Submission for
Exercise task 1
of UTU course TKO_8964-3006
Textual Data Analysis
by Botond Ortutay

Instructions:

Part 1: loading a dataset from the HF hub

Using the load_dataset function of the datasets library, load each of the following datasets in turn:

- stanfordnlp/imdb
- eriktks/conll2003
- openai/gsm8k

For each of the datasets, report the following information:

- What NLP task is the dataset intended for (e.g. syntactic analysis, toxicity detection, etc.)? (You may need to refer to the documentation of the dataset for this.)
- What parts is the dataset split into (e.g. train, test) and how many examples does each contain?
- What features (e.g. text, label) does the dataset have? (Try to understand how these relate to the NLP task the dataset is intended for.)
- What is the first item in the training set of the dataset?

Part 2: creating a dataset from your own data

You can find data collected from the Yle news RSS feed here: http://dl.turkunlp.org/TKO_8964_2023/

Download either the Finnish or English data (`news-fi-2021.jsonl` or `news-en-2021.jsonl`) using `wget` and create a datasets from the JSONL data (see https://huggingface.co/docs/datasets/loading#json). Answer the following questions:

- What NLP tasks could the dataset be used for?
- What features does the dataset have?
- How many space-separated words do the texts of the dataset contain in total?

Solutions:

Part 1:

Importing libraries & environment setup:

NOTE (to self): here we assume that this notebook is run on a kernel with all the relevant libraries installed. This is important because I sweat and bled while setting up a venv to allow my Debian laptop to have the Python libraries (I screwed up something during configuration and it took me way too long to debug.) Therefore: once again: make sure that the libraries are there before you run this (use tdavenv on your Debian laptop and if you change computers make sure to install the relevant libraries and maybe even do environment configurations.

In [1]:
```
import datasets
import random    # For getting random samples
```

Defining & documenting functions used below:

In [2]:
```
"""
A function that returns a DatasetDict object and the train split from a huggingface dataset
---
In:
path                 str            the path huggingface uses to find the dataset
mainConfigNeeded     boolean        certain datasets want you to specify a config before datasets.load_dataset, setting this as True allows you to load such a
---
Out:
(dsDict, dsTrain),
    where:
dsDict               DatasetDict    object containing all available splits in dataset as well as the splits' features and amounts
dsTrain              Dataset        contains the train split of the dataset
---
Note: we assume that all the datasets handled by this function have a split called "train"
"""
def loadDataset(path, mainConfigNeeded):
    if mainConfigNeeded:
        dsDict = datasets.load_dataset(path, "main")                     # No split was specified. Therefore: the dsDict variable should have a DatasetDict obje
        dsTrain = datasets.load_dataset(path, "main", split="train")     # Loading the train split of the dataset to take a closer peek at its contents
    else:
        dsDict = datasets.load_dataset(path)                             # No split was specified. Therefore: the dsDict variable should have a DatasetDict obje
        dsTrain = datasets.load_dataset(path, split="train")             # Loading the train split of the dataset to take a closer peek at its contents
    return dsDict, dsTrain

"""
A function that prints basic information from a huggingface dataset, such as:
split names, features & sizes
samples from an inputted split
---
In:
dsDict       DatasetDict    DatasetDict of the dataset we want to examine
dsSplit      Dataset        A split from the dataset for printing samples
splitName    str            The name of the split we print samples from within the original huggingface dataset
prints       int            The amount of samples we want printed
---
Out:

Note: we assume that all the datasets handled by this function have a split called "train"
"""
def examineDataset(dsDict, dsSplit, splitName, prints):
    print(dsDict)    # Printing the DatasetDict to get an overview of all the splits and features

    print("")
    print("---")
    print("")

    # Printing a few random samples from the dataset to look what kind of data is there
    for i in range(prints):
        print(dsSplit[random.randint(0,dsDict[splitName].num_rows)])
        print("")
```

The Stanford IMDB dataset:
code & outputs:

In [3]:
```
stanfordIMDB, stanfordIMDBTrain = loadDataset("stanfordnlp/imdb", False)
```

