TDA_exercise6

January 30, 2025

1 TEXTUAL DATA ANALYSIS Exercise 6

1.1 setup

[1]:
```
!pip3 -q install datasets openai
```

```
                                        0.0/480.6 kB
? eta -:--:--
                                    471.0/480.6
kB 13.7 MB/s eta 0:00:01
                                    480.6/480.6 kB
7.3 MB/s eta 0:00:00
                                        0.0/116.3
kB ? eta -:--:--
                                    116.3/116.3 kB
5.6 MB/s eta 0:00:00
                                        0.0/179.3
kB ? eta -:--:--
                                    179.3/179.3 kB
8.9 MB/s eta 0:00:00
                                        143.5/143.5 kB
7.6 MB/s eta 0:00:00
                                        194.8/194.8 kB
5.0 MB/s eta 0:00:00
```

```
ERROR: pip's dependency resolver does not currently take into account
all the packages that are installed. This behaviour is the source of the
following dependency conflicts.
torch 2.5.1+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cublas-cu12
12.5.3.2 which is incompatible.
torch 2.5.1+cu124 requires nvidia-cuda-cupti-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cuda-cupti-cu12
12.5.82 which is incompatible.
torch 2.5.1+cu124 requires nvidia-cuda-nvrtc-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cuda-nvrtc-cu12
12.5.82 which is incompatible.
torch 2.5.1+cu124 requires nvidia-cuda-runtime-cu12==12.4.127; platform_system
== "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-runtime-
cu12 12.5.82 which is incompatible.
torch 2.5.1+cu124 requires nvidia-cudnn-cu12==9.1.0.70; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cudnn-cu12
9.3.0.75 which is incompatible.
torch 2.5.1+cu124 requires nvidia-cufft-cu12==11.2.1.3; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cufft-cu12
11.2.3.61 which is incompatible.
torch 2.5.1+cu124 requires nvidia-curand-cu12==10.3.5.147; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-curand-cu12
10.3.6.82 which is incompatible.
torch 2.5.1+cu124 requires nvidia-cusolver-cu12==11.6.1.9; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cusolver-cu12
11.6.3.83 which is incompatible.
torch 2.5.1+cu124 requires nvidia-cusparse-cu12==12.3.1.170; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-cusparse-cu12
12.5.1.3 which is incompatible.
torch 2.5.1+cu124 requires nvidia-nvjitlink-cu12==12.4.127; platform_system ==
"Linux" and platform_machine == "x86_64", but you have nvidia-nvjitlink-cu12
12.5.82 which is incompatible.
gcsfs 2024.10.0 requires fsspec==2024.10.0, but you have fsspec 2024.9.0 which
is incompatible.
```

1.1.1 imports

[2]:
```
from google.colab import userdata # for secret access
from datasets import load_dataset
from openai import OpenAI # for the connection
from pprint import pprint
from itertools import zip_longest
import pandas as pd
import time
```

1.1.2 openai api key

[3]:
```
api_key = userdata.get('openai')
```

1.1.3 get some news texts

[4]:
```
# use the exercise 1 material as a source

!wget http://dl.turkunlp.org/TKO_8964_2023/news-en-2021.jsonl
```

```
--2025-01-30 06:53:36--  http://dl.turkunlp.org/TKO_8964_2023/news-en-2021.jsonl
Resolving dl.turkunlp.org (dl.turkunlp.org)… 195.148.30.23
Connecting to dl.turkunlp.org (dl.turkunlp.org)|195.148.30.23|:80… connected.
HTTP request sent, awaiting response… 200 OK
Length: 3385882 (3.2M) [application/octet-stream]
Saving to: ‘news-en-2021.jsonl’

news-en-2021.jsonl  100%[===================>]   3.23M  2.72MB/s    in 1.2s

2025-01-30 06:53:37 (2.72 MB/s) - ‘news-en-2021.jsonl’ saved [3385882/3385882]
```

[5]:
```
!head -n 1 news-en-2021.jsonl
```


```
{"summary": "The decisions follow a meeting of government ministers at the House
of the Estates on Thursday afternoon.", "tags": ["Kotimaan uutiset"], "text":
"Finland's government is pushing ahead with plans to introduce a Covid pass,
following a meeting of ministers at the House of the Estates in Helsinki on
Thursday afternoon. \n \"There are still many open questions that need to be
answered. At this point, it is impossible to promise that the pass will come or
when it will come,\" Prime Minister  Sanna Marin  (SDP) told the media following
the conclusion of the meeting. \n \"The government has given the green light to
the Covid pass and preparations will continue,\" Marin added. \n Minister of
Economic Affairs  Mika Lintilä  (Cen) told reporters immediately after the
meeting that there was broad agreement between the coalition parties over the
need for the certificate. \n \"It [the pass] is an important tool so that we
will not need restrictions any more,\" Lintilä said. \n The government also
```

