Keeletehnoloogia

3 lugu

3 lugu

1. "Mul nagu kodus arvutis suur hunnik tekstifaile erinevates keeltes, oleks vaja otsinguprogrammi mis suudaks kuvada kõik failid mis sisaldavad minu antud võtmesõna"

3 lugu

1. "Mul nagu kodus arvutis suur hunnik tekstifaile erinevates keeltes, oleks vaja otsinguprogrammi mis suudaks kuvada kõik failid mis sisaldavad minu antud võtmesõna"

2. "Ma ei viitsi enam netis küsimustele vastata! Kopp ees juba! Kõik küsivad nii triviaalseid asju, peaks ju kindlasti sellelaadse ülesandega ka arvuti hakkama saama!"

3 lugu

1. "Mul nagu kodus arvutis suur hunnik tekstifaile erinevates keeltes, oleks vaja otsinguprogrammi mis suudaks kuvada kõik failid mis sisaldavad minu antud võtmesõna"

2. "Ma ei viitsi enam netis küsimustele vastata! Kopp ees juba! Kõik küsivad nii triviaalseid asju, peaks ju kindlasti sellelaadse ülesandega ka arvuti hakkama saama!"

3. "

3 lugu

1. "Mul nagu kodus arvutis suur hunnik tekstifaile erinevates keeltes, oleks vaja otsinguprogrammi mis suudaks kuvada kõik failid mis sisaldavad minu antud võtmesõna"

2. "Ma ei viitsi enam netis küsimustele vastata! Kopp ees juba! Kõik küsivad nii triviaalseid asju, peaks ju kindlasti sellelaadse ülesandega ka arvuti hakkama saama!"

3. "Ma tahaks internetis surfata,

3 lugu

1. "Mul nagu kodus arvutis suur hunnik tekstifaile erinevates keeltes, oleks vaja otsinguprogrammi mis suudaks kuvada kõik failid mis sisaldavad minu antud võtmesõna"

2. "Ma ei viitsi enam netis küsimustele vastata! Kopp ees juba! Kõik küsivad nii triviaalseid asju, peaks ju kindlasti sellelaadse ülesandega ka arvuti hakkama saama!"

3. "Ma tahaks internetis surfata, aga ma saan ainult

3 lugu

1. "Mul nagu kodus arvutis suur hunnik tekstifaile erinevates keeltes, oleks vaja otsinguprogrammi mis suudaks kuvada kõik failid mis sisaldavad minu antud võtmesõna"

2. "Ma ei viitsi enam netis küsimustele vastata! Kopp ees juba! Kõik küsivad nii triviaalseid asju, peaks ju kindlasti sellelaadse ülesandega ka arvuti hakkama saama!"

3. "Ma tahaks internetis surfata, aga ma saan ainult läti keelest aru -

3 lugu

1. "Mul nagu kodus arvutis suur hunnik tekstifaile erinevates keeltes, oleks vaja otsinguprogrammi mis suudaks kuvada kõik failid mis sisaldavad minu antud võtmesõna"

2. "Ma ei viitsi enam netis küsimustele vastata! Kopp ees juba! Kõik küsivad nii triviaalseid asju, peaks ju kindlasti sellelaadse ülesandega ka arvuti hakkama saama!"

3. "Ma tahaks internetis surfata, aga ma saan ainult läti keelest aru - kas saab internetilehte

3 lugu

1. "Mul nagu kodus arvutis suur hunnik tekstifaile erinevates keeltes, oleks vaja otsinguprogrammi mis suudaks kuvada kõik failid mis sisaldavad minu antud võtmesõna"

2. "Ma ei viitsi enam netis küsimustele vastata! Kopp ees juba! Kõik küsivad nii triviaalseid asju, peaks ju kindlasti sellelaadse ülesandega ka arvuti hakkama saama!"

3. "Ma tahaks internetis surfata, aga ma saan ainult läti keelest aru - kas saab internetilehte automaatselt läti keelde tõlkida?"

Otsing tekstifailides

- ette antud otsingusõnad $w_{x1}...w_{xn}$
- otsingu algoritm:
  - iga teksti puhul
  - iga selle teksti sõna w_y puhul
  - kui mõne i puhul $w_{xi} = w_y$, kuva käesolev tekst
- miks see ei tööta?