In [4]:
```
examineDataset(stanfordIMDB, stanfordIMDBTrain, "train", 4)
```

```
DatasetDict({
    train: Dataset({
        features: ['text', 'label'],
        num_rows: 25000
    })
    test: Dataset({
        features: ['text', 'label'],
        num_rows: 25000
    })
    unsupervised: Dataset({
        features: ['text', 'label'],
        num_rows: 50000
    })
})

---

{'text': 'Why has this not been released? I kind of thought it must be a bit rubbish since it hasn\'t been. How wrong can a girl be! This film is, in a word, enthralling.<br /><br />You will be captivated. It holds your attention from the start and its pace never slows.<br /><br />The final part of the film, the "episode" as it were (not giving anything away, you saw that in the trailer) is also unmissable. You will chose a favourite, you will be shocked, you wont be able to go and make a cup of coffee because you need to find out what happens. The adrenalin rises and you cant not watch. Cudos to the actors, it\'s very believable. And it doesn\'t stop there, they have a final shock for you.<br /><br />It also makes you question reality TV and if you would watch. And how far away from this are we, really? Endemol (who make big brother) made a TV show in Holland last year offering a dying woman\'s kidney to patients in need of a transplant. The show was revealed at the end to be a hoax, ostensibly to raise awareness of organ donation, but are we getting too close for comfort?', 'label': 1}

{'text': "Okay, truthfully, I saw the previews for this movie and thought to myself, what are the producers thinking? Hutton, Jolie, and DUCHOVNY? How could the monotoned actor possibly compete with Jolie's natural power on the screen? But surprisingly, the two had the kind of chemistry that showed intense caring without a kiss. Even David's humor matched up to Jolie's spark and fire. As for Hutton, he played the psycho very well, contrasting with David's calm delivery of life threatening situations. Overall, I was very impressed with the writing and character development. I gave it 8 stars.", 'label': 1}

{'text': "and this IS a very disturbing film. I may be wrong, but this is the last film where I considered Burt Reynolds an actual actor, who transformed the role, and delivered a message.<br /><br />Jon Voight and Ned Beatty are also excellent. They are unassuming and unaware; businessmen wanting to enjoy the country. Little did they know what would happen next.<br /><br />The photography and sets are realistic and natural. This was before the days of Wes Craven.<br /><br />What is most disturbing about this film is the fact that places like this still exist. In America, country folk still detest city people; it is almost a century and a half since the Civil War.<br /><br />You will enjoy this film. It was filmed in the rural sections of South Georgia, which still exist. Just don't drive past that to Mobile, Alabama; That area still has not been repaired since Hurricane Katrina. 10/10.", 'label': 1}

{'text': 'I enjoyed this movie. Unlike like some of the pumped up, steroid trash that is passed off as action movies, Playing God is simple and realistic, with characters that are believable, action that is not over the top and enough twists and turns to keep you interested until the end.<br /><br />Well directed, well acted and a good story.', 'label': 1}
```

answers:

What NLP task is the dataset intended for?
According to the stanfordnlp/imdb dataset's README.md file the dataset is intended to be used for sentiment analysis. The author's claim that the datasets contain "highly polar" movie reviews. This is probably done so that the positive & negative feelings people have about the movies (a.k.a. the "sentiments") could be more easily assiociated with certain kinds of language.

What parts is the dataset split into and how many examples does each contain?
The dataset has a train and a test split each with 25000 members. It also has a split called "unsupervised" with 50000 members. This is probably just the data from both the train and the test splits combined in one collection for unsupervised learning.

What features does the dataset have?
The dataset consists of the text and label features. The text feature contains the movie review and the label feature just contains a 0 or a 1. 0 seems to be associated with negative reviews and 1 with positeive reviews. The model probably uses the label data to learn what kind of language is associated with positive and negative emotion

What is the first item in the training set of the dataset?
Printed below:

In [5]:
```
print(stanfordIMDBTrain[0])
```

```
{'text': 'I rented I AM CURIOUS-YELLOW from my video store because of all the controversy that surrounded it when it was first released in 1967. I also heard that at first it was seized by U.S. customs if it ever tried to enter this country, therefore being a fan of films considered "controversial" I really had to see this for myself.<br /><br />The plot is centered around a young Swedish drama student named Lena who wants to learn everything she can about life. In particular she wants to focus her attentions to making some sort of documentary on what the average Swede thought about certain political issues such as the Vietnam War and race issues in the United States. In between asking politicians and ordinary denizens of Stockholm about their opinions on politics, she has sex with her drama teacher, classmates, and married men.<br /><br />What kills me about I AM CURIOUS-YELLOW is that 40 years ago, this was considered pornographic. Really, the sex and nudity scenes are few and far between, even then it\'s not shot like some cheaply made porno. While my countrymen mind find it shocking, in reality sex and nudity are a major staple in Swedish cinema. Even Ingmar Bergman, arguably their answer to good old boy John Ford, had sex scenes in his films.<br /><br />I do commend the filmmakers for the fact that any sex shown in the film is shown for artistic purposes rather than just to shock people and make money to be shown in pornographic theaters in America. I AM CURIOUS-YELLOW is a good film for anyone wanting to study the meat and potatoes (no pun intended) of Swedish cinema. But really, this film doesn\'t have much of a plot.', 'label': 0}
```

Eriktks's conll2003 dataset:
code & outputs:

In [6]:
```
conll2003Dict, conll2003Train = loadDataset("eriktks/conll2003", False)
```

In [7]:
```
examineDataset(conll2003Dict, conll2003Train, "train", 4)
```

```
DatasetDict({
    train: Dataset({
        features: ['id', 'tokens', 'pos_tags', 'chunk_tags', 'ner_tags'],
        num_rows: 14041
    })
    validation: Dataset({
        features: ['id', 'tokens', 'pos_tags', 'chunk_tags', 'ner_tags'],
        num_rows: 3250
    })
    test: Dataset({
        features: ['id', 'tokens', 'pos_tags', 'chunk_tags', 'ner_tags'],
        num_rows: 3453
    })
})

---

{'id': '9179', 'tokens': ['Atheist', 'China', 'officially', 'bans', 'missionary', 'activities', 'but', 'often', 'turns', 'a', 'blind', 'eye', 'to', 'religious', 'activities', 'of', 'people', 'nominally', 'employed', 'as', 'foreign', 'language', 'teachers', ',', 'particularly', 'in', 'remote', 'areas', 'that', 'are', 'unable', 'to', 'attract', 'other', 'candidates', '.'], 'pos_tags': [21, 22, 30, 42, 16, 24, 10, 30, 42, 12, 16, 21, 35, 16, 24, 15, 24, 30, 40, 15, 16, 21, 24, 6, 30, 15, 16, 24, 43, 41, 16, 35, 37, 16, 24, 7], 'chunk_tags': [11, 12, 21, 22, 11, 12, 0, 21, 22, 11, 12, 12, 13, 11, 12, 13, 11, 21, 22, 13, 11, 12, 12, 0, 3, 13, 11, 12, 11, 21, 1, 21, 22, 11, 12, 0], 'ner_tags': [0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]}

{'id': '9131', 'tokens': ['It', 'also', 'offers', 'to', 'refinance', 'up', 'to', '50', 'percent', 'of', 'the', 'debt', 'held', 'by', 'state', 'banks', 'whose', 'governments', 'decide', 'to', 'keep', 'control', 'of', 'their', 'banks', '.'], 'pos_tags': [28, 30, 42, 35, 37, 15, 35, 11, 21, 15, 12, 21, 40, 15, 21, 24, 45, 24, 41, 35, 37, 21, 15, 29, 24, 7], 'chunk_tags': [11, 3, 21, 22, 22, 11, 12, 12, 12, 13, 11, 12, 21, 13, 11, 12, 11, 12, 21, 22, 22, 11, 13, 11, 12, 0], 'ner_tags': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]}

{'id': '395', 'tokens': ['Second', 'round'], 'pos_tags': [16, 21], 'chunk_tags': [11, 12], 'ner_tags': [0, 0]}

{'id': '984', 'tokens': ['37.', 'Northern', 'Ireland', '7.89'], 'pos_tags': [22, 22, 22, 11], 'chunk_tags': [11, 12, 12, 12], 'ner_tags': [0, 5, 6, 0]}
```

answers:

What NLP task is the dataset intended for?
According to the eriktks/conll2003 dataset's README.md file the dataset is intended to be used for "language-independent named entity recognition". So if I understand this correctly the aim is to teach a model to look at a text and then determine from the contex which words refer to "named entities", such as people, places and organizations etc. (The README said they were focusing on these three areas).

What parts is the dataset split into and how many examples does each contain?
The dataset consits of a train split of 14041 rows, a validation split of 3250 rows and a test split of 3453 rows.

What features does the dataset have?
The dataset has the following features: id, tokens, pos_tags, chunk_tags and ner_tags. The id feature is - surprise surprise - an ID tag and its purpose is to let any computer system reference individual rows of the dataset. Having such a feature is a common practice in any computer system. The tokens feature is the text to be analyzed. So sentences or collections of words including the named entities among other words. (Although my random sample included several rows where token was just: City name, number, number, number, number. What is up with that? Probably postal adresses or something but seems weird to have so many of these that a random sample of 4 out of 1400 caught two of them.) The ner_tags feature is a reference to a token and marks whether the token is a person, location, organization, other named entity or none of these.[source: the README file] This is important because teaching these is the goal of the dataset. The pos_tags and chunk_tag features were a bit harder to understand for me (probably easier for other people who have taken nlp courses, I haven't), but my hypotheses was that these are grammatical. I found an article that seems to confirm this. According to this article pos_tags refer to POS (Parts-of-speech) which is basically a system to categorize grammatic functions of words into groups such as nouns, verbs, pronouns, etc. This is basically exactly my hypothesis and it's in line with the README's explanation as well. According to the same article chunk_tags refer to groups ow words that refer to the same concept (for example in the sentence "The quick brown fox jumped over the lazy dog", "the quick brown fox" would be a chunk.