decided at Thursday afternoon's meeting to offer coronavirus vaccines to all 12- to 15-year-olds, starting as early as next week. \n \"Fortunately, we have received an extra batch of approximately 200,000 doses of vaccine in Finland, from which these vaccinations [for 12- to 15-year-olds] can be started without interfering with other vaccination programmes,\" Marin told Yle's A-studio on Wednesday evening. \n Restrictions for bars, restaurants in spreading regions \n Furthermore, the government will reintroduce restrictions on the opening hours and operations of bars and restaurants due to the deteriorating coronavirus situation in regions considered to be in the spreading - or most serious - phase of the epidemic. \n This means that bars and restaurants in the regions of Southwest Finland, Pirkanmaa and Kymenlaakso, as well as the Helsinki metropolitan area, will have to adapt to new regulations that are due to take effect from Sunday. \n The measures include the opening hours of bars being limited to between 7am and 10pm, while restaurants can stay open one hour later. A ban on karaoke and dancing indoors has also been reintroduced. \n There will be no changes to the opening hours of bars or restaurants in regions considered to be in the acceleration phase of the pandemic. \n Changes to external border traffic \n The government has also decided to make changes to the restrictions on Finland's external border traffic, according to the Ministry of the Interior. External border traffic refers to traffic between Finland and countries not belonging to the Schengen area. \n The regulations currently in effect will be amended, beginning from 9 August, so that entry restrictions are removed for Ukrainian residents traveling to Finland from Ukraine. \n Restrictions on entry will be restored for residents of Azerbaijan, South Korea, Japan, Moldova, Serbia and Singapore travelling from these countries to Finland. \n If a person arriving from the above-mentioned countries has not received a full series of vaccinations, the permitted entry criteria are a resident returning to Finland or other EU or Schengen countries, transit of regular scheduled flight traffic at the airport, or other essential reasons. \n A person can travel to Finland from any country by presenting an acceptable certificate of the full vaccination series. \n These new regulations aside, the restrictions that entered into force on 19 July still apply. \n The latest restrictions are in effect until 22 August. \n Protesters demand \"same rules for all\" \n A small but vocal group of protestors, representing cultural sector workers, gathered near the House of the Estates while the government meeting was ongoing to demonstrate against coronavirus restrictions, demanding a fairer distribution of measures. \n Restrictions have hit especially hard on the cultural and event industry, with many workers in the sector unable to work for the past year and a half. At the same time, the protestors pointed out, shopping malls have been allowed to operate as normal. \n \"Same rules for all,\" the protesters chanted.", "timestamp": "2021-08-05T14:58:29", "title": "Government opens vaccinations for 12-15-year-olds, gives green light to Covid pass", "url": "https://yle.fi/uutiset/12048911"}

[6]:
```
dataset = load_dataset('json', data_files='news-en-2021.jsonl')
print(dataset)
```

```
Generating train split: 0 examples [00:00, ? examples/s]
```

```
DatasetDict({
    train: Dataset({
        features: ['summary', 'tags', 'text', 'timestamp', 'title', 'url'],
        num_rows: 1059
    })
})
```

[7]:
```
print(dataset['train'][100]['text'][:300]) # only 300 first
```

```
A feature allowing gamblers to limit their losses will be installed on slot
machines operated by Finland's state gambling monopoly Veikkaus from September.
 Loss limits allow players to place a cap on the amount of money they could
potentially lose over a given period of time.
 The move is the lat
```

[8]:
```
articles = dataset['train'].select(range(10))
len(articles)
```

[8]:
```
10
```

[9]:
```
# ensure only len of 300 words

def get_words(text,n) -> str:
    words = text.split()
    return " ".join(words[:n])
```

[10]:
```
fit_art = []

for a in articles:
  msg = get_words(a['text'],300)
  print(msg)
  print(f"count of words: {len(msg.split())}")
  fit_art.append(msg)
```

```
Finland's government is pushing ahead with plans to introduce a Covid pass,
following a meeting of ministers at the House of the Estates in Helsinki on
Thursday afternoon. "There are still many open questions that need to be
answered. At this point, it is impossible to promise that the pass will come or
when it will come," Prime Minister Sanna Marin (SDP) told the media following
the conclusion of the meeting. "The government has given the green light to the
Covid pass and preparations will continue," Marin added. Minister of Economic
Affairs Mika Lintilä (Cen) told reporters immediately after the meeting that
there was broad agreement between the coalition parties over the need for the
certificate. "It [the pass] is an important tool so that we will not need
restrictions any more," Lintilä said. The government also decided at Thursday
afternoon's meeting to offer coronavirus vaccines to all 12- to 15-year-olds,
starting as early as next week. "Fortunately, we have received an extra batch of
approximately 200,000 doses of vaccine in Finland, from which these vaccinations
```

[for 12- to 15-year-olds] can be started without interfering with other vaccination programmes," Marin told Yle's A-studio on Wednesday evening. Restrictions for bars, restaurants in spreading regions Furthermore, the government will reintroduce restrictions on the opening hours and operations of bars and restaurants due to the deteriorating coronavirus situation in regions considered to be in the spreading - or most serious - phase of the epidemic. This means that bars and restaurants in the regions of Southwest Finland, Pirkanmaa and Kymenlaakso, as well as the Helsinki metropolitan area, will have to adapt to new regulations that are due to take effect from Sunday. The measures include the opening hours of bars being limited to between 7am and 10pm, while restaurants can stay open one
count of words: 300
Rental fees for non-subsidised apartments rose across most of Finland during April to June, compared to the same period a year ago, according to data from Statistics Finland. On average, rents rose by 0.9 percent during that period across the country. Timo Metsola , board chair of rental agency Vuokraturva, attributed the increase to growing demand, saying that competition clearly intensified for the most desirable properties. The sharpest rise in apartment rents during the April-June period was seen in the city of Turku, where costs rose by 1.6 percent, with the city of Tampere seeing an increase of 1.4 percent. Meanwhile in the Greater Helsinki area, rents ticked up by 0.9 percent. Among the country's municipal centres, the town of Mikkeli was the only area which saw rental fees decline. Still priciest in Helsinki area The number-crunching agency reported that the median rent for a studio apartment in central Helsinki was 809 euros per month, while they stood at 583 euros in downtown Tampere and 515 euros in the centre of Oulu. Meanwhile, larger homes, for example three-room apartments cost a median of 1,634 in downtown Helsinki, 1,070 euros in Tampere and 940 euros in Oulu. Metsola said that new students began searching for places to live towards the end of July, but noted that there were fewer young flat hunters this summer than in years before the coronavirus crisis. "The pandemic has increased distance learning, and only some educational institutions are planning to conduct in-person teaching when the academic year gets started this autumn," he said, explaining that fewer new students will start renting, compared to pre-crisis years.
count of words: 269
Double Olympic ice hockey bronze medallist Emma Terho has been elected chair of the International Olympic Committee Athletes' Commission (IOC AC). The Athletes' Commission is a majority elected body that serves as the mediator between athletes and the IOC. Terho will replace swimming champion Kirsty Coventry from Zimbabwe as AC chair after being elected ahead of Russian two-time Olympic pole vault champion, Yelena Isinbayeva . Terho will be following in the footsteps of Finnish Olympian and former sailor Peter Tallberg , who was involved in the Olympic movement for more than 40 years. "Finland's Peter Tallberg was among the most proactive members of the Athletes' Commission and ensured that the voices of athletes would play a bigger role in the IOC. Both the role and workload of the commission have grown exponentially in recent years, which is a good thing because athletes are the most important part of the Games," the 39-year-old Terho said in an Olympic Committee press release. "That is why it is both a

