej.

1. Zaimplementuj algorytm obliczania odległości edycyjnej w taki sposób, aby możliwe było określenie przynajmniej jednej sekwencji edycji (dodanie, usunięcie, zmiana znaku), która pozwala w minimalnej liczbie kroków, przekształcić jeden łańcuch w drugi.
2. Na podstawie poprzedniego punktu zaimplementuj prostą wizualizację działania algorytmu, poprzez wskazanie kolejnych wersji pierwszego łańcucha, w których dokonywana jest określona zmiana. "Wizualizacja" może działać w trybie tekstowym. Np. zmiana łańcuch "los" w "kloc" może być zrealizowana następująco: *k*los (dodanie litery k) klo*c* (zamiana s->c)
3. Przedstaw wynik działania algorytmu z p. 2 dla następujących par łańcuchów: los - kloc Łódź - Lodz kwintesencja - quintessence ATGAATCTTACCGCCTCG - ATGAGGCTCTGGCCCCTG
4. Zaimplementuj algorytm obliczania najdłuższego wspólnego podciągu dla pary ciągów elementów.
5. Korzystając z gotowego tokenizera (np. spaCy - https://spacy.io/api/tokenizer) dokonaj podziału załączonego tekstu na tokeny.
6. Stwórz 2 wersje załączonego tekstu, w których usunięto 3% losowych tokenów.
7. Oblicz długość najdłuższego podciągu wspólnych tokenów dla tych tekstów.
8. Korzystając z algorytmu z punktu 4 skonstruuj narzędzie, o działaniu podobnym do narzędzia diff, tzn. wskazującego w dwóch plikach linie, które się różnią. Na wyjściu narzędzia powinny znaleźć się elementy, które nie należą do najdłuższego wspólnego podciągu. Należy wskazać skąd dana linia pochodzi (< > - pierwszy/drugi plik) oraz numer linii w danym pliku.
9. Przedstaw wynik działania narzędzia na tekstach z punktu 6. Zwróć uwagę na dodanie znaków przejścia do nowej linii, które są usuwane w trakcie tokenizacji.

1.2 Zad 1

[1]:
```
def levenshteinDistance(text_a, text_b):
    delta = lambda x, y: 0 if x == y else 1

    len_a = len(text_a)
    len_b = len(text_b)

    dist_table = [[0 for _ in range(len_b + 1)] for _ in range(len_a + 1)]
    traceback_table = [[None for _ in range(len_b + 1)] for _ in range(len_a +␣
↪1)]

    for i in range(1, len_a + 1):
        dist_table[i][0] = i
        traceback_table[i][0] = "up"
    for j in range(1, len_b + 1):
        dist_table[0][j] = j
        traceback_table[0][j] = "left"

    for i in range(1, len_a + 1):
        for j in range(1, len_b + 1):
            x, y = text_a[i - 1], text_b[j - 1]  # current letters

            up_cost, left_cost, diag_cost = dist_table[i - 1][j] + 1,␣
↪dist_table[i][j - 1] + 1, \
                                            dist_table[i - 1][j - 1] + delta(x,␣
↪y)

            if up_cost < left_cost and up_cost < diag_cost:
                dist_table[i][j] = up_cost
                traceback_table[i][j] = "up"
            elif left_cost < diag_cost:
                dist_table[i][j] = left_cost
                traceback_table[i][j] = "left"
            else:
                dist_table[i][j] = diag_cost
                traceback_table[i][j] = "diag"

    return dist_table[len_a][len_b], traceback_table
```

1.3 Zad 2

[2]:
```
def editionVisualization(text_a, text_b):
    _, traceback_table = levenshteinDistance(text_a, text_b)
    i = len(text_a)
    j = len(text_b)
```

```
    steps = []
    curr_step = traceback_table[i][j]

    while curr_step is not None:
        steps.append(curr_step)

        if curr_step == "up":
            i -= 1
        elif curr_step == "left":
            j -= 1
        elif curr_step == "diag":
            i -= 1
            j -= 1
        else:
            raise AssertionError("Unknown Step")

        curr_step = traceback_table[i][j]

    steps.reverse()

    modified_string = text_a

    for step in steps:
        letter_a, letter_b = modified_string[i], text_b[j]
        if step == "up":
            print(f"{modified_string[:i]}\\{letter_a}/{modified_string[i+1:
↪]}\t'{letter_a}' deleted")
            modified_string = modified_string[:i] + modified_string[i+1:]
        elif step == "left":
            print(f"{modified_string[:i]}+{letter_b}+{modified_string[i:
↪]}\t'{letter_b}' added")
            modified_string = modified_string[:i] + letter_b +␣
↪modified_string[i:]
            j += 1
            i += 1
        elif step == "diag":
            if letter_a != letter_b:
                print(f"{modified_string[:i]}*{letter_b}*{modified_string[i+1:
↪]}\t'{letter_a}' -> '{letter_b}'")
                modified_string = modified_string[:i] + letter_b +␣
↪modified_string[i+1:]
            i += 1
            j += 1
```