What is the first item in the training set of the dataset?
Printed below:

In [8]:
```
print(conll2003Train[0])
```

```
{'id': '0', 'tokens': ['EU', 'rejects', 'German', 'call', 'to', 'boycott', 'British', 'lamb', '.'], 'pos_tags': [22, 42, 16, 21, 35, 37, 16, 21, 7], 'chunk_tags': [11, 21, 11, 12, 21, 22, 11, 12, 0], 'ner_tags': [3, 0, 7, 0, 0, 0, 7, 0, 0]}
```

Openai's gsm8k dataset:
code & outputs:

In [9]:
```
gsm8kDict, gsm8kTrain = loadDataset("openai/gsm8k", True)
```

In [10]:
```
examineDataset(gsm8kDict, gsm8kTrain, "train", 4)
```

```
DatasetDict({
    train: Dataset({
        features: ['question', 'answer'],
        num_rows: 7473
    })
    test: Dataset({
        features: ['question', 'answer'],
        num_rows: 1319
    })
})

---

{'question': "Viggo's age was 10 years more than twice his younger brother's age when his brother was 2. If his younger brother is currently 10 years old, what's the sum of theirs ages?", 'answer': "Twice Viggo's younger brother's age when his brother was 2 is 2*2 = <<2*2=4>>4 years.\nIf Viggo's age was 10 more than twice his younger brother's age when his brother was 2, Viggo was 10+4 = 14 years old.\nViggo is 14 years - 2 years = <<14-2=12>>12 years older than his brother\nSince Viggo's brother is currently 10 and Vigo is 12 years older, he's currently 10 + 12 = <<10+12=22>>22 years old\nTheir combined age is 22 years + 10 years = <<22+10=32>>32 years\n#### 32"}

{'question': 'Jack goes hunting 6 times a month.  The hunting season lasts for 1 quarter of the year.  He catches 2 deers each time he goes hunting and they weigh 600 pounds each.  He keeps half the weight of deer a year.  How much deer does he keep in pounds?', 'answer': 'The hunting season last 12/4=<<12/4=3>>3 months\nSo he goes hunting 3*6=<<3*6=18>>18 times a year\nThat means he gets 18*2=<<18*2=36>>36 deers per year\nSo he catches 36*600=<<36*600=21600>>21,600 pounds of deer a year\nThat means he keeps 21600/2=<<21600/2=10800>>10,800 pounds of deer a year\n#### 10800'}

{'question': "Three times as many children as adults attend a concert on Saturday. An adult ticket costs $7 and a child's ticket costs $3. The theater collected a total of $6,000. How many people bought tickets?", 'answer': 'Let X be the number of adults attending the concert. The number of children attending the concert is 3*X.\nThe theater collected a total of 7*X + 3*(3*X) = $6,000.\nMultiplying through the parentheses we get 7X + 9X = $6,000\nAdding like terms we get 16X = $6,000\nDividing both sides by 16 we get X = 375, the number of adults attending the concert\nThe number of children attending the concert is triple this number: 375 adults * 3 children/adult = <<375*3=1125>>1125 children.\nThe number of people attending the concert is 375 + 1125 = <<375+1125=1500>>1500 people.\n#### 1500'}

{'question': 'Emily bought a shirt and a coat for $600. What does the shirt cost if it is one-third the price of the coat?', 'answer': 'Let X be the price of the shirt.\nThe price of the coat is X*3.\nThe total amount is X + X*3 = $600.\nCombining like terms, we get X*4 = $600.\nDividing both sides by 4, we get X = $600 / 4 = $<<600/4=150>>150.\n#### 150'}
```

answers:

What NLP task is the dataset intended for?
According to the openai/gsm8k dataset's README.md file the dataset is intended to be used to teach transformer-based LLM systems to extract mathematical problems from written descriptions and then perform multi-step mathematical reasoning similarly to how elementary school math-word problems work.

What parts is the dataset split into and how many examples does each contain?
The dataset consits of a train split of 7473 rows and a test split of 1319 rows.

What features does the dataset have?
The dataset contains question and answer features. The question feature contains a written problem, and the answer feature contains a written answer with the logical deduction parts written down in plaintext as well as mathematical syntax.

What is the first item in the training set of the dataset?
Printed below:

In [11]:
```
print(gsm8kTrain[0])
```

```
{'question': 'Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?', 'answer': 'Natalia sold 48/2 = <<48/2=24>>24 clips in May.\nNatalia sold 48+24 = <<48+24=72>>72 clips altogether in April and May.\n#### 72'}
```

Part 2:

NOTE (to the person checking this): We assume that Jupyter is configured in such a way that bash commands can be run on here. This is due to the exercise instructions requiring the use of `wget`.

