

Korpusbaseret lemmaselektion og opdatering

Jørg Asmussen
Afdeling for Digitale Ordbøger og Tekstkorpora
Det Danske Sprog- og Litteraturselskab
www.dsl.dk

Program

1. Introduktion til DSL
2. Introduktion til projektet *ordnet.dk*
3. Lemmaselektion i DDO
4. Lemmaopdatering i *ordnet.dk*
 - a. Tekstklassifikation
 - b. Bestemmelse af nye ord
 - c. Diskussion af metoden
 - d. Konklusion

Program

1. Introduktion til DSL
2. Introduktion til projektet *ordnet.dk*
3. Lemmaselektion i DDO
4. Lemmaopdatering i *ordnet.dk*
 - a. Tekstklassifikation
 - b. Bestemmelse af nye ord
 - c. Diskussion af metoden
 - d. Konklusion

Hvad er DSL?

Hvad er DSL?

Det **D**anske **S**prog- og **L**itteraturselskab:

Hvad er DSL?

Det **D**anske **S**prog- og **L**itteraturselskab:

- Tekstudgivelser

Hvad er DSL?

Det **D**anske **S**prog- og **L**itteraturselskab:

- Tekstudgivelser
- Sproghistoriske fremstillinger

Hvad er DSL?

Det **D**anske **S**prog- og **L**itteraturselskab:

- Tekstudgivelser
- Sproghistoriske fremstillinger
- Bibliografier

Hvad er DSL?

Det **D**anske **S**prog- og **L**itteraturselskab:

- Tekstudgivelser
- Sproghistoriske fremstillinger
- Bibliografier
- Ordbøger og sprogteknologi

Program

1. Introduktion til DSL ✓
2. Introduktion til projektet *ordnet.dk*
3. Lemmaselektion i DDO
4. Lemmaopdatering i *ordnet.dk*
 - a. Tekstklassifikation
 - b. Bestemmelse af nye ord
 - c. Diskussion af metoden
 - d. Konklusion

Hvad er *ordnet.dk*?

Hvad er *ordnet.dk*?

ordnet.dk – kombinerer og udvider

Hvad er *ordnet.dk*?

ordnet.dk – kombinerer og udvider

I. Ordbog over det danske Sprog, ODS

Hvad er *ordnet.dk*?

ordnet.dk – kombinerer og udvider

1. Ordbog over det danske Sprog, ODS
2. Den Danske Ordbog, DDO

Hvad er *ordnet.dk*?

ordnet.dk – kombinerer og udvider

1. Ordbog over det danske Sprog, ODS
2. Den Danske Ordbog, DDO
3. Korpus 2000

Hvad er *ordnet.dk*?

ordnet.dk – kombinerer og udvider

1. Ordbog over det danske Sprog, ODS
2. Den Danske Ordbog, DDO
3. Korpus 2000

Det bliver tilgængeligt på webbet

Resurse I: ODS

Type: Ekscerptbaseret ordbog

Periode: 1700 – 1950

Målgruppe: Den „dannede“ læser

Udarbejdet: 1915 – 1956

Digitalt format: typografisk opmærket

Web: ods.ordnet.dk



Resurse 2: DDO

Type: Korpusbaseret ordbog

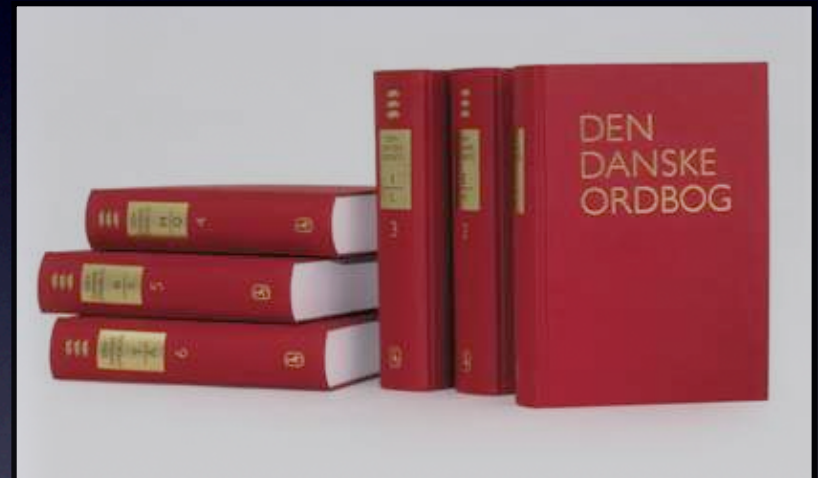
Periode: 1950 – i dag

Målgruppe: „sproginteresserede“

Udarbejdet: 1991 – 2005

Digitalt format: detaljeret XML

Web: dev.ordnet.dk/dk/ (eksperimentel)



