Språkteknologi (SV2122)
Föreläsning 4: Kategorisering av text

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

Richard Johansson

richard.johansson@svenska.gu.se

7 februari 2014

översikt

information

- föreläsning 9 och datorlaboration har bytt plats i schemat

var är vi nu?

- de första föreläsningarna har varit ganska grundläggande
- nu kommer vi i stället till tillämpningarna

kategorisering

- ett **kategoriseringsprogram** tar en text och placerar den i en kategori
- vilken typ av kategorier vi använder beror på vad vi vill göra

spamfiltrering

Clemens Cavallin är docent i religionsvetenskap/religionshistoria och kommer under seminariet att berätta om sin tid vid ett av USAs främsta liberal arts colleges, Haverford college, där han under höstterminen 2013 fick möjlighet att undervisa genom STINT och Teaching Sabbatical (f.d. Excellence in Teaching).

Goda nyheter, Du VANN! Du VANN! DU VANN!!! Du är veckans vinnare hos Karamba och du tog hem MEGA-BONUSEN! Ta reda på hur mycket du vann! Din mega-bonus innehåller:

1. 20 gratissnurr att spela dina favoritspel med
2. En chans att vinna den stora 1,000,000kr jackpotten!
3. 100% Välkomstbonus

klassificering enligt ämne

- exempelvis såsom bibliotek gör
  - SAB, LCC

A - Bok- och biblioteksväsen
B - Allmänt och blandat
C - Religion
E - Uppfostran och undervisning
G - Litteraturvetenskap
H - Skönlitteratur
I - Konst, musik, teater och film
L - Biografi med genealogi
N - Geografi
O - Samhälls- och rättsvetenskap
P - Teknik, industri och kommunikationer
R - Idrott, lek och spel
X - Musikalier

målgruppsklassificering

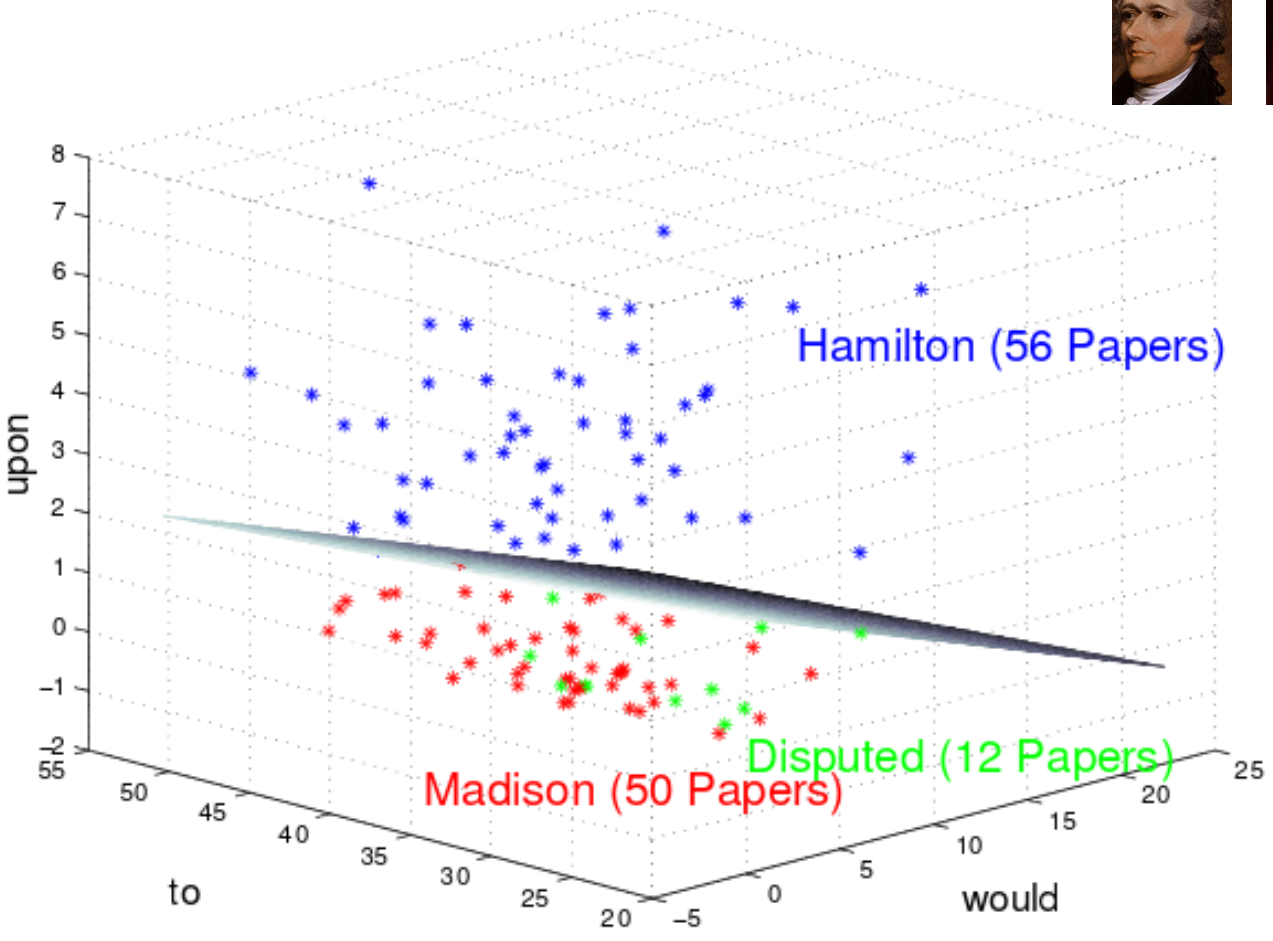
- textsamlingen MedEval innehåller medicinsk text av olika slag (Läkartidningen, etc)
  - vi kan använda ett kategoriseringsprogram för att skilja på text avsedd för patienter och för vårdpersonal

- kategorisering av text enligt svårighetsgrad, t.ex. för att hitta lämpliga texter för inlärare
  - t.ex. enligt CEFR (A1, . . . , C2)

att skilja på författare / författaregenskaper

- är det Shakespeare eller Marlowe som har skrivet den här texten?
- har författaren svenska som modersmål?
- är texten översatt från något annat språk?
- är författaren man eller kvinna?
- har författaren någon åkomma som påverkar språket? (afasi efter stroke, Alzheimer, . . .)

exempel: Federalist Papers



example: disambiguation of word meaning in context

*A woman and child suffered minor injuries after the car they were riding in crashed into a **rock** wall Tuesday morning.*

- what is the meaning of *rock* in this context?

- S: (n) **rock**, stone (a lump or mass of hard consolidated mineral matter) *"he threw a rock at me"*
- S: (n) **rock**, stone (material consisting of the aggregate of minerals like those making up the Earth's crust) *"that mountain is solid rock"; "stone is abundant in New England and there are many quarries"*
- S: (n) **Rock**, John Rock (United States gynecologist and devout Catholic who conducted the first clinical trials of the oral contraceptive pill (1890-1984))
- S: (n) **rock** ((figurative) someone who is strong and stable and dependable) *"he was her rock during the crisis"; "Thou art Peter, and upon this rock I will build my church"--Gospel According to Matthew*
- S: (n) rock candy, **rock** (hard bright-colored stick candy (typically flavored with peppermint))
- S: (n) rock 'n' roll, rock'n'roll, rock-and-roll, rock and roll, **rock**, rock music (a genre of popular music originating in the 1950s; a blend of black rhythm-and-blues with white country-and-western) *"rock is a generic term for the range of styles that evolved out of rock'n'roll."*
- S: (n) **rock**, careen, sway, tilt (pitching dangerously to one side)

översikt

automatisk analys av attityd och värdering i text

- kan ett datorprogram avgöra vilka värderingar/känslor/attityder som uttrycks i en text?
- detta är intressant i många sammanhang, t.ex.
  - fick min reklamkampanj någon mätbar effekt i vad folk säger på nätet om min produkt?
  - hur populär är Reinfeldt bland svenska skribenter på nätet?
- ämnet är en djungel när det gäller både terminologi och problemformulering:
  - "sentiment analysis"
  - "opinion mining"
  - ...
- för översikt, se t.ex. Pang och Lee (2008) eller Liu (2012)

exempel

Grekisk pizzeria
Plaka
Utlandagatan 14

Det är inte mycket som lockar till en pizzeria i dag eftersom det är oftast är bukfylla utan vidare ambitioner, men ibland finns en parallell taffel som kan glimra i hemlighet. Vi gick till Plaka som ligger i ett annars restaurangsnålt område.

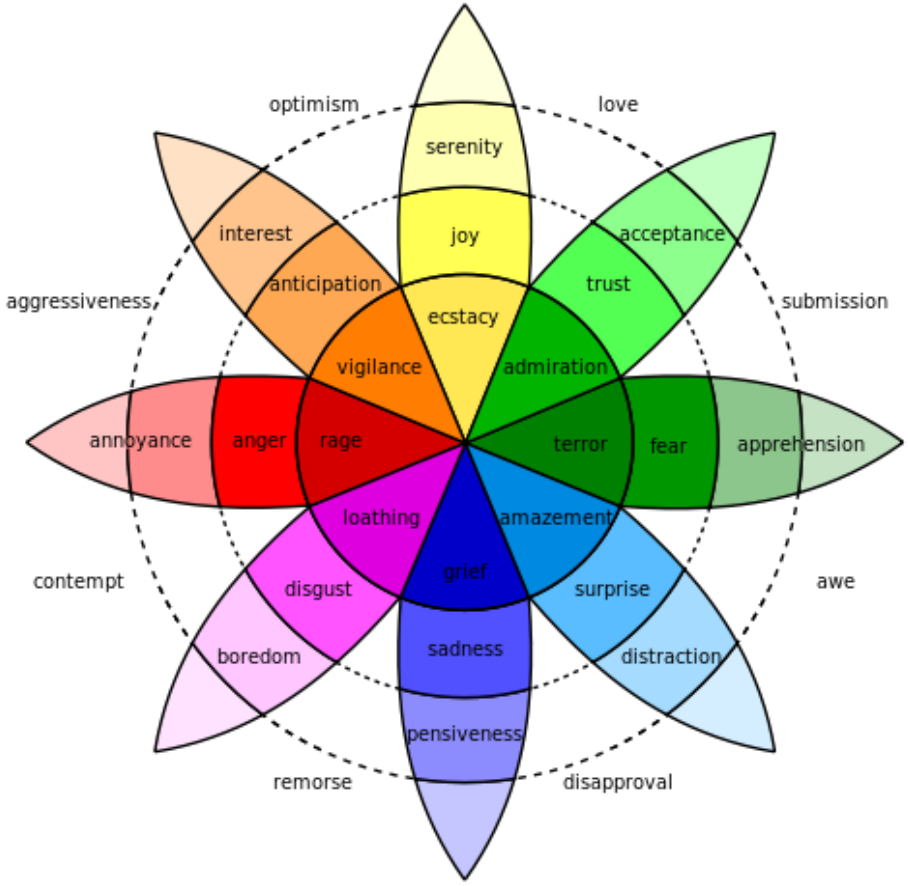
Vi anade snart att det var servitrisen som var kvällens blyga behållning – snabb, vänlig, lyhörd. Plaka har en pizzameny med hela 49 varianter mellan 85 och 95 kronor som slutar med en Gyros Pizza. Det föreföll lika äckligt som Kebab Pizza. En maträtt från helvetet som kanske borde returneras dit.

Därför sneglade vi mot den grekiska menyn med den famlande tanken om att den grekiska krisen kanske ändå inte nått ända dit. Vi beställde två standardförrätter. Grillad halloumi och friterad bläckfisk. I halvfabrikatens värld fordras inget geni direkt

en möjlig problemformulering

- givet en text, kan ett datorprogram gissa . . .
  - hur många GP-fyrar? (**värdering**)
    - bara bra / dålig eller en skala?
    - vilka slags värderingar ska vi tala om?
  - vad är det vi värderar? (**ämne**)
    - vilka saker som var bra och vilka var dåliga? (**aspekter**)
  - vem det är som tycker? (**källa**)

mer än GP-fyrar: Robert Plutchiks "wheel of emotions"

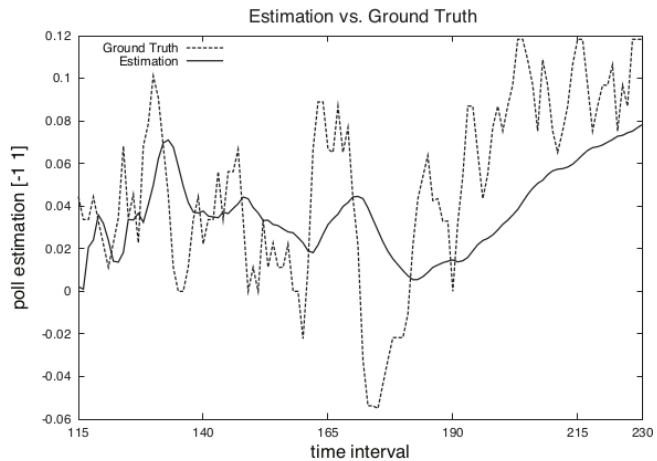


enklaste fallet: avgör om positiv eller negativ

- enklast möjliga uppgift: avgör om ett dokument (t.ex. en recension) uttrycker en positiv eller negativ värdering
- den naiva lösningen:
  - tilldela varje enskilt ord någon "poäng" oberoende av sitt sammanhang
    - poängen kan komma från ett lexikon eller en statistisk modell (Pang, Lee, Vaithyanathan, 2002)
  - summera alla poäng
  - om summan större än en given tröskel antas dokumentet uttrycka positiv värdering, annars negativ
- denna lösning är uppenbart lingvistiskt bristfällig men **oerhört svår att slå i praktiken**
  - typiskt runt 85% korrekthet för t.ex. filmrecensioner (se t.ex. Pang, Lee, Vaithyanathan, 2002; Johansson och Moschitti, 2013)
  - dock ganska genre-känslig! (Blitzer et al., 2007)

statsvetenskaplig tillämpning

- Demartini et al. (2011) använde värderingsanalysverktyg i ett stort antal bloggtexter från hösten 2008 (Obama mot McCain)
- de kunde visa på en korrelation över tiden med traditionella opiniondundersökningar



diskursstruktur och värderingar

- utsagorna i en text är inte oberoende utan de är sammanlänkade i en argumentationskedja – en **diskursstruktur**
  - exempel: kontrastrelation

    Rummet stank av cigarettrök ***men*** utsikten över Seine var fantastisk.
  - exempel: förstärkningsrelation

    Puben hade en mysig atmosfär ***och dessutom*** var inredningen smakfull.
- det är uppenbart att argumentationen spelar en viktig roll när värderingar uttrycks
- . . . och den ordbaserade modellen är teoretiskt undermålig (bl.a.) eftersom den inte tar hänsyn till detta

hjälper automatisk diskursstrukturanalys?

- diskursstruktur t.ex. enligt PDTB kan vi (till en viss del) analysera automatiskt
- . . . men de flesta försök att tillämpa detta för analys av värderingar har inte lyckats
- alldeles nyligen har det kommit fungerande modeller som använder diskurs (Lazaridou et al., 2013; Socher et al., 2013)
- detta är genre-specifika statistiska modeller som oftast inte anknyter till diskursteori: framtiden får utvisa om de kan göras mer allmänt tillämpliga

översikt

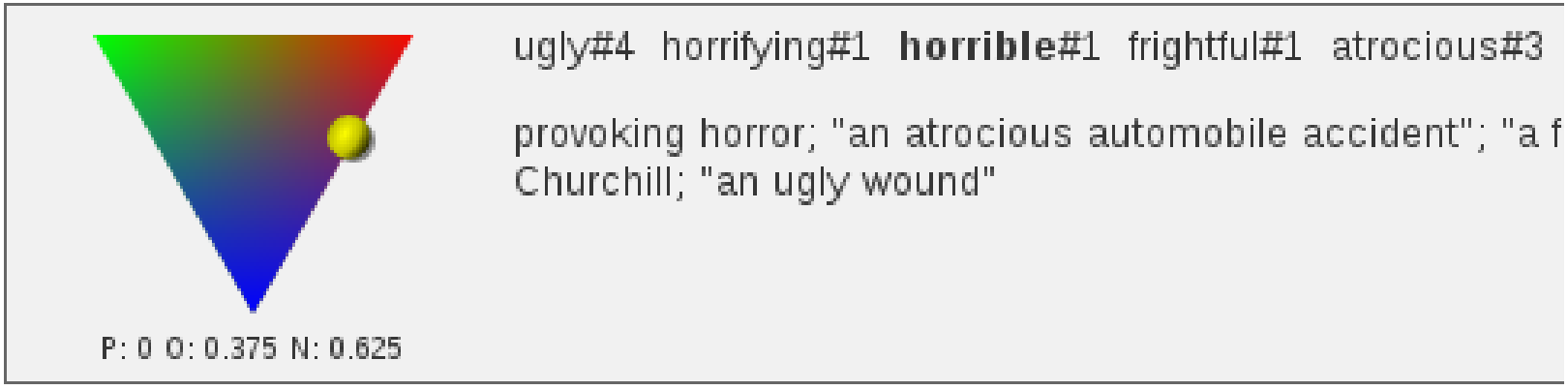
one approach: use a sentiment wordlist

- . . . for instance the MPQA list

```
type=strongsubj len=1 word1=wretchedly pos1=anypos stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=wretchedness pos1=noun stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=writhe pos1=verb stemmed1=y priorpolarity=negative
type=weaksubj len=1 word1=wrong pos1=adj stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=wrong pos1=anypos stemmed1=y priorpolarity=negative
type=weaksubj len=1 word1=wrongful pos1=adj stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=wrongly pos1=anypos stemmed1=y priorpolarity=negative
type=weaksubj len=1 word1=wrought pos1=adj stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=wrought pos1=noun stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=wry pos1=adj stemmed1=n priorpolarity=positive
type=strongsubj len=1 word1=yawn pos1=noun stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=yawn pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=yeah pos1=anypos stemmed1=y priorpolarity=neutral
type=strongsubj len=1 word1=yearn pos1=verb stemmed1=y priorpolarity=positive
type=strongsubj len=1 word1=yearning pos1=noun stemmed1=n priorpolarity=positive
type=strongsubj len=1 word1=yearningly pos1=anypos stemmed1=n priorpolarity=positive
type=strongsubj len=1 word1=yelp pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=yep pos1=anypos stemmed1=y priorpolarity=positive
type=strongsubj len=1 word1=yes pos1=anypos stemmed1=y priorpolarity=positive
type=weaksubj len=1 word1=youthful pos1=adj stemmed1=n priorpolarity=positive
type=strongsubj len=1 word1=zeal pos1=noun stemmed1=n priorpolarity=positive
type=strongsubj len=1 word1=zealot pos1=noun stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=zealous pos1=adj stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=zealously pos1=anypos stemmed1=n priorpolarity=negative
type=strongsubj len=1 word1=zenith pos1=noun stemmed1=n priorpolarity=positive
type=strongsubj len=1 word1=zest pos1=noun stemmed1=n priorpolarity=positive
```

one approach: use a sentiment wordlist

- . . . eller SentiWordNet



document sentiment by summing word scores

- store all MPQA sentiment values in a table as numerical values
- e.g. 2 points for strong positive, -1 point for weak negative
- predict the overall sentiment value of the document by summing the scores of each word occurring

```
def guess_sentiment(document, score_table):
    score = 0
    for word in document:
        score = score + score_table.get(word, 0)
    if score >= 0:
        return "pos"
    else:
        return "neg"
```

översikt

testmängder

- för att utvärdera kategoriserare använder vi en **testmängd** – ett facit
- vi kan jämföra kategoriserarens automatiska gissningar med detta facit och därefter beräkna **korrekthet** (*accuracy*):

$$A = \frac{\text{antal korrekta gissningar}}{\text{antal dokument i testmängden}}$$

exempel

facit	gissning
P	P
P	N
N	N
N	N
N	P
N	N
N	N
P	N
N	N
P	P

- i denna testmängd har vi 70% korrekthet

experiment

- we evaluate on 50% of a sentiment dataset
  `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

```
def evaluate(labeled_documents, score_table):
    ncorrect = 0
    for label, document in labeled_documents:
        guess = guess_sentiment(document, score_table)
        if guess == label:
            ncorrect = ncorrect + 1
    return float(ncorrect) / len(labeled_documents)
```

- with MPQA, we get an accuracy of 59.5%

precision och täckning

- i bland har vi uppgifter av typen “hitta nålen i höstacken”
- då är korrekthet mindre intressant (varför?)
- vi använder då i stället **precision** och **täckning** (*precision*/*recall*)

$$P = \frac{\text{antal korrekta positiva}}{\text{antal positiva gissningar}}$$

$$R = \frac{\text{antal korrekta positiva}}{\text{antal positiva i testmängden}}$$

exempel

facit	gissning
P	P
P	N
N	N
N	N
N	P
N	N
N	N
P	N
N	N
P	P

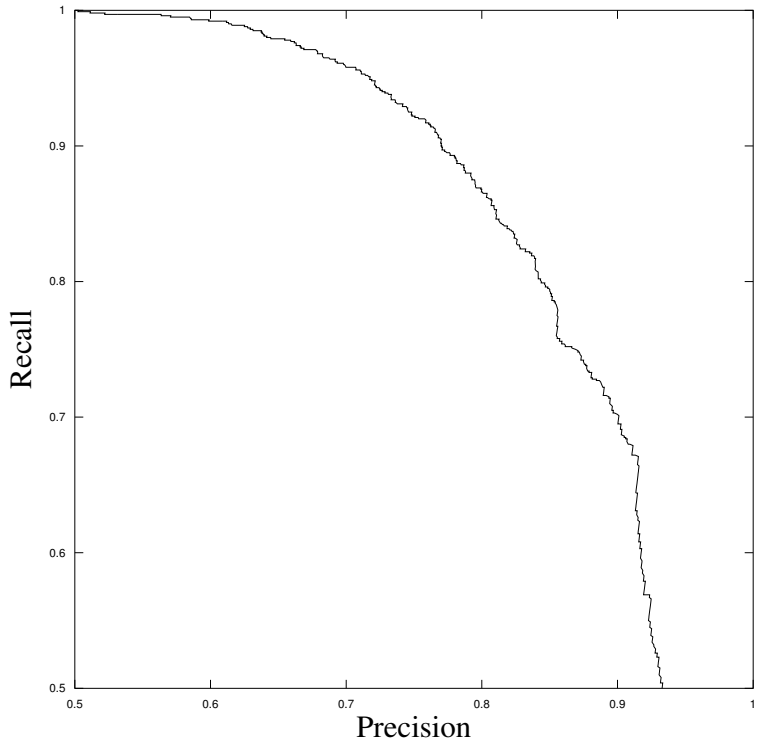
- precision = 2 / 3 = 67%
- täckning = 2 / 4 = 50%

förhållande mellan precision och täckning

- är precision eller täckning viktigast? det beror på vad vi gör:
  - döma någon för mord: hög precision
  - leta bomber på flygplatsen: hög täckning
- ofta kan man göra en avvägning mellan precision och täckning
  - t.ex. genom att ändra tröskelvärdet till något annat än 0

förhållande mellan precision och täckning: exempel

- exempel: hitta positiva recensioner



- varför är det lätt att få 100% täckning men svårare att få 100% precision?

översikt

can we do better?

- it's hard to set the word scores in the table
- what if we don't even have a resource such as MPQA or SentiWordNet?
- can we set the word scores automatically?

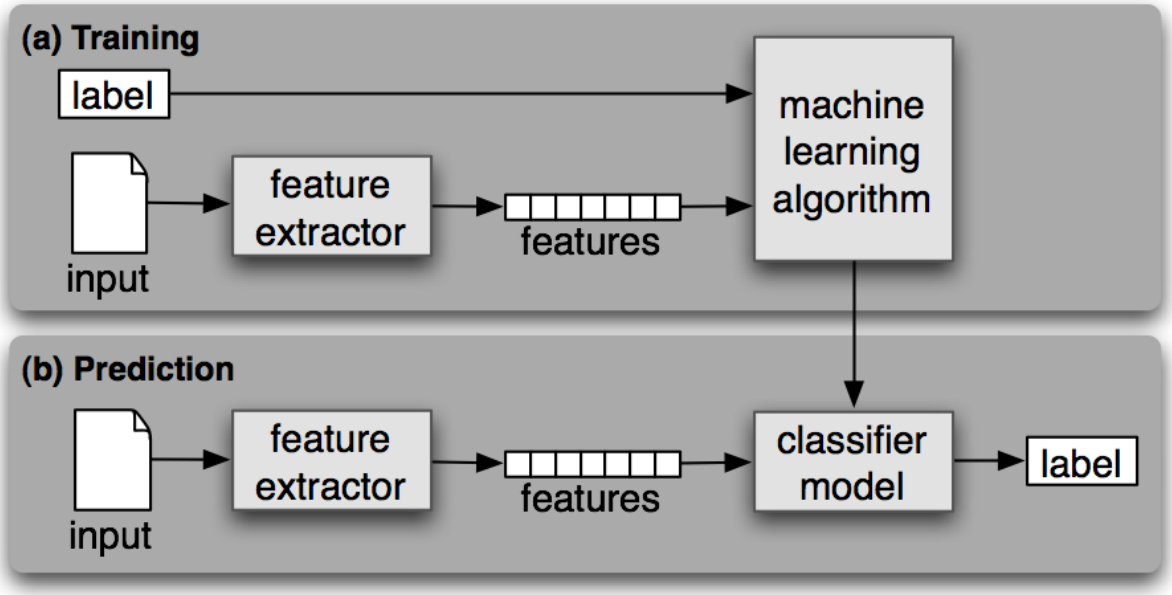
“träning” av kategoriseringssystem

- we are given a set of **examples** (e.g. reviews)
- each example comes with a positive or negative label **added manually**
- we then use these examples to “train” a categorizer
- . . . can then be used to classify reviews we haven’t seen before

features for categorization

- to be able to classify an document, we must describe its properties: **features**
- useful information that we believe helps us tell the classes apart
- this is an art more than a science
- examples:
  - in document classification, typically the **words**
  - . . . but also stylistic or structural features such as sentence length, word variation, syntactic complexity

översikt av träning



an idea for setting the scores in the table automatically

- start with an empty word score table (instead of using MPQA)
- classify documents according to the current table
- each time we misclassify, change the table a bit
  - if a positive document was misclassified, add 1 to the score of each word in the document
  - and conversely . . .

```
def train_by_errors(labeled_documents, number_iterations):
    score_table = {}
    for iteration in range(number_iterations):
        for label, document in labeled_documents:
            guess = guess_sentiment(document, score_table)
            if label == "pos" and guess == "neg":
                for word in document:
                    score_table[word] = score_table.get(word, 0) + 1
            elif label == "neg" and guess == "pos":
                for word in document:
                    score_table[word] = score_table.get(word, 0) - 1
    return score_table
```

new experiment

- we compute the score table using 50% of the sentiment data and test on the other half
- the accuracy is 81.4%, up from the 59.5% we had when we used the MPQA
- `train_by_errors` is called the **perceptron** algorithm and is one of the most widely used machine learning algorithms

examples of the word scores

amazing 171	waste -175
easy 124	worst -168
perfect 109	boring -154
highly 108	poor -134
five 107	‘ -130
excellent 104	unfortunately -122
enjoy 93	horrible -118
job 92	ok -111
question 90	disappointment -109
wonderful 90	unless -108
performance 83	called -103
those 80	example -100
r&b 80	bad -100
loves 79	save -99
best 78	bunch -98
recommended 77	talk -96
favorite 77	useless -95
included 76	author -94
medical 75	effort -94
america 74	oh -94

tillämpning Twitter-meddelanden

- min masterstudent utvecklade ett värderingsanalysverktyg för engelska Twitter-meddelanden (Günther, 2013)
- exempel för svenska (samarbete med inst. för statsvetenskap)

```
hehe: 8.10221666194292
glad: 7.352553519116041
nice: 7.295245127519536
fint: 7.050338664221716
trevligt: 6.990588423141296
haha: 6.20050756063502
roligt: 5.783331192129283
bäst: 5.553355942717432
gillar: 5.46573713292561
älskar: 5.397436760706412
0011101101100: 5.208547439882868
great: 5.1106956158441
gott: 5.072610362259834
/ @: 5.047366648930172
snyggt: 5.000351705216223
thanks: 4.955087233838285
bästa: 4.934822139250934
good: 4.933667007970653
trevlig: 4.900781046982166
grattis: 4.74212010418576
happy: 4.611786541918463
0011101101110: 4.596190625930094
hahaha: 4.526232478489328
0011101101111: 4.524749438728765
coolt: 4.480482070619616
haha ,: 4.441341640333299
: @: 4.412427150745611
absolut: 4.37189271825136
grymt: 4.357319347565992
```

```
tyvärr: -10.22318679483943
tråkigt: -9.78336129305762
trist: -9.615096375296792
sad: -8.719382038303367
sorry: -7.862172199124354
ledsen: -7.664423909954896
00111111101010: -7.516806324583317
synd: -7.480396928270459
missar: -7.429289981635725
kan inte: -7.28691959310897
krya: -6.919040650242962
sjuk: -6.863803876545991
missade: -6.844046465977916
stackars: -6.662102403285653
fail: -6.571241352077368
1011110010: -6.234619681580576
usch: -6.202110297168491
ont: -6.094625132706771
saknar: -5.943933909034524
jag som: -5.853490362531544
inte bra: -5.778972275531888
pga: -5.717375928269563
sucks: -5.674802773015557
... ): -5.575717750002824
miss: -5.548370276590176
? !: -5.540807513845234
sorgligt: -5.416686839298432
0110000111: -5.403851537862249
kram: -5.336685239489895
```

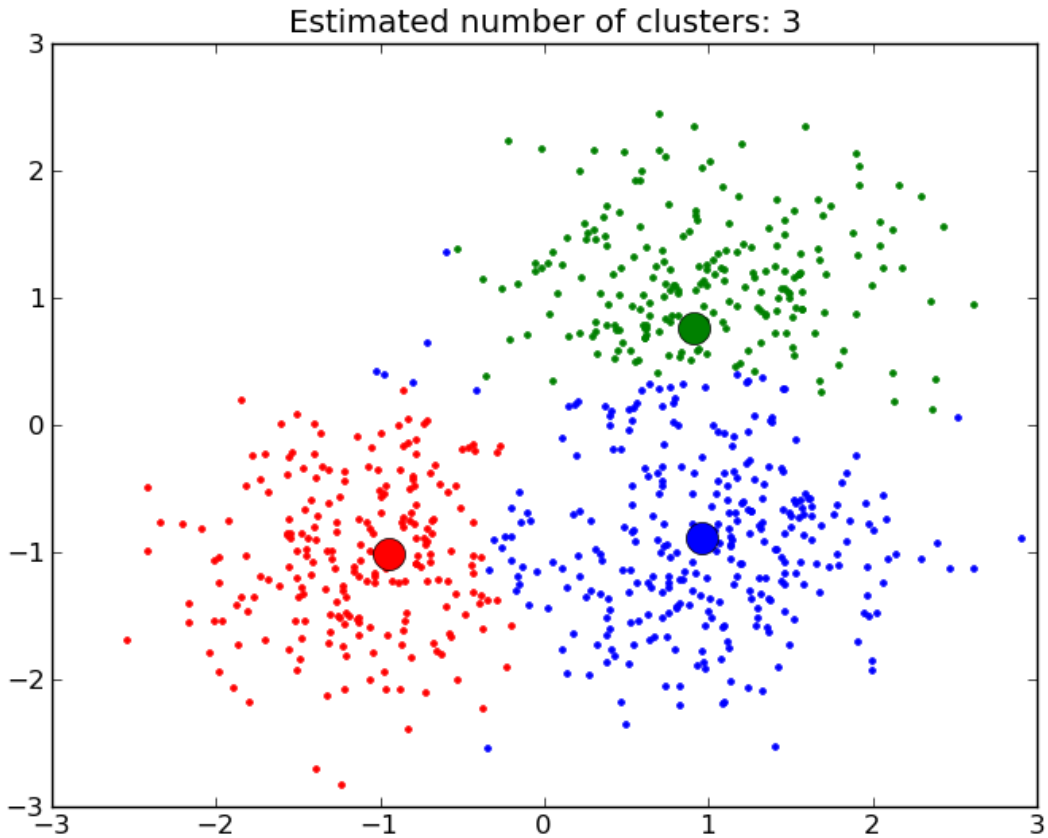
översikt

om kategorierna är okända

- vi har nu sett hur man kan kategorisera dokument enligt ett givet kategorisystem
  - positiv/negativ, svår/lätt, ämneskategori, . . .
- i bland vill vi studera en korpus av dokument utan att på förhand veta hur kategorierna är fördelade, eller ens vilka kategorier som finns
- metoder att tillämpa i sådana sammanhang
  - **klustring**: dokumentet hamnar i en avgränsad kategori
  - **topic modeling**: dokumentet ses som en blandning av kategorier

klustring

- **klustring** används för att gruppera objekt (t.ex. dokument) i olika grupper
- idén är att liknande dokument ska hamna i samma grupp

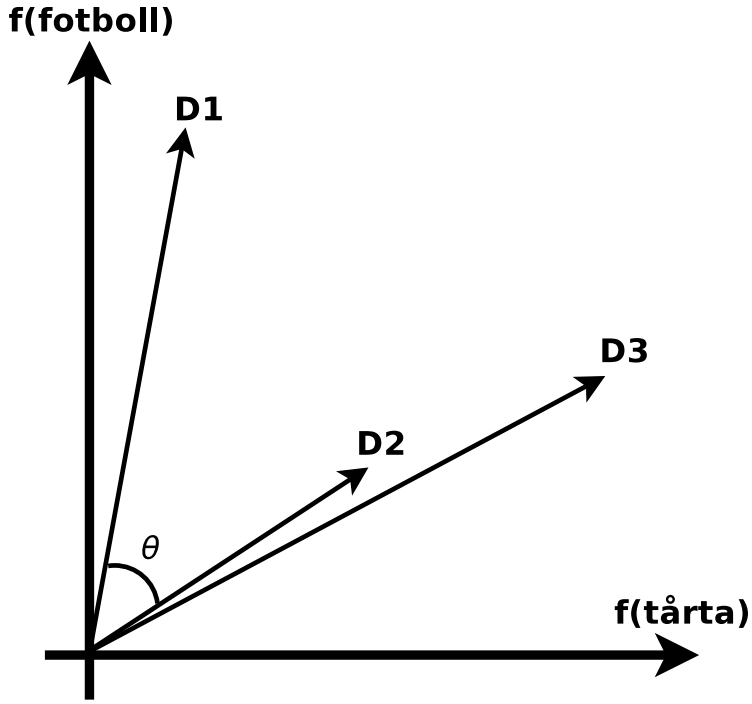


likhetsmått

- vi sa att klustringen samlar ihop dokument som är “lika”, men vad betyder det?
- klustringen behöver använda ett **likhetsmått** som säger hur pass lika två dokument är
- hur likhetsmåttet fungerar beror på vad vi är intresserade av
  - t.ex. om vi vill gruppera dokumenten efter ämne, så vill vi förmodligen avlägsna grammatiska funktionsord och liknande
  - å andra sidan, funktionsorden kan vara informativa om vi vill gruppera på något annat sätt

geometrisk jämförelse av två texter

	D1	D2	D3
fotboll	5	2	3
tårta	1	3	4



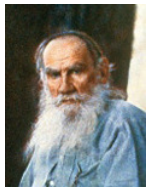
$$\text{likhet} = \frac{5 \cdot 2 + 1 \cdot 3}{\sqrt{5 \cdot 5 + 1 \cdot 1} \cdot \sqrt{2 \cdot 2 + 3 \cdot 3}}$$

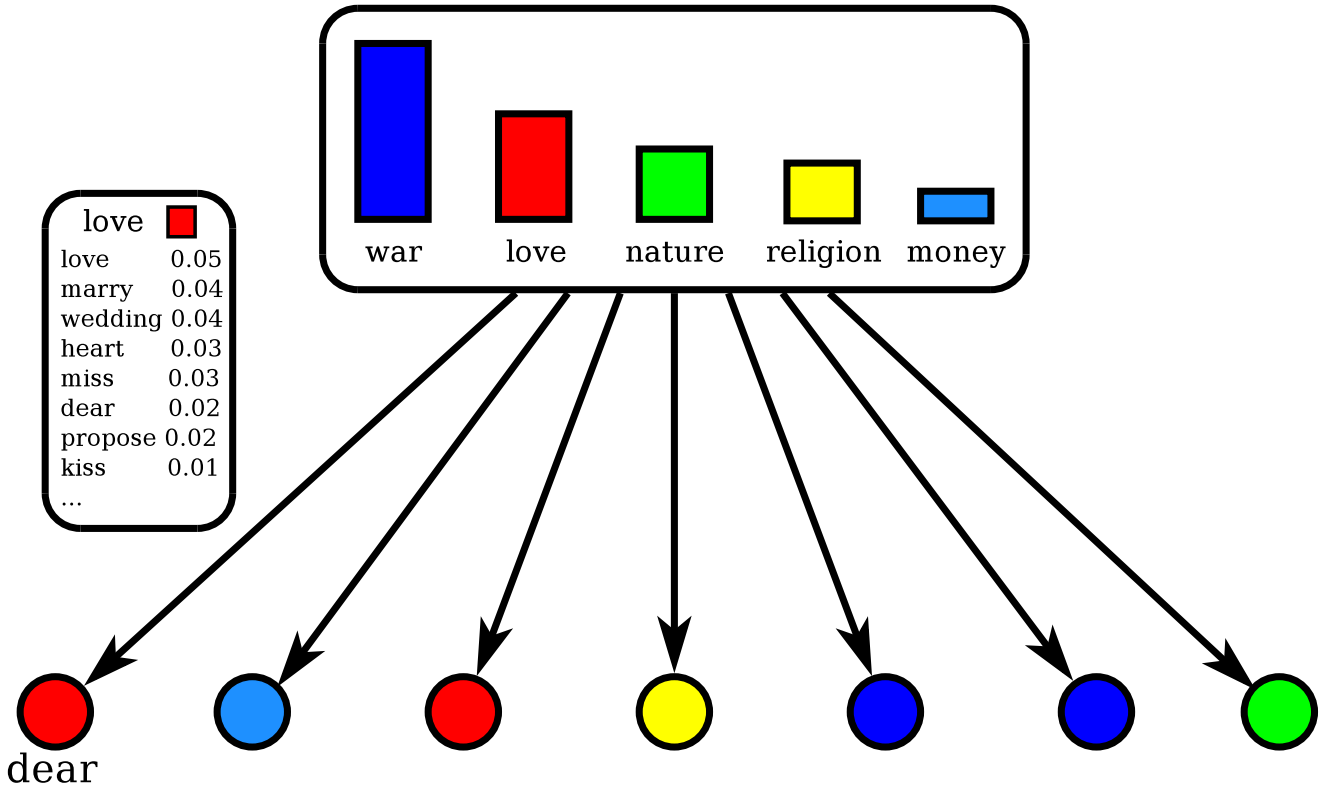
$$= 0.71$$

Topic modeling

- For documents, clustering is sometimes too simple
- A document contains more than one "topic"
  - Example: a camera review has a camera topic and a sentiment topic
- **Topic modeling** for a corpus of documents:
  1. Find the "topics" in the corpus
     - A topic is a probability distribution over words
  2. Analyze documents as composed by the topics
- The most popular topic model is called Latent Dirichlet Allocation (**LDA**)
  - `http://www.cs.princeton.edu/~blei/topicmodeling.html`
  - Mallet
  - Mahout

example of topic model analysis of *War and Peace*





ytterligare ett exempel

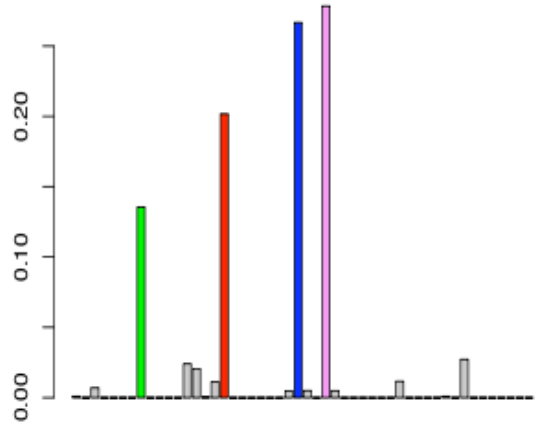
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

översikt

fortsättningen

- nästa föreläsning: morfologi och ordklassanalys
- tillbaka i humanisthuset (**C452**)

referenser

- J. Blitzer, M. Dredze, F. Pereira. *Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification*. ACL, 2007.
- G. Demartini, S. Siersdorfer, S. Chelaru, W. Nejdl. *Analyzing Political Trends in the Blogosphere*. AAAI workshop on Weblogs and Social Media, 2011.
- S. Ghosh, R. Johansson, G. Riccardi, S. Tonelli. *Shallow Discourse Parsing with Conditional Random Fields*. IJCNLP, 2011.
- S. Ghosh, S. Tonelli, R. Johansson. *Mining Fine-grained Opinion Expressions with Shallow Parsing*. RANLP, 2013.
- T. Günther. *Sentiment Analysis of Microblogs*. Examensarbete handlett av R. Johansson, MLT-programmet, GU, 2013.
- R. Johansson, A. Moschitti. *Relational Features in Fine-grained Opinion Analysis*. Computational Linguistics 39(3), 2013.
- A. Lazaridou, I. Titov, C. Sporleder. *A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations*. ACL, 2013.
- B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.
- B. Pang. *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval 2(1-2), 2008.
- B. Pang, L. Lee, S. Vaithyanathan. *Thumbs up? Sentiment Classification using Machine Learning Techniques*. EMNLP, 2002.
- R. Prasad, et al. *The Penn Discourse Treebank 2.0*. LREC, 2008.
- R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, C. Potts *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. EMNLP, 2013.
- J. Wiebe, T. Wilson, C. Cardie. *Annotating Expressions of Opinions and Emotions in Language*. Language Resources & Evaluation, 39(2-3), 2005.