

Natural Language Processing
SoSe 2016



Text Classification

Dr. Mariana Neves

June 13th, 2016

Outline

- Text Classification
 - Motivation
 - Task
 - Machine Learning Approach
 - Evaluation

Outline

- Text Classification
 - Motivation
 - Task
 - Machine Learning Approach
 - Evaluation

Motivation: 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>

Motivation: News Classification

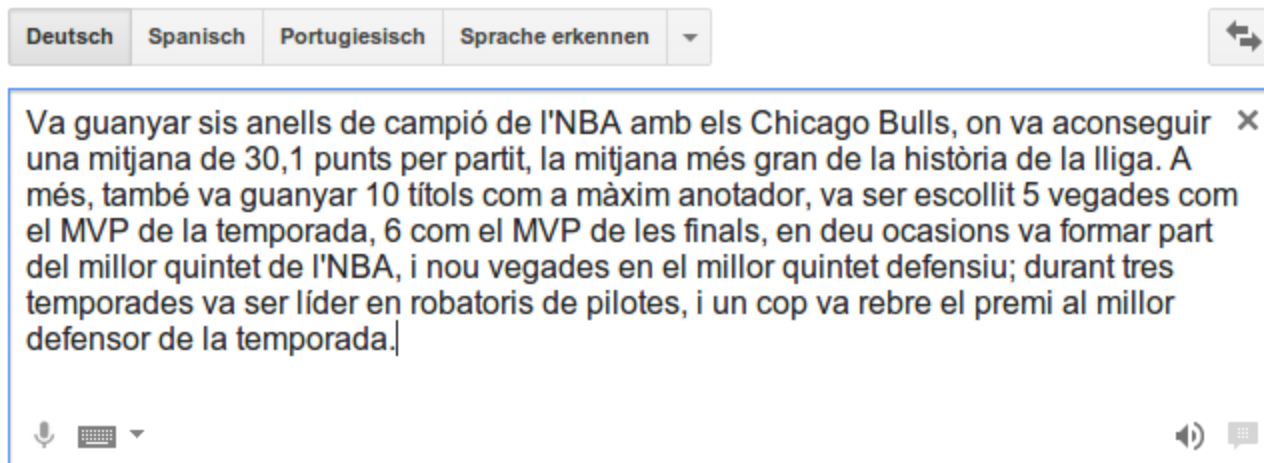
- Multi-class, multi-label

Personalize Google News

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

Add any news topic

Motivation: Language Identification



The screenshot shows a web interface for language identification. At the top, there is a navigation bar with buttons for 'Deutsch', 'Spanisch', 'Portugiesisch', and 'Sprache erkennen' (with a dropdown arrow). To the right of this bar is a refresh icon. Below the navigation bar is a large text box containing the following Catalan text: '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.' The text box has a close button (X) in the top right corner. At the bottom of the text box, there are icons for a microphone, a keyboard, a speaker, and a chat bubble.

Ausgangssprache: [Katalanisch](#)

Motivation: Sentiment Analysis

Customer Reviews Speech and Language Processing, 2nd Edition



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

- Text Classification
 - Motivation
 - **Task**
 - Machine Learning Approach
 - Evaluation

Variations for the Task

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>

The image shows a 'Personalize Google News' interface. It features a list of news topics, each with a slider control for adjusting the amount of news shown. The topics listed are World, U.S., Business, Technology, Entertainment, Sports, Science, and Health. Below the list is a text input field labeled 'Add any news topic' and a plus sign button.

Topic	Slider Control
World	Slider with minus and plus signs
U.S.	Slider with minus and plus signs
Business	Slider with minus and plus signs
Technology	Slider with minus and plus signs
Entertainment	Slider with minus and plus signs
Sports	Slider with minus and plus signs
Science	Slider with minus and plus signs
Health	Slider with minus and plus signs