1.4 Zad 3

[3]:
```
editionVisualization("los", "kloc")
print()

editionVisualization( "Łódź", "Lodz")
print()

editionVisualization( "kwintesencja", "quintessence")
print()

editionVisualization("ATGAATCTTACCGCCTCG", "ATGAGGCTCTGGCCCCTG")
```

```
+k+los  'k' added
klo*c*  's' -> 'c'

*L*ódź  'Ł' -> 'L'
L*o*dź  'ó' -> 'o'
Lod*z*  'ź' -> 'z'

*q*wintesencja  'k' -> 'q'
q*u*intesencja  'w' -> 'u'
quinte+s+sencja 's' added
quintessenc\j/a 'j' deleted
quintessenc*e*  'a' -> 'e'

ATGA*G*TCTTACCGCCTCG    'A' -> 'G'
ATGAG*G*CTTACCGCCTCG    'T' -> 'G'
ATGAGGCT+C+TACCGCCTCG   'C' added
ATGAGGCTCT*G*CCGCCTCG   'A' -> 'G'
ATGAGGCTCTG*G*CGCCTCG   'C' -> 'G'
ATGAGGCTCTGGC*C*CCTCG   'G' -> 'C'
ATGAGGCTCTGGCCCCT\C/G   'C' deleted
```

1.5 Zad 4

[4]:
```
def lcsequence(seq_a, seq_b):
    len_a = len(seq_a)
    len_b = len(seq_b)

    lcs = [[0 for _ in range(len_b + 1)] for _ in range(len_a + 1)]
    traceback = [[None for _ in range(len_b + 1)] for _ in range(len_a + 1)]

    for i in range(1, len_a + 1):
        lcs[i][0] = 0
        traceback[i][0] = "up"
    for j in range(1, len_b + 1):
        lcs[0][j] = 0
```

```
        traceback[0][j] = "left"

    for i in range(1, len_a + 1):
        for j in range(1, len_b + 1):
            x, y = seq_a[i - 1], seq_b[j - 1]  # current elements in sequences

            if x == y:
                lcs[i][j] = lcs[i - 1][j - 1] + 1
                traceback[i][j] = "diag"
            elif lcs[i - 1][j] >= lcs[i][j - 1]:
                lcs[i][j] = lcs[i - 1][j]
                traceback[i][j] = "up"
            else:
                lcs[i][j] = lcs[i][j - 1]
                traceback[i][j] = "left"

    return lcs[len_a][len_b], traceback
```

[5]:
```
longest, traceback = lcsequence(["A", "B", "C", "B", "D", "A", "B"], ["B", "D",␣
↪"C", "A", "B", "A"])
print(longest)
```

```
4
```