Resurse 3: Korpus 2000

Type: Referencekorpus, 2 · 28 mio. tekstord

Periode: 1983 – 92 (Korpus 90), 1998 – 2002 (Korpus 2000)

Målgruppe: „sproginteresserede“

Udarbejdet: 2000 – 2002

Digitalt format: variabelt

Web: korpus2000.dk



Korpus 2000

ordnet.dk endnu engang!

ordnet.dk – kombinerer og udvider

1. Ordbog over det danske Sprog, ODS
2. Den Danske Ordbog, DDO
3. Korpus 2000

ordnet.dk endnu engang!

ordnet.dk – kombinerer og udvider

1. Ordbog over det danske Sprog, ODS
2. Den Danske Ordbog, DDO
3. Korpus 2000

ordnet.dk endnu engang!

ordnet.dk – kombinerer og udvider

1. Ordbog over det danske Sprog, ODS

2. Den Danske Ordbog, DDO

3. Korpus 2000

ordnet.dk endnu engang!

ordnet.dk – kombinerer og **udvider**

1. Ordbog over det danske Sprog, ODS

2. Den Danske Ordbog, DDO

3. Korpus 2000

Nye ord og
betydninger

ordnet.dk endnu engang!

ordnet.dk – kombinerer og udvider

1. Ordbog over det danske Sprog, ODS

2. Den Danske Ordbog, DDO

3. Korpus 2000

Nyt tekstmateriale

Nye ord og
betydninger

ordnet.dk endnu engang!

ordnet.dk – kombinerer og udvider

1. Ordbog over det danske Sprog, ODS

2. Den Danske Ordbog, DDO

3. Korpus 2000

Nyt tekstmateriale

Nye ord og
betydninger

Program

1. Introduktion til DSL ✓
2. Introduktion til projektet *ordnet.dk* ✓
3. Lemmaselektion i DDO
4. Lemmaopdatering i *ordnet.dk*
 - a. Tekstklassifikation
 - b. Bestemmelse af nye ord
 - c. Diskussion af metoden
 - d. Konklusion

Lemmakilder for DDO

- DDO's korpus, 40 mio. tekstord, 1983 – 92
- *Retskrivningsordbogen*
- Blinkenberg & Høybye: *Dansk-fransk ordbog*
- Vinterberg & Bodelsen: *Dansk-engelsk ordbog*
- Dansk Sprognævns register

Lemmaselektion i DDO

Lemmaselektion i DDO

I. „Manuel“ korpusbaseret: *gal* – *greb* (2%)

Lemmaselektion i DDO

1. „Manuel“ korpusbaseret: *gal – greb* (2%)
2. „Manuel“ ordbogsbaseret: bogstav A (5%)

Lemmaselektion i DDO

1. „Manuel“ korpusbaseret: *gal* – *greb* (2%)
2. „Manuel“ ordbogsbaseret: bogstav A (5%)
3. Computersimulering af (1) og (2)

Lemmaselektion i DDO

1. „Manuel“ korpusbaseret: *gal* – *greb* (2%)
2. „Manuel“ ordbogsbaseret: bogstav A (5%)
3. Computersimulering af (1) og (2)
4. Automatisk udvælgelse af resten

Selektionskriterier

Selektionskriterier

- Hovedkriteriet: Mindst 5 korpusforekomster

Selektionskriterier

- Hovedkriteriet: Mindst 5 korpusforekomster
- Konsekvens: *adjunktur, afhentningspris, amnesi* kommer ikke med

Selektionskriterier

- Hovedkriteriet: Mindst 5 korpusforekomster
- Konsekvens: *adjunktur, afhentningspris, amnesi* kommer ikke med
- Tillægskriterium: Repræsentation i mindst 3 af de 4 ordbøger

Selektionskriterier

- Hovedkriteriet: Mindst 5 korpusforekomster
- Konsekvens: *adjunktur, afhentningspris, amnesi* kommer ikke med
- Tillægskriterium: Repræsentation i mindst 3 af de 4 ordbøger
- Vi endte med ca. 30 kriterier

Program

1. Introduktion til DSL ✓
2. Introduktion til projektet *ordnet.dk* ✓
3. Lemmaselektion i DDO ✓
4. Lemmaopdatering i *ordnet.dk*
 - a. Tekstklassifikation
 - b. Bestemmelse af nye ord
 - c. Diskussion af metoden
 - d. Konklusion

Nye ord – hvorfra?