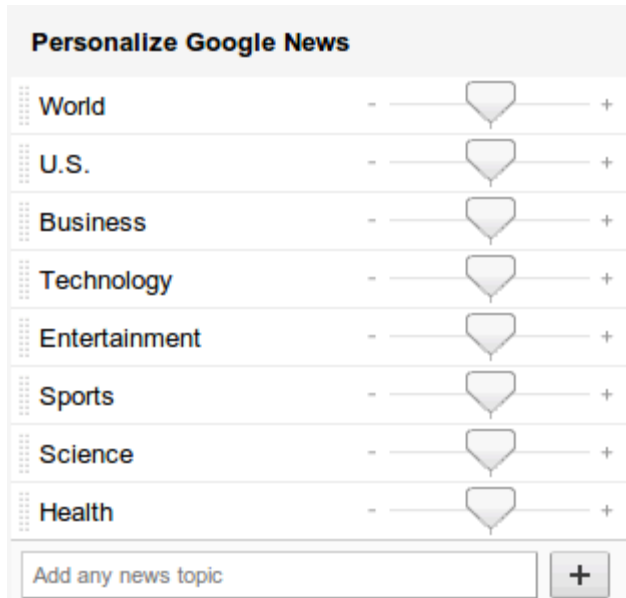
Additional features: A text input field 'Add any news topic' and a '+' button.

Variations for the Task

Flat

vs.

Hierarchical



[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 \[A05.810.890\]](#)

Variations for the Task

Hard

vs.

Soft (Multi-label)

Deutsch Spanish Portugiesisch Sprache erkennen ▾

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](#)

Personalize Google News

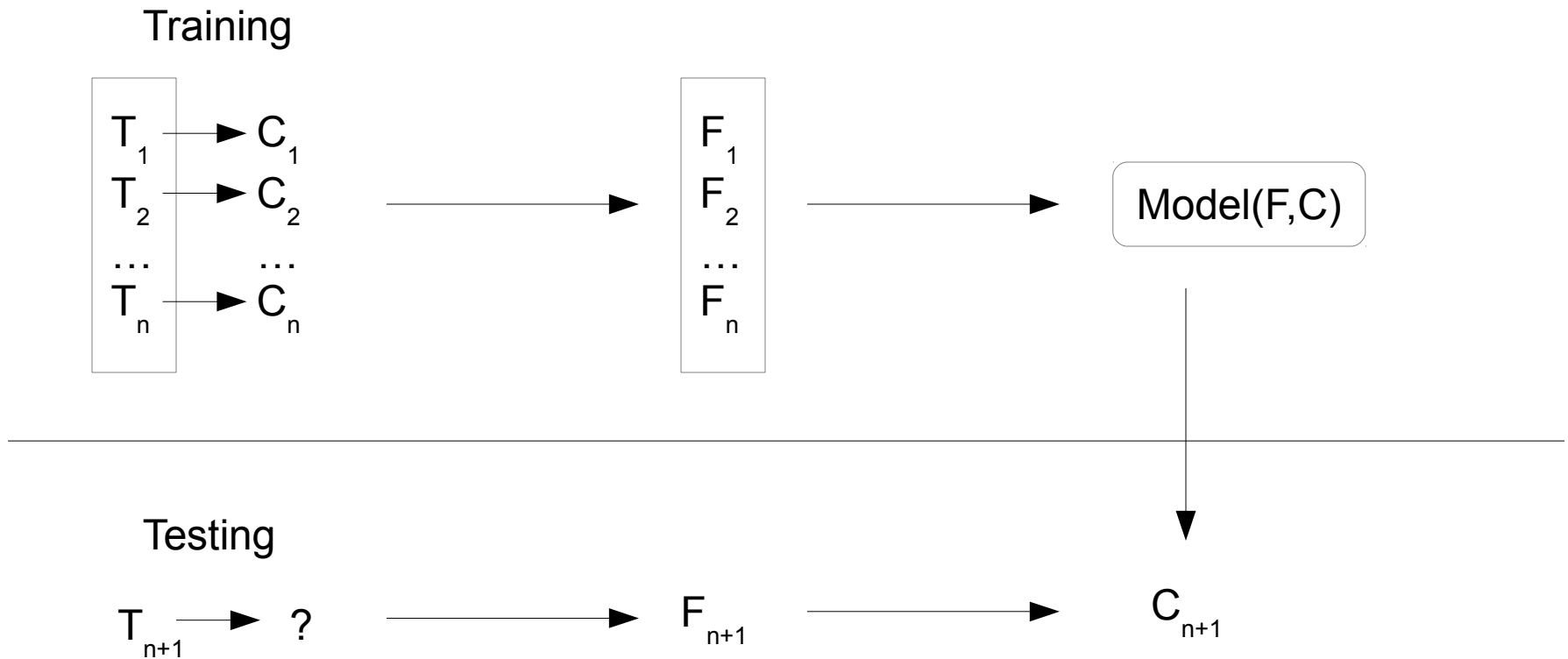
World	-	+
U.S.	-	+
Business	-	+
Technology	-	+
Entertainment	-	+
Sports	-	+
Science	-	+
Health	-	+