1.6 Zad 5

[6]:
```
# instalowanie pakietu w obecnym wirtualnym środowisku condy
# import sys
# !conda install -c conda-forge --yes --prefix {sys.prefix} spacy
```

```
Collecting package metadata (current_repodata.json): …working… done
Solving environment: …working… done

## Package Plan ##

  environment location: C:\Users\pawel\anaconda3\envs\hahaha

  added / updated specs:
    - spacy


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    aiohttp-3.7.4              |   py38h294d835_0         596 KB  conda-forge
    boto-2.49.0                |             py_0         838 KB  conda-forge
    boto3-1.17.57              |     pyhd8ed1ab_0          70 KB  conda-forge
```

```
    botocore-1.20.57           |     pyhd8ed1ab_0         4.6 MB  conda-forge
    brotlipy-0.7.0             |py38h294d835_1001         368 KB  conda-forge
    catalogue-2.0.3            |   py38haa244fe_0          31 KB  conda-forge
    chardet-4.0.0              |   py38haa244fe_1         224 KB  conda-forge
    cryptography-3.4.7         |   py38hd7da0ea_0         706 KB  conda-forge
    cymem-2.0.5                |   py38h885f38d_1          40 KB  conda-forge
    cython-blis-0.7.4          |   py38h347fdf6_0         5.6 MB  conda-forge
    google-crc32c-1.1.2        |   py38h554a69a_0          27 KB  conda-forge
    googleapis-common-protos-1.53.0|   py38h5b57dd5_0         125 KB  conda-
forge
    grpcio-1.37.0              |   py38he5377a8_0         2.0 MB  conda-forge
    idna-2.10                  |     pyh9f0ad1d_0          52 KB  conda-forge
    intel-openmp-2021.2.0      |     h57928b3_616         2.6 MB  conda-forge
    libblas-3.9.0              |           8_mkl          3.9 MB  conda-forge
    libcblas-3.9.0             |           8_mkl          3.9 MB  conda-forge
    liblapack-3.9.0            |           8_mkl          3.9 MB  conda-forge
    mkl-2020.4                 |     hb70f87d_311       172.4 MB  conda-forge
    multidict-5.1.0            |   py38h294d835_1          63 KB  conda-forge
    murmurhash-1.0.5           |   py38h885f38d_0          26 KB  conda-forge
    numpy-1.20.2               |   py38h09042cb_0         5.3 MB  conda-forge
    pathy-0.5.2                |     pyhd8ed1ab_0          37 KB  conda-forge
    preshed-3.0.5              |   py38h885f38d_0          97 KB  conda-forge
    protobuf-3.15.8            |   py38h885f38d_0         261 KB  conda-forge
    pydantic-1.7.3             |   py38h294d835_1         164 KB  conda-forge
    pyopenssl-20.0.1           |     pyhd8ed1ab_0          48 KB  conda-forge
    pysocks-1.7.1              |   py38haa244fe_3          28 KB  conda-forge
    pytz-2021.1                |     pyhd8ed1ab_0         239 KB  conda-forge
    requests-2.25.1            |     pyhd3deb0d_0          51 KB  conda-forge
    s3transfer-0.4.2           |     pyhd8ed1ab_0          55 KB  conda-forge
    spacy-3.0.6                |   py38h2f20550_0         9.2 MB  conda-forge
    spacy-legacy-3.0.5         |     pyhd8ed1ab_0          14 KB  conda-forge
    srsly-2.4.1                |   py38h885f38d_0         517 KB  conda-forge
    thinc-8.0.3                |   py38h2f20550_1         936 KB  conda-forge
    tqdm-4.60.0                |     pyhd8ed1ab_0          79 KB  conda-forge
    urllib3-1.26.4             |     pyhd8ed1ab_0          99 KB  conda-forge
    win_inet_pton-1.1.0        |   py38haa244fe_2           8 KB  conda-forge
    yarl-1.6.3                 |   py38h294d835_1         136 KB  conda-forge
    zlib-1.2.11                |    h62dcd97_1010         126 KB  conda-forge
    ------------------------------------------------------------
                                           Total:       219.3 MB

The following NEW packages will be INSTALLED:

  aiohttp            conda-forge/win-64::aiohttp-3.7.4-py38h294d835_0
  async-timeout      conda-forge/noarch::async-timeout-3.0.1-py_1000
  boto               conda-forge/noarch::boto-2.49.0-py_0
  boto3              conda-forge/noarch::boto3-1.17.57-pyhd8ed1ab_0
  botocore           conda-forge/noarch::botocore-1.20.57-pyhd8ed1ab_0
```

```
  brotlipy           conda-forge/win-64::brotlipy-0.7.0-py38h294d835_1001
  bz2file            conda-forge/noarch::bz2file-0.98-py_0
  cachetools         conda-forge/noarch::cachetools-4.2.1-pyhd8ed1ab_0
  catalogue          conda-forge/win-64::catalogue-2.0.3-py38haa244fe_0
  chardet            conda-forge/win-64::chardet-4.0.0-py38haa244fe_1
  click              conda-forge/noarch::click-7.1.2-pyh9f0ad1d_0
  cryptography       conda-forge/win-64::cryptography-3.4.7-py38hd7da0ea_0