great and huge responsibility to be in this role and also to continue Peter's legacy. I also believe that this is a valuable post and an opportunity to bring forward the values that I consider important as a Finn," she added. Terho has been involved in the Athletes' Commission and the IOC since the 2018 Winter Olympics in Pyeongchang.
count of words: 226
The Regional State Administrative Agency of Southern Finland (Avi) has announced that coronavirus restrictions in the Helsinki Metropolitan Area will not be tightened, or at least not yet. In a statement released on Friday morning, Avi said that it has requested expert analysis on the development of the coronavirus epidemic in the capital region and any new measures will be brought into force after changes to the opening hours and operations of bars and restaurants begin from Sunday 8 August. Therefore, any possible tightening of restrictions will be discussed and decided upon next week, Avi added. New restrictions for Kymenlaakso region However, there are updated restrictions for the Kymenlaakso region, which entered the spreading - or most serious - phase of the epidemic on 4 August. The new measures mostly affect how many people can attend public and private events in the region, with organisers responsible for ensuring that participants can maintain a distance of two metres from each other at all times, both for indoor and outdoor events. The regulations will come into force on 14 August.
count of words: 178
A small percentage of Finland's healthcare workers are yet to receive a coronavirus vaccine, an Yle survey has revealed. This means that unvaccinated carers are currently tending to elderly patients in some care homes across the country, despite a law outlining that the use of vaccinated staff is prioritised when it comes to the treatment of vulnerable groups. However, unlike with influenza, the Communicable Diseases Act does not require healthcare staff to get the coronavirus vaccine. Yle emailed a questionnaire enquiring into the vaccination status of staff at all 20 hospital districts. Replies were received from 13 districts, and the survey found that between 85 and 95 percent of doctors and nurses have been vaccinated with at least one dose of a coronavirus vaccine. The figure is however just an estimate as precise health information is covered by patient confidentiality, even in the case of healthcare workers. The hospital district estimates were based on voluntary reports. The Helsinki and Uusimaa Hospital District (HUS), for example, reported that the capital's caregivers' vaccination coverage is estimated at 95 percent, with that figure expected to rise. "Our goal is more than 95 percent vaccination coverage," HUS Communications Manager Niina Kauppinen said. Unvaccinated younger workers causing concern Although the number of unvaccinated staff is low and is most commonly detected among younger workers, the issue is causing concern for the Kymenlaakso Hospital District, or Kymsote, according to Chief Administrative Officer Kari Kristeri . "They are [typically] socially active young adults who get the virus from trips and gatherings," Kristeri said. Risto Pietikäinen , Kymsote's chief physician, said that although some of the workers tend to patients in their homes, chains of infections among the elderly population have been rare. The infectious diseases specialist added that young people have found themselves at the tail of

count of words: 300
Dozens of coronavirus infections have been diagnosed among a group of berry pickers from Thailand working in Lapland. Mass testing organised by the Lapland Hospital District of 110 seasonal workers in the region found that 40 had contracted the virus. The testing was carried out after news reports earlier this week revealed that 44 berry pickers in the Kainuu region had tested positive for the virus. Markku Broas , Chief Physician of Infectious Diseases of the Lapland Hospital District (LSHP) told Yle that the workers had been tested as required by regulations when they entered Finland and on the third day after entry. He said that this case proves that no system is one hundred percent effective. "Even if the test is done on the third day, the virus can appear and become contagious later. It has apparently happened here that the infection has come after the third day," he said. Mass testing set to continue in Lapland Broas pointed out that as the berry pickers are living and working together in the same groups, there is a high risk that the virus will be transmitted from one person to another. "Especially in the case of the Delta virus variant, this is the situation," he said. A total of 250 berry pickers are set to be tested in the Lapland region over Friday and the weekend. "We also go through with the companies very carefully about what their isolation and quarantine practices are. So far, the companies have done a good job in terms of safety. Preliminary testing has been done and bubble practices have been pursued," Broas said. The infected berry pickers have been isolated and are not currently working. Quarantined workers are tested every two days. "If they are healthy, they can pick berries. But if a negative
count of words: 300
The City of Lahti, Europe's Green Capital of 2021, has brought a climate action art installation entitled the Indisputable Case of Emergency (ICE) to the centre of Helsinki. The nine large blocks containing 13 cubes of ice each are intended to serve as a reminder of the ongoing climate crisis, as the melting ice symbolises the rapid progress of climate change and subsequent rise in sea levels, which will pose a severe threat to coastal cities over the coming decades. The installation, unveiled on the capital's Kansalaistori square at noon on Friday, also features infographics related to climate change and an ICE musical composed by Cecilia Damström , which premiered at the opening ceremony. "Ever since childhood I have wanted to make the world a better place. I feel that I can influence it by composing as well. The work is inspired by the melting of ice, the music depicts winter and ice," Damström said of the musical arrangement. The ice cubes were initially formed in the Ylläs region of Lapland last winter, after which they were placed into cold storage in Kivikko, Helsinki, the installation's architect Erkko Aarti told Yle. "The theme, climate change, is a massive one. For a long time, we wondered how to put it into practice. Then we ended up with ice and wood because they are environmentally friendly and durable materials. The goal was an impressive piece of work that provokes discussion," Aarti explained. The installation also includes a wooden pavilion, where visitors can learn more about Lahti's status as European Green Capital for 2021. The pavilion will be open until August 15, but the ice cubes are not expected to last that long. "The ice cubes seem to be attracting attention, but this is a huge space where everything looks small. But the