- Især avismateriale via www.infomedia.dk



Nye ord – hvorfra?

- Især avismateriale via www.infomedia.dk



Nye ord – hvorfra?

- Især avismateriale via www.infomedia.dk



Opdateringsprocessen

Opdateringsprocessen

Nye ord opdeles i emnegrupper (domæner),
før de beskrives i *ordnet.dk*:

Opdateringsprocessen

Nye ord opdeles i emnegrupper (domæner),
før de beskrives i *ordnet.dk*:

- I. Hver tekst tilordnes et domæne

Opdateringsprocessen

Nye ord opdeles i emnegrupper (domæner), før de beskrives i *ordnet.dk*:

1. Hver tekst tilordnes et domæne
2. Påfaldende ord udtrækkes fra teksten

Opdateringsprocessen

Nye ord opdeles i emnegrupper (domæner), før de beskrives i *ordnet.dk*:

1. Hver tekst tilordnes et domæne
2. Påfaldende ord udtrækkes fra teksten
3. Disse ord er kandidater til optagelse

Program

1. Introduktion til DSL ✓
2. Introduktion til projektet *ordnet.dk* ✓
3. Lemmaselektion i DDO ✓
4. Lemmaopdatering i *ordnet.dk* ✓

a. Tekstklassifikation

- b. Bestemmelse af nye ord
- c. Diskussion af metoden
- d. Konklusion

Forudsætninger

Forudsætninger

I. En brugbar domæneklassifikation

Forudsætninger

I. En brugbar domæneklassifikation

a) granularitet: *antal domæner?*

Forudsætninger

I. En brugbar domæneklassifikation

a) granularitet: *antal domæner?*

b) indhold: *afgrænsning af et domæne?*

Forudsætninger

I. En brugbar domæneklassifikation

a) granularitet: *antal domæner?*

b) indhold: *afgrænsning af et domæne?*

➤ Decimalklassifikationssystemet DK5

Forudsætninger

Forudsætninger

2. En klassifikationsprocedure

Forudsætninger

2. En klassifikationsprocedure

➤ Kvantitativ og heuristisk

Forudsætninger

2. En klassifikationsprocedure

- Kvantitativ og heuristisk
- Baseret på DDO's korpus

Forudsætninger

2. En klassifikationsprocedure

- Kvantitativ og heuristisk
- Baseret på DDO's korpus
 - 🌀 DK-klassifikation med 66 domæner

Forudsætninger

2. En klassifikationsprocedure

- Kvantitativ og heuristisk
- Baseret på DDO's korpus
 - 👁️ DK-klassifikation med 66 domæner
 - 👁️ 89% af teksterne er klassificeret

Forudsætninger

2. En klassifikationsprocedure

- Kvantitativ og heuristisk
- Baseret på DDO's korpus
 - 👁 DK-klassifikation med 66 domæner
 - 👁 89% af teksterne er klassificeret
 - 66 domænespecifikke vokabularer

Domænevokabularer

Domænevokabularer

Sådan laves domænespecifikke vokabularer:

Domænevokabularer

Sådan laves domænespecifikke vokabularer:

- I. Opbygning af domænespecifikke subkorpora

Domænevokabularer

Sådan laves domænespecifikke vokabularer:

1. Opbygning af domænespecifikke subkorpora
2. Opstilling af frekvensprofiler

Domænevokabularer

Sådan laves domænespecifikke vokabularer:

1. Opbygning af domænespecifikke subkorpora
2. Opstilling af frekvensprofiler
3. Sammenligning af frekvensprofiler

Domænevokabularer

66

forskellige domænekoder
i korpus

Sådan laves domænespecifikke vokabularer:

1. Opbygning af domænespecifikke subkorpora
2. Opstilling af frekvensprofiler
3. Sammenligning af frekvensprofiler

Domænevokabularer

66

forskellige domænekoder
i korpus

for

hele korpus og
for hvert af de 66
subkorpora

Sådan laves domænespecifikke vokabularer:

1. Opbygning af domænespecifikke subkorpora
2. Opstilling af frekvensprofiler
3. Sammenligning af frekvensprofiler

Domænevokabularer

66

forskellige domænekoder
i korpus

for

hele korpus og
for hvert af de 66
subkorpora

Sådan laves domænespecifikke vokabularer:

1. Opbygning af domænespecifikke subkorpora
2. Opstilling af frekvensprofiler
3. Sammenligning af frekvensprofiler

hvert

af de 66 frekvensprofiler
sammenlignes med profilen for
hele korpus. Signifikanstest: log
likelihood ($p \geq 0,99$)

Tre domænevokabularer

Edb	Filosofi	Økonomi
<i>data</i>	<i>mennesket</i>	<i>kr</i>
<i>programmer</i>	<i>kierkegaard</i>	<i>X,X</i>
<i>computer</i>	<i>moral</i>	<i>pct</i>
<i>computeren</i>	<i>løgstrup</i>	<i>procent</i>
<i>edb</i>	<i>aristoteles</i>	<i>kroner</i>
<i>computere</i>	<i>filosofi</i>	<i>rente</i>
<i>ibm</i>	<i>fornuft</i>	<i>offentlige</i>
<i>pc</i>	<i>platon</i>	<i>økonomiske</i>
<i>kan</i>	<i>kierkegaards</i>	<i>bank</i>
<i>mb</i>	<i>tim</i>	<i>X</i>
<i>apple</i>	<i>den</i>	<i>økonomi</i>
<i>amiga</i>	<i>menneskets</i>	<i>vil</i>
<i>commodore</i>	<i>filosof</i>	<i>mia</i>

Tre domænevokabularer

Edb	Filosofi	Økonomi
<i>data</i> <i>programmer</i> <i>computer</i> <i>computeren</i> <i>edb</i> <i>computere</i> <i>ibm</i> <i>pc</i> <i>kan</i> <i>mb</i> <i>apple</i> <i>amiga</i> <i>commodore</i>	<i>mennesket</i> <i>kierkegaard</i> <i>moral</i> <i>løgstrup</i> <i>korpusset</i> <i>indeholder overvejende</i> <i>tekster fra 1980'erne</i> <i>kierkegaards</i> <i>tim</i> <i>den</i> <i>menneskets</i> <i>filosof</i>	<i>kr</i> <i>X,X</i> <i>pct</i> <i>procent</i> <i>kroner</i> <i>rente</i> <i>offentlige</i> <i>økonomiske</i> <i>bank</i> <i>X</i> <i>økonomi</i> <i>vil</i> <i>mia</i>

Tre domænevokabularer

Edb	Filosofi	Økonomi
<i>data</i> <i>programmer</i> <i>computer</i> <i>computeren</i> <i>edb</i> <i>computere</i> <i>ibm</i> <i>pc</i> <i>kan</i> <i>mb</i> <i>apple</i> <i>amiga</i> <i>commodore</i>	<i>mennesket</i> <i>kierkegaard</i> <i>moral</i> <i>løgstrup</i> <i>aristoteles</i> <i>filosofi</i> <i>fornuft</i> <i>platon</i> <i>kierkegaards</i> <i>tim</i> <i>den</i> <i>menneskets</i> <i>filosof</i>	<i>kr</i> <i>X,X</i> <i>frekvente</i> <i>ord fra én tekst kan</i> <i>snige sig ind</i> <i>offentlige</i> <i>økonomiske</i> <i>bank</i> <i>X</i> <i>økonomi</i> <i>vil</i> <i>mia</i>

Tre domænevokabularer

Edb	Filosofi	Økonomi
<i>data</i> <i>programmer</i> <i>computer</i> <i>computeren</i> <i>edb</i> <i>computere</i> <i>ibm</i> <i>pc</i> <i>kan</i> <i>mb</i> <i>apple</i> <i>amiga</i> <i>commodore</i>	<i>mennesket</i> <i>kierkegaard</i> <i>moral</i> <i>løgstrup</i> <i>aristoteles</i> <i>filosofi</i> <i>fornuft</i> <i>platon</i> <i>kierkegaards</i> <i>tim</i> <i>den</i> <i>menneskets</i> <i>filosof</i>	<i>kr</i> <i>X,X</i> <i>pct</i> <i>procent</i> <i>kroner</i> <i>rente</i> <i>offentlige</i> <i>økonomiske</i> <i>bank</i> <i>X</i> <i>økonomi</i> <i>vil</i> <i>mia</i>

