Natural Language Processing
SoSe 2015

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Text Classification

Dr. Mariana Neves

June 8th, 2015

(based on the slides of Dr. Saeedeh Momtazi)

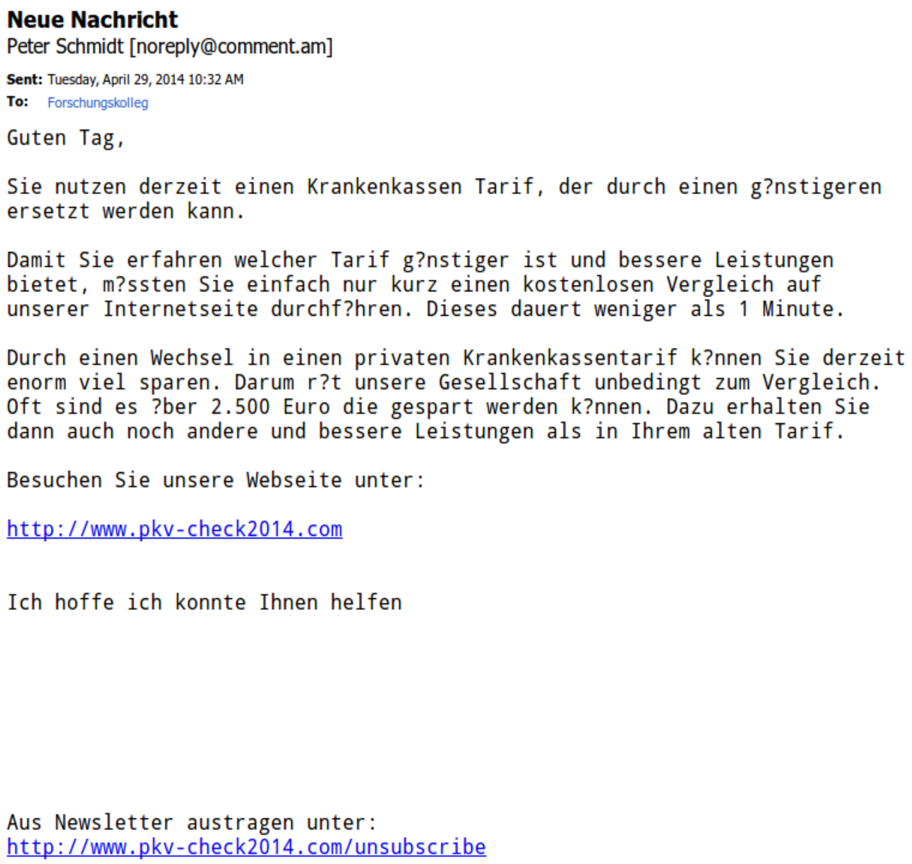
Outline

- Applications
- Task
- Naïve Bayes Classification
  - Smoothing
  - Language Modeling
- Evaluation

Outline

- Applications
- Task
- Naïve Bayes Classification
  - Smoothing
  - Language Modeling
- Evaluation

Spam Mail Detection

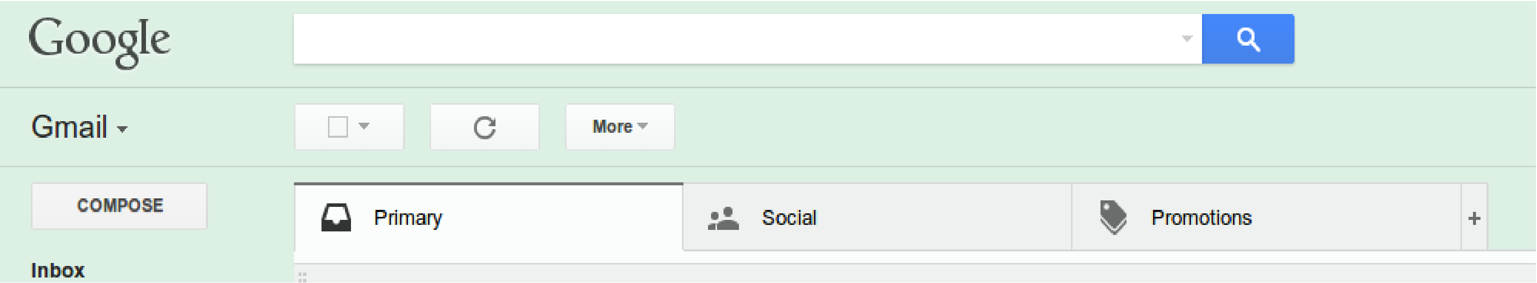


Neue Nachricht
Peter Schmidt [noreply@comment.am]