count of words: 300
The Union of Upper Secondary School Students has issued a demand that pop-up coronavirus vaccination points be set up in high schools in an effort to secure on-site learning and increase the vaccine coverage of young people. The pop-up stations could facilitate the vaccination of students without an appointment, for example during a break or a gap between lessons, the union said, adding that it believes vaccine appointments should also be an acceptable reason to be away from class. "We see that this would be a much needed opportunity with a low threshold to increase vaccine coverage for young people, seeing that a large number of young people have not yet received the first dose of the vaccine," President of the Association of High School Students, Emilia Uljas said. The opportunity to get the vaccine spontaneously at school would make it easier for younger groups to get vaccinated, Uljas said, adding that "young people would then not need to plan their schedule around a vaccination appointment." Coronavirus infections are currently on the rise, especially among young adults. However, Uljas told Yle it is unfair that discussions around the surge in cases typically point the finger of blame at younger age groups. "Young people were left last in the vaccination queue and now people complain that not all young people have gotten the vaccine. Of course they have not, as they haven't even had the opportunity to get the vaccine for as long as older people have been able to," the union president said, but added that she is pleased to see people taking a stand in favour of contact teaching since the emergence of rumours that the new school year will see distance learning resume. "Young people have suffered terribly from this pandemic. Being able to go to school is
count of words: 300
A gang of three masked individuals have stolen or attempted to steal pay terminals at four locations in different parts of Finland over the past week. Police say they are still working to identify the three suspects. The first theft took place on Wednesday morning at a petrol station in Tervakoski near the municipality of Janakkala in southern Finland. Häme police said that a vehicle previously stolen from the Parola area of Hattula was used in the robbery. Story continues after the photo. Häme police say the car used by the gang was a 2013 brown-coloured Kia Ceed. Poliisi The car is a 2013 brown Kia Ceed station wagon with a white ski box on the roof, as seen in the above image from surveillance footage captured at the scene. The gang made their getaway after obtaining a small sum of money. Gang moves east The following evening, Thursday, police believe the gang struck again in southeast Finland, hitting petrol stations in Virolahti, Miehikkälä and Luumäki. In both Miehikkälä and Luumäki, the gang's attempts to steal from the machines were unsuccessful. They were interrupted in Miehikkälä and quickly fled the scene, but police are unsure why the Luumäki attempt failed. "The machine was broken, but for one reason or another it was not taken," Chief Inspector Jarmo Rahkola from the Southeast Finland Police Department said. The department confirmed on Friday that the gang had successfully stolen a machine from an ABC filling station in Virolahti, Kymenlaakso. The machine has not yet been recovered and the amount of cash that was in the terminal at the time of the robbery is not currently known. Footage of the gang was captured by surveillance cameras in the Kymenlaakso region as well, and police believe the machine was removed from the Virolahti station by

```
count of words: 300
Helsinki District Court has sentenced a 43-year-old taxi driver to six years and
three months in prison for a total of 18 different crimes, most of which were
related to the drugging and raping of four female victims. The court heard that
from 2019 the defendant had been driving a car with a taxi sign on the roof and
picking up unaccompanied women from taxi ranks. Police believe he provided his
victims with alcohol and possibly other intoxicants while pretending to bring
them to their intended destination. However, he instead brought the passengers
to an apartment outside Helsinki, where the court heard he raped the women. Some
of the sexual offences also took place inside the taxi. One of the victims was
only 16 years old. The man was convicted on charges of rape, aggravated rape and
coercion into sexual activity, as well as other offences including deprivation
of liberty, illicit observation, distribution of sexually obscene images, drug
offences, and violation of an order banning the use of a psychoactive substance.
The defendant had denied most of the charges. More suspected victims have
emerged during the course of the preliminary investigation , and police have
already brought new charges against the man for suspected sexual offences with
investigations into other cases continuing. Friday's ruling by the Helsinki
District Court only concerned the earlier cases.
count of words: 224
```

