

Stockholms universitet

KTH

Text Mining

Sumithra Velupillai
DSV-SU-KTH
kontakt: sumithra@dsv.su.se

Sumithra Velupillai 1

Stockholms universitet

KTH

Text Mining - Överblick

- Introduktion
- Relationer och metoder
- Tillämpningar
- Kort sammanfattning

Sumithra Velupillai 2

Stockholms universitet

KTH

Text Mining – Introduktion

- Vad är Text Mining?
 - Upptäcka ny, *tidigare okänd* information från textresurser
 - Hitta mönster och ny kunskap i textmängder
 - → Knowledge Discovery

Sumithra Velupillai 3

Stockholms universitet

KTH

Text Mining – Introduktion

- Varför Text Mining?
 - ”Låta (data)mängden tala för sig själv”
 - Informationsrymden stor, en stor del är text
 - Nya upptäckter baserad på den information som finns

Sumithra Velupillai 4

Stockholms universitet

KTH

Text Mining – Relationer

```

graph TD
    DM[Data Mining] --> TM[Text Mining]
    DM --> WM[Web Mining]
    TM --> WM
  
```

Sumithra Velupillai 5

Stockholms universitet

KTH

Text Mining – Metoder

- Knowledge Discovery – vad innebär det?
 - Samla ihop (stor!) datamängd (texter)
 - Skapa lämplig beskrivning/representation av mängden
 - Formulera hypotes från beskrivningen (hitta ny information)
 - Testa hypotesen (evaluering)

Sumithra Velupillai 6

Stockholms universitet KTH

Text Mining – Metoder

- Text Mining → "Text **data** mining" – metoder från data mining
 - Statistik
 - Maskininlärning
 - IR/IE
 - Visualisering
 - etc
- Anpassat till textuell data

Sumithra Velupillai 7

Stockholms universitet KTH

Text Mining – Metoder

```

graph TD
    DM[Data Mining] --> TM[Text Mining]
    DM --> WM[Web Mining]
    TM --> WM
    TM --> ST[Språkteknologi]
    WM --> B[Statistik, Databaser, Maskininlärning, IR/IE, Visualisering, etc.]
  
```

Sumithra Velupillai 8

Stockholms universitet KTH

Text Mining – Metoder

- Mål: skapa "bild" av datamängden
 - Beskrivande (deskriptiv)
 - Klustering
 - Sammanfattning
 - etc
 - Prediktiv
 - Kategorisering/klassifikation
 - Beslutsträd
 - etc

Sumithra Velupillai 9

Stockholms universitet KTH

Text Mining vs Data Mining

- Data Mining:
 - Analysera strukturerade datamängder – "siffror"!
- Text Mining:
 - Analysera ostrukturerade datamängder – "texter"!
 - Representation av datamängden: "texter" → "siffror"
 - Strukturerade det ostrukturerade
 - Vektormodellen tex

Sumithra Velupillai 10

Stockholms universitet KTH

Text Mining vs Web Mining

- Web Mining:
 - Utnyttja all information på webben
 - En stor del av informationsmängden är text → Text Mining

Sumithra Velupillai 11

Stockholms universitet KTH

Text Mining vs IR/IE

- IR/IE:
 - Hitta/ta fram information som finns och matchar ett informationsbehov
- (Text) Mining:
 - Hitta ny, tidigare okänd information
 - Hitta mönster
 - Utnyttja IR/IE → utforskning

Sumithra Velupillai 12



Text Mining vs Språkteknologi

- Språkteknologi:
 - Empirisk/statistisk språkteknologi
- Text Mining:
 - Utnyttja språkteknologiska metoder
 - Stemming/ordklasstaggning/syntax/disambiguering etc.
- Hjälper språkteknologi Text Mining?

Sumithra Velupillai
15

Text Mining – Tillämpningar

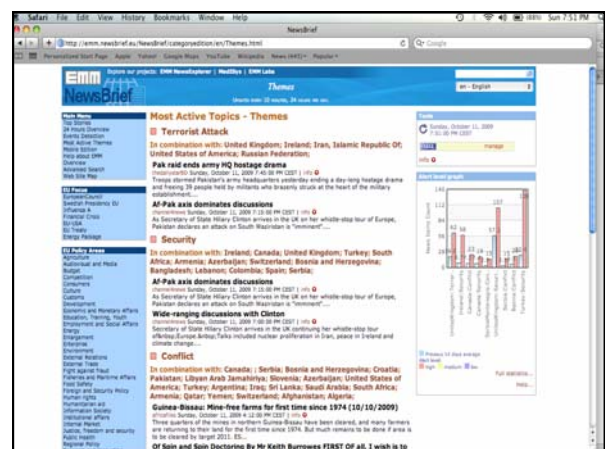
- Marknadsföring: kundprofilering, marknadssegmentering:
 - ”Vem skulle kunna vilja köpa min produkt?”
 - ”Vill denna kund få förslag på fler produkter?”
 - ”Vilken marknad ska vi rikta in oss på för denna produkt?”
- Hitta via texter: webben, enkäter etc.

Sumithra Velupillai
16

Text Mining – Tillämpningar

- Beteendemönster (på tex webben)
 - Loggar
 - Klickmönster
 - ”Är detta en kvinna eller man?”
 - Grooming: ”Följer potentiella förövare vissa mönster?”
 - Försäljningsmönster
 - etc.

Sumithra Velupillai
17



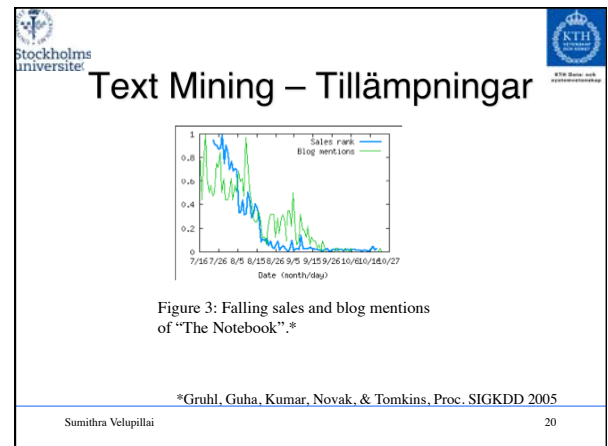


Figure 3: Falling sales and blog mentions of "The Notebook" *

*Gruhl, Guha, Kumar, Novak, & Tomkins, Proc. SIGKDD 2005

Sumithra Velupillai

20

Text Mining – Tillämpningar

- Bioinformatik
 - Stor mängd information i textform
 - Artikeldatabaser
 - Ex PubMed
 - Gendatabaser
 - Etc
 - Bas för att kommunicera vetenskapliga upptäckter

Sumithra Velupillai

21

Text Mining – Tillämpningar

- Bioinformatik
 - Domänspecifikt språk
 - Tekniska termer
 - Ambiguösa termer
 - Akronymer
 - Ex LTP: Long-term potentiation
 - Nya termer kontinuerligt
 - Typografisk variation
 - ex gener: TNF-alpha, TNF alpha

Sumithra Velupillai

22

Text Mining – Tillämpningar

- Bioinformatik
 - Hitta nya kopplingar via litteraturen
 - Swanson – ARROWSMITH
 - Univ. Of Illinois, Chicago: http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html
 - Länka information från litteratur (Medline) → nya relationer
 - A är länkat till B men inte till C, B är dock länkat till C, är A länkat till C?
 - Undersöka manuellt

Sumithra Velupillai

23

Text Mining – Tillämpningar

- stress is associated with migraines
- stress can lead to loss of magnesium
- calcium channel blockers prevent some migraines
- magnesium is a natural calcium channel blocker
- spreading cortical depression (SCD) is implicated in some migraines
- high levels of magnesium inhibit SCD
- migraine patients have high platelet aggregability
- magnesium can suppress platelet aggregability*

* Från Hearst, *Untangling Text Data Mining*, Proceedings of ACL '99: <http://people.ischool.berkeley.edu/~hears/papers/acl99/acl99-tdm.html>

Sumithra Velupillai

24

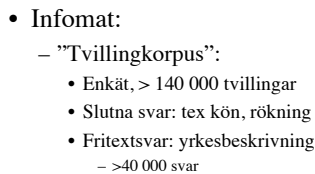
Stockholm
universitet

KTH Sverige AB

- From a complex of relationship network, Chilibot suggested that CREB may have a relationship (e.g. regulation) with synaptotagmin, based on 1) there is no known relationship between CREB and synaptotagmin; and 2) they are both connected to many gene/protein and are both involved in similar biological process (e.g. LTP). The original search included one keyword ("LTP") and 22 other genes. CREB was first selected by clicking on the graph. Hypothesis was generated by clicking on the "CREB might be related to..." button.

26

Stockholms
universitet



28

Stockholms
universitet

- Infomat
 - Klustrering av de öppna svaren
 - Vilka grupperingar (kluster) verkar skilja sig markant från 71%?
 - Manuellt skapade sammanfattande namn på klustren med utgångspunkt i de vanligaste orden/kluster

30



Stockholms universitet



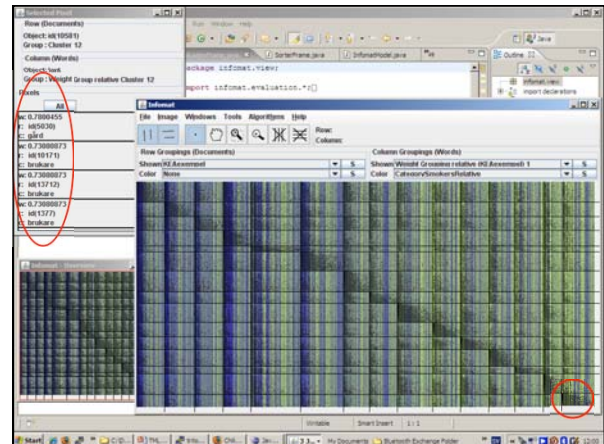
KTH

Text Mining – Tillämpningar

- Infomat
 - Folk som jobbar inom jordbruk, ca 78%
 - vanliga ord: lant, kreatur, skörd, bond, odla, mjölka
 - Denna grupp verkar vara icke-rökare
 - Evaluering: sökning efter studier som bekräftar/avfärdar

Sumithra Velupillai

31





Stockholms universitet



KTH

Text Mining – Kort sammanfattning

- Identifiera ny, tidigare okänd information från stora mängder data via helhetsbild
- Ungt ämne – mycket kvar att upptäcka!
 - [Wikipedia 2007 vs. 2009](#)
- Litteratur:
 - Marti Hearst *Untangling Text Data Mining*:
<http://people.ischool.berkeley.edu/~hears/papers/ac199/ac199-tdm.html>

Sumithra Velupillai

33