Using `wget` to install the data:

In [12]:
```
#NOTE: bash kernel needed !!!
!echo Downloading data from TurkuNLP
!echo
!wget http://dl.turkunlp.org/TKO_8964_2023/news-en-2019.jsonl
!echo
!echo Printing all files in current directory to check data has downloaded
!echo
!ls
```

```
Downloading data from TurkuNLP

--2025-01-20 14:17:40--  http://dl.turkunlp.org/TKO_8964_2023/news-en-2019.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: 7855444 (7,5M) [application/octet-stream]
Saving to: ‘news-en-2019.jsonl’

news-en-2019.jsonl  100%[===================>]   7,49M  5,16MB/s    in 1,5s

2025-01-20 14:17:42 (5,16 MB/s) - ‘news-en-2019.jsonl’ saved [7855444/7855444]


Printing all files in current directory to check data has downloaded

empty.ipynb  exercise_task_1.ipynb  news-en-2019.jsonl  tdavenv
```

Using `datasets.load_dataset()` to load the downloaded json file to python as a dataset:

In [13]:
```
newsEn = datasets.load_dataset("json", data_files="news-en-2019.jsonl")
```

Poking around our new dataset (bash):

In [14]:
```
#NOTE: bash kernel needed !!!
# Printing a few random samples from the dataset to look what kind of data is there
!shuf -n 4 news-en-2019.jsonl

!echo
!echo ---
!echo

!echo -n Line count:
!wc -l news-en-2019.jsonl
```


```
{"summary": "Funds from the campaign will be channeled into conservation projects aimed at reducing the impact of destructive algae.", "tags": ["Baltia", "Itämeri", "Muumi-kirjat", "Pohjoismainen kirjallisuus", "Tove Jansson", "kuvittajat", "lasten- ja nuortenkirjallisuus", "levät", "meret", "muumit", "sarjakuvataiteilijat", "scifi", "sinilevät", "suomenkielinen kirjallisuus", "vesistöt"], "text": "If Moomin creator  Tove Jansson  could see the Baltic Sea in 2019, she would do her best to improve it. \n That's the view of Moomin Characters' current artistic director, Tove Jansson's niece  Sophia Jansson . \n Moomin Characters Ltd, the limited liability company that controls the image and licensing rights of the Moomin characters, plans to raise one million euros next year to protect the Baltic Sea. Company CEO  Roleff Kråkström  first proposed the idea of using Moomins as patrons, when he witnessed the effect of blue-green algae on the water. \n Since then, more than 50 companies and organisations have become involved, including the John Nurminen Foundation, known for its protection campaigns in the Baltic Sea. \n Story continues after photo \n Moomin Characters' Artistic Director Sophia Jansson and Managing Director Roleff Kråkström believe that Moomins can inspire people to protect the Baltic Sea. Lehtikuva \n The Baltic Sea campaign will launch in 2020, and will coincide with the 75th anniversary of the publication of the first Moomin story,  The Moomins and The Great Flood . According to Sophia Jansson, the sea was an important element for Tove Jansson. \n \"The sea is present in all of the Moomin stories, and was also a very large part of Tove's personality, production and heart. She lived by the sea and in the archipelago all her life,\" explains Sophia Jansson. \n Nutrients from agriculture lead to algae \n The biggest problem in the Baltic Sea is caused by eutrophication, which occurs when a body of water becomes overly enriched with minerals and nutrients, creating algae. When dead algae descends to the bottom, it consumes oxygen. \n In the Baltic Sea eutrophication has resulted in excessive growth of algae and record-breaking levels of oxygen depletion. \n \"We are in a vicious circle that makes the eutrophication situation in the Baltic Sea difficult, despite the fact that much has been done to improve the situation on land,\" says  Seppo Knuuttila , Senior Research Scientist at Finnish Research Institute SYKE. \n Knuuttila cites the 60 percent drop in phosphorus loads in the Gulf of Finland since 2000, which he says is largely due to Russia's improved wastewater treatment, as evidence that much is being achieved. However, the phosphorus loads from agriculture are still a major problem. \n None of the Baltic Sea states have as yet significantly reduced their agricultural nutrient loads. \n \"If this fails, HELCOM's (the Baltic Marine Environment Protection Commission's) goal of improving the ecological status of the sea will not be achieved,\" Knuuttila says. \n The effects of the blue-green cyanobacteria on the Baltic Sea were clearly evident in the summer of 2018. While the level of algae in the water this summer has not been on the same scale, Knuuttila warns that the conditions for cyanobacteria are still rife. \n \"In terms of nutrition this summer, the conditions for vigorous cyanobacterial bloom are in place. The weather conditions have not been favourable for heavy cyanobacterial blooms so far, but the weather of the coming weeks may change that,\" says Knuuttila. \n \"No time to wait\" \n The money raised through the Moomin campaign will be channeled into conservation projects at the John Nurminen Foundation. According to the Foundation's agent  Annamari Arrakoski-Engardt , one way to reduce nutrient loads is by increasing the use of gypsum treatment. \n \"It (gypsum treatment) can significantly reduce the nutrient loads in fields. This should be introduced in Finland in all suitable areas, and also tested in Poland and Sweden, which are major contributors to agricultural nutrient loads,\" says Arrakoski-Engardt, also adding that she believes the state of the Baltic Sea can be improved. \n \"I know that the Baltic Sea can be saved. I am concerned about whether the measures are adequate and whether the pace is fast enough. Climate change has put pressure on marine protection, and we should do more. There is no time to wait,\" says Arrakoski-Engardt. \n Sophia Jansson believes that the Moomin characters can raise a great deal of interest in the Baltic Sea. \n \"Moomins usually solve things in a very harmonious and good way, and with that positive attitude, I believe that a lot can be accomplished,\" says Jansson.", "timestamp": "2019-07-18T19:28:59", "title": "Moomins want to raise a million euros to protect the Baltic Sea", "url": "https://yle.fi/uutiset/10884516"}