[11]:
```
fit_art[0]
```

[11]:
```
'Finland\'s government is pushing ahead with plans to introduce a Covid pass,
following a meeting of ministers at the House of the Estates in Helsinki on
Thursday afternoon. "There are still many open questions that need to be
answered. At this point, it is impossible to promise that the pass will come or
when it will come," Prime Minister Sanna Marin (SDP) told the media following
the conclusion of the meeting. "The government has given the green light to the
Covid pass and preparations will continue," Marin added. Minister of Economic
Affairs Mika Lintilä (Cen) told reporters immediately after the meeting that
there was broad agreement between the coalition parties over the need for the
certificate. "It [the pass] is an important tool so that we will not need
restrictions any more," Lintilä said. The government also decided at Thursday
afternoon\'s meeting to offer coronavirus vaccines to all 12- to 15-year-olds,
starting as early as next week. "Fortunately, we have received an extra batch of
approximately 200,000 doses of vaccine in Finland, from which these vaccinations
[for 12- to 15-year-olds] can be started without interfering with other
vaccination programmes," Marin told Yle\'s A-studio on Wednesday evening.
Restrictions for bars, restaurants in spreading regions Furthermore, the
government will reintroduce restrictions on the opening hours and operations of
bars and restaurants due to the deteriorating coronavirus situation in regions
considered to be in the spreading - or most serious - phase of the epidemic.
This means that bars and restaurants in the regions of Southwest Finland,
Pirkanmaa and Kymenlaakso, as well as the Helsinki metropolitan area, will have
to adapt to new regulations that are due to take effect from Sunday. The
measures include the opening hours of bars being limited to between 7am and
```

```
10pm, while restaurants can stay open one'
```

1.1.4 connect to OpenAI API

[12]:
```
client = OpenAI(
    api_key=api_key
)
```

1.1.5 TEST CALL

Write code to access the API, and retrieve results. Debug this with one short request until you know it works!

test with single word

[13]:
```
chat_completion = client.chat.completions.create(
    messages=[
        {
            "role": "user",
            "content": "Say this is a test",
        }
    ],
    model="gpt-4o-mini", # MUST BE 4o-mini >:(
)
```

[14]:
```
print(chat_completion,end="\n\n")
print(chat_completion.choices, end="\n\n")
print(chat_completion.choices[0].message.content, end="\n\n")
```

```
ChatCompletion(id='chatcmpl-AvIii9MFhMYY2tNzY7P0LDBsO1tcf',
choices=[Choice(finish_reason='stop', index=0, logprobs=None,
message=ChatCompletionMessage(content='This is a test. How can I assist you
further?', refusal=None, role='assistant', audio=None, function_call=None,
tool_calls=None))], created=1738220020, model='gpt-4o-mini-2024-07-18',
object='chat.completion', service_tier='default',
system_fingerprint='fp_72ed7ab54c', usage=CompletionUsage(completion_tokens=13,
prompt_tokens=12, total_tokens=25,
completion_tokens_details=CompletionTokensDetails(accepted_prediction_tokens=0,
audio_tokens=0, reasoning_tokens=0, rejected_prediction_tokens=0),
prompt_tokens_details=PromptTokensDetails(audio_tokens=0, cached_tokens=0)))

[Choice(finish_reason='stop', index=0, logprobs=None,
message=ChatCompletionMessage(content='This is a test. How can I assist you
further?', refusal=None, role='assistant', audio=None, function_call=None,
tool_calls=None))]

This is a test. How can I assist you further?
```

1.2 TASKS

Write a prompt to extract named entities from a news article. Your prompt can either focus on one entity type (in this case, discard other types), or extract multiple entity types in the same prompt. Do not repeat the extraction separately for each entity type (let's save quota here!).

1.2.1 algos

actual function to call the api

[15]:
```
def api_call(message:str) -> str:
  chat_completion = client.chat.completions.create(
      messages=[
          {
              "role": "user",
              "content": f"Make NER analysis using BIO tags, reply only *word:tag* and newline format: {message}",
          }
      ],
      model="gpt-4o-mini", # MUST BE 4o-mini >:(
  )
  return chat_completion.choices[0].message.content
```

harvest the NER tags

[16]:
```
def harvester(message:str) -> list:
  NER_TAGS = []
  message = message.replace(" ", "") # we know this is the format inside the␣
↪api_call function
  message = message.splitlines()
  for i,m in enumerate(message):
    ml = m.split(":")
    if len(ml) < 2:
      continue
    NER_TAGS.append((ml[0],ml[1]))
  return NER_TAGS
```

formatted printing

[17]:
```
def formatted_printing(harvested_list:list) -> None:
  for row in harvested_list:
    print(f"{row[0]:<20} \t {row[1]}")
```

time it to var

[18]:
```
def time_it_to_var(st:int, et:int) -> int:
  exe_time = et - st
  return exe_time
```

1.2.2 TEST NER

[19]:
```
st = time.time()

reply_message = api_call(fit_art[0])

et = time.time()
```

[20]:
```
exe_time = time_it_to_var(st,et)
print(f'runtime of one call was {int(exe_time // 60)} minutes and {int(exe_time␣
↪% 60)} seconds')
```

```
runtime of one call was 0 minutes and 23 seconds
```