Otsing tekstifailides

- ette antud otsingusõnad $w_{x1}...w_{xn}$
- otsingu algoritm:
  - iga teksti puhul
  - iga selle teksti sõna w_y puhul
  - kui mõne i puhul $w_{xi} = w_y$, kuva käesolev tekst
- miks see ei tööta?
  - otsing "otsin **töökohta**" ei leia dokumenti, mis sisaldab fraasi "linnas X on palju vabu **töökohti**"

Otsing tekstifailides

- ette antud otsingusõnad $w_{x1}...w_{xn}$
- otsingu algoritm:
  - iga teksti puhul
  - iga selle teksti sõna w_y puhul
  - kui mõne i puhul $w_{xi} = w_y$, kuva käesolev tekst
- miks see ei tööta?
  - otsing "otsin **töökohta**" ei leia dokumenti, mis sisaldab fraasi "linnas X on palju vabu **töökohti**"
  - otsing "mul on töökohta vaja" leiab esimese kahe sõna järgi ("**mul**" ja "**on**") peaaegu kõike (eestikeelseid) tekste

Otsing tekstifailides

- kuidas teha kindlaks et "töökohta" ja "töökohti" on sama sõna?

Otsing tekstifailides

- kuidas teha kindlaks et "töökohta" ja "töökohti" on sama sõna?
  - **Morfoloogiline analüüs!**
    - töökohta -> töökoht (osastav, ainsus)
    - töökohti -> töökoht (osastav, mitmus)

Otsing tekstifailides

- kuidas teha kindlaks et "töökohta" ja "töökohti" on sama sõna?
  - **Morfoloogiline analüüs!**
    - töökohta -> töökoht (osastav, ainsus)
    - töökohti -> töökoht (osastav, mitmus)
- Kuidas tuvastada et "mul" pole otsingu jaoks tähtis, ning "töökoht" - on?

Otsing tekstifailides

- kuidas teha kindlaks et "töökohta" ja "töökohti" on sama sõna?
  - **Morfoloogiline analüüs!**
    - töökohta -> töökoht (osastav, ainsus)
    - töökohti -> töökoht (osastav, mitmus)
- Kuidas tuvastada et "mul" pole otsingu jaoks tähtis, ning "töökoht" - on?
  - **Sõnaliigid!**
    - mul -> ma -> asesõna (neid on piiratud arv, igas tekstis neid leidub palju, otsingu jaoks pole olulised)

Otsing tekstifailides

- kuidas teha kindlaks et "töökohta" ja "töökohti" on sama sõna?
  - **Morfoloogiline analüüs!**
    - töökohta -> töökoht (osastav, ainsus)
    - töökohti -> töökoht (osastav, mitmus)
- Kuidas tuvastada et "mul" pole otsingu jaoks tähtis, ning "töökoht" - on?
  - **Sõnaliigid!**
    - mul -> ma -> asesõna (neid on piiratud arv, igas tekstis neid leidub palju, otsingu jaoks pole olulised)
- Aga kui kirjutada "jobs" või "looking" siis eestikeelne morf. analüüs hakkama ei saa

Otsing tekstifailides

- kuidas teha kindlaks et "töökohta" ja "töökohti" on sama sõna?
  - **Morfoloogiline analüüs!**
    - töökohta -> töökoht (osastav, ainsus)
    - töökohti -> töökoht (osastav, mitmus)
- Kuidas tuvastada et "mul" pole otsingu jaoks tähtis, ning "töökoht" - on?
  - **Sõnaliigid!**
    - mul -> ma -> asesõna (neid on piiratud arv, igas tekstis neid leidub palju, otsingu jaoks pole olulised)
- Aga kui kirjutada "jobs" või "looking" siis eestikeelne morf. analüüs hakkama ei saa
  - **Keele äratundmine!**
    - jobs -> *inglise* , töökoht -> *eesti* , работа -> *vene*

Infootsingu demo

Netis küsimustele vastaja

- Täpset algoritmi raske slaidi peal näidata
- Kasutab samasuguseid elemente nagu tekstifailides otsing:
  - morfoloogiline analüüs
    - "kas kinos **cinamon** toimub täna midagi?"
    - "mida **cinamonis** saab täna vaadata?"
  - nimetuste tuvastamine
    - "mis on **ekraani** kinokava?"
    - "mis on **cinamoni** ekraanide suurus?"

http://www.dialoogid.ee/teatriagent/

Automaatne tõlkimine

- Algoritmi ei saagi nii täpselt kirjeldada
- On kasu samasugustest elementidest
  - morfoloogiline analüüs
  - süntaktiline analüüs
  - nimetuste tuvastamine
  - sõnaliikide määramine
  - ...

http://masintolge.ut.ee/

Keeletehnoloogia

- Tegeleb igasuguse teksti- ja keeletöötlusega
  - masintõlge
  - õigekirja kontroll
  - komavigade tuvastamine
  - dialoogisüsteemid
  - automaatne sisukokkuvõtte koostamine
  - kõnetuvastus
  - kõneleja tuvastus
  - ...