{"summary": "The combination service will initially be available as an app while a travel card will be introduced later on.", "tags": ["Baltian maat", "Harjumaa", "Helsingin seudun liikenne (HSL)", "Helsinki", "Lähimaksaminen", "Matkakortit ja -liput", "Pääkaupunkiseudun joukkoliikenne (Suomi)", "Tallinna", "Viro", "joukkoliikenne", "matkailu", "matkustajaliikenne", "mobiilisovellus", "pääkaupunkiseudut", "virolaiset"], "text": "Starting next spring, commuters will be able to use Helsinki Regional Transport’s travel card app for public transportation in Tallinn, the Estonian capital. \n It's part of a long-planned combination travel service for use in Helsinki and Tallinn and will first be available as a mobile app. \n \"We have a goal that next spring we will be able to offer tickets for public transportation in Tallinn here [in Helsinki] via the HSL app and vice versa as well,\" HSL division head  Mari Flink  said. \n The biggest winners under the proposed system will be Estonians travelling in Helsinki as they will be able to use a travel app purchased via the Estonian service pilet.ee to move around on HSL routes. Currently, they have to download the HSL app separately and link it to a bank card - or purchase tickets from automated ticket vending machines. \n Standing outside the Western Harbour in Helsinki, Estonian  Triin O’Brock  said she was excited to hear of the new combined travel service. She had downloaded the Discover Helsinki app on her smartphone in the mistaken belief that she would also be able to use it to buy tickets to travel on public transportation. \n \"Well it didn’t work so now I’m queuing at this ticket machine. I missed one tram already so hopefully I’ll catch the next one,\" she added \n Estonian cards a step ahead \n It will be some time however, before a combination travel card can be used, because travel cards in Helsinki and Tallinn are very different. Although they appear to be similar at a glance, they are based on different platforms. \n The technical implementation of the Estonian cards is more modern, so Tallinn public transportation officials will have to wait for HSL to catch up and design a card that works according to the same logic. \n \"I don’t think we are talking about years,\" Flink said about the waiting period. Her colleague  Tiit Laikso  from Tallinn was a bit more optimistic. \n \"We are aiming for the combination card to be available in 2021,\" he commented. \n HSL is also looking into the possibility of using contactless payment options used already in Tallinn public transportation in the HSL fleet. But there is no fixed timetable for that project. \n \"There a strong will to come up with some kind of timetable for implementing contactless payments in public transportation but that is a major step. We have more than 2,000 vehicles so that’s a large number of card readers,\" Flink pointed out. \n Card readers currently installed in public transportation vehicles do not offer a contactless payment option. HSL said that there would be no point in upgrading them, but noted that it is looking into the possibility of adding the feature via a separate add-on device.", "timestamp": "2019-10-16T11:11:04", "title": "Helsinki, Tallinn plan combination travel card by 2021", "url": "https://yle.fi/uutiset/11022854"}
{"summary": "Helsingin Sanomat reports on new efforts to reorganize and streamline Finland's social and health care services.", "tags": ["Darude", "Euroopan parlamentti", "Euroviisut", "Helsingin Sanomat", "Pohjois-Eurooppa", "Pohjoismaat", "Ruotsi", "Suomi", "Terveydenhuollon palvelut", "aluehallinto", "politiikka", "sote-uudistus", "vaalit", "yhteiskunnalliset tapahtumat", "äänestäminen"], "text": "The daily  Helsingin Sanomat asks the question  whether Finland's next government will be able to assure the nation's residents better access to healthcare services. \n According to the paper, the five parties involved in government formation talks are now working on a revamped project to reform Finland's social and healthcare system. \n The model, it says, is based on the  same \"sote\" reform  of the previous government that was intended to shift responsibility for services away from municipalities to 18 elected regional authorities. \n Helsingin Sanomat reports that its sources say that the issue under discussion at this stage is only the integration of services, not an expansion of the mandate of the regional authorities. \n One of the Centre Party's non-negotiable demands in the formation talks is that the reform sticks to the 18-region scheme and that regional authorities are given a wide range of responsibilities. \n As HS points out, the last government's overhaul plan failed in large part because of a dispute over the inclusion of allowing people to choose public services or publicly funded services provided by private and third-sector operators. \n Now, some of the parties in government formation talks are those who oppose opening public healthcare to private companies. \n HS says that the biggest questions, though, centre on financing the reform and the right of regional authorities to levy taxes. \n Talks on a policy programme for social and health care services are being carried out by a broad-based group led by former Social Democratic Party minister  Krista Kiuru . \n EU advance voting starts \n Oulu's  Kaleva  was among the papers reminding readers that advance voting in elections to the European Parliament starts on Wednesday, and continues through Tuesday, 21 May. \n Election day is Sunday, 26 May. \n For the purposes of the European Parliament elections, Finland is a single election district. \n When going to the polls, voters need to be able to present an ID, such an identification card issued by the police, a passport, driver's licence, or other official document including a personal photo. \n Citizens of other EU countries living in Finland  may vote in Finland's European Parliamentary election  as long they have signed up to the voting register in advance. \n In a related item, the Kuopio-based  Savon Sanomat  reports that veteran politician  Paavo Väyrynen  has announced that if he wins an MEP seat in the European Parliament elections, he would be willing to accept the chairmanship of Finland's Centre Party next year. \n Väyrynen claims that he has received feelers from within the Centre about taking over the party's top spot, a position he held between 1980 and 1990. Only this week, Väyrynen announced  he was leaving the party he founded less than a year ago , the Seven Star Movement. \n He says, though, that his availability to chair the Centre is conditional on being