[21]:
```
harvested_tags = harvester(reply_message)
```

[22]:
```
formatted_printing(harvested_tags)
```

```
Finland                 B-LOC
's                      O
government              O
is                      O
pushing                 O
ahead                   O
with                    O
plans                   O
to                      O
introduce               O
a                       O
Covid                   B-MISC
pass                    O
following               O
a                       O
meeting                 O
of                      O
ministers               O
at                      O
the                     O
House                   B-ORG
of                      O
the                     O
Estates                 I-ORG
in                      O
Helsinki                B-LOC
on                      O
Thursday                B-DATE
afternoon               O
.                       O
```

```
"There                  O
are                     O
still                   O
many                    O
open                    O
questions               O
that                    O
need                    O
to                      O
be                      O
answered                O
.                       O
At                      O
this                    O
point                   O
,                       O
it                      O
is                      O
impossible              O
to                      O
promise                 O
that                    O
the                     O
pass                    O
will                    O
come                    O
or                      O
when                    O
it                      O
will                    O
come                    O
,                       O
"                       O
Prime                   B-PER
Minister                I-PER
Sanna                   B-PER
Marin                   I-PER
(                       O
SDP                     B-ORG
)                       O
told                    O
the                     O
media                   O
following               O
the                     O
conclusion              O
of                      O
the                     O
```

```
meeting                 O
.                       O
"The                    O
government              O
has                     O
given                   O
the                     O
green                   O
light                   O
to                      O
the                     O
Covid                   B-MISC
pass                    O
and                     O
preparations            O
will                    O
continue                O
,                       O
"                       O
Marin                   B-PER
added                   O
.                       O
Minister                B-PER
of                      O
Economic                B-MISC
Affairs                 I-MISC
Mika                    B-PER
Lintilä                 I-PER
(                       O
Cen                     B-ORG
)                       O
told                    O
reporters               O
immediately             O
after                   O
the                     O
meeting                 O
that                    O
there                   O
was                     O
broad                   O
agreement               O
between                 O
the                     O
coalition               O
parties                 O
over                    O
the                     O
```

```
need                    O
for                     O
the                     O
certificate             O
.                       O
"It                     O
[                       O
the                     O
pass                    O
]                       O
is                      O
an                      O
important               O
tool                    O
so                      O
that                    O
we                      O
will                    O
not                     O
need                    O
restrictions            O
any                     O
more                    O
,                       O
"                       O
Lintilä                 B-PER
said                    O
.                       O
The                     O
government              O
also                    O
decided                 O
at                      O
Thursday                B-DATE
afternoon               O
's                      O
meeting                 O
to                      O
offer                   O
coronavirus             B-MISC
vaccines                O
to                      O
all                     O
12-                     B-AGE
to                      O
15                      I-AGE
-                       O
year                    O
```

```
-                       O
olds                    O
,                       O
starting                O
as                      O
early                   O
as                      O
next                    O
week                    O
.                       O
"                       O
Fortunately             O
,                       O
we                      O
have                    O
received                O
an                      O
extra                   O
batch                   O
of                      O
approximately           O
200,000                 O
doses                   O
of                      O
vaccine                 O
in                      O
Finland                 B-LOC
,                       O
from                    O
which                   O
these                   O
vaccinations            O
[                       O
for                     O
12-                     B-AGE
to                      O
15                      I-AGE
-                       O
year                    O
-                       O
olds                    O
]                       O
can                     O
be                      O
started                 O
without                 O
interfering             O
with                    O
```

```
other                   O
vaccination             O
programmes              O
,                       O
"                       O
Marin                   B-PER
told                    O
Yle                     B-ORG
's                      O
A-studio                B-ORG
on                      O
Wednesday               B-DATE
evening                 O
.                       O
Restrictions            O
for                     O
bars                    O
,                       O
restaurants             O
in                      O
spreading               O
regions                 O
Furthermore             O
,                       O
the                     O
government              O
will                    O
reintroduce             O
restrictions            O
on                      O
the                     O
opening                 O
hours                   O
and                     O
operations              O
of                      O
bars                    O
and                     O
restaurants             O
due                     O
to                      O
the                     O
deteriorating           O
coronavirus             B-MISC
situation               O
in                      O
regions                 O
considered              O
```

```
to                      O
be                      O
in                      O
the                     O
spreading               O
-                       O
or                      O
most                    O
serious                 O
-                       O
phase                   O
of                      O
the                     O
epidemic                O
.                       O
This                    O
means                   O
that                    O
bars                    O
and                     O
restaurants             O
in                      O
the                     O
regions                 O
of                      O
Southwest               B-LOC
Finland                 I-LOC
,                       O
Pirkanmaa               B-LOC
and                     O
Kymenlaakso             B-LOC
,                       O
as                      O
well                    O
as                      O
the                     O
Helsinki                B-LOC
metropolitan            O
area                    O
,                       O
will                    O
have                    O
to                      O
adapt                   O
to                      O
new                     O
regulations             O
that                    O
```

```
are                     O
due                     O
to                      O
take                    O
effect                  O
from                    O
Sunday                  B-DATE
.                       O
The                     O
measures                O
include                 O
the                     O
opening                 O
hours                   O
of                      O
bars                    O
being                   O
limited                 O
to                      O
between                 O
7am                     B-TIME
and                     O
10pm                    B-TIME
,                       O
while                   O
restaurants             O
can                     O
stay                    O
open                    O
one                     O
```

works...

Take 10 news articles from the same news data collection (Finnish or English), verify that the selected articles are not extremely long (should be less than 300 words each), discard or truncate longer documents.

[23]:
```
# located inside fit_art variable

print(f'count of articles: {len(fit_art)}') # count of articles
for i,a in enumerate(fit_art):
  print(f'count of words in article {i+1}: {len(a.split())}') # count words
```

```
count of articles: 10
count of words in article 1: 300
count of words in article 2: 269
count of words in article 3: 226
count of words in article 4: 178
count of words in article 5: 300
```

```
count of words in article 6: 300
count of words in article 7: 300
count of words in article 8: 300
count of words in article 9: 300
count of words in article 10: 224
```

collect tags for all 10 articles

run time for one call was

[24]:
```
print(f'runtime of one call was {int(exe_time // 60)} minutes and {int(exe_time␣
  ↪% 60)} seconds')
print(f'so for {len(fit_art)} calls we can expect 10*exe_time')
ten_exe = 10 * exe_time
print(f'runtime for {len(fit_art)} calls estimated {int(ten_exe // 60)} minutes␣
  ↪and {int(ten_exe % 60)} seconds', end="\n\n")
```

```
runtime of one call was 0 minutes and 23 seconds
so for 10 calls we can expect 10*exe_time
runtime for 10 calls estimated 3 minutes and 50 seconds
```