  cymem              conda-forge/win-64::cymem-2.0.5-py38h885f38d_1
  cython-blis        conda-forge/win-64::cython-blis-0.7.4-py38h347fdf6_0
  dataclasses        conda-forge/noarch::dataclasses-0.8-pyhc8e2a94_1
  google-api-core    conda-forge/noarch::google-api-core-1.26.2-pyhd8ed1ab_0
  google-auth        conda-forge/noarch::google-auth-1.28.0-pyh44b312d_0
  google-cloud-core  conda-forge/noarch::google-cloud-core-1.5.0-pyhd3deb0d_0
  google-cloud-stor~ conda-forge/noarch::google-cloud-storage-1.19.0-py_0
  google-crc32c      conda-forge/win-64::google-crc32c-1.1.2-py38h554a69a_0
  google-resumable-~ conda-forge/noarch::google-resumable-
media-1.2.0-pyhd3deb0d_0
  googleapis-common~ conda-forge/win-64::googleapis-common-
protos-1.53.0-py38h5b57dd5_0
  grpcio             conda-forge/win-64::grpcio-1.37.0-py38he5377a8_0
  idna               conda-forge/noarch::idna-2.10-pyh9f0ad1d_0
  intel-openmp       conda-forge/win-64::intel-openmp-2021.2.0-h57928b3_616
  jmespath           conda-forge/noarch::jmespath-0.10.0-pyh9f0ad1d_0
  libblas            conda-forge/win-64::libblas-3.9.0-8_mkl
  libcblas           conda-forge/win-64::libcblas-3.9.0-8_mkl
  libcrc32c          conda-forge/win-64::libcrc32c-1.1.1-h0e60522_2
  liblapack          conda-forge/win-64::liblapack-3.9.0-8_mkl
  libprotobuf        conda-forge/win-64::libprotobuf-3.15.8-h7755175_0
  mkl                conda-forge/win-64::mkl-2020.4-hb70f87d_311
  multidict          conda-forge/win-64::multidict-5.1.0-py38h294d835_1
  murmurhash         conda-forge/win-64::murmurhash-1.0.5-py38h885f38d_0
  numpy              conda-forge/win-64::numpy-1.20.2-py38h09042cb_0
  pathy              conda-forge/noarch::pathy-0.5.2-pyhd8ed1ab_0
  preshed            conda-forge/win-64::preshed-3.0.5-py38h885f38d_0
  protobuf           conda-forge/win-64::protobuf-3.15.8-py38h885f38d_0
  pyasn1             conda-forge/noarch::pyasn1-0.4.8-py_0
  pyasn1-modules     conda-forge/noarch::pyasn1-modules-0.2.7-py_0
  pydantic           conda-forge/win-64::pydantic-1.7.3-py38h294d835_1
  pyopenssl          conda-forge/noarch::pyopenssl-20.0.1-pyhd8ed1ab_0
  pysocks            conda-forge/win-64::pysocks-1.7.1-py38haa244fe_3
  python_abi         conda-forge/win-64::python_abi-3.8-1_cp38
  pytz               conda-forge/noarch::pytz-2021.1-pyhd8ed1ab_0
  requests           conda-forge/noarch::requests-2.25.1-pyhd3deb0d_0
  rsa                conda-forge/noarch::rsa-4.7.2-pyh44b312d_0
  s3transfer         conda-forge/noarch::s3transfer-0.4.2-pyhd8ed1ab_0
  shellingham        conda-forge/noarch::shellingham-1.4.0-pyh44b312d_0
  smart_open         conda-forge/noarch::smart_open-2.2.1-pyh9f0ad1d_0
  spacy              conda-forge/win-64::spacy-3.0.6-py38h2f20550_0
```

```
  spacy-legacy       conda-forge/noarch::spacy-legacy-3.0.5-pyhd8ed1ab_0
  srsly              conda-forge/win-64::srsly-2.4.1-py38h885f38d_0
  thinc              conda-forge/win-64::thinc-8.0.3-py38h2f20550_1
  tqdm               conda-forge/noarch::tqdm-4.60.0-pyhd8ed1ab_0
  typer              conda-forge/noarch::typer-0.3.2-pyhd8ed1ab_0
  typing-extensions  conda-forge/noarch::typing-extensions-3.7.4.3-0
  typing_extensions  conda-forge/noarch::typing_extensions-3.7.4.3-py_0
  urllib3            conda-forge/noarch::urllib3-1.26.4-pyhd8ed1ab_0
  wasabi             conda-forge/noarch::wasabi-0.8.2-pyh44b312d_0
  win_inet_pton      conda-forge/win-64::win_inet_pton-1.1.0-py38haa244fe_2
  yarl               conda-forge/win-64::yarl-1.6.3-py38h294d835_1
  zlib               conda-forge/win-64::zlib-1.2.11-h62dcd97_1010

The following packages will be UPDATED:

  certifi            pkgs/main::certifi-2020.12.5-py38haa9~ --> conda-
forge::certifi-2020.12.5-py38haa244fe_1

The following packages will be SUPERSEDED by a higher-priority channel:

  ca-certificates    pkgs/main::ca-certificates-2021.4.13-~ --> conda-forge::ca-
certificates-2020.12.5-h5b45459_0
  openssl              pkgs/main::openssl-1.1.1k-h2bbff1b_0 --> conda-
forge::openssl-1.1.1k-h8ffe710_0



Downloading and Extracting Packages

intel-openmp-2021.2. | 2.6 MB    |            |   0%
intel-openmp-2021.2. | 2.6 MB    |            |   1%
intel-openmp-2021.2. | 2.6 MB    | #8         |  18%
intel-openmp-2021.2. | 2.6 MB    | #########2 |  92%
intel-openmp-2021.2. | 2.6 MB    | ########## | 100%

srsly-2.4.1          | 517 KB    |            |   0%
srsly-2.4.1          | 517 KB    | #########2 |  93%
srsly-2.4.1          | 517 KB    | ########## | 100%

catalogue-2.0.3      | 31 KB     |            |   0%