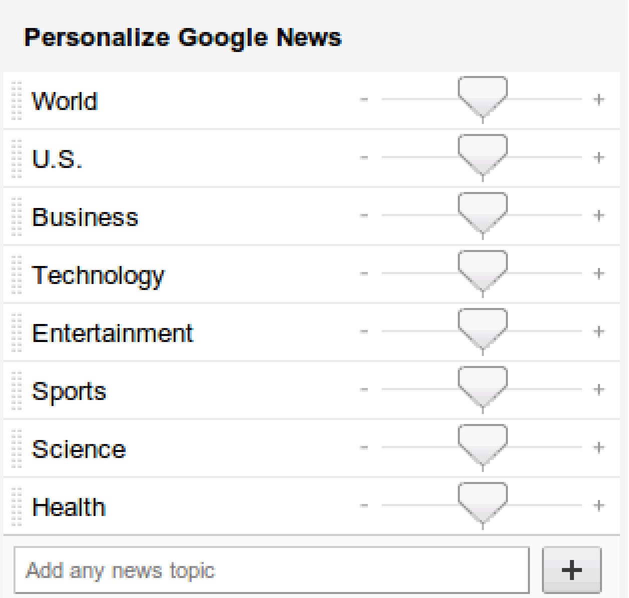
Sent: Tuesday, April 29, 2014 10:32 AM
To: Forschungskolleg

```
Guten Tag,

Sie nutzen derzeit einen Krankenkassen Tarif, der durch einen g?nstigeren
ersetzt werden kann.

Damit Sie erfahren welcher Tarif g?nstiger ist und bessere Leistungen
bietet, m?ssten Sie einfach nur kurz einen kostenlosen Vergleich auf
unserer Internetseite durchf?hren. Dieses dauert weniger als 1 Minute.

Durch einen Wechsel in einen privaten Krankenkassentarif k?nnen Sie derzeit
enorm viel sparen. Darum r?t unsere Gesellschaft unbedingt zum Vergleich.
Oft sind es ?ber 2.500 Euro die gespart werden k?nnen. Dazu erhalten Sie
dann auch noch andere und bessere Leistungen als in Ihrem alten Tarif.

Besuchen Sie unsere Webseite unter:

http://www.pkv-check2014.com


Ich hoffe ich konnte Ihnen helfen





Aus Newsletter austragen unter:
http://www.pkv-check2014.com/unsubscribe
```

Email Foldering



News Classification



Personalize Google News

World

U.S.