Add any news topic +

Outline

- Text Classification
 - Motivation
 - Task
 - Machine Learning Approach
 - Evaluation

Machine Learning Approach



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
 - ...

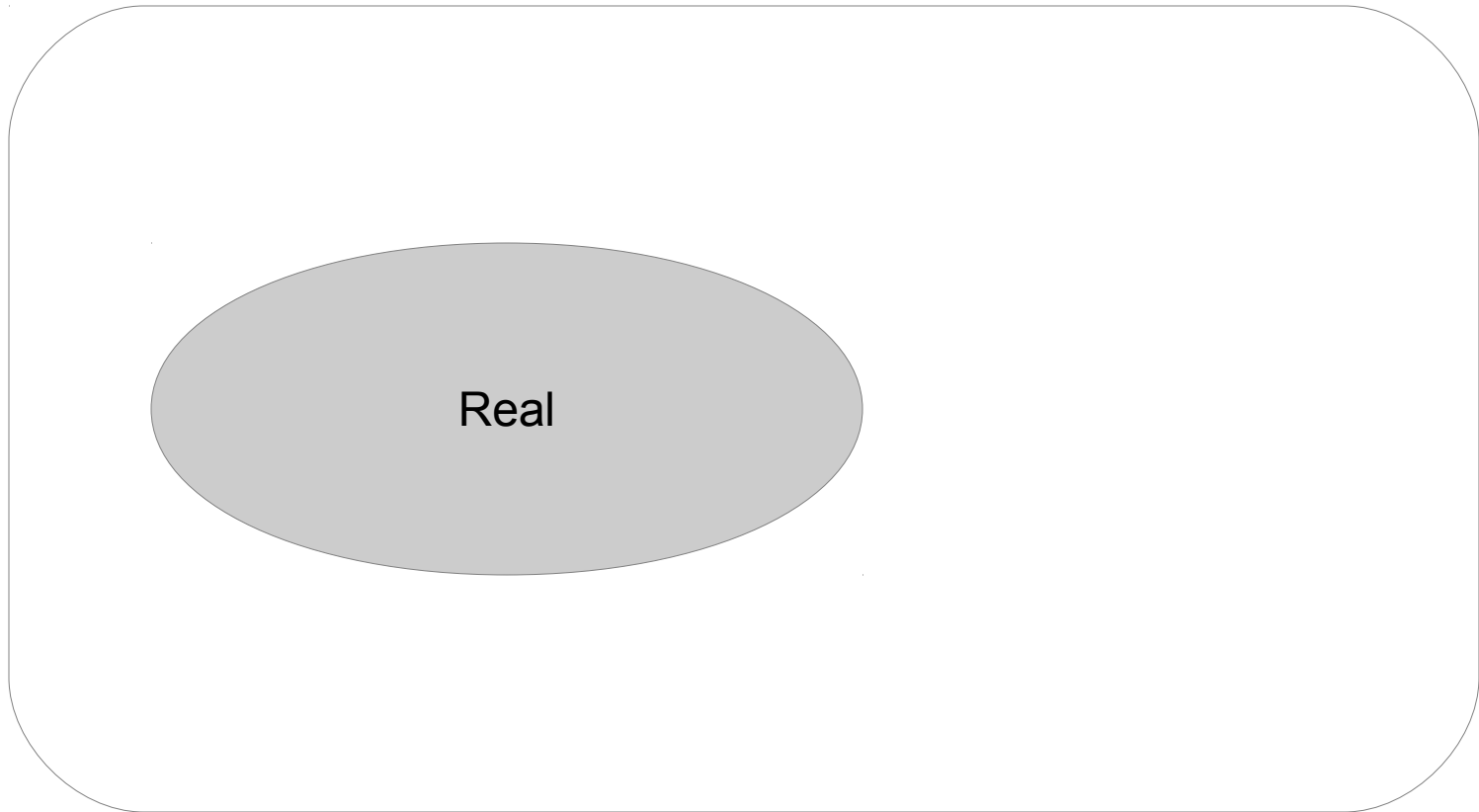
Supervised Categorization

- Features
 - Bag-of-words
 - Stopword removal
 - Stemming or lemmatization
 - TF-IDF scores
 - Named entities