catalogue-2.0.3      | 31 KB     | ########## | 100%
catalogue-2.0.3      | 31 KB     | ########## | 100%

pytz-2021.1          | 239 KB    |            |   0%
pytz-2021.1          | 239 KB    | ########## | 100%
pytz-2021.1          | 239 KB    | ########## | 100%

google-crc32c-1.1.2  | 27 KB     |            |   0%
```

```
google-crc32c-1.1.2  | 27 KB     | ########## | 100%
google-crc32c-1.1.2  | 27 KB     | ########## | 100%

tqdm-4.60.0          | 79 KB     |            |   0%
tqdm-4.60.0          | 79 KB     | ########## | 100%
tqdm-4.60.0          | 79 KB     | ########## | 100%

yarl-1.6.3           | 136 KB    |            |   0%
yarl-1.6.3           | 136 KB    | ########## | 100%
yarl-1.6.3           | 136 KB    | ########## | 100%

urllib3-1.26.4       | 99 KB     |            |   0%
urllib3-1.26.4       | 99 KB     | ########## | 100%
urllib3-1.26.4       | 99 KB     | ########## | 100%

mkl-2020.4           | 172.4 MB  |            |   0%
mkl-2020.4           | 172.4 MB  |            |   0%
mkl-2020.4           | 172.4 MB  | 3          |   3%
mkl-2020.4           | 172.4 MB  | 5          |   6%
mkl-2020.4           | 172.4 MB  | 7          |   7%
mkl-2020.4           | 172.4 MB  | 8          |   9%
mkl-2020.4           | 172.4 MB  | #          |  10%
mkl-2020.4           | 172.4 MB  | #1         |  12%
mkl-2020.4           | 172.4 MB  | #3         |  14%
mkl-2020.4           | 172.4 MB  | #5         |  15%
mkl-2020.4           | 172.4 MB  | #7         |  17%
mkl-2020.4           | 172.4 MB  | #9         |  20%
mkl-2020.4           | 172.4 MB  | ##1        |  22%
mkl-2020.4           | 172.4 MB  | ##3        |  24%
mkl-2020.4           | 172.4 MB  | ##7        |  27%
mkl-2020.4           | 172.4 MB  | ##9        |  30%
mkl-2020.4           | 172.4 MB  | ###2       |  33%
mkl-2020.4           | 172.4 MB  | ###5       |  36%
mkl-2020.4           | 172.4 MB  | ###8       |  39%
mkl-2020.4           | 172.4 MB  | ####2      |  42%
mkl-2020.4           | 172.4 MB  | ####5      |  46%
mkl-2020.4           | 172.4 MB  | #####1     |  51%
mkl-2020.4           | 172.4 MB  | #####4     |  55%
mkl-2020.4           | 172.4 MB  | #####7     |  58%
mkl-2020.4           | 172.4 MB  | ######2    |  62%
mkl-2020.4           | 172.4 MB  | ######5    |  66%
mkl-2020.4           | 172.4 MB  | #######1   |  71%
mkl-2020.4           | 172.4 MB  | #######5   |  75%
mkl-2020.4           | 172.4 MB  | #######9   |  79%
mkl-2020.4           | 172.4 MB  | ########3  |  83%
mkl-2020.4           | 172.4 MB  | ########7  |  88%
mkl-2020.4           | 172.4 MB  | #########2 |  93%
mkl-2020.4           | 172.4 MB  | #########7 |  98%
```

```
mkl-2020.4           | 172.4 MB  | ########## | 100%

spacy-3.0.6          | 9.2 MB    |            |   0%
spacy-3.0.6          | 9.2 MB    | 9          |  10%
spacy-3.0.6          | 9.2 MB    | ########## | 100%
spacy-3.0.6          | 9.2 MB    | ########## | 100%

spacy-legacy-3.0.5   | 14 KB     |            |   0%
spacy-legacy-3.0.5   | 14 KB     | ########## | 100%

thinc-8.0.3          | 936 KB    |            |   0%
thinc-8.0.3          | 936 KB    | ########## | 100%
thinc-8.0.3          | 936 KB    | ########## | 100%

s3transfer-0.4.2     | 55 KB     |            |   0%
s3transfer-0.4.2     | 55 KB     | ########## | 100%

cryptography-3.4.7   | 706 KB    |            |   0%
cryptography-3.4.7   | 706 KB    | ########## | 100%
cryptography-3.4.7   | 706 KB    | ########## | 100%

botocore-1.20.57     | 4.6 MB    |            |   0%
botocore-1.20.57     | 4.6 MB    | ###1       |  31%
botocore-1.20.57     | 4.6 MB    | ########## | 100%
botocore-1.20.57     | 4.6 MB    | ########## | 100%

liblapack-3.9.0      | 3.9 MB    |            |   0%
liblapack-3.9.0      | 3.9 MB    | ###3       |  33%
liblapack-3.9.0      | 3.9 MB    | ########## | 100%
liblapack-3.9.0      | 3.9 MB    | ########## | 100%

protobuf-3.15.8      | 261 KB    |            |   0%
protobuf-3.15.8      | 261 KB    | ########## | 100%
protobuf-3.15.8      | 261 KB    | ########## | 100%

libcblas-3.9.0       | 3.9 MB    |            |   0%
libcblas-3.9.0       | 3.9 MB    | ##7        |  27%
libcblas-3.9.0       | 3.9 MB    | ########## | 100%
libcblas-3.9.0       | 3.9 MB    | ########## | 100%

pyopenssl-20.0.1     | 48 KB     |            |   0%
pyopenssl-20.0.1     | 48 KB     | ########## | 100%

aiohttp-3.7.4        | 596 KB    |            |   0%
aiohttp-3.7.4        | 596 KB    | ########## | 100%
aiohttp-3.7.4        | 596 KB    | ########## | 100%

boto-2.49.0          | 838 KB    |            |   0%
```