tal (cifre)
generaliseres

Tre domænevokabularer

Edb	Filosofi	Økonomi
<i>data</i> <i>programmer</i> <i>computer</i> <i>computeren</i> <i>edb</i> <i>computere</i> <i>ibm</i> <i>pc</i> <i>kan</i> <i>mb</i> <i>apple</i> <i>amiga</i> <i>commodore</i>	<i>mennesket</i> <i>kierkegaard</i> <i>moral</i> <i>løgstrup</i> <i>aristoteles</i> <i>filosofi</i> <i>generelt</i> <i>højfrekvente ord</i> <i>kommer delvis også</i> <i>med</i> <i>den</i> <i>menneskets</i> <i>filosof</i>	<i>kr</i> <i>X,X</i> <i>pct</i> <i>procent</i> <i>kroner</i> <i>rente</i> <i>offentlige</i> <i>økonomiske</i> <i>bank</i> <i>X</i> <i>økonomi</i> <i>vil</i> <i>mia</i>

Metodiske problemer

Metodiske problemer

I. Signifikansniveauet ($p \geq 0,99$) er arbitrært

Metodiske problemer

I. Signifikansniveauet ($p \geq 0,99$) er arbitrært

- indvirker på domænevokabularets omfang

Metodiske problemer

1. Signifikansniveauet ($p \geq 0,99$) er arbitrært
 - indvirker på domænevokabularets omfang
2. Domænekorporaene er forskelligt store

Metodiske problemer

1. Signifikansniveauet ($p \geq 0,99$) er arbitrært

- indvirker på domænevokabularets omfang

2. Domænekorporaene er forskelligt store

- indvirker på domænevokabularets omfang

Metodiske problemer

1. Signifikansniveauet ($p \geq 0,99$) er arbitrært

- indvirker på domænevokabularets omfang

2. Domænekorporaene er forskelligt store

- indvirker på domænevokabularets omfang

Domæne	Antal typer
Folklore	1957
Sport	16022
SNIT	7256

Metodiske problemer

1. Signifikansniveauet ($p \geq 0,99$) er arbitrært

- indvirker på domænevokabularets omfang

2. Domænekorporaene er forskelligt store

- indvirker på domænevokabularets omfang

3. Højfrekvente ord optræder som signifikante

Metodiske problemer

1. Signifikansniveauet ($p \geq 0,99$) er arbitrært

- indvirker på domænevokabularets omfang

2. Domænekorporaene er forskelligt store

- indvirker på domænevokabularets omfang

3. Højfrekvente ord optræder som signifikante

- indvirker på domænetilordningen

Metodiske problemer

1. Signifikante ord (p < 0.05) optræder
 - indvirker på domænetilordningen
2. Domænetilordning
 - indvirker på domænetilordningen
3. Højfrekvente ord optræder som signifikante
 - indvirker på domænetilordningen

Type	“Typisk”
<i>kan</i>	Edb
<i>den</i>	Filosofi
<i>vil</i>	Økonomi

Tekstklassifikationen

Tekstklassifikationen



Grund-idé:

Tekstklassifikationen



Grund-idé:

- Største antal type-overensstemmelser mellem et domænevokabular D og vokabularet T i den tekst der skal klassificeres

Tekstklassifikationen



Grund-idé:

- Største antal type-overensstemmelser mellem et domænevokabular D og vokabularet T i den tekst der skal klassificeres
- Mere formelt: Find den største fællesmængde $|D \cap T|$

Tekstklassifikationen

Vokabular-overlapning



Grund-idé:

- > Største antal type-overensstemmelser mellem et domænevokabular D og vokabularet T i den tekst der skal klassificeres
- > Mere formelt: Find den største fællesmængde $|D \cap T|$

Tekstklassifikationen

Største vokabular-overensstemmelse?

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 1

En typeoverensstemmelse mellem tekst og domænevokabular tæller altid kun én selvom den er højfrekvent i teksten

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 1

En typeoverensstemmelse mellem tekst og domænevokabular tæller altid kun én selvom den er højfrekvent i teksten



Løsning

Tæl i stedet overensstemmelser mellem tekst-tokens og typer i domænevokabularerne

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 1

En typeoverensstemmelse mellem tekst og domænevokabular tæller altid kun én selvom den er højfrekvent i teksten



Løsning

Tæl i stedet overensstemmelser mellem tekst-tokens og typer i domænevokabularerne

Type eller token?

