

Automatisk extraktion av nyckelord

Johan Eklund

Bibliotekshögskolan
Högskolan i Borås



HÖGSKOLAN I BORÅS
VETENSKAP FÖR PROFESSION

Om LIVA-projektet

- Forsknings- och utvecklingsprojekt mellan Bibliotekshögskolan, BTJ och Bibliotekscentrum samt fem projektbibliotek
- Delvis finansierat under 2005-2007 av KK-stiftelsen
- Mål: att applicera moderna metoder och resultat från forskningen inom biblioteks- & informationsvetenskap på urvalsmängder av bibliografiska data och visa hur befintliga system kan förbättras genom tillämpning av dessa rön



Indexering

- Indexering är processen att skapa en formell beskrivning av ett dokument innehåll genom tilldelning av deskriptorer
- Klassiskt, manuellt tillvägagångssätt:
 - vi har tillgång till en uppsättning indextermer
 - ett *urval* (t ex 5 st) av dessa tilldelas dokumentet
- Automatiserat tillvägagångssätt:
 - vi har tillgång till en uppsättning indextermer
 - *samtliga* av dessa tilldelas dokumentet tillsammans med ett numeriskt värde (en s k **vikt**) som anger graden av association mellan dokument och term



Termviktning

- Ett mått på associationen mellan en term **t** och ett dokument **d** bör svara på frågan: ”om vi slumpmässigt väljer ett dokument som indexerats av **t**, hur sannolikt är det att detta dokument är **d**?”
- Detta skrivs formellt $P(d | t)$, vilket utläses ”sannolikheten för **d** givet **t**”
- Hur beräknar vi denna sannolikhet? Vi använder oss av ett knep som går under benämningen **Bayes teorem**



Bayes teorem

- **Bayes teorem**, uppkallat efter prästen och matematikern Thomas Bayes (1702-1761) har en stor betydelse för beräkning av s k konditionala sannolikheter inom statistisk inferens (= slutledning)



Thomas Bayes

- $P(d|t) = P(t|d) P(d) / P(t)$
- $P(d|t) \propto P(t|d) / P(t)$

↑
Läses "är proportionell mot"



HÖGSKOLAN I BORÅS
VETENSKAP FÖR PROFESSION

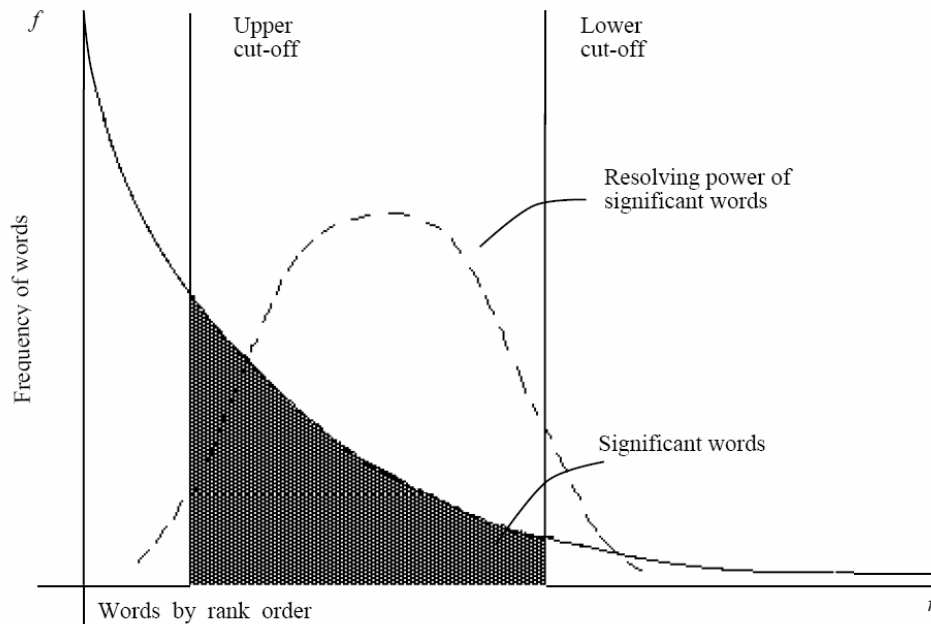
Termviktning

- Vi låter $P(t|d)$ vara den relativa frekvensen av term t i dokument d och kallar detta mått **tf**
- Vi låter $1/P(t)$ vara inversen av den relativa frekvensen av term t i hela dokumentkollektionen och kallar detta mått **idf**
- Måttet **tf-idf** = **tf** \times **idf** har länge tillämpats inom automatisk indexering och bygger på en kombination av två principer:
 - en term är betydelsefull i ett dokument om den förekommer många gånger i dokumentet
 - en term är betydelsefull i ett dokument om den förekommer få gånger i kollektionen



Indexering med termviktning

- Redan på 1950-talet publicerades idéer om hur termernas frekvenser i dokumenten kan användas för att mäta termernas diskrimineringsförmåga
- Luhn, H. (1958). *The automatic creation of literature abstracts*.



Exempel

	kamp	seger	bröllop	straff	räddning
d_1	0,8	0,5	0,0	0,7	0,6
d_2	0,1	0,1	0,7	0,1	0,2
d_3	0,3	0,3	0,5	0,4	0,6
d_4	0,8	0,9	0,4	0,5	0,8



Termviktning

- Värdet 0,5 i relationen (eg. funktionen)
 $(d_1, \text{seger}) \rightarrow 0,5$
kallas termens **vikt** i dokumentet
- Ju högre vikt desto större relation mellan dokument och term
- Enkel termviktning: 0 eller 1 (binär)
- Termviktning baserad på termfrekvens



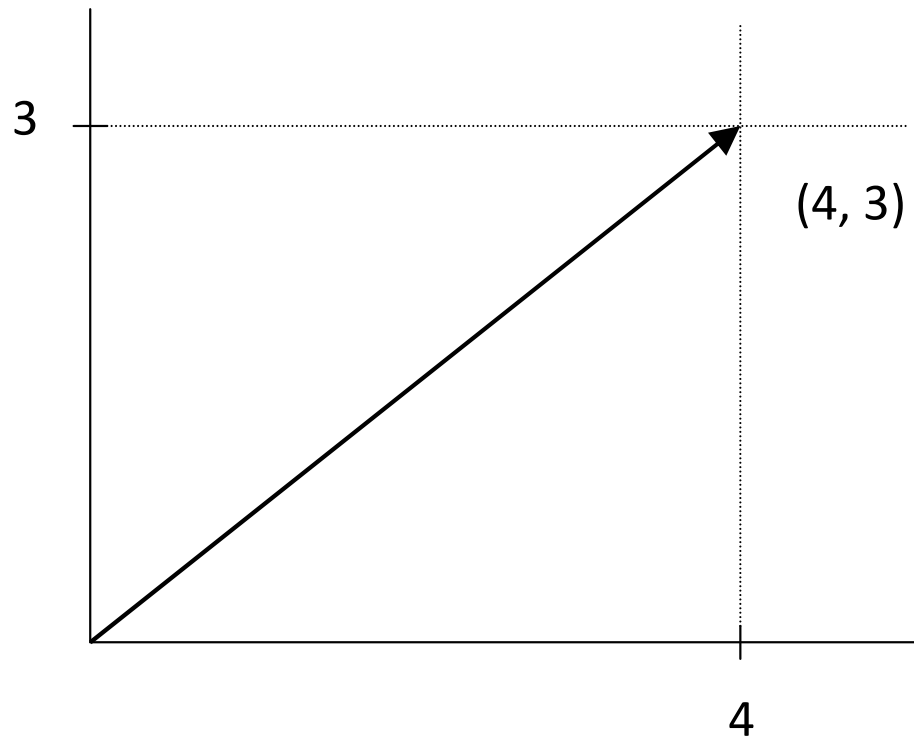
Vektorrymdsmodellen

- Gerard Salton, Cornell University
- SMART retrieval system, sent 1960-tal
- Dokument och sökfrågor representeras som **vektorer** (\approx punkter) i ett högdimensionellt rum
- Antal dimensioner i rummet bestäms av antalet unika indextermer som används i kollektionen
- Koordinaterna hos en dokumentvektor utgörs av dess termvikter

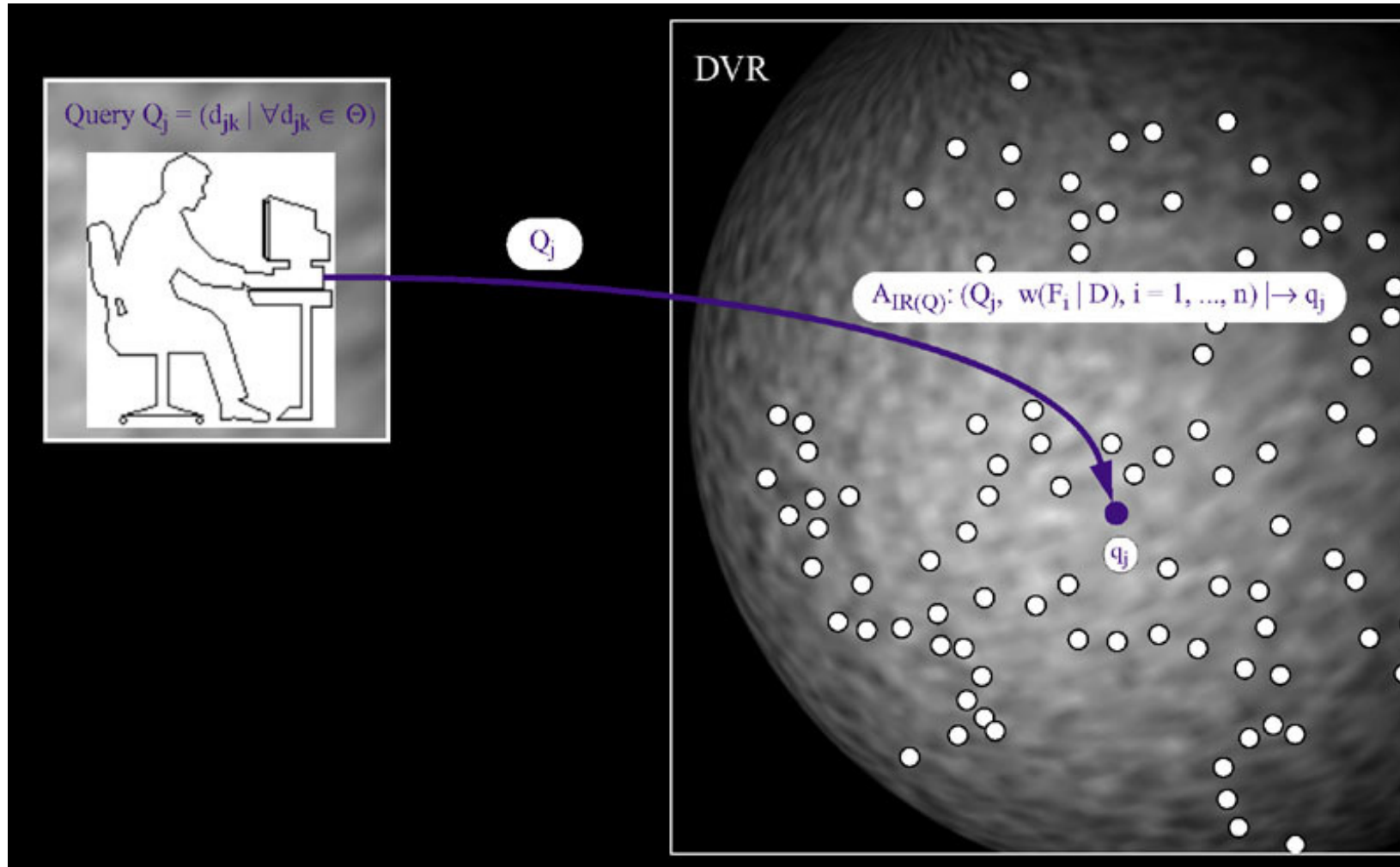


Vektorer

- En vektor är en matematisk struktur som kan beskrivas som:
 - en sekvens av numeriska värden
 - en entitet med en storlek och en riktning i ett rum



Vektorrymsmodellen



Fördelar med vektorrymdsmodellen

- Vi kan geometriskt mäta likheten mellan dokument och sökfråga istället för att som ifråga om boolsk sökning svara ”uppfyller / uppfyller inte sökfrågan” för respektive dokument
- Med vektorrymdsmodellen kan vi alltså beräkna **i vilken utsträckning** dokumentet verkar motsvara sökfrågans innehåll
- Dokument kan rankas utifrån mått på likhet med sökfrågan, t ex genom att mäta cosinus för vinkeln (!) mellan dokument och sökfråga



Problem vid sökning

- Vanligen utför ett informationssystem en **lexikal matchning** vilket innebär att sökfrågans ord måste matcha dokumentrepresentationens ord för att dokumentet skall återvinnas
- I en miljö av naturligt, okontrollerat, språk är denna matchningsstrategi ofta ett problem
- Fulltextindex och manuellt skapade index befinner sig vanligen på olika semantiska nivåer



Exempel

- Jag söker information om engelska bilar från 1960-talet
- Sökfrågan "**engelska bilar**" **AND 1960-talet** misslyckas med att återvinna dokument som inte representeras av dessa termer utan av **Bentley** respektive **1962**.
- Bentley är en engelsk bil (hyponymi)
- 1962 är en del av 1960-talet (meronymi)



Query expansion

- Processen att utöka en sökfråga i avsikt att förbättra dess infångningsförmåga
- Nackdel: resultatet blir ofta större, med fler intressanta dokument, men också med fler ointressanta dokument
- Högre recall på bekostnad av lägre precision
- Interaktiv query expansion: användaren får förslag på sökfrågetermer av systemet och väljer vilka av dessa som skall användas



Latent semantisk analys

- Dokumenten analyseras statistiskt för att hitta **mönster** för deras **samförekomster**
- Ord som "tränare" och "spelare" tenderar att förekomma i **samma texter**
- Ord som "respektera" och "högakta" tenderar att förekomma i **samma kontexter**
- Vår metod är att koda texter numeriskt och tillämpa verktyg för klustring för att detektera sådana mönster



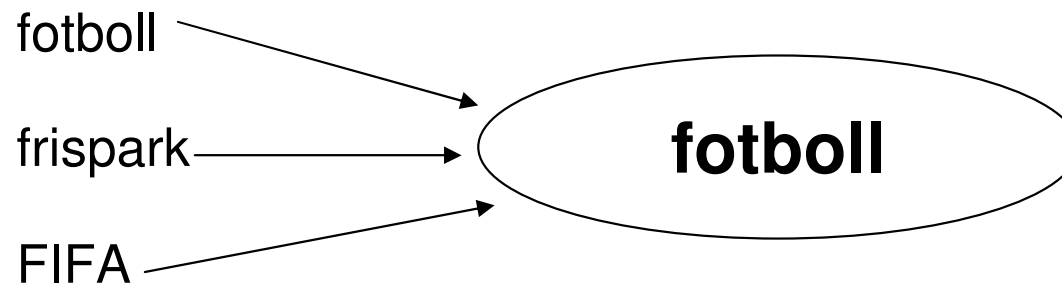
Latent semantisk analys

- Enligt språkfilosofer som Ludwig Wittgenstein är det problematiskt att tala om fixa lexikala betydelser hos ord
- Ords betydelser formas av hur de används
- För att uttrycka det matematiskt: betydelsen hos ett ord är en funktion av ordets kontext
- Två ord kan misstänkas ha en snarlik betydelse om deras kontexter är snarlika
- Vi säger isåfall att orden har liknande **samförekomstmönster**



Latent semantisk analys

- En familj av metoder som statistiskt detekterar en betydelselighet mellan termer utifrån hur de används i dokumenten
- Termer med liknande **samförekomstmönster** avbildas på en gemensam representation



Klustring av termer med LSI

- Stresshantering Hjärtinfarkt Rehabilitering Livsstil Livskvalitet Hjärtrehabilitering Könsskillnader_sjukvård
- Djurförsök Försöksdjur
- Hypertoni Högt_blodtryck Blodtryck
- Mongolism Medicinsk_genetik Genetik_medicin Utvecklingsstörda Psykiskt_utvecklingsstörda Downs_syndrom Förståndshandikappade Medicin_historia
- Miljöförstöring Hälsorisker Gifter Miljögifter
- Hjärt-kärlsjukdomar Kolesterol Blodfett Hjärtsjukdomar Kärlsjukdomar Blodfettsänkande_behandling
- Fingerborgsblomma Digitalis_purpurea Medicinalväxter
- UV-strålning Solbränna Ultraviolet_tstrålning Solskyddsmedel Cancer Hudcancer Malignt_melanom Melanom
- Sårbehandling Fotvård Skavsår



Morfologisk normalisering

- De nordiska språken har en rik morfologi, dvs en stor mängd ordformer (såväl grammatiska som derivativa)
- En rik morfologi utgör ett problem när ord jämförs som strängar, exempelvis vid sökning
- Sträng = sekvens av tecken (fotboll = f·o·t·b·o·l·l)
- {fotboll, fotbollen, fotbollar, fotbollsspelare, fotbollslag, ...} är exempel på morfologiska varianter på fotboll

Morfologisk normalisering

- Ord i fulltext kan normaliseras till reducerad form i index genom
 - stemming: ordet reduceras till en ordstam, t ex pojkar → pojk
 - lemmatisering: ordet normaliseras till ett lemma (lexikal grundform), t ex gick → gå, pojkar → pojke
 - sammansättningsuppdelning: ordet delas upp i lexikala beståndsdelar (fotboll → fot + boll)
- Samtliga metoder kan bidra till ökad recall (täckning) men riskerar lägre precision
- Stemming kräver minst implementeringsresurser



Ordklasstagging och frasindexering

- En del nyckelord behöver kvalificeras för att "säga hela sanningen" – det är m a o viktigt att kunna identifiera fraser (sekvenser av minst två ord) som deskriptorer
- Ta som exempel "rött vin", "södra Sverige"
- För att implementera frasindexering tillämpar vi tf-idf-viktning av sekvenser av ord tillsammans med ordklasstagging
- Ordklasstagging (även part-of-speech tagging) är en datalingvistisk teknik för att tilldela ordklasser (substantiv, adjektiv, verb...)
- Som fras-kandidat räknas en sekvens av adjektiv och substantiv, jfr exemplen ovan



Levenshtein-avstånd

- Ett mått för jämförelser av ord på strängnivå
- Värdet på måttet anger antalet ändringar som behöver utföras för att förändra ett ord till ett annat
- Ändringar: tillägg, borttag eller utbyte av en bokstav
- Exempel: skaft → skotta : 3 ändringar
- Kan användas för att implementera stavningsförslag i en söktjänst



Nyckelordsextraktion: tillvägagångssätt

- Kandidater till egennamn märks upp i texten
- Heuristisk princip: ett ord, eller en sekvens av ord, i en svensk text är troligen ett egennamn om
 - begynnelsebokstaven är stor
 - ordet står inte först i en mening
- Texten konverteras till gemener och interpunktionstecken tas bort
- Numeriska värden tas bort
- Ett ord identifieras som en enhet av tecken som föregås och efterföljs av blanktecken eller tom sträng



Tillvägagångssätt forts

- Texten tokeniseras (delas upp i ord) och resulterande ord behandlas med stemming för att sammanföra olika ordformer till samma stam
- {gård, gården, gårdar} → gård
- Ordfrekvens beräknas för dokument resp. kollektion och tf-idf-värden tilldelas samtliga termer
- Orden rankas i fallande ordning efter tf-idf
- Ord som antas vara egennamn särskiljs från övriga ord men tilldelas och rankas efter tf-idf



Exempel på utdata

[2006] [Kampen om kärnkraften]

KW: kärnkraftens, folkomröstningen, vpk, center, folkomröstning, energiformen, kärnkraftsomröstningen, omröstningen, avvecklingen, folkpartiet

PN: Linje, Harrisburg, Barsebäcks, Three Mile Island, Ola Ullsten, Tage Danielssons, Thorbjörn Fälldin, I Sverige, Joakim Thelander

[2006] [Malmö 1939]

KW: hälsingborg, hamnfotot, kylslaget, gemensamhetsbad, avsnörpt, lamslå, hamnbassängen, parveln, flygbåtar, passagerarkön, ångfärjan

PN: Skeppsbron, Malmö, Jungfru Marie, Kockumskranen, Fotoatelier Otto, Turning Torso, Hjälmarekajen, Ohm, Mölle

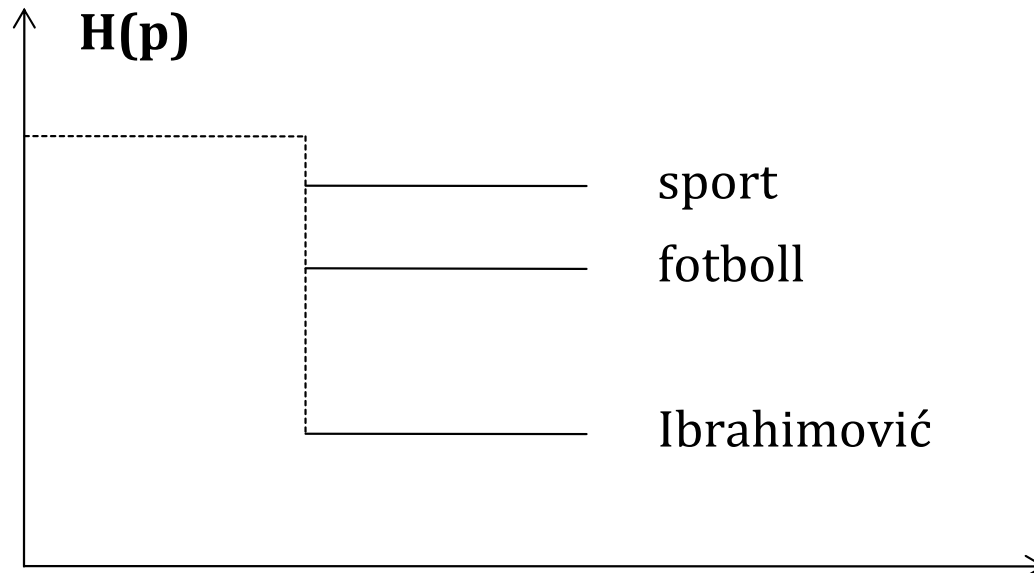
Informationsteori och osäkerhet

- Claude Shannon definierade mängden information i ett meddelande som **graden av osäkerhet** associerad med meddelandets innehåll; ju mer osäkerhet desto mer information
- Låt π vara ett problem med N olika svar; π kan exempelvis vara frågan "vad handlar dokument \mathbf{d} om?"
- $H(\pi)$ är mängden information (osäkerhet) i problemet π
- Antag att vi vet att dokument \mathbf{d} är indexerat av term \mathbf{t}
- Den nya osäkerheten skriver vi $H(\pi | \mathbf{t})$
- Ju större skillnaden $H(\pi) - H(\pi | \mathbf{t})$, desto mer information har vi fått om problemet



Information gain

- **Information gain** är ett mått som anger hur mycket osäkerheten kring ett problem minskar då vi inför en faktor (som t ex en indexterm)



Termevaluering av index

- En terms diskrimineringsförmåga avgör hur väl den skiljer mellan dokument av olika innehåll
- Vi har tillämpat **information gain**, genom att för en given term t beräkna hur mycket information t rymmer om en eller flera specifika ämneskategorier
- Termen "frispark" är en mer signifikant term för ämnet "fotboll" än t ex termen "eftermiddag"
- För att mäta informationsmängden i en term tillämpar vi idéer från informationsteori; mängden information i en signal kvantifieras i termer av hur mycket signalen **bidrar** till tidigare kunskap

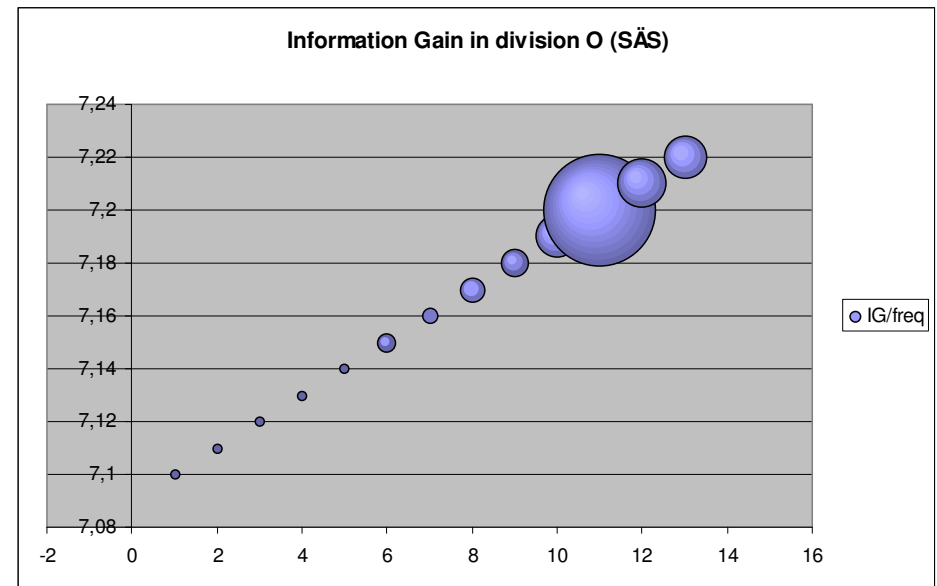
$$H(\Omega) = - \sum_{\omega_i \in \Omega} p(\omega_i) \log_2 p(\omega_i)$$



Utvärdering av index

glesbygd	7,221207
handikappade ungdomar	7,221207
biståndsbedömning	7,221207
långtidssjuka	7,221207
invandrarbarn	7,221207

hälso- och sjukvård	6,806173
barn	6,783628
statistik	6,731118
facklitteratur	6,446311
sverige	6,122083



Information Gain hos
indextermer i avdelning O
(urval från SÄS)



HÖGSKOLAN I BORÅS
VETENSKAP FÖR PROFESSION

Automatisk textsammanfattning

- Ett alternativ/komplement till att skapa en lista med deskriptorer är att generera en sammanfattning av texten
- Extraktion – abstraktion
- Enkel metod för textsammanfattning:
 1. Dela upp texten i meningar.
 2. Tilldela varje mening ett värde baserat på en viktad summa av faktorerna "första mening", "innehåller fetstil", "innehåller numeriska värden", "innehåller nyckelord".
 3. Ranka meningarna i fallande ordning efter tilldelat värde.
 4. Skapa en ny text av de $n\%$ första meningarna i listan enligt 3.



Exempel på utdata

En kraftig brand har under dagen härjat i Stratford i östra London.

- Det är ett tjockt svart moln som syns överallt, säger Hanna Elin Åhman, som bor i östra London, till SvD.se och fortsätter: - Tanken slog mig att det kunde ha något med terror att göra, men jag blev inte speciellt oroad.

Vittnen beskrev röken som lika kraftig som när World Trade Center i New York utsattes för terrorattacker 2001, men polisen uppger nu att de inte tror att det rör sig om ett terrordåd.