Business

Technology

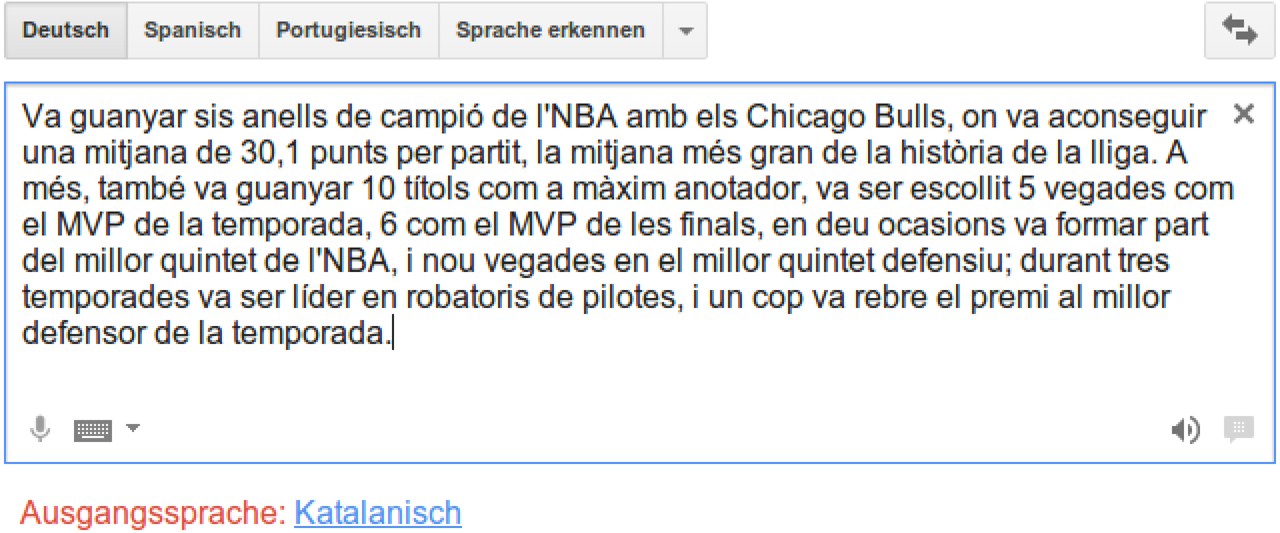
Entertainment

Sports

Science

Health

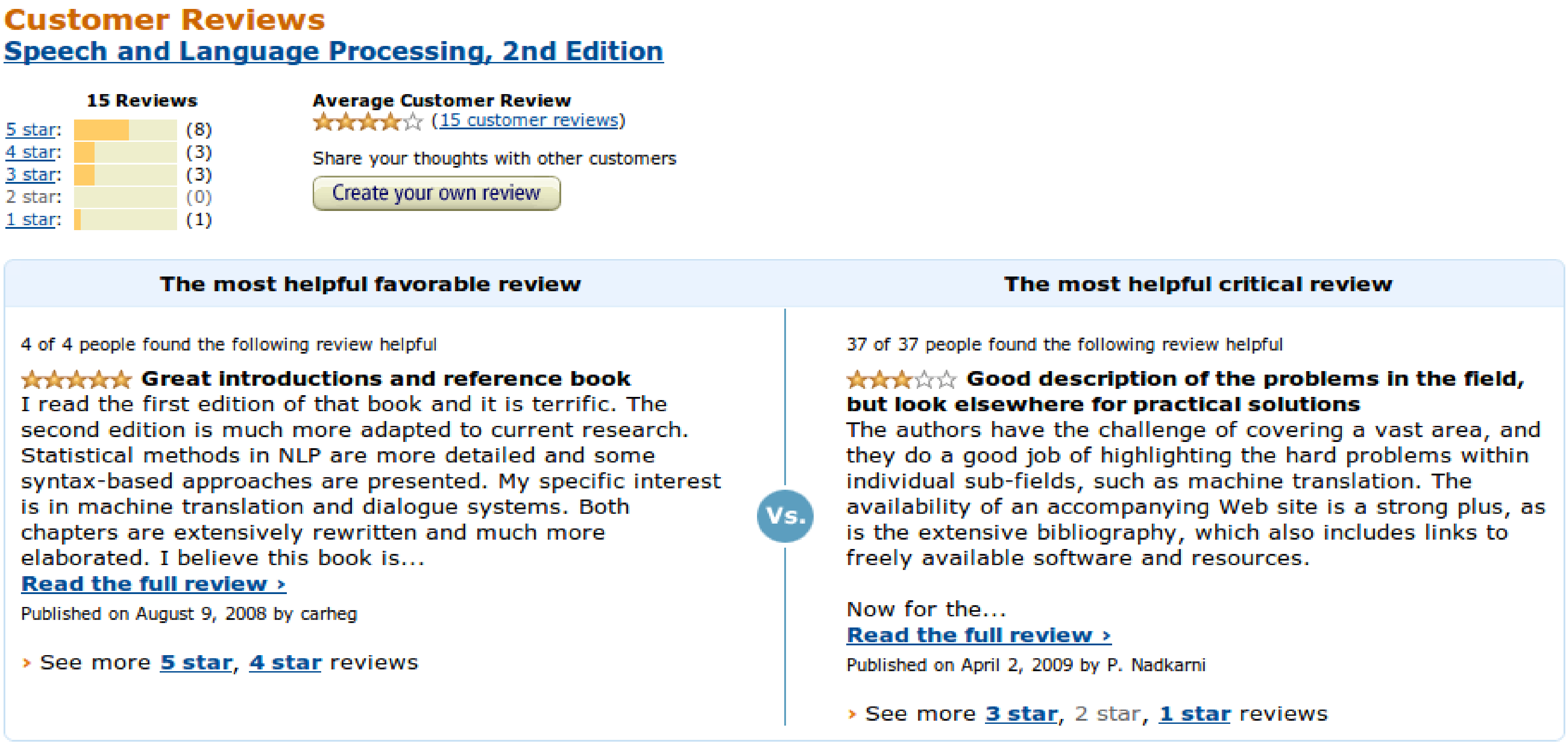
Language Identification



Va guanyar sis anells de campió de l'NBA amb els Chicago Bulls, on va aconseguir una mitjana de 30,1 punts per partit, la mitjana més gran de la història de la lliga. A més, també va guanyar 10 títols com a màxim anotador, va ser escollit 5 vegades com el MVP de la temporada, 6 com el MVP de les finals, en deu ocasions va formar part del millor quintet de l'NBA, i nou vegades en el millor quintet defensiu; durant tres temporades va ser líder en robatoris de pilotes, i un cop va rebre el premi al millor defensor de la temporada.

Ausgangssprache: Katalanisch

Sentiment Analysis



Customer Reviews
Speech and Language Processing, 2nd Edition

15 Reviews

5 star:	(8)
4 star:	(3)
3 star:	(3)
2 star:	(0)
1 star:	(1)

Average Customer Review
★★★★☆ (15 customer reviews)

Share your thoughts with other customers

Create your own review

The most helpful favorable review

4 of 4 people found the following review helpful

★★★★★ **Great introductions and reference book**
I read the first edition of that book and it is terrific. The second edition is much more adapted to current research. Statistical methods in NLP are more detailed and some syntax-based approaches are presented. My specific interest is in machine translation and dialogue systems. Both chapters are extensively rewritten and much more elaborated. I believe this book is...
Read the full review ›
Published on August 9, 2008 by carheg

› See more **5 star**, **4 star** reviews

Vs.

The most helpful critical review

37 of 37 people found the following review helpful

★★★☆☆ **Good description of the problems in the field, but look elsewhere for practical solutions**
The authors have the challenge of covering a vast area, and they do a good job of highlighting the hard problems within individual sub-fields, such as machine translation. The availability of an accompanying Web site is a strong plus, as is the extensive bibliography, which also includes links to freely available software and resources.