Tekstklassifikationen

Største vokabular-overensstemmelse?

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 2

„Funktionsord“ kan få for høj vægt

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 2

„Funktionsord“ kan få for høj vægt



Løsning

Tag højde for antallet af domænevokabularer
som et givet token matcher

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 2

„Funktionsord“ kan få for høj vægt



Løsning

Tag højde for antallet af domænevokabularer
som et givet token matcher

Unikhed

Tekstklassifikationen

Største vokabular-overensstemmelse?

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 3

Domæner med store vokabularer
vil have lettere ved at score højt

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 3

Domæner med store vokabularer
vil have lettere ved at score højt



Løsning

Tag højde for størrelsen
af de enkelte domænespecifikke vokabularer

Tekstklassifikationen

Største vokabular-overensstemmelse?



Problem 3

Domæner med store vokabularer
vil have lettere ved at score højt



Løsning

Tag højde for størrelsen
af de enkelte domænespecifikke vokabularer

Domænestørrelse

Beregning af score

Vokabular-overlapning

+

Type eller token?

$$S_D =$$

Beregning af score

Vokabular-overlapning

+

Type eller token?

Lad hvert token t fra teksten W
som matcher en type i domænevokabularet D
addere en bestemt værdi w til scoren

$$S_D =$$

Beregning af score

Vokabular-overlapning

+

Type eller token?

Lad hvert token t fra teksten W
som matcher en type i domænevokabularet D
addere en bestemt værdi w til scoren

$S_D =$

$$\sum_{t \in D \cap W} w_t$$

Beregning af score

Unikhed

$$S_D = \sum_{t \in D \cap W} w_t$$

Beregning af score



Unikhed

Værdien w skal være omvendt proportional til
antallet af domæner d
i hvis vokabularer tekst-tokenet optræder

$$S_D = \sum_{t \in D \cap W} w_t$$

Beregning af score

Unikhed

Værdien w skal være omvendt proportional til
antallet af domæner d
i hvis vokabularer tekst-tokenet optræder

$$S_D = \sum_{t \in D \cap W} w_t$$

$$w = \frac{1}{d} \text{ hvor}$$

$$d = \sum_i |t \cap D_i|$$

Beregning af score

Domænestørrelse

$$S_D = \sum_{t \in D \cap W} w_t$$

Beregning af score

Domænestørrelse

Beregn en vægt v for den samlede score der skal være omvendt proportional til omtrent størrelsen af domænevokabularet D

$$S_D = \sum_{t \in D \cap W} w_t$$

Beregning af score

Domænestørrelse

Beregn en vægt v for den samlede score der skal være omvendt proportional til omtrent størrelsen af domænevokabularet D

$$S_D = \sum_{t \in D \cap W} w_t \cdot v \quad \text{hvor} \quad v = \frac{1}{\sqrt{|D|}}$$

Beregning af score

Eksperimentel forbedring: „Kendthed“

$$s_D = \sum_{t \in D \cap W} w_t \cdot v \cdot$$

Beregning af score

Eksperimentel forbedring: „Kendthed“

Beregn en vægt for den samlede score
som tager højde for forholdet mellem
de tekst-tokens som optræder i et domænevokabular,
og dem som ikke gør

$$S_D = \sum_{t \in D \cap W} w_t \cdot v \cdot$$

Beregning af score

Eksperimentel forbedring: „Kendthed“

Beregn en vægt for den samlede score
som tager højde for forholdet mellem
de tekst-tokens som optræder i et domænevokabular,
og dem som ikke gør

$$S_D = \sum_{t \in D \cap W} w_t \cdot v \cdot \frac{k}{u}$$

Beregning af score

Relativering af scoren

$$S_D = \sum_{t \in D \cap W} w_t \cdot v \cdot \frac{k}{u} \cdot$$

Beregning af score

Relativering af scoren

Gør scoren **relativ til** tekstlængden i antal tokens

$$S_D = \sum_{t \in D \cap W} w_t \cdot v \cdot \frac{k}{u} \cdot$$

Beregning af score

Relativering af scoren

Gør scoren **relativ til** tekstlængden i antal tokens

$$S_D = \sum_{t \in D \cap W} w_t \cdot v \cdot \frac{k}{u} \cdot \frac{1}{n}$$

Program

1. Introduktion til DSL ✓
2. Introduktion til projektet *ordnet.dk* ✓
3. Lemmaselektion i DDO ✓
4. Lemmaopdatering i *ordnet.dk* ✓
 - a. Tekstklassifikation ✓
 - b. Bestemmelse af nye ord
 - c. Diskussion af metoden
 - d. Konklusion

Bestemmelse af nye ord

Bestemmelse af nye ord

I. Sammenlign frekvensprofiler vha. en statistisk test (log likelihood)

- nyt domænespecifikt materiale
- DDO's korpus

Bestemmelse af nye ord

1. Sammenlign frekvensprofiler vha. en statistisk test (log likelihood)

- nyt domænespecifikt materiale
- DDO's korpus

2. Fremtrædende ord i det nye materiale er kandidater til optagelse i ordbogen

Eksempel: tekst

Du skal bruge en diskette til installationen. På et tidspunkt bliver du spurgt om du vil lave en bootdiskette. Erfaringen siger at det godt kan betale sig at formatere en diskette i forvejen med tjek for dårlige sektorer. Før du installerer Linux, skal der være en partition til rådighed, der er stor nok til at rumme det hele (samt en swap-partition). I løbet af Linux-installationen vil der blive lejlighed til at repartitionere så meget, du har behov for, inden for den plads, der nu er blevet til rådighed.

Eksempel: procedure

Exempel: procedure

I. Klassifikation

Eksempel: procedure

I. Klassifikation

✓ Teksten klassificeres som edb-tekst

Eksempel: procedure

1. Klassifikation

✓ Teksten klassificeres som edb-tekst

2. Sammenligning

Eksempel: procedure

1. Klassifikation

✓ Teksten klassificeres som edb-tekst

2. Sammenligning

🌀 Tekstens beskedne størrelse forvrænger

Eksempel: procedure

1. Klassifikation

- ✓ Teksten klassificeres som edb-tekst

2. Sammenligning

- 👁 Tekstens beskedne størrelse forvrænger
- ✓ List tekstens fremtrædende ord

Eksempel: procedure

1. Klassifikation

- ✓ Teksten klassificeres som edb-tekst

2. Sammenligning

- 👁 Tekstens beskedne størrelse forvrænger
- ✓ List tekstens fremtrædende ord
- ✓ Tilføj DDO-domænekoder til listen

Nye ord?

Type	<i>f</i> i DDOC	<i>f</i> i teksten	DDO-fag
<i>diskette</i>	78	2	edb
<i>bootdiskette</i>	0	1	artikel mangler
<i>formatere</i>	0	1	edb
<i>linux</i>	0	1	artikel mangler
<i>linux-installationen</i>	0	1	artikel mangler
<i>partition</i>	0	1	artikel mangler
<i>repartitionere</i>	0	1	artikel mangler
<i>swap-partition</i>	0	1	artikel mangler

Nye ord?

Type	<i>f</i> i DDOC	<i>f</i> i teksten	DDO-fag
<i>diskette</i>	78	2	edb
<i>bootdiskette</i>	0	1	artikel mangler
<i>formatere</i>	0	1	edb
<i>linux</i>	0	1	artikel mangler
<i>linux-installationen</i>	0	1	artikel mangler
<i>partition</i>	0	1	artikel mangler
<i>repartitionere</i>	0	1	artikel mangler
<i>swap-partition</i>	0	1	artikel mangler

Nye betydninger?

Type	<i>f</i> i DDOC	<i>f</i> i teksten	DDO-fag
<i>rådighed</i>	1730	2	alment
<i>installerer</i>	16	1	alment teknik
<i>du</i>	143798	5	alment
<i>installationen</i>	34	1	teknik kunst militær
<i>tjek</i>	100	1	alment
<i>sektorer</i>	112	1	samfund politik matematik

Nye betydninger?

Type	<i>f</i> i DDOC	<i>f</i> i teksten	DDO-fag
<i>rådighed</i>	1730	2	alment
<i>installerer</i>	16	1	alment teknik
<i>du</i>	143798	5	alment
<i>installationen</i>	34	1	teknik kunst militær
<i>tjek</i>	100	1	alment
<i>sektorer</i>	112	1	samfund politik matematik

Program

1. Introduktion til DSL ✓
2. Introduktion til projektet *ordnet.dk* ✓
3. Lemmaselektion i DDO ✓
4. Lemmaopdatering i *ordnet.dk* ✓
 - a. Tekstklassifikation ✓
 - b. Bestemmelse af nye ord ✓
 - c. Diskussion af metoden
 - d. Konklusion

Diskussion af metoden

Diskussion af metoden

- Opgave: Bestem nye domæne-ord til leksikografisk beskrivelse

Diskussion af metoden

- Opgave: Bestem nye domæne-ord til leksikografisk beskrivelse
- Fremgangsmåde:
 1. Korpus $\Rightarrow$ domæne-vokabularer
 2. Domæne-vokabularer $\Rightarrow$ tekstklassifikation
 3. Klassificeret materiale $\leftrightarrow$ korpus
 4. Fremtrædende ord $\Rightarrow$ nye ord/betydninger

Beslutninger

Beslutninger

1. DDO-korpussets domæneklassifikation
2. Signifikanstest
3. Klassifikationsproceduren

I. Domæneklassifikation

I. Domæneklassifikation

- Stort antal domæner (66!)

I. Domæneklassifikation

- Stort antal domæner (66!)
 - Færre domæner?

I. Domæneklassifikation

- Stort antal domæner (66!)
 - Færre domæner?
- Stor forskel på mængden af tekstmateriale for hvert domæne (fra 1957 til 16022 ord)

I. Domæneklassifikation

- Stort antal domæner (66!)
 - Færre domæner?
- Stor forskel på mængden af tekstmateriale for hvert domæne (fra 1957 til 16022 ord)
 - Mindre mængdeforskel?

2. Signifikanstest

2. Signifikanstest

- Log likelihood

2. Signifikanstest

- Log likelihood
 - „Arbitrært“ valg

2. Signifikanstest

- Log likelihood
 - „Arbitrært“ valg
 - Bedre egnede tests?

2. Signifikanstest

- Log likelihood
 - „Arbitrært“ valg
 - Bedre egnede tests?
 - Hvordan forholder de sig til fænomenet?

3. Procedure

3. Procedure

- Bør afspejle egenskaber ved teksten
 - Token-overlapning
 - Vokabularstørrelse
 - Unikhed
 - Domænestørrelse
 - „Kendthed“

3. Procedure

- Bør afspejle egenskaber ved teksten
 - Token-overlapning
 - Vokabularstørrelse
 - Unikhed
 - Domænestørrelse
 - „Kendthed“

Andre egenskaber?
Intuitive karakteristika!
Passende kvantificering?

Testning

Testning

- Komplekse gensidige afhængigheder
 - Test af forskellige alternerende parametre

Testning

- Komplekse gensidige afhængigheder
 - Test af forskellige alternerende parametre
- Mulig test:
 1. Del DDOC op i 2 dele med samme relative andel tekst fra hvert domæne
 2. Del 1 $\Rightarrow$ domænevokabularer
 3. Del 2 $\Rightarrow$ testning

Program

1. Introduktion til DSL ✓
2. Introduktion til projektet *ordnet.dk* ✓
3. Lemmaselektion i DDO ✓
4. Lemmaopdatering i *ordnet.dk* ✓
 - a. Tekstklassifikation ✓
 - b. Bestemmelse af nye ord ✓
 - c. Diskussion af metoden ✓
 - d. Konklusion

Konklusion

Konklusion

- Godt:

Konklusion

- Godt:
 - Metoden er brugbar til opgaven

Konklusion

- Godt:
 - Metoden er brugbar til opgaven
- Skidt:

Konklusion

- Godt:
 - Metoden er brugbar til opgaven
- Skidt:
 - Metoden giver ingen svar på ...

Konklusion

- Godt:
 - Metoden er brugbar til opgaven
- Skidt:
 - Metoden giver ingen svar på ...
 - Hvad gør et ord eller en tekst *domæne-specifik*?

Konklusion

- Godt:
 - Metoden er brugbar til opgaven
- Skidt:
 - Metoden giver ingen svar på ...
 - Hvad gør et ord eller en tekst *domæne-specifik*?
 - Hvad gør et ord til et *nyt ord*?

Program

1. Introduktion til DSL ✓
2. Introduktion til projektet *ordnet.dk* ✓
3. Lemmaselektion i DDO ✓
4. Lemmaopdatering i *ordnet.dk* ✓
 - a. Tekstklassifikation ✓
 - b. Bestemmelse af nye ord ✓
 - c. Diskussion af metoden ✓
 - d. Konklusion ✓

Tak for jeres interesse!