[25]:
```
%%time

ner_tags_for_10_articles = {}

for i,a in enumerate(fit_art): # remember this is the truncated list
  reply = api_call(a) # call OpenAI API with the message
  harvested_tags = harvester(reply) # harvest the tags
  ner_tags_for_10_articles[f'article {str(i+1)}'] = harvested_tags
```

```
CPU times: user 1.33 s, sys: 226 ms, total: 1.55 s
Wall time: 5min 2s
```

[26]:
```
its = []

for v in ner_tags_for_10_articles.values():
  its.append(v)
```

[29]:
```
df = pd.DataFrame([x for x in ner_tags_for_10_articles.values()])
df = df.T
df.columns=[y for y in ner_tags_for_10_articles.keys()]
```

[33]:
```
df.head(50)
```

[33]:

	article 1	article 2	article 3	\
0	(Finland, B-LOC)	(Rental, O)	(Double, O)	
1	('s, O)	(fees, O)	(Olympic, O)	
2	(government, O)	(for, O)	(ice, O)	

3	(is, O)	(non-subsidised, O)	(hockey, O)
4	(pushing, O)	(apartments, O)	(bronze, O)
5	(ahead, O)	(rose, O)	(medallist, O)
6	(with, O)	(across, O)	(Emma, B-PER)
7	(plans, O)	(most, O)	(Terho, I-PER)
8	(to, O)	(of, O)	(has, O)
9	(introduce, O)	(Finland, LOC)	(been, O)
10	(a, O)	(during, O)	(elected, O)
11	(Covid, B-MISC)	(April, O)	(chair, O)
12	(pass, O)	(to, O)	(of, O)
13	(,, O)	(June, O)	(the, O)
14	(following, O)	(,, O)	(International, B-ORG)
15	(a, O)	(compared, O)	(Olympic, I-ORG)
16	(meeting, O)	(to, O)	(Committee, I-ORG)
17	(of, O)	(the, O)	(Athletes', B-ORG)
18	(ministers, O)	(same, O)	(Commission, I-ORG)
19	(at, O)	(period, O)	((IOC, O)
20	(the, O)	(a, O)	(AC), O)
21	(House, B-ORG)	(year, O)	(., O)
22	(of, O)	(ago, O)	(The, O)
23	(the, O)	(,, O)	(Athletes', O)
24	(Estates, I-ORG)	(according, O)	(Commission, O)
25	(in, O)	(to, O)	(is, O)
26	(Helsinki, B-LOC)	(data, O)	(a, O)
27	(on, O)	(from, O)	(majority, O)
28	(Thursday, B-DATE)	(Statistics, ORG)	(elected, O)
29	(afternoon, O)	(Finland, LOC)	(body, O)
30	(., O)	(., O)	(that, O)
31	("There, O)	(On, O)	(serves, O)
32	(are, O)	(average, O)	(as, O)
33	(still, O)	(,, O)	(the, O)
34	(many, O)	(rents, O)	(mediator, O)
35	(open, O)	(rose, O)	(between, O)
36	(questions, O)	(by, O)	(athletes, O)
37	(that, O)	(0.9, O)	(and, O)
38	(need, O)	(percent, O)	(the, O)
39	(to, O)	(during, O)	(IOC, O)
40	(be, O)	(that, O)	(., O)
41	(answered, O)	(period, O)	(Terho, B-PER)
42	(., O)	(across, O)	(will, O)
43	(At, O)	(the, O)	(replace, O)
44	(this, O)	(country, O)	(swimming, O)
45	(point, O)	(., O)	(champion, O)
46	(,, O)	(Timo, PERSON)	(Kirsty, B-PER)
47	(it, O)	(Metsola, PERSON)	(Coventry, I-PER)
48	(is, O)	(,, O)	(from, O)
49	(impossible, O)	(board, O)	(Zimbabwe, B-LOC)

	article 4	article 5	article 6	\
0	(The, O)	(A, O)	(Dozens, O)	
1	(Regional, B-ORG)	(small, O)	(of, O)	
2	(State, I-ORG)	(percentage, O)	(coronavirus, B-ILL)	
3	(Administrative, I-ORG)	(of, O)	(infections, I-ILL)	
4	(Agency, I-ORG)	(Finland, B-LOC)	(have, O)	
5	(of, I-ORG)	('s, O)	(been, O)	
6	(Southern, B-LOC)	(healthcare, O)	(diagnosed, O)	
7	(Finland, I-LOC)	(workers, O)	(among, O)	
8	((Avi), B-ORG)	(are, O)	(a, O)	
9	(has, O)	(yet, O)	(group, O)	
10	(announced, O)	(to, O)	(of, O)	
11	(that, O)	(receive, O)	(berry, B-ORG)	
12	(coronavirus, B-MISC)	(a, O)	(pickers, I-ORG)	
13	(restrictions, O)	(coronavirus, B-MISC)	(from, O)	
14	(in, O)	(vaccine, O)	(Thailand, B-LOC)	
15	(the, O)	(,, O)	(working, O)	
16	(Helsinki, B-LOC)	(an, O)	(in, O)	
17	(Metropolitan, I-LOC)	(Yle, B-ORG)	(Lapland, B-LOC)	
18	(Area, I-LOC)	(survey, O)	(., O)	
19	(will, O)	(has, O)	(Mass, O)	
20	(not, O)	(revealed, O)	(testing, O)	
21	(be, O)	(., O)	(organised, O)	
22	(tightened, O)	(This, O)	(by, O)	
23	(or, O)	(means, O)	(the, O)	
24	(at, O)	(that, O)	(Lapland, B-LOC)	
25	(least, O)	(unvaccinated, O)	(Hospital, I-ORG)	
26	(not, O)	(carers, O)	(District, I-ORG)	
27	(yet, O)	(are, O)	(of, O)	
28	(In, O)	(currently, O)	(110, O)	
29	(a, O)	(tending, O)	(seasonal, O)	
30	(statement, O)	(to, O)	(workers, O)	
31	(released, O)	(elderly, O)	(in, O)	
32	(on, O)	(patients, O)	(the, O)	
33	(Friday, B-DATE)	(in, O)	(region, O)	
34	(morning, I-DATE)	(some, O)	(found, O)	
35	(Avi, B-ORG)	(care, O)	(that, O)	
36	(said, O)	(homes, O)	(40, O)	
37	(that, O)	(across, O)	(had, O)	
38	(it, O)	(the, O)	(contracted, O)	
39	(has, O)	(country, O)	(the, O)	
40	(requested, O)	(,, O)	(virus, B-ILL)	
41	(expert, B-MISC)	(despite, O)	(., O)	
42	(analysis, I-MISC)	(a, O)	(The, O)	
43	(on, O)	(law, O)	(testing, O)	
44	(the, O)	(outlining, O)	(was, O)	

