# Tasks

## What is the difference between multi-class and multi-label text classification?

Class and label are similar and given a text collection for multi-class or multi-label classification tasks, all texts have at least 3 classes or labels.

The difference is that for the multi-class text classification tasks, each text belongs to one class only, while for the multi-label text classification, each text has 0 or more labels.

## Why is named entity recognition typically modelled as a sequence labelling task?

# Challenges of text data

## What is a long tail distribution?

Given a text collection, there is a large number of words in the text occurring infrequently, while there is a small number of words in the given text occurring frequently.

## (3 points) What is sparseness in the context of the vector space model?

Sparseness means that there are many zero elements and a small number of non-zero elements in the feature vectors.

## (4 points) Why does the bag-of-words model lead to a sparse feature space?

The first step for the bag-of-words model is to create a set of vocabulary that contains all unique words from a text collection given. Therefore, when a text is represented as a feature, there are many words or terms from the vocabulary not occurring in this text such that almost each text feature contains many 0.

# 3. Text pre-processing

## 3.1 (4 points) Does lemmatization increase or decrease the dimensionality of feature space for text classification? Explain why.

1. decrease

2.1 Lemmatization: think for ‘thinking’, ‘thought’ or ‘thinks’.

2.2 Stemming: comput for ‘computer’, ‘computing’ or ‘computation’.

## 3.2. (3 points) Is stemming a language-dependent task? Motivate your answer.

Yes. Stemming is to extract stems from words or terms and uses dictionary, therefore, it depends on language.

## 3.3 (5 points) Compute the Levenshtein distance between ‘alone’ and ‘galore’. Show your computation. (Showing the full matrix is allowed but not required; the table with edits suffices.)

1.

|  |  |  |  |
| --- | --- | --- | --- |
| input | output | Operation | Cost |
| \* | g | insertion | 1 |
| a | a | copy | 0 |
| l | L | copy | 0 |
| o | o | copy | 0 |
| n | r | substitution | 1 |
| e | e | copy | 0 |
|  |  |  | Total 2 |

2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| / | / | g | a | l | o | r | e |
| / | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| a | 1 | 1 | 1 | 2 | 3 | 4 | 5 |
| l | 2 | 2 | 2 | 1 | 2 | 3 | 4 |
| o | 3 | 3 | 3 | 2 | 1 | 2 | 3 |
| n | 4 | 4 | 4 | 3 | 2 | 2 | 3 |
| e | 5 | 5 | 5 | 4 | 3 | 3 | 2 |

# 4

## 4.1 (7 points) Compute Cohen’s Kappa for this agreement table. Show your computation. (You can keep the last fraction of your computation as it is, without estimating the decimal numbers.)

Pr(a) = (25+25+10) / (25+10+5+0+25+15+5+5+10) = 60/100 = 0.6

Pr(e, positive) = 30/100 \* 40/100 = 0.3\*0.4 = 0.12

Pr(e, negative) = 40/100 \* 40/100 = 0.4\*0.4 = 0.16

Pr(e, neutral) = 20/100 \* 30/100 = 0.2\*0.3 = 0.06

Pr(e) = 0.12+0.16+0.06 = 0.34

K = (Pr(a)-Pr(e))/(1-Pr(e)) = (0.6-0.34) / (1-0.34) = 0.26/0.66 = 13/33 (0.393939)

## 4.2 (3 points) What is the meaning of Kappa = 0?

Pr(a) = Pr(e), i.e., actual (measured) agreement equal to expected (chance) agreement.

# 5. Text classification

## 5.1 (3 points) What is the prior probability of the ‘relevant’ class?

P(relevant) = 1/4

## 5.2. (3 points) What is the vocabulary size of the training set? Assume that we do not remove stop words.

12

## 5.3 (4 points) Estimate P(‘make’, not relevant) using the maximum likelihood estimate on the train set.

(2+1) / (15+12) = 3/27 = 1/9

## 5.4 (4 points) Why is add-one smoothing needed when we estimate the probability of an unseen document? Provide an example test document given the toy training set for which add-one smoothing is needed.

1. In some cases that there exist a few terms in a text to be predicted, not occurring in the training set, the outcome of multiplication of probabilities for all terms will be zero if add-one smoothing is not applied.

2. E.g.,

|  |  |  |
| --- | --- | --- |
| Doc id | Content | Class |
| 5 (example) | We make the worst plans | Not relevant |

According to the example shown above, P (worst | not relevant) = 0/15, therefore

P (not relevant | We, make, the, worst, plan) = 0 if add-one is not introduced.

If add-one smoothing is introduced, then:

P (not relevant | We, make, the, worst, plan) = P (not relevant) \* P (We, make, the, worst, plan | not relevant) = 3/4 \* (3+1) / (15+12) \* (2+1)/(15+12) \* (2+1)/(15+12) \* (0+1)/(15+12) \* (1+1)/(15+12) = 3/4 \* 13/27 = 13/36 > 0

# 6. Neural NLP and transfer learning

## 6.1 (4 points) How can transfer learning alleviate the limited data problem?

Pre-trained language model is applied in transfer learning. Then, the knowledge or weights trained on unlabeled data is transferred from pre-trained model to this NER tasks.

## 6.2 (3 points) What resource(s) would be needed for transfer learning for this task?

# 7. Evaluation

## 7.1 (6 points) Compute (please show the fractions): i) the recall for the A class ii) the precision for the A class

TP (A) = 2; TN (A) = 3; FP (A) = 2; FN (A) = 3.

therefore,

Recall = TP / (TP+FN) = 2/5 = 0.4

Precision = TP / (TP+FP) = 2/4 = 0.5

## 7.2 (3 points) Which of the three labels I, O, B would you disregard in the evaluation of sequence labelling methods, and why?

1. O

2.1 not relevant since they are outside any entity and do not give too much informativeness.

2.2 class distribution is not balanced since the number of class O tend to greatly outnumber other classes.

# 11．Domain-specific search

## 11.1 (3 points) What is a controlled vocabulary?

Controlled vocabulary is an arrangement of words and phrases used to index and retrieve content. It typically describes a specific domain.