i

Front page

IN4080 Natural Language Processing

Monday 25 November
9:00 AM - 01:00 PM (4 hours)

All questions should be answered!
Each question is assigned a weight which is indicated.
The maximum number of points for the whole set is 100 points.

Permitted materials: None
An on-screen calculator is available.

You may answer in English, Norwegian, Danish or Swedish.

1(a)

Part-of-speech tags

- What do we mean by “part-of-speech”?
- What does it mean that a text is tagged with part-of-speech?
- Why is it useful in natural language processing to part-of-speech-tag a text?

Fill in your answer here

Format

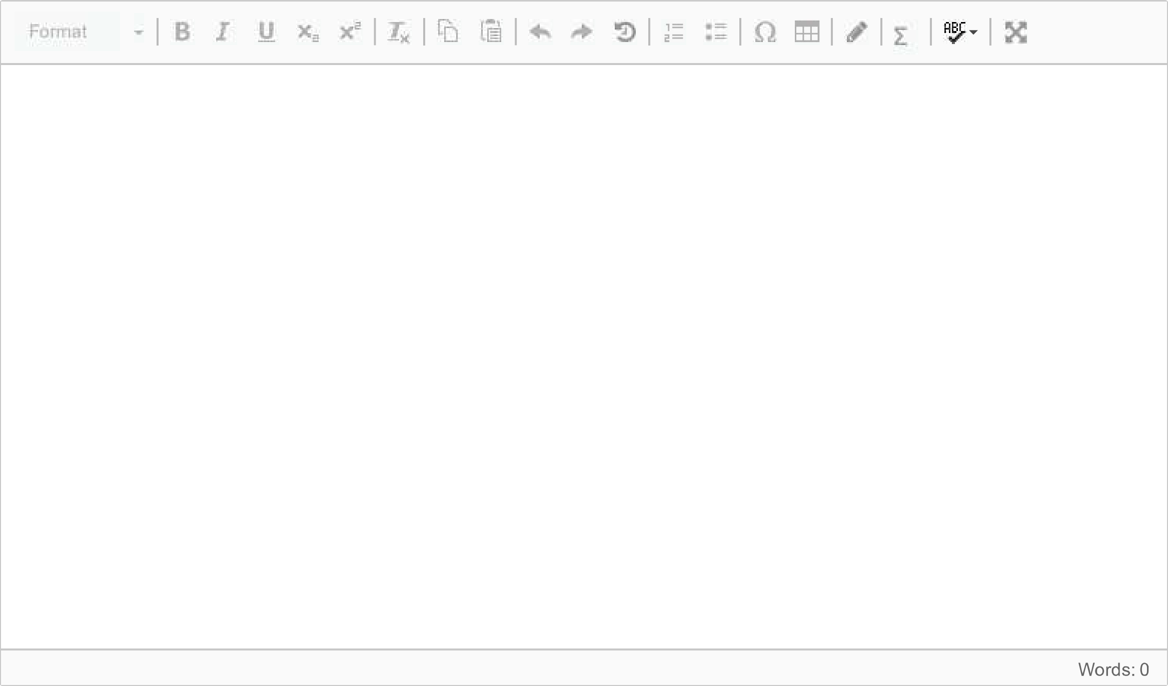
Words: 0

Maximum marks: 5

1(b) Hidden Markov models

- Given a sentence $w_1, w_2, \ldots, w_n$ and a tag set, what is the goal of a part-of-speech tagger?
- How does a hidden Markov model use the Bayes formula to solve this task?
- Which simplifying assumptions are made by the hidden Markov model?
- Give the formula for the hidden Markov model!

Fill in your answer here



Maximum marks: 10

1(c) Training and applying an HMM-tagger

We will train an HMM-tagger and use it to tag the sentence

1) *they gave her flowers*

As training data, we will use the Brown corpus with the universal POS-tag set. All the relevant counts for sentence (1) can be found in the enclosed pdf-file. How will the HMM-tagger go about to tag the sentence, and which tag sequence will it choose?

Fill in your answer here

Format

Words: 0

Maximum marks: 10

2(a) Test sets

In the mandatory assignments, we used two different test sets, one during development, and another one for final testing? Why did not we use only one and the same test set?