[26]:
```
from spacy.tokenizer import Tokenizer
from spacy.lang.pl import Polish

nlp = Polish()
```

[27]:
```
nlp
```

[27]:
```
<spacy.lang.pl.Polish at 0x1a56e294910>
```

[28]:
```
tokenizer = Tokenizer(nlp.vocab)
```

[96]:
```
PATH = "romeo-i-julia-700.txt"
with open(PATH, "r", encoding="UTF-8") as file:
    text = file.readlines()
```

1.7 Zad 6

[103]:
```
import random

def remove_random_tokens(text, path, percent):
    pipe = tokenizer.pipe(text)
    with open(path, "w", encoding="UTF-8") as file:
        for doc in pipe:
            for token in doc:
                if random.random() >= percent/100:
                    file.write(token.text_with_ws)

remove_random_tokens(text, "romeo_trimmed_1.txt", 3)
remove_random_tokens(text, "romeo_trimmed_2.txt", 3)
```

1.8 Zad 7

[104]:
```
def get_list_of_tokens(text):
    pipe = tokenizer.pipe(text)
    result = []
    for doc in pipe:
        for token in doc:
            result.append(token.text)
    return result

with open("romeo_trimmed_1.txt", "r", encoding="UTF-8") as file:
    list_of_tokens_1 = get_list_of_tokens(file.readlines())
with open("romeo_trimmed_2.txt", "r", encoding="UTF-8") as file:
    list_of_tokens_2 = get_list_of_tokens(file.readlines())
```

```
lcs_length, _ = lcsequence(list_of_tokens_1, list_of_tokens_2)
print(f'Liczba tokenów 1. tekstu: {len(list_of_tokens_1)}')
print(f'Liczba tokenów 2. tekstu: {len(list_of_tokens_2)}')
print(f'Długość najdłuższego wspólnego podciągu: {lcs_length}')
```

```
Liczba tokenów 1. tekstu: 2568
Liczba tokenów 2. tekstu: 2545
Długość najdłuższego wspólnego podciągu: 2485
```

1.9 Zad 8

[107]:
```
def diff(seq_a, seq_b):
    _, traceback = lcsequence(seq_a, seq_b)
    line_a, line_b = len(seq_a), len(seq_b)

    differences = []
    while line_a > 0 and line_b > 0:
        if traceback[line_a][line_b] == "diag":
            line_a, line_b = line_a - 1, line_b - 1
        elif traceback[line_a][line_b] == "up":
            differences.append(f'< [{line_a}] {seq_a[line_a-1]} ')
            line_a -= 1
        else:
            differences.append(f'> [{line_b}] {seq_b[line_b-1]}')
            line_b -= 1

    while line_a > 0:
        differences.append(f'< [{line_a}] {seq_a[line_a]}')
        line_a -= 1
    while line_b > 0:
        differences.append(f'> [{line_b}] {seq_b[line_b]}')
        line_b -= 1

    differences.reverse()
    return differences
```

1.10 Zad 9

[109]:
```
def get_lines(path):
    with open(path, "r", encoding="UTF-8") as file:
        return [line.strip() for line in file.readlines()]

lines_text_1 = get_lines("romeo_trimmed_1.txt")
lines_text_2 = get_lines("romeo_trimmed_2.txt")


result = diff(lines_text_1, lines_text_2)
for line in result:
```

```
    print(line)
```

```
> [12] * PARYS - młody Weroneńczyk szlachetnego rodu, krewny księcia
< [12] * PARYS - młody Weroneńczyk szlachetnego rodu, krewny
> [14] * STARZEC - brat Kapuleta
> [15] * ROMEO - syn Montekiego * MERKUCJO - krewny księcia
< [14] * STARZEC - stryjeczny brat Kapuleta
< [15] * ROMEO - syn Montekiego
< [16] * MERKUCJO - krewny księcia
> [18] * LAURENTY - ojciec
> [19] * JAN - brat z tegoż zgromadzenia
< [19] * LAURENTY - ojciec franciszkanin
< [20] * - brat z tegoż zgromadzenia
> [28] * PANI MONTEKI - małżonka
< [29] * PANI MONTEKI - małżonka Montekiego
> [50] Z łon tych dwu wzięło bowiem życie,
< [51] Z łon tych dwu wrogów wzięło bowiem życie,
> [60] Które otoczcie cierpliwymi względy,
> [61] Jest w nim co złego, my usuniem błędy…
< [61] Które otoczcie względy,
< [62] Jest w nim złego, my usuniem błędy…
> [72] / Plac publiczny. Wchodzą Samson i Grzegorz uzbrojeni tarcze i miecze. /
< [73] / Plac publiczny. Wchodzą Samson i Grzegorz uzbrojeni w tarcze i miecze.
/
> [76]
> [93]
> [111]
> [112] Rozruchać się tyle znaczy co ruszyć się z miejsca; być walecznym jest to
stać nieporuszenie: pojmuję więc, że skutkiem rozruchania się twego będzie -
drapnięcie.
< [111] Rozruchać się tyle znaczy co ruszyć się z być walecznym jest to stać
nieporuszenie: pojmuję więc, że skutkiem się twego będzie - drapnięcie.
> [117] Te psy z domu Montekich rozruchać mię mogą tylko do stania na miejscu.
Będę jak mur każdego mężczyzny i każdej kobiety z tego
< [115] Te psy z domu Montekich rozruchać mię mogą tylko do stania na miejscu.
Będę jak mur dla każdego i każdej kobiety z tego domu.
> [126]
< [125]
> [131] jest tylko między naszymi panami i między nami, ich ludźmi.
< [129] Spór jest tylko między naszymi panami i między nami, ich ludźmi.
< [135]
> [145] Nie inaczej: wtłoczę miecz w każdą po kolei. Wiadomo, że się do lwów
liczę.
< [144] Nie inaczej: wtłoczę miecz w każdą po kolei. Wiadomo, że się do liczę.
> [150] Tym lepiej, że się do zwierząt; bo gdybyś się liczył do ryb, to byłbyś
pewnie sztokfiszem. Weź no się za instrument, bo oto nadchodzi dwóch domowników
Montekiego.
< [149] Tym lepiej, że się liczysz zwierząt; bo gdybyś się liczył do ryb, to
```