Now for the...
Read the full review ›
Published on April 2, 2009 by P. Nadkarni

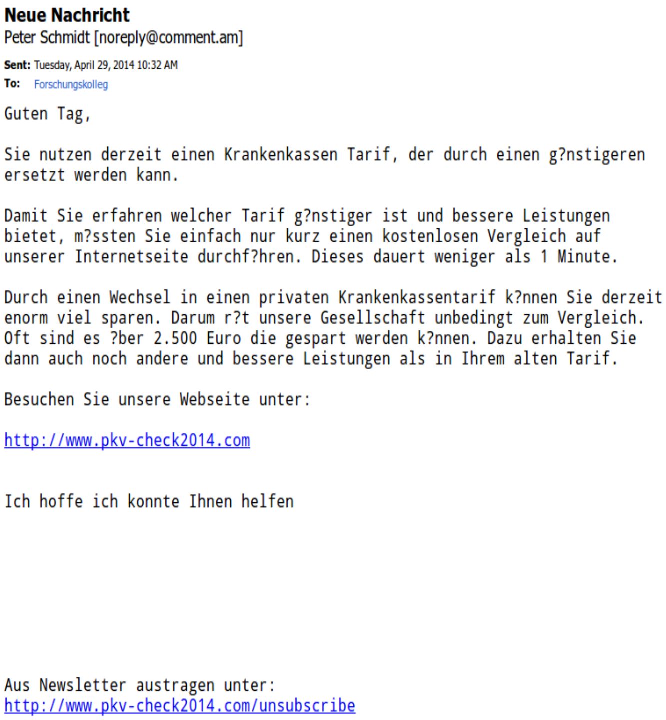
› See more **3 star**, 2 star, **1 star** reviews

Outline

- Applications
- Task
- Naïve Bayes Classification
  - Smoothing
  - Language Modeling
- Evaluation

Variations

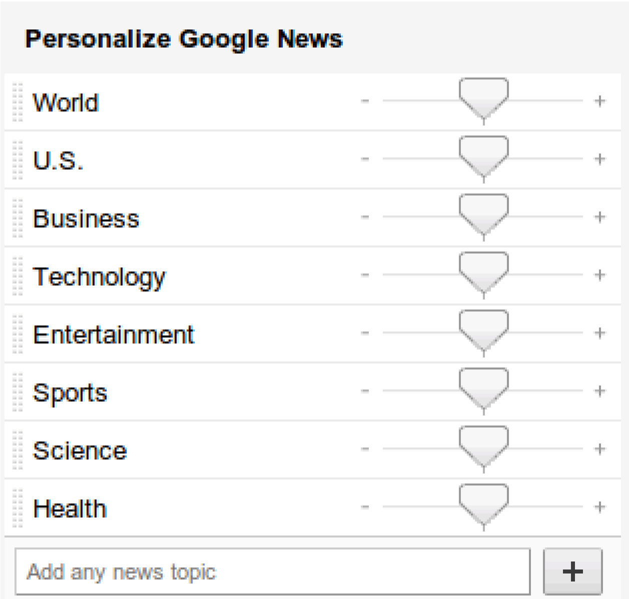
Binary vs. Multiclass



Neue Nachricht
Peter Schmidt [noreply@comment.am]

Sent: Tuesday, April 29, 2014 10:32 AM
To: Forschungskolleg

```
Guten Tag,

Sie nutzen derzeit einen Krankenkassen Tarif, der durch einen g?nstigeren
ersetzt werden kann.

Damit Sie erfahren welcher Tarif g?nstiger ist und bessere Leistungen
bietet, m?ssten Sie einfach nur kurz einen kostenlosen Vergleich auf
unserer Internetseite durchf?hren. Dieses dauert weniger als 1 Minute.

Durch einen Wechsel in einen privaten Krankenkassentarif k?nnen Sie derzeit
enorm viel sparen. Darum r?t unsere Gesellschaft unbedingt zum Vergleich.
Oft sind es ?ber 2.500 Euro die gespart werden k?nnen. Dazu erhalten Sie
dann auch noch andere und bessere Leistungen als in Ihrem alten Tarif.

Besuchen Sie unsere Webseite unter:

http://www.pkv-check2014.com


Ich hoffe ich konnte Ihnen helfen




Aus Newsletter austragen unter:
http://www.pkv-check2014.com/unsubscribe
```

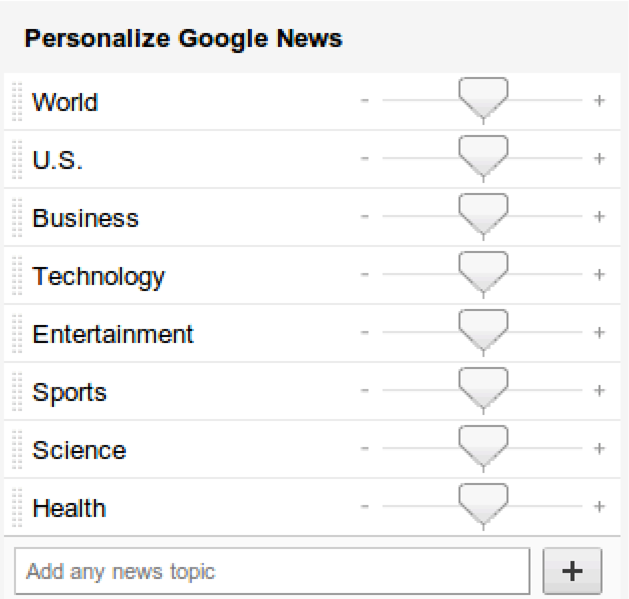


Personalize Google News

- World
- U.S.
- Business
- Technology
- Entertainment
- Sports
- Science
- Health