Fill in your answer here

Format

Words: 0

Maximum marks: 3

2(b) N-fold cross-validation

Explain the principles for *n*-fold cross-validation, e.g. 10-fold cross-validation.

Fill in your answer here

Format

Words: 0

Maximum marks: 6

2(c) Motivation

- Why does one apply cross-validation?
- What do you think are the reasons we used cross-validation in mandatory assignment 2A text classification, but not in mandatory assignment 2B tagging?

Fill in your answer here

Format

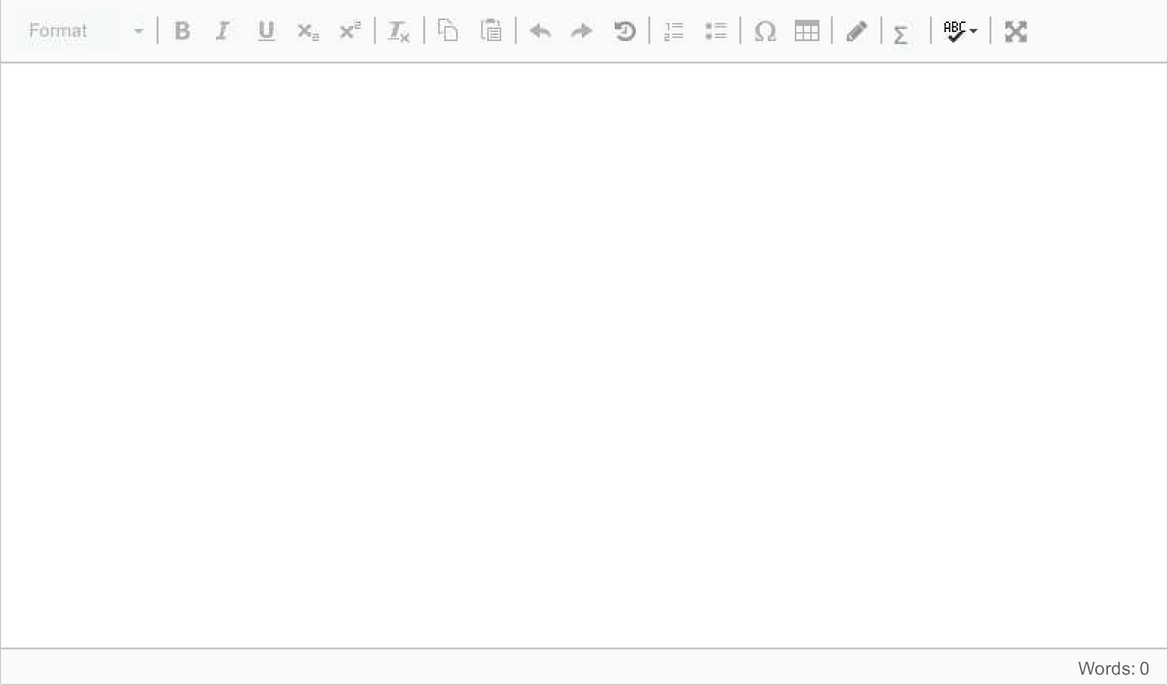
Words: 0

Maximum marks: 6

3 Generative and discriminative

- What is the difference between a generative and a discriminative classifier?
- Which of the following are generative and which of them are discriminative?
  - Bernoulli Naive Bayes for text classification
  - Multinomial Naive Bayes for text classification
  - Logistic regression for text classification
  - An hidden Markov model (HMM) tagger
  - A maximum entropy (logistic regression) tagger