byłbyś pewnie sztokfiszem. Weź no się za instrument, bo oto nadchodzi dwóch domowników Montekiego.
> [153]
> [156]
> [157] Mój giwer już dobyty: zaczep ich, ja z tyłu.
< [155] Mój giwer już dobyty: zaczep ich, ja stanę tyłu.
< [165]
> [171] Ja bym się miał bać z twojej przyczyny!
< [169] Ja bym się miał bać z twojej
> [176] Miejmy prawo sobą, niech oni zaczną.
< [174] Miejmy za sobą, niech oni zaczną.
> [191] Skrzywiłeś na nas, mości panie?
< [189] Skrzywiłeś się nas, mości panie?
< [190]
> [200] Czy nas się skrzywiłeś, mości panie?
< [199] Czy na nas się skrzywiłeś, mości panie?
< [203]
> [206] Będziemy-ż mieli prawo za sobą, jak powiem: tak jest?
< [206] Będziemy-ż mieli prawo za sobą, jak powiem: jest?
> [216] Nie, mości panie; nie skrzywiłem się na was, skrzywiłem się tak sobie.
< [216] Nie, mości panie; nie skrzywiłem się na was, tylko skrzywiłem się tak sobie.
< [217]
> [220] do Abrahama /
< [221] / do Abrahama /
> [257]
< [260] ABRAHAM
> [264] SAMSON
< [265]
> [275] Wchodzi Tybalt. /
< [276] / Wchodzi Tybalt. /
> [297] / Walczą. Nadchodzi kilku przyjaciół obu partii i się do zwady; wkrótce potem wchodzą mieszczanie z pałkami. /
< [298] / Walczą. Nadchodzi kilku przyjaciół obu partii i mieszają się do zwady; wkrótce potem wchodzą mieszczanie z pałkami. /
< [299]
> [302] Precz z Montekimi, precz z
> [303]
< [304] Precz z Montekimi, precz z Kapuletami!
< [308] KAPULET
< [310] Co za hałas? Podajcie mi długi
< [311] Mój miecz! hej!
> [309] Co za hałas? Podajcie mi długiMój miecz! hej!
> [313] Raczej kulę; co ci z miecza?
< [316] kulę; co ci z
> [321] / Wchodzą i Pani Monteki. /
< [324] / Wchodzą Monteki i Pani Monteki. /
< [328]

```
> [347] Z dłoni skrwawionych tę broń buntowniczą
< [351] Z skrwawionych tę broń buntowniczą
> [350] Domowe starcia, z marnych słów zrodzone
< [354] Domowe starcia, z marnych słów
> [353] Tak że poważni i zasługą
< [357] Tak że poważni wiekiem i zasługą
> [362] Ty, Kapulecie, pójdziesz ze mną razem;
> [363] Ty Monteki, przyjdziesz po południu
< [366] Kapulecie, pójdziesz ze mną razem;
< [367] Ty zaś, Monteki, przyjdziesz po południu
> [369] / Książę z orszakiem wychodzi. Podobnież Kapulet, Pani Kapulet,
obywatele i /
< [373] / Książę z orszakiem wychodzi. Podobnież Kapulet, Pani Kapulet, Tybalt,
obywatele i słudzy. /
> [374] Kto wszczął nową zwadę? Mów, synowcze,
< [378] Kto wszczął tę zwadę? Mów, synowcze,
> [380] Nieprzyjaciela naszego pachołcy
> [381] I wasi już się bili, kiedym nadszedł;
< [384] Nieprzyjaciela pachołcy
< [385] I już się bili, kiedym nadszedł;
> [385] Jął się wywijać nim i siec powietrze,
< [389] Jął się nim i siec powietrze,
> [394] Lecz gdzież Romeo? Widział go dzisiaj?
< [398]
< [399] Lecz gdzież Romeo? Widział żeś go dzisiaj?
> [403] W sykomorowy ów gaj, co się ciągnie
> [404] Ku południowi od naszego miasta.
> [405] już tak rano, syn się przechadzał.
< [408] W sykomorowy ów gaj, co się ciągnieKu południowi od naszego miasta.
< [409] Tam, już tak rano, syn wasz się przechadzał.
> [407] Lecz on, spostrzegłszy mię, skrył natychmiast
> [408] I w najciemniejszej ukrył się gęstwinie.
< [411] Lecz on, spostrzegłszy mię, skrył się natychmiast
< [412] I w najciemniejszej ukrył się
> [412] Nie mu w jego dumaniach
< [416] Nie przeszkadzałem mu jego dumaniach
> [414] Stroniąc od tego, co rad mnie unikał.
< [418] Stroniąc od co rad mnie unikał.
< [422]
> [419] Łzami poranną mnożącego rosę,
< [424] Łzami mnożącego rosę,
> [425] Co tchu zamykał w swoim pokoju;
> [426] Zasłaniał okna przed jasnym dnia blaskiem
< [430] Co tchu zamykał się w swoim pokoju;
< [431] Zasłaniał przed jasnym dnia blaskiem
> [429] Jeśli na to lekarstwo nie znajdzie.
< [434] Jeśli się na to lekarstwo nie znajdzie.
> [434] Szanowny stryju, znasz-że powód
```