Variations

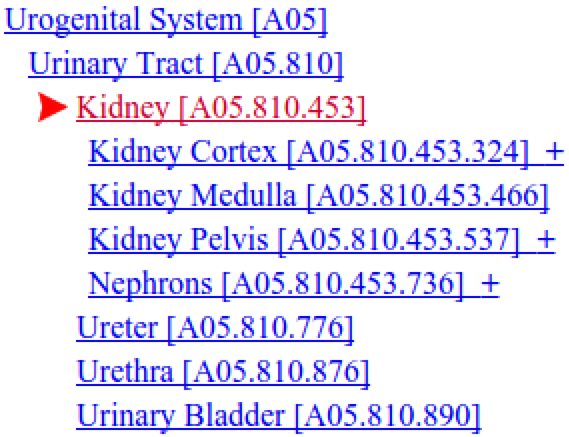
Flat vs. Hierarchical



Personalize Google News

- World
- U.S.
- Business
- Technology
- Entertainment
- Sports
- Science
- Health

Add any news topic

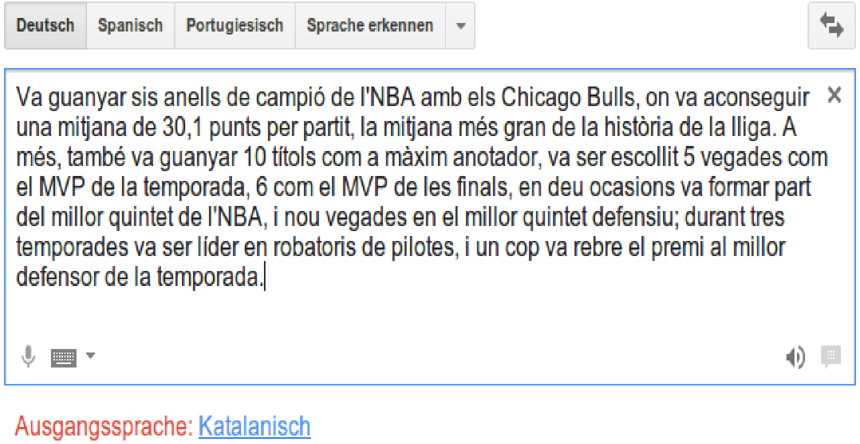


- Urogenital System [A05]
  - Urinary Tract [A05.810]
    - ➤ Kidney [A05.810.453]
      - Kidney Cortex [A05.810.453.324] +
      - Kidney Medulla [A05.810.453.466]
      - Kidney Pelvis [A05.810.453.537] +
      - Nephrons [A05.810.453.736] +
    - Ureter [A05.810.776]
    - Urethra [A05.810.876]
    - Urinary Bladder [A05.810.890]

(https://www.nlm.nih.gov/cgi/mesh/2015/MB_cgi)

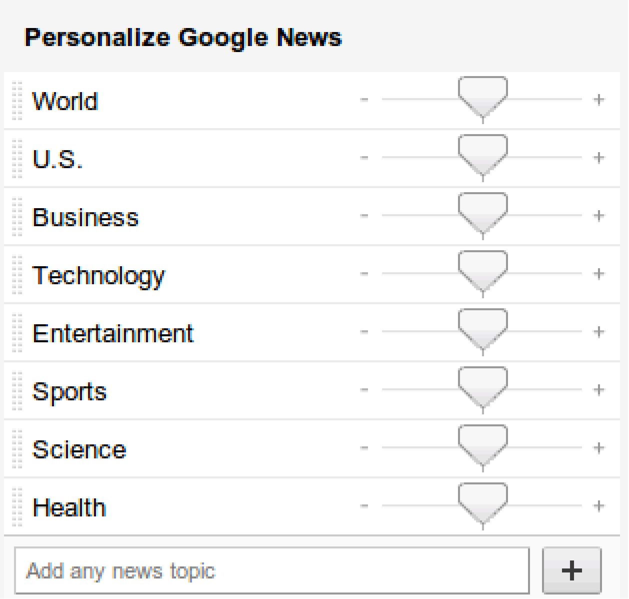
Variations

Hard vs. Soft (Multi-label)



Va guanyar sis anells de campió de l'NBA amb els Chicago Bulls, on va aconseguir una mitjana de 30,1 punts per partit, la mitjana més gran de la història de la lliga. A més, també va guanyar 10 títols com a màxim anotador, va ser escollit 5 vegades com el MVP de la temporada, 6 com el MVP de les finals, en deu ocasions va formar part del millor quintet de l'NBA, i nou vegades en el millor quintet defensiu; durant tres temporades va ser líder en robatoris de pilotes, i un cop va rebre el premi al millor defensor de la temporada.

Ausgangssprache: Katalanisch



Supervised Categorization

- Using a training set of m manually labeled documents
  - $d_1 \rightarrow c_1$
  - $d_2 \rightarrow c_2$
  - …
  - $d_m \rightarrow c_m$

Supervised Categorization

- Applying any kinds of classifiers
  - K Nearest Neighbor
  - Support Vector Machines
  - Naïve Bayes
  - Maximum Entropy
  - Logistic Regression
  - ...

Outline

- Applications
- Task
- Naïve Bayes Classification
  - Smoothing
  - Language Modeling
- Evaluation

Naïve Bayes

- Selecting the class with highest probability

  ⇒ Minimizing the number of items with wrong labels

$$\hat{c} = argmax_{c_i} P(c_i|d)$$

$$\hat{c} = argmax_{c_i} \frac{P(d|c_i) \cdot P(c_i)}{P(d)}$$

$$\hat{c} \approx argmax_{c_i} P(d|c_i) \cdot P(c_i)$$

Naïve Bayes

$$\hat{c} \approx argmax_{c_i} \underbrace{P\left(d|c_i\right)}_{\text{Likelihood probability}} \cdot \underbrace{P\left(c_i\right)}_{\text{Prior probability}}$$

Prior Probability

$$P(c_i)$$

- How much the class c_i is important disregarding the document?

$$P(c_i)=\frac{\#(c_i)}{N}$$

Likelihood Probability

$$P(d|c_i)$$

How likely the document d is selected, if we know c_i is the correct class?