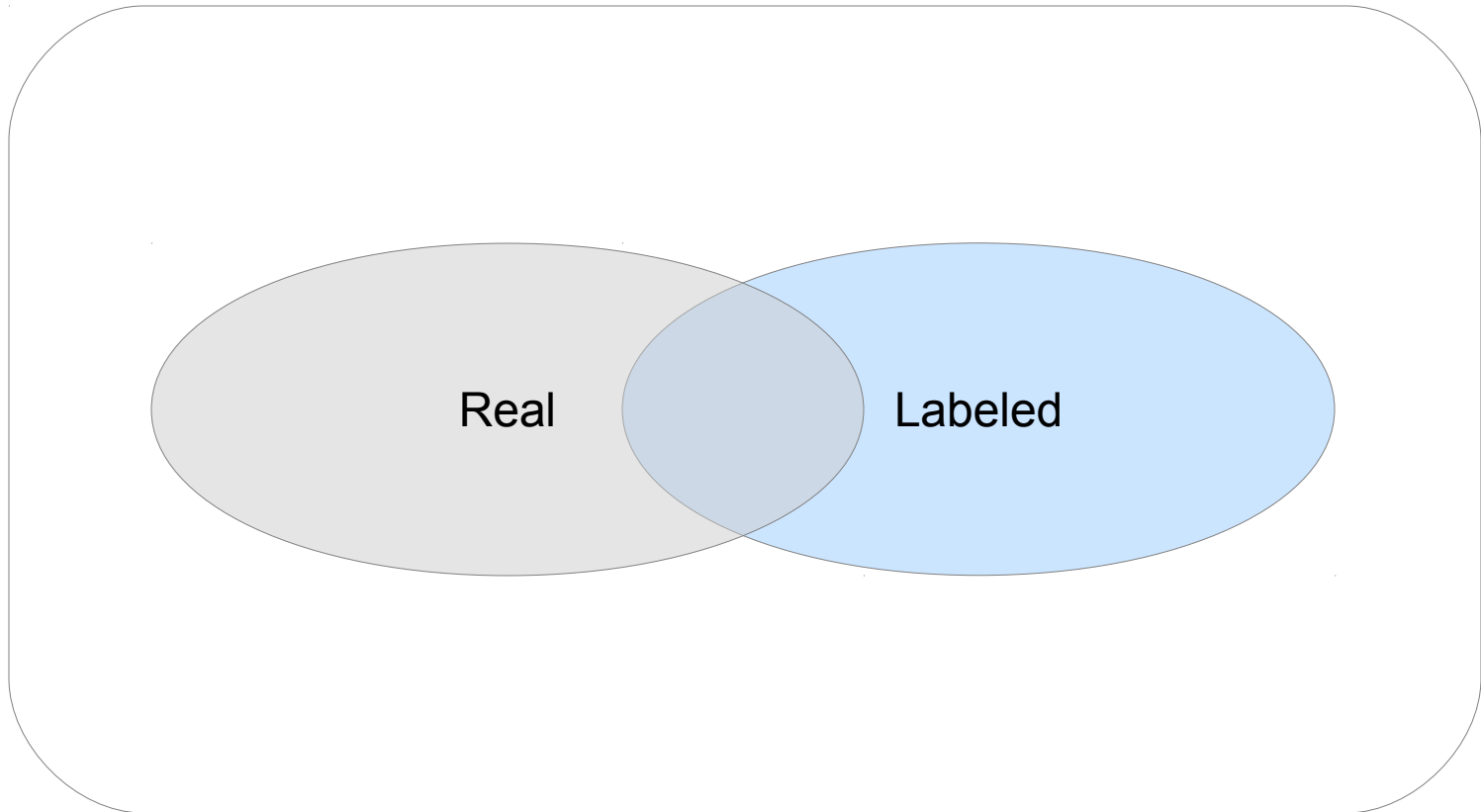
Outline

- Text Classification
 - Motivation
 - Task
 - Machine Learning Approach
 - Evaluation