45	(development, O)	(that, O)	(carried, O)
46	(of, O)	(the, O)	(out, O)
47	(the, O)	(use, O)	(after, O)
48	(coronavirus, B-MISC)	(of, O)	(news, O)
49	(epidemic, I-MISC)	(vaccinated, O)	(reports, O)

	article 7	article 8	article 9 \
0	(The, O)	(The, O)	(A, B-ORG)
1	(City, B-LOC)	(Union, ORG)	(gang, I-ORG)
2	(of, I-LOC)	(of, O)	(of, I-ORG)
3	(Lahti, I-LOC)	(Upper, O)	(three, B-MISC)
4	(,, O)	(Secondary, O)	(masked, I-MISC)
5	(Europe, B-LOC)	(School, O)	(individuals, I-MISC)
6	('s, O)	(Students, ORG)	(have, O)
7	(Green, B-MISC)	(has, O)	(stolen, O)
8	(Capital, I-MISC)	(issued, O)	(or, O)
9	(of, O)	(a, O)	(attempted, O)
10	(2021, B-DATE)	(demand, O)	(to, O)
11	(,, O)	(that, O)	(steal, O)
12	(has, O)	(pop-up, O)	(pay, O)
13	(brought, O)	(coronavirus, O)	(terminals, O)
14	(a, O)	(vaccination, O)	(at, O)
15	(climate, B-MISC)	(points, O)	(four, B-MISC)
16	(action, I-MISC)	(be, O)	(locations, I-MISC)
17	(art, B-MISC)	(set, O)	(in, O)
18	(installation, I-MISC)	(up, O)	(different, O)
19	(entitled, O)	(in, O)	(parts, O)
20	(the, O)	(high, O)	(of, O)
21	(Indisputable, B-MISC)	(schools, O)	(Finland, B-LOC)
22	(Case, I-MISC)	(in, O)	(over, O)
23	(of, I-MISC)	(an, O)	(the, O)
24	(Emergency, I-MISC)	(effort, O)	(past, O)
25	((ICE), O)	(to, O)	(week, O)
26	(to, O)	(secure, O)	(Police, B-ORG)
27	(the, O)	(on-site, O)	(say, O)
28	(centre, B-LOC)	(learning, O)	(they, O)
29	(of, I-LOC)	(and, O)	(are, O)
30	(Helsinki, I-LOC)	(increase, O)	(still, O)
31	(., O)	(the, O)	(working, O)
32	(The, O)	(vaccine, O)	(to, O)
33	(nine, O)	(coverage, O)	(identify, O)
34	(large, O)	(of, O)	(the, O)
35	(blocks, O)	(young, O)	(three, B-MISC)
36	(containing, O)	(people, O)	(suspects, I-MISC)
37	(13, O)	(The, O)	(The, O)
38	(cubes, O)	(pop-up, O)	(first, O)
39	(of, O)	(stations, O)	(theft, O)

40	(ice, B-MISC)	(could, O)	(took, O)
41	(each, O)	(facilitate, O)	(place, O)
42	(are, O)	(the, O)	(on, O)
43	(intended, O)	(vaccination, O)	(Wednesday, B-DATE)
44	(to, O)	(of, O)	(morning, O)
45	(serve, O)	(students, O)	(at, O)
46	(as, O)	(without, O)	(a, O)
47	(a, O)	(an, O)	(petrol, B-LOC)
48	(reminder, O)	(appointment, O)	(station, I-LOC)
49	(of, O)	(for, O)	(in, O)

	article 10
0	(Helsinki, B-LOC)
1	(District, I-LOC)
2	(Court, I-LOC)
3	(has, O)
4	(sentenced, O)
5	(a, O)
6	(43-year-old, O)
7	(taxi, O)
8	(driver, O)
9	(to, O)
10	(six, O)
11	(years, O)
12	(and, O)
13	(three, O)
14	(months, O)
15	(in, O)
16	(prison, O)
17	(for, O)
18	(a, O)
19	(total, O)
20	(of, O)
21	(18, O)
22	(different, O)
23	(crimes, O)
24	(most, O)
25	(of, O)
26	(which, O)
27	(were, O)
28	(related, O)
29	(to, O)
30	(the, O)
31	(drugging, O)
32	(and, O)
33	(raping, O)
34	(of, O)

```
35          (four, O)
36        (female, O)
37       (victims, O)
38             (., O)
39           (The, O)
40         (court, O)
41         (heard, O)
42          (that, O)
43          (from, O)
44          (2019, O)
45           (the, O)
46     (defendant, O)
47           (had, O)
48          (been, O)
49       (driving, O)
```