Automatisk textsammanfattning

Hercules Dalianis
NADA-KTH
100 44 Stockholm
email:hercules@nada.kth.se
tel 08-790 91 05

Automatisk textsammanfattning

- Automatisk textsammanfattning är tekniken där en dator sammanfattar en text. En text matas in i datorn och ut kommer en sammanfattad text.
- Tekniken har utvecklats under 30 år, men först idag med Internet och Webben har den fått större användning. Microsoft Word 97, 98 och Word 2000 har en sammanfattare av dokument.

- Automatisk textsammanfattning bygger på statistiska, lingvistiska och heuristiska metoder där sammanfattnings-systemet tex beräknar hur ofta vissa nyckelord förekommer i texten, i vilka meningar, och var meningarna befinner sig i texten

Automatisk textsammanfattning

1) Vilka är nyckelorden?
2) Vad handlar texten om?
3) Skapa sammanfattningen.

•Två metoder är
  –Textabstraktion eller textextraktion

(Textabstraktion görs av människor och är än så länge för svårt för en dator)

Textextraktion

- Stoppa in en text i datorn och ut kommer ett icke-redundant extrakt av orginaltexten

Textsammanfattning (extraktion)

Statistiska, lingvistiska och heuristiska metoder

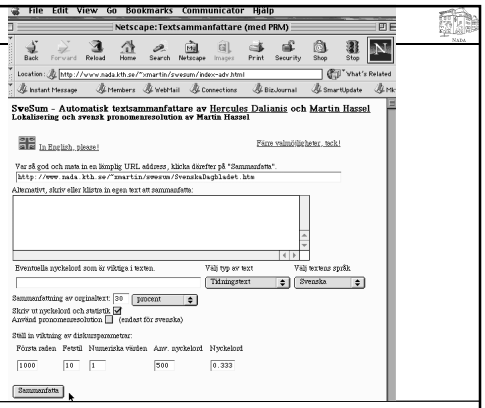
- Meningspositioner (Nyheter/Rapporter)
- Titelord
- Fetstil,Stycken,
- Citat
- Ordfrekvenser (Nyckelord-använd lexikon)
- Nyckelord som förekommer i alla andra meningar.

- Användaranpassning
  - Nyckelordsval
- Kombinationsfunktion av alla beräkningar med olika vikter räknar fram meningarnas rangordning.

SweSum

- Den första textsammanfattaren för svenska
- Sammanfattar svensk tidningstext i HTML/text format på webben
- Använder ett svenskt lexikon på 40.000 nyckelord och deras 700.000 böjningar.
- Under sammanfattningen produceras 5-10 nyckelord- en minisammanfattning.
- Träffsäkerhet 84% vid 40% sammanfattning av 181 ords nyhetstexter (medelvärde)

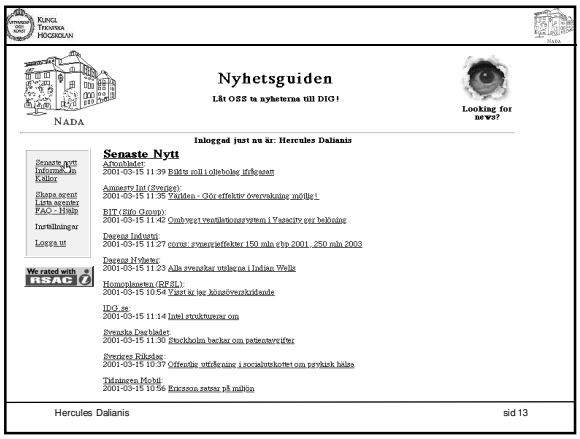
- Text sammanfattaren finns även för engelska,tyska och spanska texter
- Swesum använder Perl-CGI script, det finns även en standalone version för text/html.
- http://www.nada.kth.se/~xmartin/ seasum/index.html

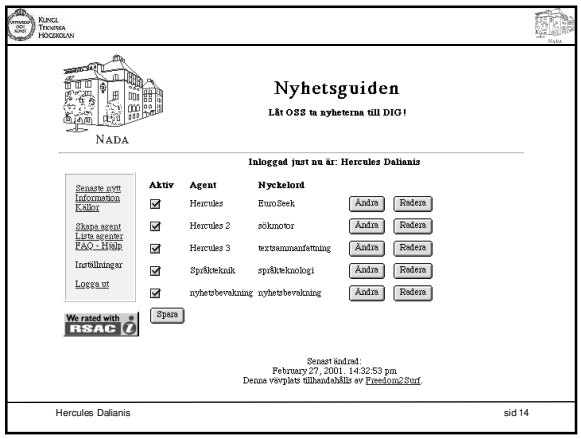


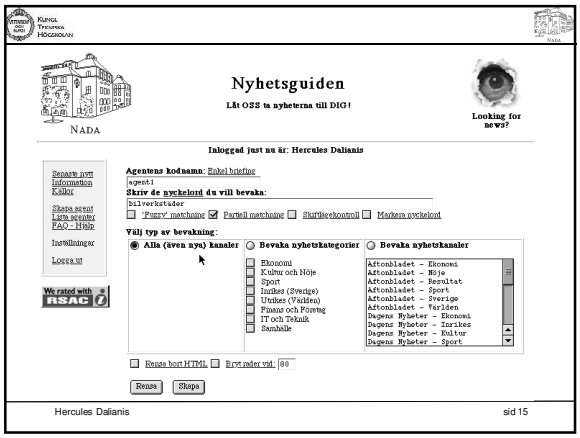
- Cell exemplet
- Prova

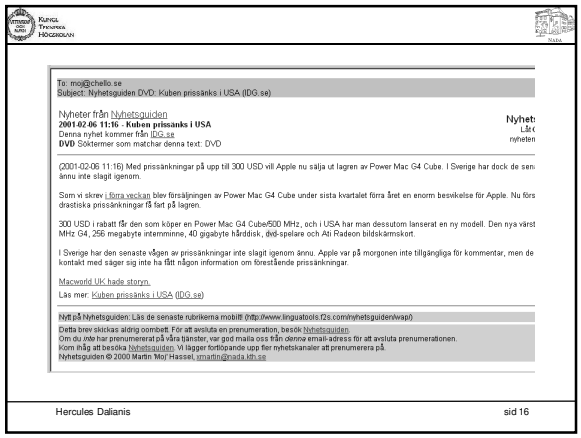
Nyhetsguiden

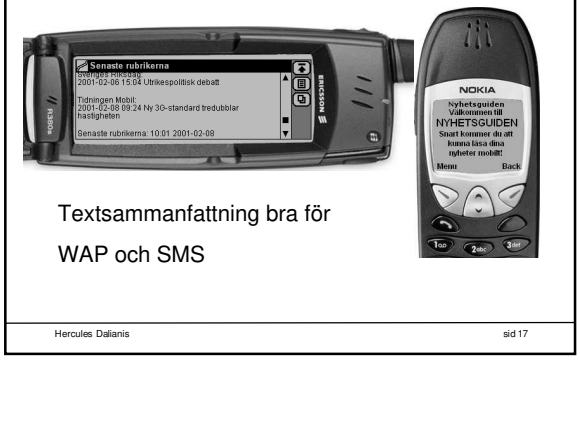
- Agent bevakar 12 stycken nyhetskanaler
- Användaren talar om vilka nyckelord han/hon och vilka nyhetskanaler han/hon vill bevaka.
- Ordet/orden eller del av ordet skall finnas i nyheten, dvs stemming tillämpas.
- Om detta uppfylls så mailas nyheten sammanfattad eller som vanligt till användaren.











Textsammanfattning bra för

WAP och SMS

TDT-Topic Detection Tracking

- TDT, innebär att med datorer automatiskt detektera nyheter, extrahera och översätta dessa.
- Engelska och kinesiska nyheter på radio, TV och vanliga textmedia bevakas med taligenkänning och statistiska och lingvistiska metoder.

Fem olika uppgifter i TDT.

- Segmentation - att kunna urskilja olika nyheter i nyhetsfloden,
- Tracking - att kunna följa nyheter med samma ämne/topic.
- Detection - att kunna gruppera nyheter efter ämne,
- Link detection- har två nyheter ngn koppling ämnesmässigt och slutligen det svåraste och intressantaste
- First Story Detection kunna upptäcka en helt ny nyhet i nyhetsfloden.
- Man lyckas hyfsat bra med de fyra första uppgifterna, man gör 33-25 % fel där, men med First Story Detection hamnar man på 70% fel.

Tillämpningar

- Sammanfattning av nyhetstext till SMS eller WAP format för mobiltelefon
- Att låta en dator läsa upp sammanfattade webbsidor per telefon
- I sökmotorer för presentera sökresultaten komprimerat (jmfr Google)
- Nyckelordsstyrt abonnemang av nyheter som presenteras sammanfattade med Pushteknik

- Omvärldbevakning (Business Intelligence) av tidningstext (för journalister, mediabevakare, omvärldsbevakare, etc).
- MUST - Söker på främmande språk, sammanfattar och översätter.

Framtid

- NE name entity recognition
  - —Namn-Plats-Organisation-Tid-igenk nning
- Textsammanfattning med NE
- Multitextsammanfattning
  Sammanfatta flera liknade texter till en icke-redundant text
- TDT-Topic Detection Tracking

Textsammanfattare

http://www.nada.kth.se/~hercules/HDbookmarks.html

- MuST Prototype
- Automated Text Summarization (SUMMARIST)
- Autonomy
- Datahammer (Glucose Development Corporation)

- EchoSearch Search and Summarize
- Extractor: Text Summarization Software
- Intelligent Miner for Text - Summarization tool
- Inxight (XEROX)
- Infogist
- SweSum (KTH)

- Mitre
- OracleContext
- ProSum . . . Text Summarisation For Word
- Microsoft Office 97 AutoSummarize
- Websumm
- WebCompass