Fill in your answer here



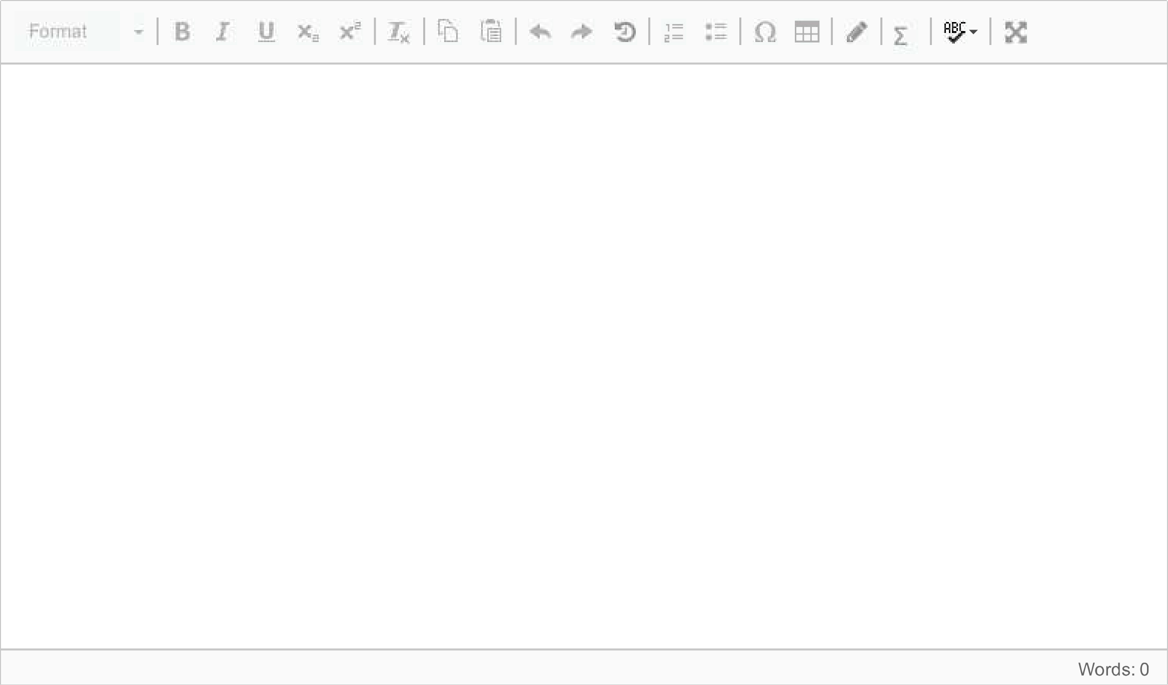
Maximum marks: 5

4(a) Unlabeled scores

The goal of this exercise is to evaluate a named-entity recognizer. The pdf-document shows the results of a named-entity system trained and tested on the Spanish data from the conll2002 shared task, distributed with NLTK. The second column shows the tokens and the third column shows POS-tags. The POS-tags were used by the NER-system but are without interest for this exercise. The fourth column shows the gold IOB-tags and the fifth column shows the predicted IOB- tags. These two columns are the basis for the evaluation.

There are two ways to evaluate the NER-system, labeled or unlabeled. We will first consider the unlabeled scores. With the unlabeled score, one only evaluates whether the system have localized the named entity spans correctly, e.g. the span (15, 16) in sentence 554 is counted as a true positive. What are the unlabeled recall, precision and f-measure for the named entity spans? Explain how you find the numbers.