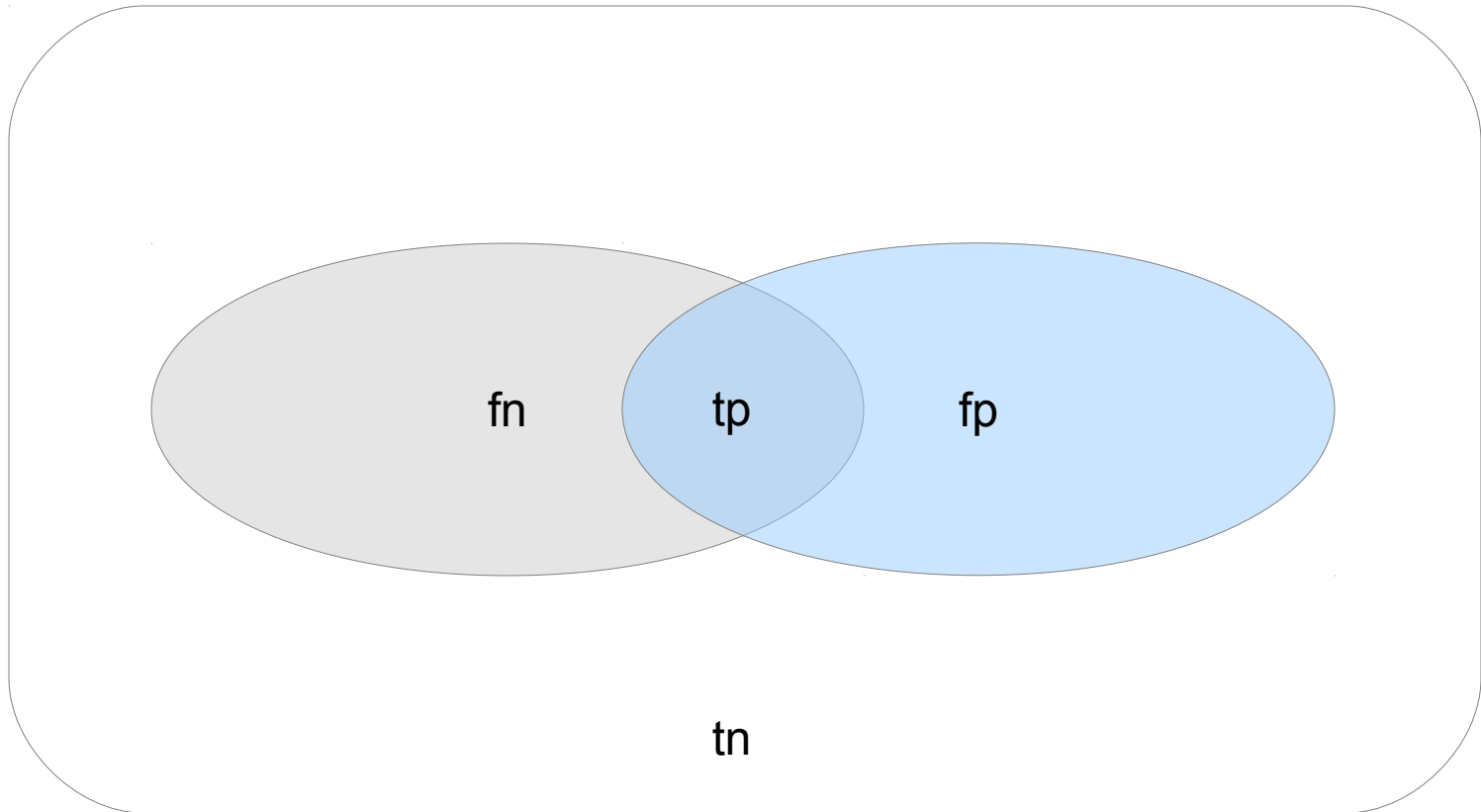
Precision and Recall



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

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

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