< [439] Szanowny stryju, znasz-że powód tego?
> [441] BENWOLIO
> [443] Wybadywał żeś go jakim sposobem?
< [447] BENWOLIO
< [448] Wybadywał żeś go sposobem?
> [448] Wybadywałem i sam, przez drugich,
< [453] Wybadywałem i sam, i przez drugich,
> [453] Nim światu wonny swój kielich
< [458] Nim światu wonny swój kielich roztoczył
> [468]
> [469] Obyś w tej sprawie, co nam serce
> [470] być szczęśliwszym od nas! Pójdźmy, pani.
> [472] / Wychodzą Monteki i Pani Monteki. /
< [474] Obyś tej sprawie, co nam serce rani,
< [475] Mógł być szczęśliwszym od nas! Pójdźmy, pani.
< [477] Wychodzą Monteki i Pani Monteki. /
< [480] BENWOLIO
< [486]
> [484] BENWOLIO
< [490]
> [491] Jak nudnieWloką się chwile. Moi-ż to rodzice
< [497] Jak nudnie
< [498] Wloką się chwile. Moi-ż to rodzice
< [500]
> [496] jest. Lecz cóż tak chwile twoje dłuży?
< [504] Tak jest. Lecz cóż tak chwile twoje dłuży?
> [501] Nieposiadanie co je skraca.
< [509] Nieposiadanie tego, co je skraca.
> [505]
> [515] Jak to? brak
< [522]
< [523] Jak to? brak miłości?
> [525] Niestety! Czemuż, zdając niebianką,
< [533] Niestety! Czemuż, zdając się niebianką,
> [532] Miłość na oślep zawsze cel swój goni!
< [540] Miłość na oślep zawsze swój goni!
> [535] W grze tu nienawiść lecz i miłość.
> [536] O! wy sprzeczności niepojęte dziwa!Szorstka miłości! nienawiści tkliwa!
< [543] grze tu nienawiść wielka, lecz i miłość.
< [544] O! wy sprzeczności niepojęte dziwa!
< [545] Szorstka miłości! nienawiści tkliwa!
> [543] Czy się nie śmiejesz?
< [552] Czy się nie
> [565] Miłości nawet odbitkę działa?
< [574] Miłości nawet przez odbitkę działa?
> [575] Żółcią trawiącą i zbawczym
< [584] Żółcią i zbawczym balsamem.
> [581] BENWOLIO

```
< [590]
> [584] Gdybyś mą przyjaźń z kwitkiem zostawił.
< [593] Gdybyś mą przyjaźń z kwitkiem tak zostawił.
> [590] To nie Romeo, co z tobą.
< [599] To nie Romeo, co rozmawia z tobą.
< [603]
> [598]
> [607] to kochasz? Powiedz.
< [616] Kogóż to kochasz? Powiedz.
> [616]
> [621] Gdym to pomyślał, nimeś mi powierzył.
< [629] Gdym to pomyślał, nimeś mi
> [627] piękna.
> [628]
< [635] Jest piękna.
< [638] BENWOLIO
> [637] A właśnieś chybił. Niczym tu kołczany
< [645] A właśnieś chybił. Niczym kołczany
> [639] Pod twardą zbroją wstydliwości swojej
< [647] Pod twardą wstydliwości swojej
> [646] bogactwo, którego tak skąpi.
< [654] Całe bogactwo, którego tak skąpi.
> [660] Zbyt mądrze piękna: stąd jest głazem.
< [668] Zbyt mądrze piękna: stąd istnym jest głazem.
> [663]
```

[ ]: