

Koolitus „Andmeteadus on Popp“

Masinnägemine ja tekstitöötlus

November 2021

Kristjan Eljand



Euroopa Liit
Euroopa Sotsiaalfond



Eesti
tuleviku heaks

Masinnägemine

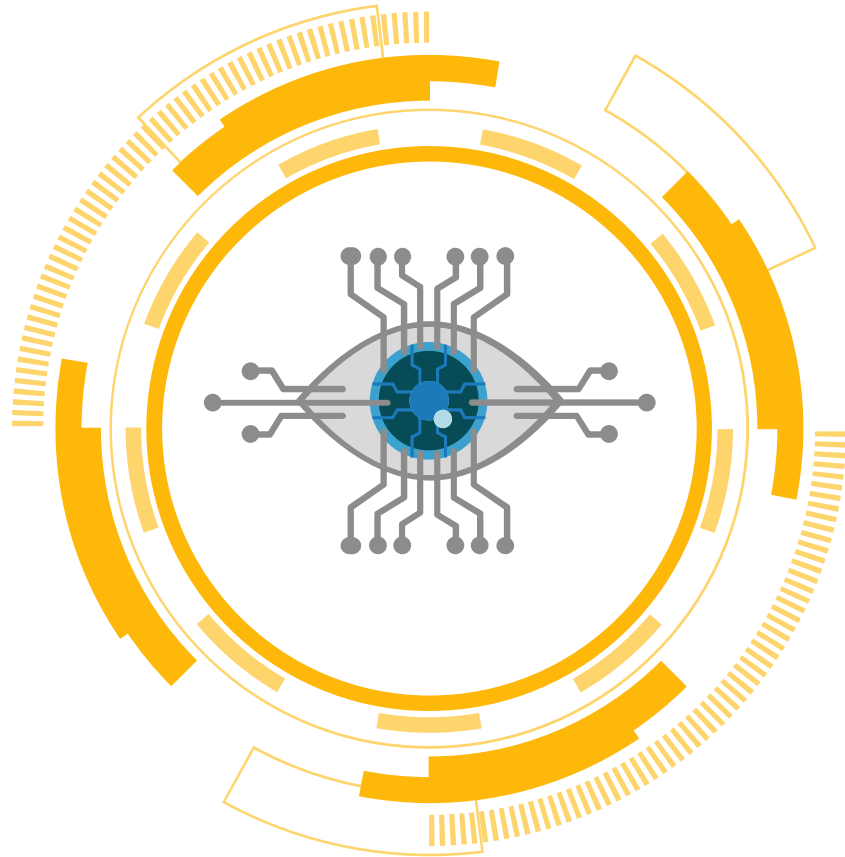


Euroopa Liit
Euroopa Sotsiaalfond



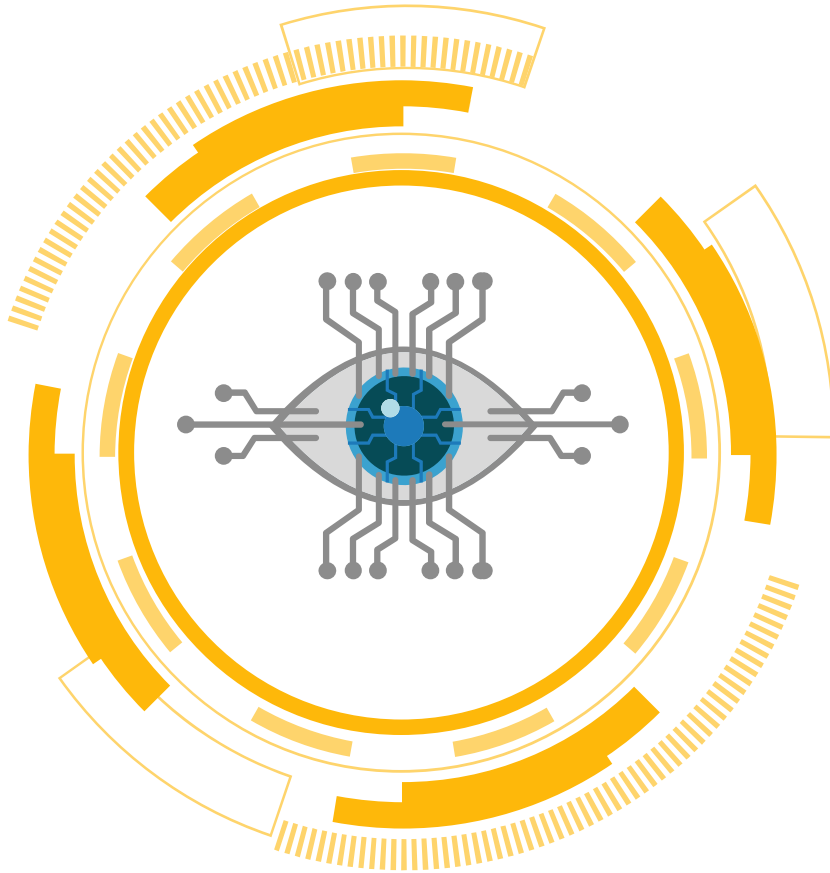
Eesti
tuleviku heaks

Mis on masinnägemine?



Masinnägemine on tehisintellekti alamvaldkond, mille eesmärk on eraldada informatsiooni digitaalsetelt piltidelt ja videodelt.

Mille jaoks saab masinnägemist kasutada!?



Masinnägemist saab kasutada:

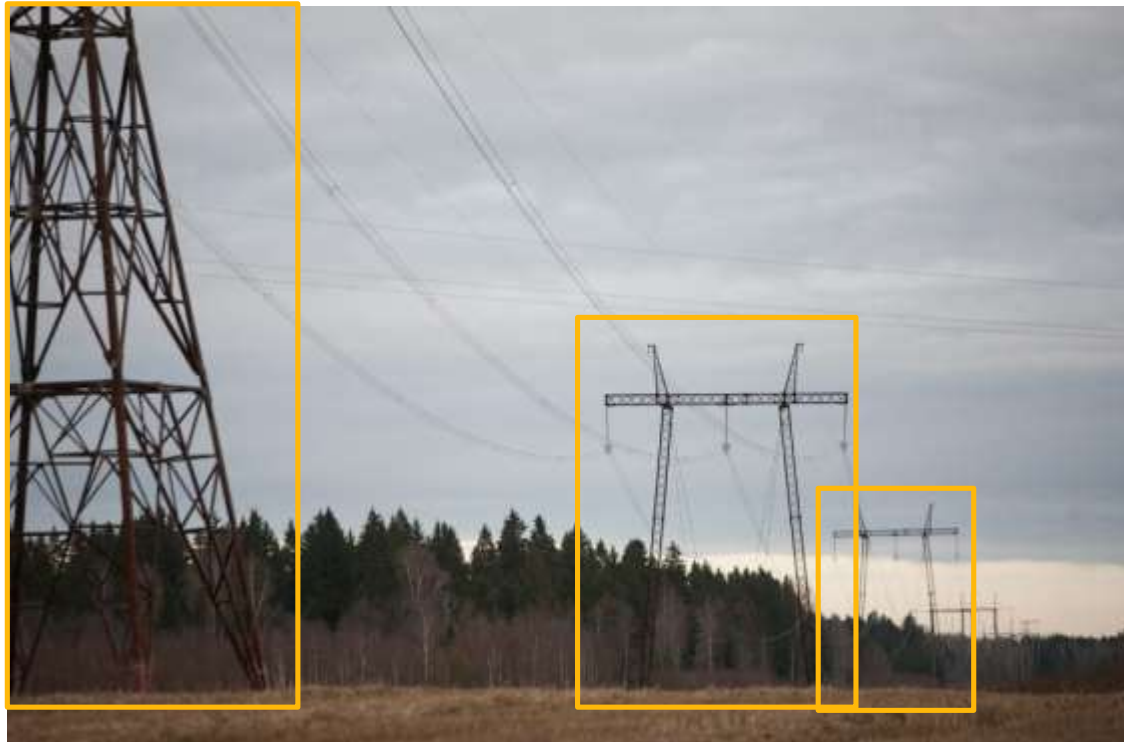
1. Objektide tuvastamiseks ja klassifitseerimiseks;
2. Teksti eraldamiseks piltidelt;
3. Nägude ja emotsioonide tuvastamiseks.

Objektide tuvastamine



Mis on pildil?

Objektide tuvastamine



Mis on pildil?



* *Google Vision AI tulemused*

Objektide klassifitseerimine

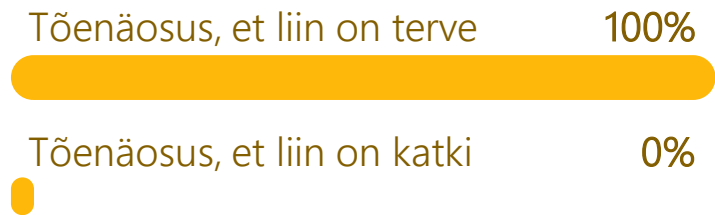


Kas elektriliin on terve?

Objektide klassifitseerimine



Kas elektriliin on terve?



Azure Custom Vision Engine

Näiteid võimalikest klassifitseerimisülesannetest

1. **Optiline järelvalve** -- Videokaamera tuvastab automaatselt suitsu ja lahtise tule, teavitab omanikku ja kutsub päästeameti.
2. **Optiline ärivõimaluste tuvastamine** -- Pildituvastus leiab satelliidipiltidelt ja aerofotodelt automaatselt kodud, kellele võiks pakkuda päikeseelektri lahendust.
3. **Optiline defektide tuvastamine** -- Regulaarselt tehtavate droonipiltide põhjal tuvastatakse automaatselt tuulikute hooldusvajadus: tuuliku labadele kogunev mustus, surnud putukad, sool (mereäärsetel tuulikutel); mõrad vundamendis või tornis.
4. **Optiline ennustav hooldus** -- Elektriliinide või transpordimasinate hooldusvajaduse automaatne tuvastamine.

Tekstituvastus



Mis tekst on pildil?

Tekstituvastus



Mis tekst on pildil?

Blokk 1

Eesti Energia

Blokk 2

TESTONIA KAEVANDUS
Killustiku ja aheraine müük

* *Google Vision AI tulemused*

Ideid tekstituvastuseks

1. **Dokumentide digitaliseerimine** -- Prinditud dokumentide digitaliseerimine.
2. **Dokumentide automaattöötlus** -- .pdf kujul olevatest tellimus/tarnedokumentidest tabelite ja väärtuste automaatne eraldamine.
3. **Optiline tekstituvastus** -- Elektrimõõdikute näidu lugemine pildilt.

Nägude ja emotsioonide tuvastamine



Näod ja emotsioonid!?

Nägude ja emotsioonide tuvastamine



Näod ja emotsioonid!?

Face 1: confidence 70%

Joy		Likely
Anger		Very Unlikely
Surprise		Very Unlikely

Face 2: confidence 92%

Joy		Likely
Anger		Very Unlikely
Surprise		Very Unlikely

Face 3: confidence 52%

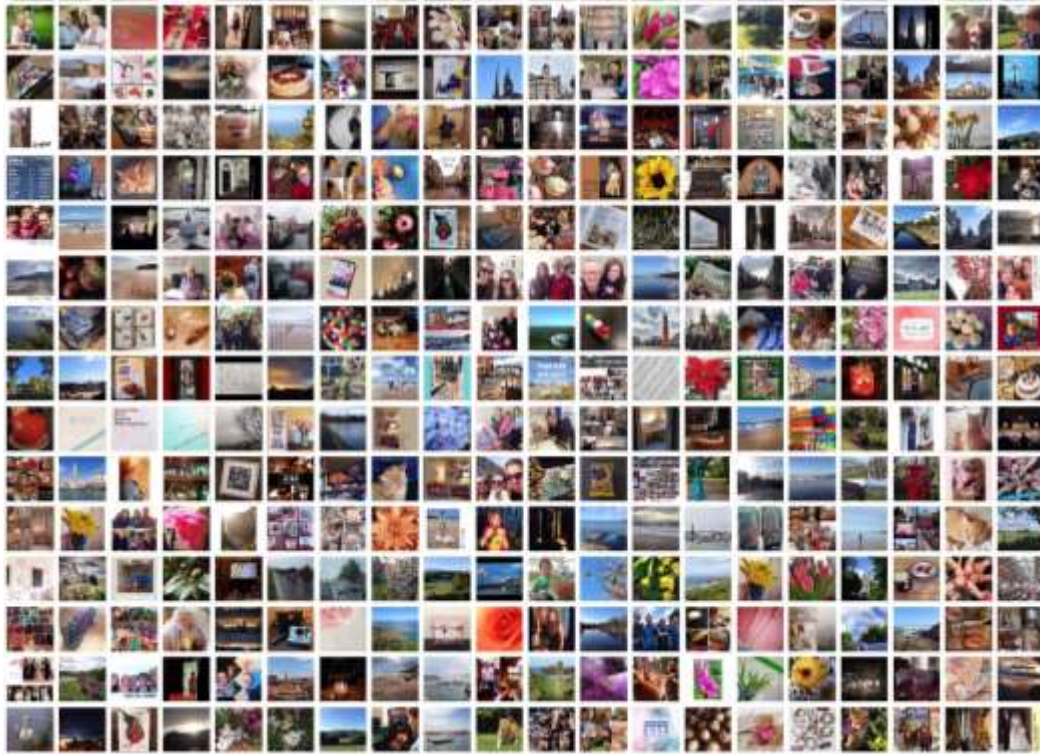
Joy		Unlikely
Anger		Very Unlikely
Surprise		Very Unlikely

* *Google Vision AI tulemused*

Ideid nägude ja pooside tuvastuseks

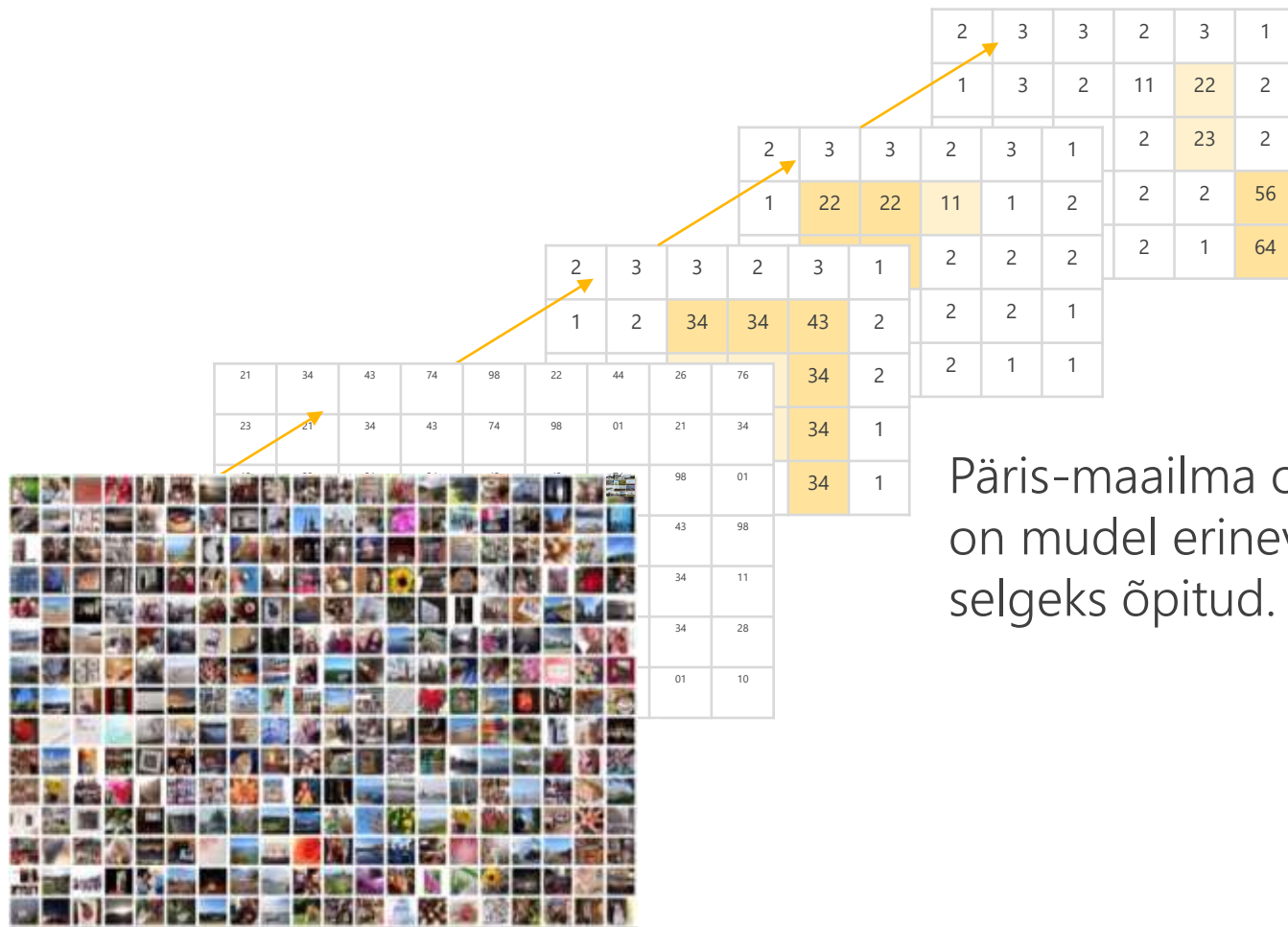
1. **Personaliseerimine** -- Videokaamera tuvastab automaatselt kodus viibijad ning reguleerib targad süsteemid vastavalt nende tavapärastele eelistustele.
2. **Optiline heaolu kontroll** -- Arvuti veebikaamera jälgib töötaja tööasendit ning tööruumi valgustaset ning annab sõbraliku soovitusel, kui asendi või tööruumiga võivad kaasnevad terviseriskid.
3. **Optiline isikutuvastus** -- Kaevandusse sisenevate töötajate automaatne tuvastamine ja isiku verifitseerimine.

Eeltreenitud mudelite kasutamine – mis toimub taustal



Azure ja Google on treeninud oma masinnägemise mudelid miljonite piltide baasil.

Eeltreenitud mudelite kasutamine – mis toimub taustal



Päris-maailma objektide tuvastamise on mudel erinevatel kihtidel juba selgeks õpitud.

<https://cloud.google.com/vision>

Google Vision AI tutvustus

Reaalsete probleemide lahendamine eeltreenitud mudelitega



- Meil on näited katkistest ja tervetest elektriliinidest
- Eesmärk on luua süsteem, mis ütleks, kas liin on katki või terve.

Demo

<https://teachablemachine.withgoogle.com>

Katkise elektriliini tuvastamine

Ülesanne

Loo masinnägemise mudel – mõtle ise välja, mida peaks mudel ennustama.

<https://teachablemachine.withgoogle.com>



Euroopa Liit
Euroopa Sotsiaalfond



Eesti
tuleviku heaks

Tekstitöötlus

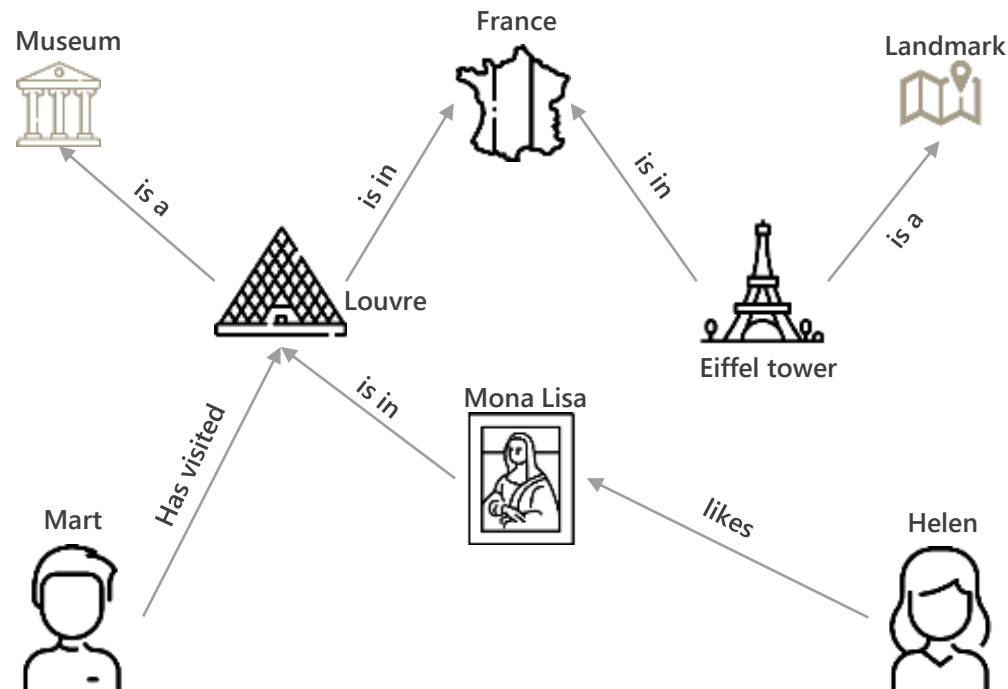


Euroopa Liit
Euroopa Sotsiaalfond



Eesti
tuleviku heaks

Struktureerimata tekstiandmete analüüs – miks?



- Struktureerimata andmed (e-mailid, tekstidokumendid, töökirjeldused, jne) võivad moodustada enam kui 80% kõigist ettevõttes tekkivatest andmetest.
- Samas, otsuseid tehakse valdavalt vaid struktureeritud (mõtle SQL) andmete baasil - s.o vaid 20% kogu infost!!
- Proovime ülejäänud 80% ka kasutusele võtta !

Tõlkimine

↗ Translation

Kui kaua ma pean ootama, et te mu ühenduse ära parandak

Compute

Computation time on cpu: 0.3924 s

How long do I have to wait for you to fix my connection?

↗ Translation

Elekter on inimkonna üks suurimaid saavutusi

Compute

Computation time on cpu: 0.2988 s

Electricity is one of mankind's greatest achievements

Eeltreenitud tekstitöötamise AI on valdavalt ingliskeelne. Seega on paljude ülesannete puhul hea mõte tõlkida eestikeelne tekst esmalt ingliskeelseks.



Euroopa Liit
Euroopa Sotsiaalfond



Eesti
tuleviku heaks

Semantiline otsing

Ask your question **Co-worker search**
Who knows crypto?

Name Hidden
Energy Trade ...

Certainty score: **83.28 %**

About a year ago, I started to **invest** in **shares**. I made my first **deal** back then. Before making a decision, I made a thorough preliminary work: I read books on this topic, took part in trainings and kept an eye on stock news every day. Why? As Estonian education system doesn't almost teach handling the **money**, so about five years ago I understood that I need to start saving because it's not expedient to spend all the **money** earned. Step by step I started to examine the world of investments and opportunities of making my savings grow. Today the popularity of **currency** or **cryptocurrency** is growing....

Name Hidden
Software developer

Certainty score: **82.84 %**

... My reading recommendation comes to those **interested** in **investing** – "Intelligent Investor"! Most of the books on the subject (e.g. "Rikas isa, vaene isa" (rich dad, poor dad)), create an **interest** in people in **investing** and constantly there is talk about one's success stories, which create the impression that **investing** is easy and profits easy to come. This book however highlights in detail, how to successfully **invest** and what effort is required, so that **investing** truly becomes **profitable**. Successful investments!

Semantilise otsinguga saame otsida struktureerimata tekstidest struktureerimata küsimustega?

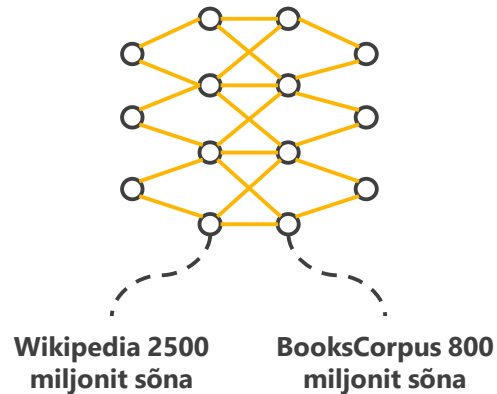
Mudel proovib aru saada, mida me oma küsimusega „mõtleme“ ja väljastada selle põhjal vasted.

Semantiline otsing: Mis toimub taustal

1. Inimeste kirjeldused struktureerimata tekstina

Mart [..While in France, I visited Louvre and Eiffel tower..]
Hele [...I like art and visit art exhibitions regularly..]
n
... [Description ...]

2. Eeltreenitud mudel



3. Inimeste kirjeldused kolmesaja numbrina

Mart [0.8, 0.4, 0.6, 0.3, 0.7, ..., n]
Helen [0.3, 0.5, 0.6, 0.8, 0.4, ..., n]
... [0.2, 0.7, 0.6, 0.3, 0.7, ..., n]

Saame tulemuseks iga inimese kohta 300. numbrit, mis peaks seda inimest iseloomustama 😊!???

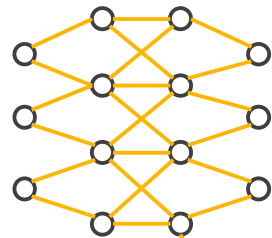
Semantiline otsing: Otsinguprotsess

Question
Who has visited France?

Mart on küsimuse
tähtsusele lähemal kui
Helen.

Answer
Mart (99% certainty)

Eeltreenitud mudel



Küsimus numbrivektorina

Mart [0.8, 0.4, 0.6, 0.3, 0.7, ..., n]

Helen [0.3, 0.5, 0.6, 0.8, 0.4, ..., n]

Q: [0.7, 0.5, 0.6, 0.3, 0.7, ..., n]

Semantiline otsing: Otsinguprotsess

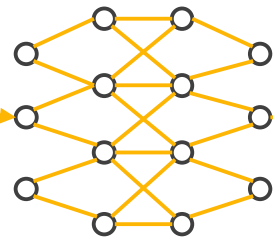
Question

Who has visited France?

Answer

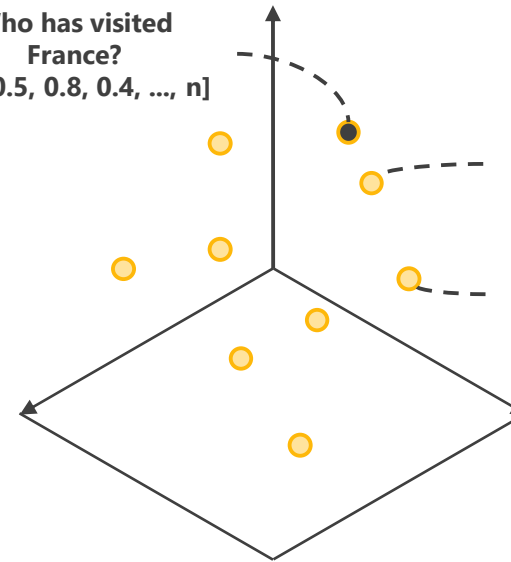
Mart (99% certainty)

Eeltreenitud mudel



Küsimus
numbrivektorina

Who has visited
France?
[0.3 0.5, 0.8, 0.4, ..., n]



Mart
[0.8, 0.4, 0.6, 0.3, ...,
n]

Helen
[0.7, 0.5, 0.6, 0.3, ..., n]

Mart on küsimuse
tähtsusele lähemal kui
Helen.



Euroopa Liit
Euroopa Sotsiaalfond



Eesti
tuleviku heaks

Meelsuse analüüs

Text Classification

How long do I have to wait until you fix my connection?

Compute

Computation time on cpu: 0.0348 s

NEGATIVE

0.999

POSITIVE

0.001

Text Classification

Electricity is one of greatest achievements of humankind

Compute

Computation time on cpu: 0.0312 s

NEGATIVE

0.000

POSITIVE

1.000

Meelsuse analüüsi

eesmärk on klassifitseerida tekst selles sisalduva meelsuse järgi.

Nimed ja objektide tuvastamine

Token Classification

Examples



This training series is organized by the Ministry of economic affairs and communications and carried out by Andmeteadu

Compute

Computation time on cpu: 0.1152 s

This training series is organized by the Ministry of economic affairs and communications **ORG** and carried out by Andmeteadus OÜ **ORG**.

This sentence is written by Kris **PER** t **PER** jan Eljand **PER**.

Nimedega ja objektide tuvastamise (*named entity recognition*) eesmärk on leida tekstist viited inimestele, asukohtadele ja organisatsioonidele.

Teksti kokkuvõtete loomine

Summarization

Examples

We operate in the markets for electricity and gas sales in the Baltic States, Finland and Poland, as well as on the international market for liquid fuels. We create energy solutions from the production of electricity, heat and fuels to innovative sales, client services and energy related additional services. Our ambition is to offer our clients useful and convenient energy solutions and to produce energy

Compute

Computation time on cpu: 6.5724 s

The aim of Eesti Energia is to ensure the profitability of the group regardless of the shocks on the world economy. We operate in the markets for electricity and gas sales in the Baltic States, Finland and Poland, as well as on the international market for liquid fuels. Our ambition is to offer our clients useful and convenient energy solutions and to produce energy ourselves, in an increasingly environment conserving way.

Kokkuvõtete tegemise eesmärgiks on saada aru pika teksti mõttest ja esitada see lühemal kujul.



Euroopa Liit
Euroopa Sotsiaalfond



Eesti
tuleviku heaks

Küsimustele vastamine teksti põhjal

Question Answering

How much was the profit for the second quarter?

Compute

Context

The sales revenues of Eesti Energia Group amounted to EUR 241.1 million in the second quarter of 2021 (+43.5% year-on-year).

Group EBITDA was EUR 40.8 million (-25.5% year-on-year).

Group's net loss for the second quarter was EUR 10.0 million (-242.9% year-on-year).

Computation time on cpu: 0.32880000000000004 s

EUR 10.0 million

0.764

Meil on tekst ja soovime vastata antud teksti põhjal küsimustele.

Küsimustele vastamine tabeli põhjal

Table Question Answering

How much revenue did Toomas generate?

Compute

Computation time on cpu: cached

2 matches : 200 300 SUM

Client	Product	Amount	Revenue
Mart	Elekter	1	100
Toomas	Gaas	2	200
Mari	Päike	3	300
Toomas	Elekter	1	300

Meil on tabel ja soovime vastata küsimustele tabeli põhjal

Kasutusvaldkondade kokkuvõte

- **Semantiline otsing** – Leiame tekstid, milles käsitletakse meid huvitavat kontseptsiooni.
- **Meelsuse analüüs** – leiame tekstidest meelsuse (positiivne, neutraalne, negatiivne).
- **Nimede ja objektide tuvastamine** tekstidest.
- **Teksti kokkuvõtete tegemine** – muudame pikad tekstid lühemateks kokkuvõteteks.
- Küsimustele vastamine **teksti põhjal** või **tabeli põhjal**.

Koolitus „Andmeteadus on Popp“

Masinnägemine ja tekstitöötlus

November 2021

Kristjan Eljand



Euroopa Liit
Euroopa Sotsiaalfond



Eesti
tuleviku heaks