IN1140

Prøveeksamen H2019

Hjelpemidler

Ingen.

Introduction

The exam consists of 11 assignments. We have removed two assignments from last year's exam, as these topics were not covered in the lectures this year.

If something is unclear, you are free to formulate your own assumptions as long as these are clearly described in your answer.

---

1 Regular expressions

Which of the alternatives cannot be recognized by the following regular expression:

```
[1-9][0-9]*\s(cent(s)?|dollar(s)?\s*([1-9][0-9]*\scent(s)?))
```

Velg ett alternativ:

- 1 dollar 35 cents
- 35 dollars
- 99 cents
- 99 dollars 1 cent

2 Tokenisation (10 poeng)

Tokenisation is an important task in language technology systems and involves dividing a text into running words. Consider the following sentence and answer the questions below.

```
On the barn sits the Santa with his Christmas porridge , so good and sweet
, so good and sweet .
```

1. We often distinguish between tokens and types when counting word occurrences in a text. What is the difference between these? How many tokens and types does the sample sentence consist of?
2. A very simple tokenizer will split a text on white spaces as well as exclude all punctuation as done in the example sentence above. However, such an approach could lead to some problems. Give at least two examples of different uses of punctuation that such a tokenizer will not process correctly.

3 Ordklasser (7 poeng)

OBS: Flervalgsoppgave (paring)
Here we will work with the following sentence:

```
The film has exciting twists and turns that take you to great places in Oslo
and Vienna
```

Given the word classes in Table 1 below, assign word classes to all the words in the sentence. You must select one option for each word.

NOUN	Nouns
VERB	Verb
ADJ	Adjective
PREP	Preposition
PRON	Pronoun
CONJ	Conjunction
ADV	Adverb
SUBJN	Subjunction
DET	Determiner

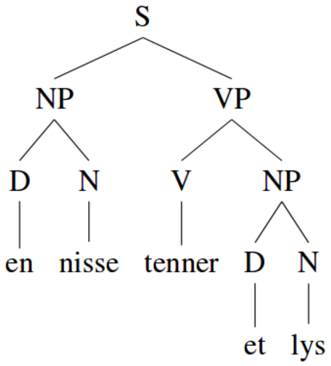
Table 1

4 Ordklassetagging (5 poeng)

Part of speech tagging algorithms fall into two main categories, which ones? Briefly explain what characterizes these and what is the difference between them.

5 Kontekstfri grammatikk (10 poeng)

Start from the attached syntactic tree and answer the following questions:



1. Give the phrase structure rules corresponding to the enclosed tree. You must include rules for both non-terminal and terminal nodes.

2. Show how you can extend the grammar from question 1. to provide a syntactic analysis for the following sentences:

   - Et barn spiser en pepperkake
   - En nisse synger

3. In Norwegian, we have a conjunction between determiners and nouns in noun phrases. Show how you can extend the grammar to exclude non-grammatical noun phrases like the ones from the grammar.

   - *et nisse
   - *en barn

6 Språkmodeller (10 poeng)

The attached table specifies the formula for a language model (a so-called biggram model) and a small text corpus.

- Which biggrams occur in the corpus?

- How do we calculate the probability of a word given the previous word ($P(w_i|w_{i-1})$) from a corpus?

- You now have to use the biggram model as well as the text corpus to calculate the probability of the sentence "`<s> Jeg foretrekker kinesisk mat <\s>`". Show which probabilities you need and how they are calculated from the corpus. You do not need to calculate the total probability of the sentence.

Formel:

$$P(w_1 \ldots w_k) = \prod_{i=1}^{k} P(w_i | w_{i-1})$$

Tekskorpus:

```
<s> Jeg liker indisk mat <\s>
<s> Gina liker kinesisk mat <\s>
<s> Thomas foretrekker kinesisk mat <\s>
<s> Jeg foretrekker indisk mat <\s>
```

Table 2: Formel og tekstkorpus.

7 Leksikale relasjoner (6 poeng)

	Synonymi	Antonymi	Hyponymi	Hypernymi	Meronymi	Homonymi
father - man						
bryter (athlete) - bryter (switch)						
fast – quick						
old - young						
bicycle - mountain bike						
member - association						

Table 3

8 Semantiske roller (5 poeng)

	INSTRUMENT	SOURCE	AGENT	THEME	EXPERIENCER	GOAL
Santa is stirring the porridge						
Santa is stirring with a *spoon*						
The *porridge* is in the barn						
Santa travels from the *North*						
The *kid* smells the Christmas porridge						

9 Bayes regel (5 poeng)

Show how Bayes rule is derived from the following conditional probability:

$$P(A|B) = P(A, B)/P(B)$$

10 Setningssemantikk (7 poeng)

What does it mean that two sentences are in an entailment relationship? Illustrate your answer with at least one example.

11 Named Entity Recognition (NER)

Den vanligste måten å løse NER oppgaven er ved ord-for-ord klassifisering, såkalt BIO-klassifisering, der data representeres vanligvis ved trekk (features).

- Forklar kort hva BIO står får.
- Gi eksempler på 4 type trekk som kan brukes for å løse oppgaven.

The most common way to solve the NER problem is by using word-by-word classification, so-called BIO classification, where the words are represented by features.

- Briefly explain what BIO stands for.
- Give examples of 4 types of features that can be used to solve this task.