⇒ How likely each of the words from document d will be selected if we know c_i is the correct class?

$$P(d|c_i) = \prod_{w \in d} P(w|c_i)$$

$$P(w|c_i) = \frac{\#(w, c_i)}{\sum_{w'} \#(w', c_i)}$$

Outline

- Applications
- Task
- Naïve Bayes Classification
  - Smoothing
  - Language Modeling
- Evaluation

Smoothing

$$P(d|c_i) = \prod_{w \in d} P(w|c_i)$$

$$P(w|c_i) = \frac{\#(w, c_i)}{\sum_{w'} \#(w', c_i)}$$

- Shortcomings
  - Words that are not available in the training data produce zero probability
  - Even one zero probability makes the whole result zero
- Solution
  - Using a smoothing method to avoid zero probability

Smoothing

$$P(d|c_i) = \prod_{w \in d} P(w|c_i)$$

$$P(w|c_i) = \frac{\#(w, c_i)}{\sum_{w'} \#(w', c_i)}$$

- Laplace (add-one) smoothing

$$P(w|c_i) = \frac{\#(w, c_i) + 1}{\sum_{w'} \#(w', c_i) + |V|}$$

Smoothing

$$P(d|c_i) = \prod_{w \in d} P(w|c_i)$$

$$P(w|c_i) = \frac{\#(w, c_i)}{\sum_{w'} \#(w', c_i)}$$

- Advanced smoothing methods
  - Bayesian smoothing with Dirichlet prior
  - Absolute discounting
  - Kneser-Ney smoothing

Outline

- Applications
- Task
- Naïve Bayes Classification
  - Smoothing
  - Language Modeling
- Evaluation

Naïve Bayes Classifier

$$P(d|c_i) = \prod_{w \in d} P(w|c_i)$$

- Using words of a document as a bag-of-word model
- Similar to the unigram model in language modeling

Naïve Bayes Classifier

$$P(d|c_i) = \prod_{w \in d} P(w|c_i)$$

- Shortcoming
  - Considering no dependencies between words
- Solution
  - Using higher order n-grams

Naïve Bayes Classifier

- Unigram

$$P(d|c_i) = \prod_{j=1}^{n} P(w_j|c_i)$$

$$P(w|c_i) = \frac{\#(w_j, c_i)}{\sum_{w'} \#(w', c_i)}$$

Naïve Bayes Classifier

- Bigram

$$P(d|c_i) = \prod_{j=1}^{n} P(w_j|w_{j-1}, c_i)$$

$$P(w_j|w_{j-1}, c_i) = \frac{\#(w_{j-1} w_j, c_i)}{\#(w_{j-1}, c_i)}$$

Naïve Bayes Classifier

- Trigram

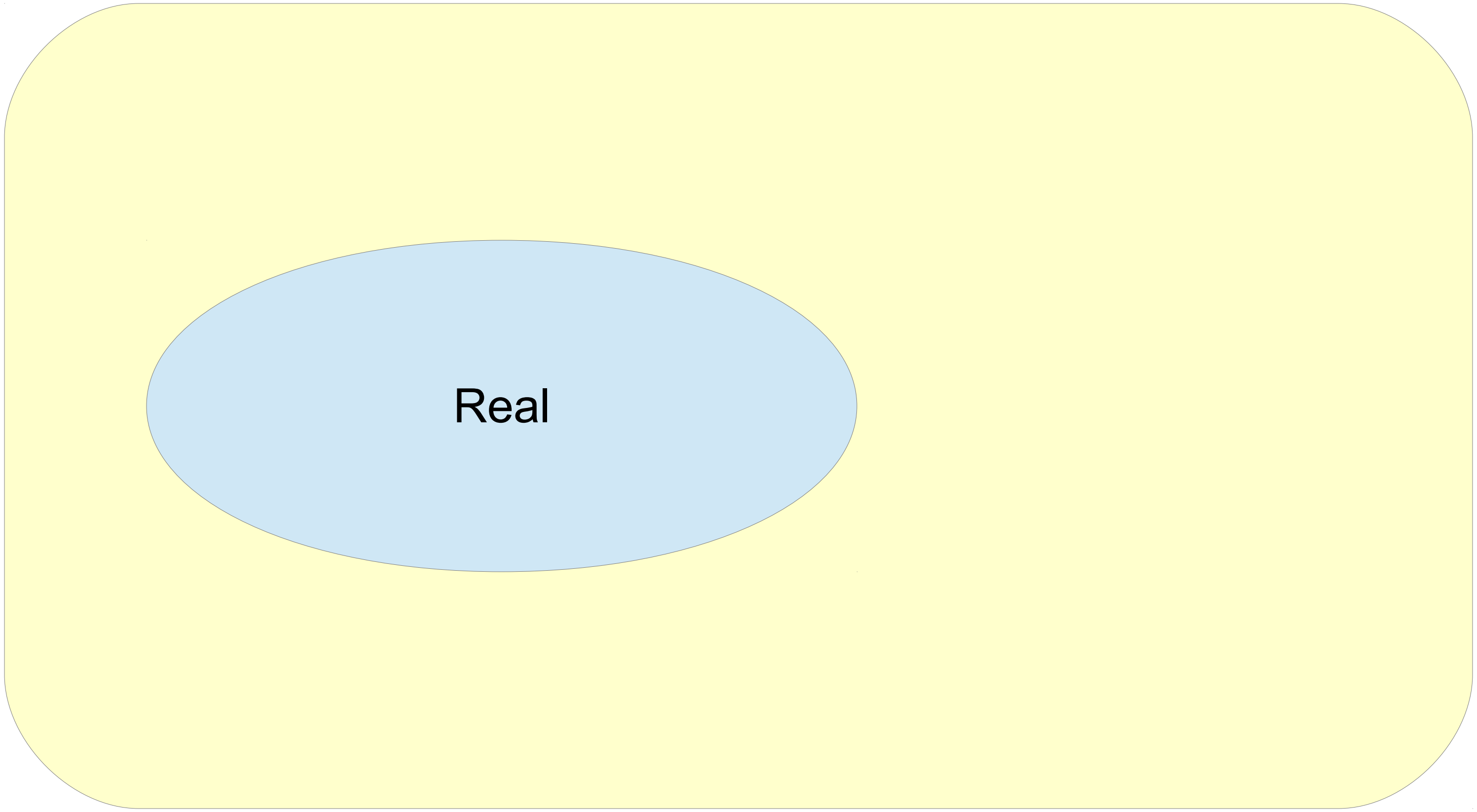
$$P(d|c_i) = \prod_{j=1}^{n} P(w_j | w_{j-2} w_{j-1}, c_i)$$