Keeletehnoloogia

- Teisiti öeldes, tegeleb loomuliku keele modelleerimisega
- s.t. otsib erinevate loomuliku keele protsesside mudeleid, ehk toimumise printsiipide kirjeldusi või selgitusi
- Kuidas neid mudeleid defineeritakse
  - keeleteadus/lingvistika
  - korpuslingvistika
  - juhendatud masinõpe
  - juhendamata masinõpe

Lingvistiline keelemodelleerimine

- Keeleteadlased täpselt kirjeldavad kuidas keel "töötab"
  - sõnamuutused (morfoloogia)
  - lause struktuus (süntaks)
  - sõnade ja fraaside tähendused (semantika)
  - jms.
- Realiseerime neid kirjeldusi algoritmiliselt, ja saavutame õnne!
- Lingvistika printsiipide algoritmiline realiseerimine -> **arvutuslingvistika**
- Kas nii saab keelt hästi modelleerida?

Korpuslingvistika

- Korpus - suur tekst või tekstide hulk
- http://www.cl.ut.ee/
- Selle asemel et teoretiseerida keele kohta, vaatame sellele otsa ning kontrollime, kas tõepoolest keel toimib nii nagu keeleteadlased kirjeldavad
- Osutub et mitte alati
- Korpuse peale otse vaadates saab inimene koostada uusi reegleid, mis on ka kooskõlas reaalsete andmetega

Loomuliku keele masinõpe

- Selle asemel et inimene peaks teksti vaadates reegleid koostama, võiks tuletada tüüpilised mustrid algoritmiliselt
- **Masinõpe!**
- Andmehulk: sisend-väljund paarid
  - sisend: sõna, väljund: selle sõnaliik
  - sisend: sõna, väljund: selle morf. analüüs
  - sisend: eestikeelne lause, väljund: inglisekeelne lause
- Printsiip: määrame, mida me tahame väljundis näha, las masin/algoritm ise õpib kuidas seda optimaalselt ja võimalikult täpselt kätte saada
- Nõuab suurt andmehulka mudelite treenimiseks
  - tavaliselt inimene korjab teksti ja märgistab vastavalt vajadusele

Keele juhendamata õpe

- Kui kasutame keele (juhendatud) masinõpet kasutades, määrame, **mida me tahame väljundis näha**
  - laud - nimisõna
  - hüppa - tegusõna
  - mina - asesõna
  - roheline - omadussõna
  - ...
- Juhendamata masinõpe puhul määrame, **kuidas me tahame et algoritm õpiks**
  - sõnad koosnevad morfeemidest
    - laulu|de|ga
    - suure|pärase|lt
  - ülesanne: leida teksti sõnade parim segmenteerimine
  - vajab ainult teksti ilma märgendamiseta!

Juhendamata sõnaehituse leidmine

emale	kassile	laual
emaga	kassiga	lauaga
emata	kassita	lauata
emadele	kassidele	lauadel
emadega	kassidega	lauadega
emadeta	kassideta	lauadeta

Juhendamata sõnaehituse leidmine

ema+le	kassi+le	laua+le
ema+ga	kassi+ga	laua+ga
ema+ta	kassi+ta	laua+ta
ema+de+le	kassi+de+le	laua+de+le
ema+de+ga	kassi+de+ga	laua+de+ga
ema+de+ta	kassi+de+ta	laua+de+ta

ema+, kassi+, laua+
+de+
+le, +ta, +ga

Keeletehnoloogia uurimisalana

- Miks on keeletehnoloogilised ülesanded huvitavad uurija (loe: bakatöö kirjutaja) seisukohast?

Keeletehnoloogia uurimisalana

- Miks on keeletehnoloogilised ülesanded huvitavad uurija (loe: bakatöö kirjutaja) seisukohast?
  - lingvistiline modelleerimine
  - korpuslingvistika
  - masinõpe
  - juhendamata masinõpe
  - nende kombinatsioonid

Keeletehnoloogia uurimisalana

- Miks on keeletehnoloogilised ülesanded huvitavad uurija (loe: bakatöö kirjutaja) seisukohast?
  - lingvistiline modelleerimine
  - korpuslingvistika
  - masinõpe
  - juhendamata masinõpe
  - nende kombinatsioonid
- **Ükski ei tööta 100% kindlalt**

Keeletehnoloogia uurimisalana

- Miks on keeletehnoloogilised ülesanded huvitavad uurija (loe: bakatöö kirjutaja) seisukohast?
  - lingvistiline modelleerimine
  - korpuslingvistika
  - masinõpe
  - juhendamata masinõpe
  - nende kombinatsioonid
- **Ükski ei tööta 100% kindlalt**
- Loomulik keel elab, muutub, areneb
- Keeletehnoloogia on seega loodusteadus, nagu füüsika, bioloogia või keemia - erinevalt matemaatikast või informaatikast
- sest see uurib ja kirjeldab olemasolevat loomuliku nähtust

TÜ Keeletehnoloogia töörühm

http://www.cs.ut.ee/~koit/KT/