Exercise 1

Assume we are classifying sentences into two classes, sentiment bearing and non-sentiment bearing and get the following confusion matrix.

		Gold	
		sentiment	non-sentiment
Predicted	sentiment	50	50
	Non-sentiment	150	750

a) What is the accuracy of this classifier?

$$acc = \frac{tp}{tp + fp + fn + tp} = \frac{50 + 750}{50 + 50 + 150 + 750} = \frac{800}{1000} = 0.8$$

b) What is the recall, precision and f-score for the sentiment bearing class?

$$p = \frac{tp}{tp + fp} = \frac{50}{50 + 50} = 0.5$$

$$r = \frac{tp}{tp + fn} = \frac{50}{50 + 150} = 0.25$$

$$f1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 0.5 \times 0.25}{0.5 + 0.25} \approx 0.333$$

c) What is the recall, precision and f-score for the non-sentiment bearing class?

$$p = \frac{tn}{tn + fn} = \frac{750}{750 + 150} \approx 0.833$$

$$r = \frac{tn}{tn + fp} = \frac{750}{750 + 50} = 0.9375$$

$$f1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 0.833 \times 0.9375}{0.833 + 0.9375} \approx 0.8824$$

Exercise 3

When applying P, R and F, we should keep in mind that there are two different kinds of scenario. They differ in particular when it comes to micro averaging.

- In the general case, as we see e.g., in information retrieval and span-evaluation for chunking or NER, the total number of predicted items and the total number of gold items may differ.
- In the special case of classification, where a classifier puts each item into exactly one class, there are equally many predicted items as gold items. This was the case in exercise 1 and 2 above.

To see the difference, consider the following. We are evaluating a NER, which considers only three classes, and we count the following numbers.

	True pos.	False pos.	False neg.
Person	720	180	80
Organization	180	20	60
Location	60	0	20

a) Calculate P, R and F for each class

Person:

$$p = \frac{tp}{tp + fp} = \frac{720}{720 + 180} = 0.8$$

$$r = \frac{tp}{tp + fn} = \frac{720}{720 + 80} = 0.9$$

$$f1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 0.8 \times 0.9}{0.8 + 0.9} \approx 0.847$$

Organization:

$$p = \frac{tp}{tp + fp} = \frac{180}{180 + 20} = 0.9$$

$$r = \frac{tp}{tp + fn} = \frac{180}{180 + 60} = 0.75$$

$$f1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 0.9 \times 0.75}{0.9 + 0.75} \approx 0.818$$

Location:

$$p = \frac{tp}{tp + fp} = \frac{60}{60 + 0} = 1.0$$

$$r = \frac{tp}{tp + fn} = \frac{60}{60 + 20} = 0.75$$

$$f1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 1.0 \times 0.75}{1.0 + 0.75} \approx 0.857$$

b) Calculate macro- and micro-averaged P, R and F

Micro-average:

1) Calculate tp, fp, fn
   tp: 720 + 180 + 60 = 960
   fp: 180 + 20 + 0 = 200
   fn: 80 + 60 + 20 = 160
2) Calculate p, r, and f1

$$p = \frac{tp}{tp + fp} = \frac{960}{960 + 200} \approx 0.828$$

$$r = \frac{tp}{tp + fn} = \frac{960}{960 + 160} \approx 0.857$$

$$f1 = \frac{2 \times p \times r}{p + r} = \frac{2 \times 0.828 \times 0.857}{0.828 + 0.857} \approx 0.842$$

Macro-average:

Average over the individual p, r, f1 of the three classes, which have been calculated.

$$p = \frac{p_{per} + p_{org} + p_{loc}}{3} = \frac{0.8 + 0.9 + 1.0}{3} = 0.9$$

$$r = \frac{r_{per} + r_{org} + r_{loc}}{3} = \frac{0.9 + 0.75 + 0.75}{3} = 0.8$$

$$f1 = \frac{f1_{per} + f1_{org} + f1_{loc}}{3} = \frac{0.847 + 0.818 + 0.857}{3} = 0.841$$

Part 2: Language models

Corpus 1

This film is funny.
I enjoyed the book.
The film was entertaining.
The book is good.
The game is not bad.
It is not boring.
This is a good book.

We are training an unsmoothed bigram language model (LM) on this corpus. We assume the strings are tokenized by splitting on white space and making punctuation a separate token.

Exercise 1

Which probability will the language model ascribe to the following sequence? Explain how it is calculated.

a) *The film is good.*

1) Pad all sentences with "<s>" and "</s>"
   <s>This film is funny.</s>
   <s>I enjoyed the book.</s>
   <s>The film was entertaining. </s>
   <s>The book is good. </s>
   <s>The game is not bad. </s>
   <s>It is not boring. </s>
   <s>This is a good book.</s>
2) Lower-case the texts, obtain the unique bigrams and count the occurrences

Bigram	Occurrences	Bigram	Occurrences
<s> this	1	entertaining .	1
this film	1	book is	1
film is	1	is good	1
is funny	1	good .	1
funny .	1	the game	1
. </s>	7	game is	1
<s> i	1	is not	2
I enjoyed	1	not bad	1
enjoyed the	1	bad .	1
the book	2	<s> it	1
book .	2	it is	1
<s> the	3	not boring	1
the film	1	boring .	1
film was	1	this is	1
was entertaining	1	is a	1
		a good	1
		good book	1

3) Calculate the probability of “<s> The film is good. </s>”

$$P = P(the| < s >) \times P(film|the) \times P(is|film) \times P(good|is) \times P(.|good) \times P(</s > |.) = \frac{3}{7} \times \frac{1}{4} \times \frac{1}{2} \times \frac{1}{5} \times \frac{1}{2} \times \frac{7}{7} \approx 0.00536$$

Tip: $P(film|the) = \frac{\#(the,film)}{\#(the,film)+\#(the,book)+\#(the\ game)} = \frac{1}{1+2+1} = \frac{1}{4}$

Exercise 2

Which problems does this model face if in ascribing a probability to the following sequence?

b) The film is not good.
The whole product will be zero, because “not good” never occurs in the corpus.

Exercise 3

Modify the model by applying add-one-smoothing and compute the adjusted probabilities for sentence (1) and (2).

Hint: Despite the bigrams from the corpus itself, we take the unique tokens to generate the combinations of all possible bigrams and then expand the bigram occurrence table. The rest is the same. Therefore, for exercise 2, “not good” will have the occurrence of 1, instead of zero.