$$P(w_j | w_{j-2} w_{j-1}, c_i) = \frac{\#(w_{j-2} w_{j-1} w_j, c_i)}{\#(w_{j-2} w_{j-1}, c_i)}$$

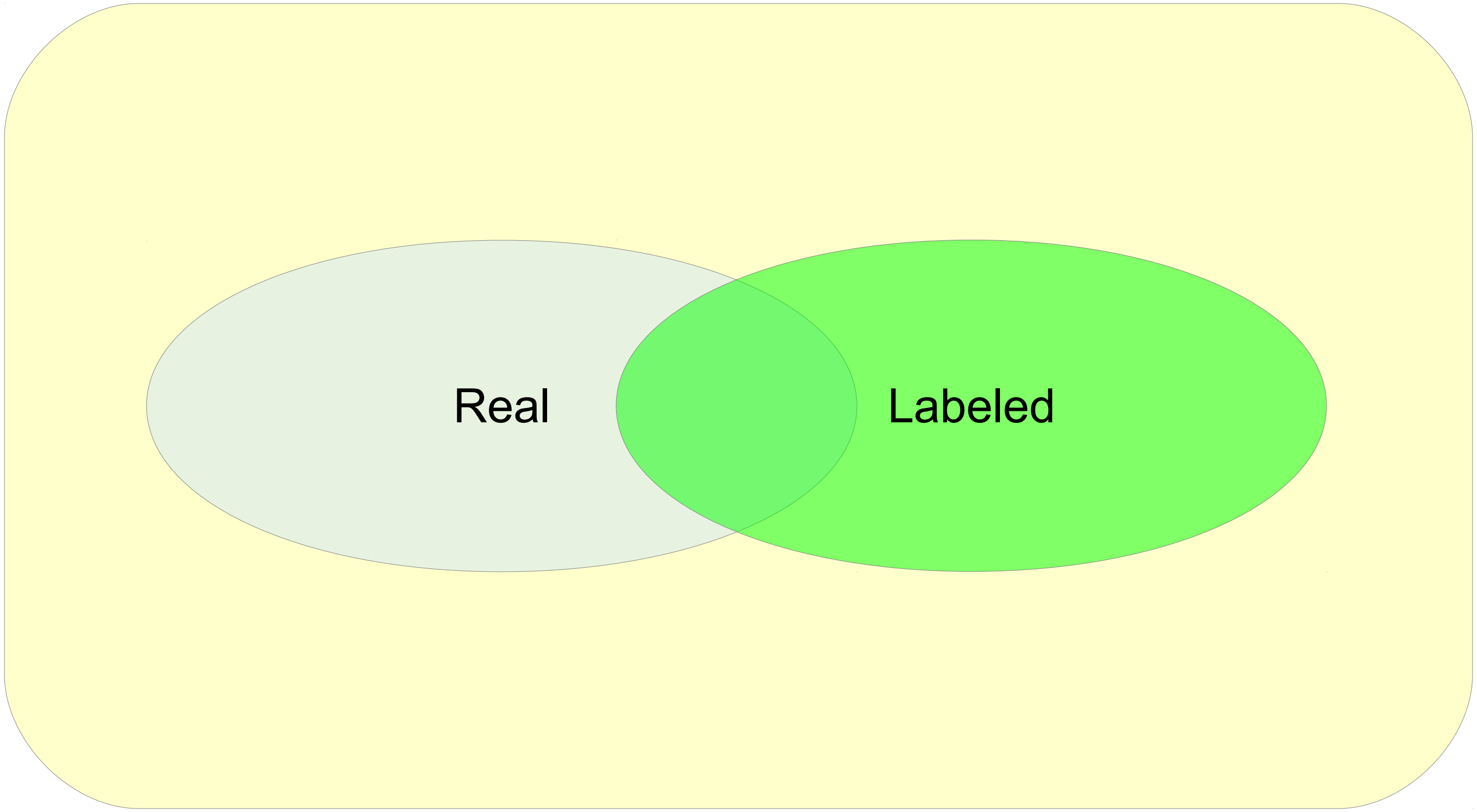
Outline

- Applications
- Task
- Naïve Bayes Classification
  - Smoothing
  - Language Modeling
- Evaluation