Fill in your answer here



Words: 0

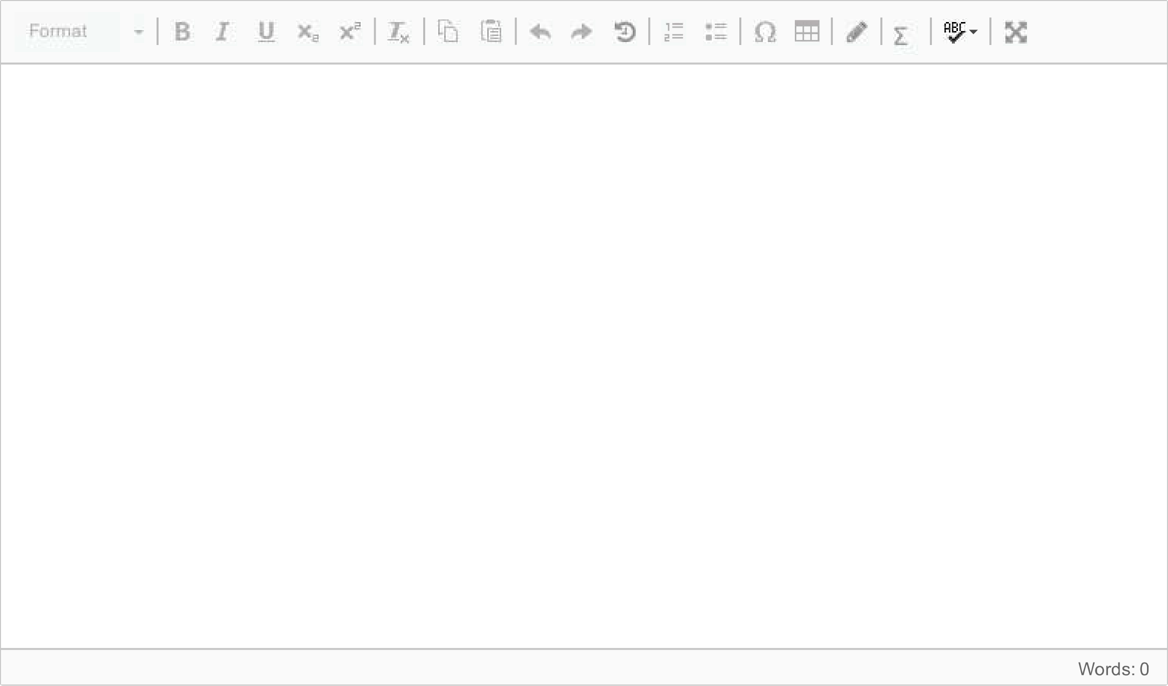
Maximum marks: 10

4(b) Labeled scores

We will calculate the labeled scores from the same four example sentences. With labeled scores, the span (15, 16) in sentence 554 is not counted as a true positive since it gets different labels in the gold set and the predicted set. Count the true positives, false positives and false negatives for each of the four classes of named entities, and report in a table!

Calculate the precision, recall and f-measure for each of the four classes and explain how you find them.

Fill in your answer here

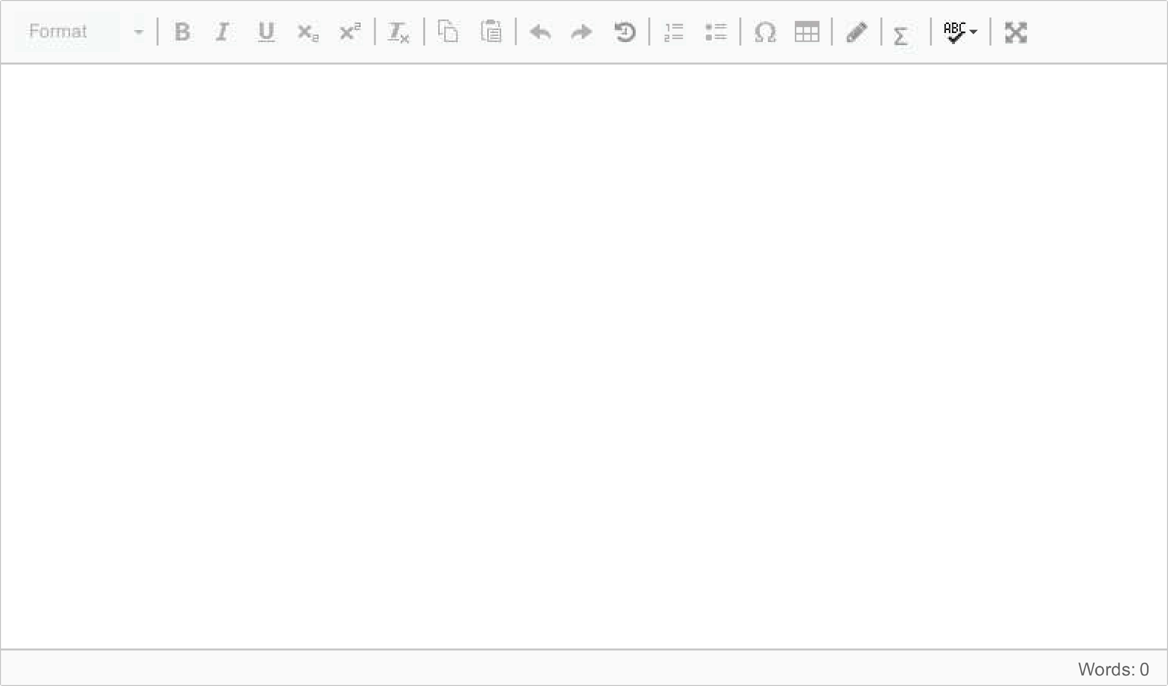


Maximum marks: 9

4(c) Micro and macro scores

Calculate the macro and micro recall, precision and f-measure across the four classes and show how you find them!

Fill in your answer here



Maximum marks: 6

5(a) Fundamental frequency

What is the fundamental frequency (F0) in acoustics? And why might it be useful for a spoken dialogue system to measure it?

Fill in your answer here

Format

Words: 0

Maximum marks: 5

5(b) IR-based chatbots

Assume you wish to develop an IR-based chatbot. To make simple, you decide to use cosine similarity over TF-IDF-weighted vectors. You collect a small corpus of 10 utterances:

1 *Hi Charles!*
2 *Hello Elsa!*
3 *How are you?*
4 *Fine, and you?*
5 *A bit tired.*
6 *Why?*
7 *Busy at work.*
8 *But you are doing great!*
9 *Thanks.*
10 *See you later!*

1) Compute the TF-IDF values for the tokens in the utterances 3, 4 and 8. Don't forget to include the punctuation in your computations.

2) Now assume that you are given the following user input:

How are you doing?

What will be the answer selected by the chatbot trained on the corpus above? Describe your calculation steps.

Fill in your answer here

Format

Words: 0

Maximum marks: 15

5(c) MDP

1) What is a Markov Decision Process (MDP)? Give a formal definition.
2) How can MDPs be used in dialogue management?

Fill in your answer here

Format

Words: 0

Maximum marks: 10

```
Absolute frequencies of words
=============================
flowers     57
gave       285
her       3036
they      3620


Absolute frequencies of tags
============================
DET       137019
NOUN      275558
PRON       49334
VERB      182750


Absolute frequencies of words with their tags
=============================================
         DET NOUN PRON VERB
flowers    0   57    0    0
   gave    0    0    0  285
    her 1929    0 1107    0
   they    0    0 3620    0


Bigram tag frequencies
======================
       DET  NOUN  PRON  VERB  <\s>
 <s> 12238  8093  9157  2588     0
 DET   809 85838  1358  8861    18
NOUN  4270 41144  5460 43763   914
PRON   864   437   404 34838     5
VERB 29784 17819 10058 33667   102
```

For example, the tag NOUN is followed by the tag DET 4270 times.

```
Sentence nr: 525
  0   Gari                    VMI   B-PER       B-PER
  1   Kasparov                AQ    I-PER       I-PER
  2   (                       Fpa   O           O
  3   RUS                     NC    B-ORG       B-LOC
  4   )                       Fpt   O           O
  5   4,5                     Z     O           O
  6   .3                      Z     O           O
  7   .                       Fp    O           O
Sentence nr: 528
  0   Michael                 VMI   B-PER       B-PER
  1   Adams                   AQ    I-PER       I-PER
  2   (                       Fpa   O           O
  3   ING                     NP    B-ORG       B-ORG
  4   )                       Fpt   O           O
  5   3,5                     Z     O           O
  6   .6                      Z     O           O
  7   .                       Fp    O           O
Sentence nr: 554
  0   En                      SP    O           O
  1   los                     DA    O           O
  2   próximos                AQ    O           O
  3   meses                   NC    O           O
  4   serán                   VSI   O           O
  5   concedidos              VMP   O           O
  6   los                     DA    O           O
  7   cinco                   DN    O           O
  8   premios                 NC    O           O
  9   restantes               AQ    O           O
 10   ,                       Fc    O           O
 11   Letras                  NC    B-MISC      O
 12   ,                       Fc    O           O
 13   Artes                   NC    B-MISC      B-PER
 14   ,                       Fc    O           O
 15   Cooperación             NC    B-MISC      B-PER
 16   Internacional           AQ    I-MISC      I-PER
 17   ,                       Fc    O           O
 18   Concordia               NC    B-MISC      O
 19   y                       CC    O           O
 20   Deportes                NC    B-MISC      O
 21   .                       Fp    O           O
Sentence nr: 668
  0   El                      DA    O           O
  1   Madrid                  NC    B-ORG       B-LOC
  2   se                      P0    O           O
  3   refugia                 VMI   O           O
  4   en                      SP    O           O
  5   Versalles               NC    B-LOC       O
  6   .                       Fp    O           O
```