Assignment 4

Information Retrieval and Text Mining 17/18

2017-12-21; to be submitted 2018-01-16

Roman Klinger, Florian Strohm

- **Deadline:** This assignment will be discussed on 2018-01-18. Submissions are accepted via Ilias until end of day 2018-01-16.

Task 1 (Naïve Bayes) 8 points

Train a Naïve Bayes (the version of the model which we discussed in class) given the following documents annotated with classes c_1 and c_2. Use Add-One-Smoothing. Provide all parameters for a full model specification.

c_1 "new year"

c_1 "holiday year"

c_2 "work again"

c_2 "never work"

Given the document

- "never holiday"

Which class is assigned by the model?

Task 2 (Maximum Entropy Classification) 8 points

Given the following features (without making a difference between upper and lower case) and documents:

weight	feature
$\lambda_1 = 0.2$	$f_1(y,x) = 1$ if "$" in x and y = SPAM
$\lambda_2 = -0.1$	$f_2(y,x) = 1$ if "$" in x and y = HAM
$\lambda_3 = 0.5$	$f_3(y,x) = 1$ if "Nigerian" in x and y = SPAM
$\lambda_4 = -0.2$	$f_4(y,x) = 1$ if "Nigerian" in x and y = HAM
$\lambda_5 = -0.1$	$f_5(y,x) = 1$ if "you" in x and y = SPAM
$\lambda_6 = 0.4$	$f_6(y,x) = 1$ if "you" in x and y = HAM
$\lambda_7 = 0.1$	$f_7(y,x) = 1$ if y = SPAM
$\lambda_8 = 0.0$	$f_7(y,x) = 1$ if y = HAM

Class y	document
SPAM	x_1 = $1 million from Nigerian defense minister
SPAM	x_2 = Please contact Nigerian finance minister
SPAM	x_3 = You won $30,000!
SPAM	x_4 = Buy these Ginsu knifes now.
HAM	x_5 = You should send the Nigerian wildlife report.
HAM	x_6 = Thanks for great dinner. I owe you $20.

Subtask 2.1, 4 points

What is $p(\text{SPAM}|x_1)$ given the maximum entropy classifier with the specified features and weights?

Subtask 2.2, 4 points

Calculate the partial derivative of the log-likelihood of all documents with respect to λ_6!

Task 3 (Evalulation) 4 points

Explain in your own words what the difference between macro and micro averaging is, when calculating the F measure!

Please make an example with 10 instances and three different classes which shows a lower micro average F score than macro average F score.

It is sufficient to list ten combinations of gold and predicted classes. Explain why your solution leads to the proposed relationship between micro and macro F score.

Programming Task 4 (10 points)

The assignment data contains two files `games-train.csv` and `games-test.csv`. These are German app reviews for games (a subset of the data described in `http://www.lrec-conf.org/proceedings/lrec2016/pdf/59_Paper.pdf`).

The files are formatted as follows:

- Column 1: Title of game
- Column 2: Class of review (good or bad)
- Column 3: Title of review
- Column 4: Review text

Title and review texts can be empty.

Subtask 1, 6 points

Implement a Naive Bayes classifier from scratch which predicts the class (good, bad) stated in column 2. You can use all information from the training file to build your classifier. You are free in chosing the meta-parameters (smoothing, stop-word deletion, stemming, preprocessing). Which terms have the highest probability to occur in a good review?

Subtask 2, 4 points

Implement an evaluation system to Subtask 1. What is your precision, recall, and F to predict the class good and what is your precision, recall, and F to predict the class bad? Also report the numbers of TP, FP, FN. Discuss your results.