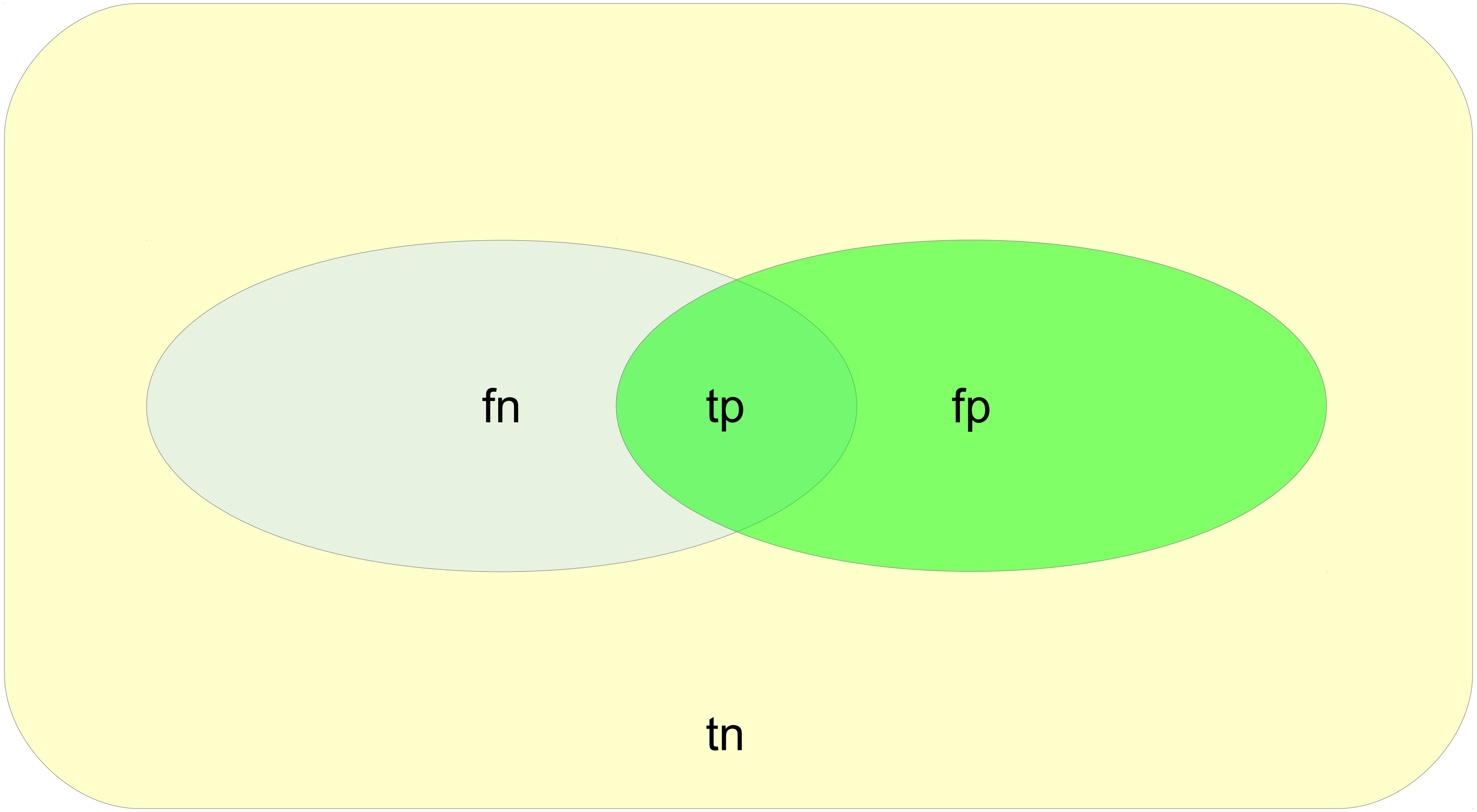
Precision and Recall



Precision and Recall



Precision and Recall



Precision and Recall

- Precision:
  - Amount of labeled items which are correct

$$Precision = \frac{tp}{tp + fp}$$

- Recall:
  - Amount of correct items which have been labeled

$$Recall = \frac{tp}{tp + fn}$$

Precision and Recall

- There is a strong anti-correlation between precision and recall
- Having a trade off between these two metrics
- Using F-measure to consider both metrics together
- F -measure is a weighted harmonic mean of precision and recall

$$F = \frac{(\beta^2 + 1) P R}{\beta^2 P + R}$$

Precision and Recall

- β < 1 gives a higher priority to precision
- β > 1 gives higher priority to recall
- β = 1 gives the same priority to both precision and recall

$$F_1 = \frac{2 P R}{P + R}$$