elected to the European Parliament, which he would consider an endorsement of his future political plans. \n Swedes ready to defend Finland \n Turun Sanomat carries a syndicated  Lännen Media repor t on a Swedish defence plan published Tuesday that includes a full brigade of some 5,000 soldiers earmarked for operations in Finland, in the event of a conflict or war. \n Antti Pihlajamaa , an instructor in strategic studies at Finland's National Defence University, is quoted as saying that although the Swedish proposal is clear and straightforward, it is still only a proposal and that no far-reaching conclusions should yet be drawn. It will take another decade, he said, until it is seen what Sweden's overall defence plan means. \n Last year, Finland and Sweden signed a Memorandum of Understanding  outlining the objectives of their defence co-operation . \n According to Pihlajamaa, the importance of Sweden's plan to offer troops would depend to a large extent on the nature of any crisis, and what kind of troops would be provided. \n Disappointment in Tel Aviv \n Probably without exception, morning newspapers reported that Finland's hopes at the Eurovision Song Contest were dashed Tuesday evening when  the Finnish entry , \"Look Away\" featuring artists  Darude  and  Sebastian Rejman , was eliminated in the first semi-final in Tel Aviv. \n The Finnish entry was composed by Darude (Ville Virtanen) an award-winning producer/DJ best known for his iconic 1999 trance instrumental \"Sandstorm\" and featured Rejman on vocals. \n Writing for the tabloid  Iltalehti , reporter  Mari Pudas  attributed the loss not to the quality of the music, but to the performance's lack of exciting visual elements.", "timestamp": "2019-05-15T06:59:32", "title": "Wednesday's papers: Healthcare reform, Swedish defence offer, Eurovision disappointment", "url": "https://yle.fi/uutiset/10783502"}
{"summary": "An expert says Huawei's troubles could still be ironed out, but one operator says it likely won't sell the firm's new phones.", "tags": ["Google", "Huawei", "Internet-palveluntarjoaja", "Matkapuhelimet ja mobiililaitteet", "Puhelinvalmistajat", "Radio-, televisio- ja tietoliikennevälineiden valmistus", "Suomi", "Verkkovalmistajat", "matkaviestinverkot", "mobiiliteknologia", "puhelimet", "teleoperaattorit", "telepalvelut", "tietoliikenne"], "text": "Mobile phone retailers around the world -- including in Finland -- are unsure about their relationship with Chinese network and smartphone firm Huawei following news of Google suspending its relations with the company. \n Google has said it would follow orders from US President Donald Trump to stop supplying Huawei with its operating system Android. \n For some time now, US officials have claimed that Huawei products - including its devices and 5G networking gear - could be so-called Trojan Horses for  Chinese intelligence agencies . \n Huawei is widely thought to be owned or controlled by the Chinese state and the Trump administration has said it considers the company to be a \"risk to national security,\" and has encouraged other nations to blacklist the firm. \n Going forward, this will likely mean that the Chinese company's new phones could still feature open source versions of the OS, but would lack the ability to use popular Google-branded apps like YouTube and the Chrome web browser. \n Some Finnish telecoms companies were cautious about the possible future of Huawei, while another, like mobile networks firm DNA have said it is unlikely it will sell new handsets released by the Chinese company going forward. \n In the meantime, however, DNA's CEO  Sami Aavikko  said the firm will continue to sell Huawei phones that are already on store shelves. \n Story continues after photo. \n DNA CEO Sami Aavikko. Mikko Ahmajärvi / Yle \n \"We'll continue selling the old models since the company has promised to continue providing security updates for the anticipated lifetimes of the phones. The situation regarding updates for the Android operating system is still uncertain,\" Aavikko said. \n Finnish telecoms provider Elisa also sells Huawei handsets, but reported via email that the company has not decided whether it will sell new models in the future. \n Mobile operator Telia declined to comment on what it plans to do with Huawei products on its store shelves, but did update its customers, according to the company's communications director  Camilla Ekholm . \n \"We understand that people are uncertain, because it is difficult to say what impact this situation will have on consumers. That is why we have encouraged our sales staff to inform [customers] about Huawei's unsure situation,\" Ekholm said. \n Current owners \"shouldn't worry\" \n Journalist at Swedish tech site Sweclockers,  Andreas Eklöv , said people who own Huawei smartphones will still be able to use Google apps like Gmail, Docs and Hangouts. \n \"The phones will continue to function like they normally do. The big difference is that Huawei cannot guarantee its customers that their handsets will receive OS updates, or when they will get them,\" Eklöv explained. \n \"It's not unsafe for people to continue using Huawei phones that they've already bought. Outside the US, consumers will still receive security updates from Huawei and Google,\" he said. \n As long as Google's restrictions are in effect, Eklöv said that Huawei will be forced to manually fetch updates from Android's Open Source Project and distribute them to Huawei smartphones themselves, rather than automatically receive official updates directly from Google. \n This situation, Eklöv explained, will likely mean Huawei handsets will not be updated as quickly as other Android-based phones on the market. \n Eklöv said people wondering whether to buy a Huawei handset should probably wait. \n \"If there's no rush, it would be a good idea to wait to see what happens with the embargo. The situation could still be resolved by the autumn and then new Huawei phones would receive new versions of Android. But before we know how long the embargo will last, I would not recommend people to buy Huawei phones because we don't know how long they will receive updates,\" Eklöv said.", "timestamp": "2019-05-24T12:27:35", "title": "Telecoms in Finland cautious about future with Huawei", "url": "https://yle.fi/uutiset/10800033"}
```


```
---

Line count:2481 news-en-2019.jsonl
```

Poking around our new dataset (python):

In [15]:
```
print(newsEn)
```

```
DatasetDict({
    train: Dataset({
        features: ['summary', 'tags', 'text', 'timestamp', 'title', 'url'],
        num_rows: 2481
    })
})
```

Written answers for the questions in Part 2:

What NLP tasks could the dataset be used for?
This is a huge amount of data and it could be used for all kinds of purposes. The first thing that came to mind was LLM training, although using this material for that has several problems. Firstly all the material is news media, meaning that it's not generic enough to train an LLM; All generated text would be news-related... Furthermore this data isn't in a conversation format, so the text-promts it'd need to generate new text wouldn't feel conversational, which would make it harder to use. I tried looking around the net for similar kinds of datasets and found this one where they collected over 5000 news articles from different newspapers in Nigeria. They used it to perform news categorization, which makes sense and could absolutely be a valid application for our news dataset. However it is a bit funny how our tags (which would be the feature to do news categorization by) are in Finnish whereas everything else in the dataset is in English... I suppose I could've chosen the Finnish dataset and then this problem wouldn't exist...

What features does the dataset have?
The dataset has the features: summary, tags, text, timestamp, title and url. I believe these are self-explanatory enough that I won't need to explain them.

How many space-separated words do the texts of the dataset contain in total?
That depends on one's definition of "space-separated word". However I'll just give the simplest answer and let the following bash code count them for me:

In [16]:
```
#NOTE: bash kernel needed !!!
!echo -n Word count in news-en-2019.jsonl as counted by the wc command:
!wc -w news-en-2019.jsonl
```

```
Word count in news-en-2019.jsonl as counted by the wc command:1193770 news-en-2019.jsonl
```