IN4080 2018

1

Language models

In this exercise we will explore various properties of language models.

i

Front page

IN4080 Natural Language Processing

Wednesday 5 December
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: On-screen calculator

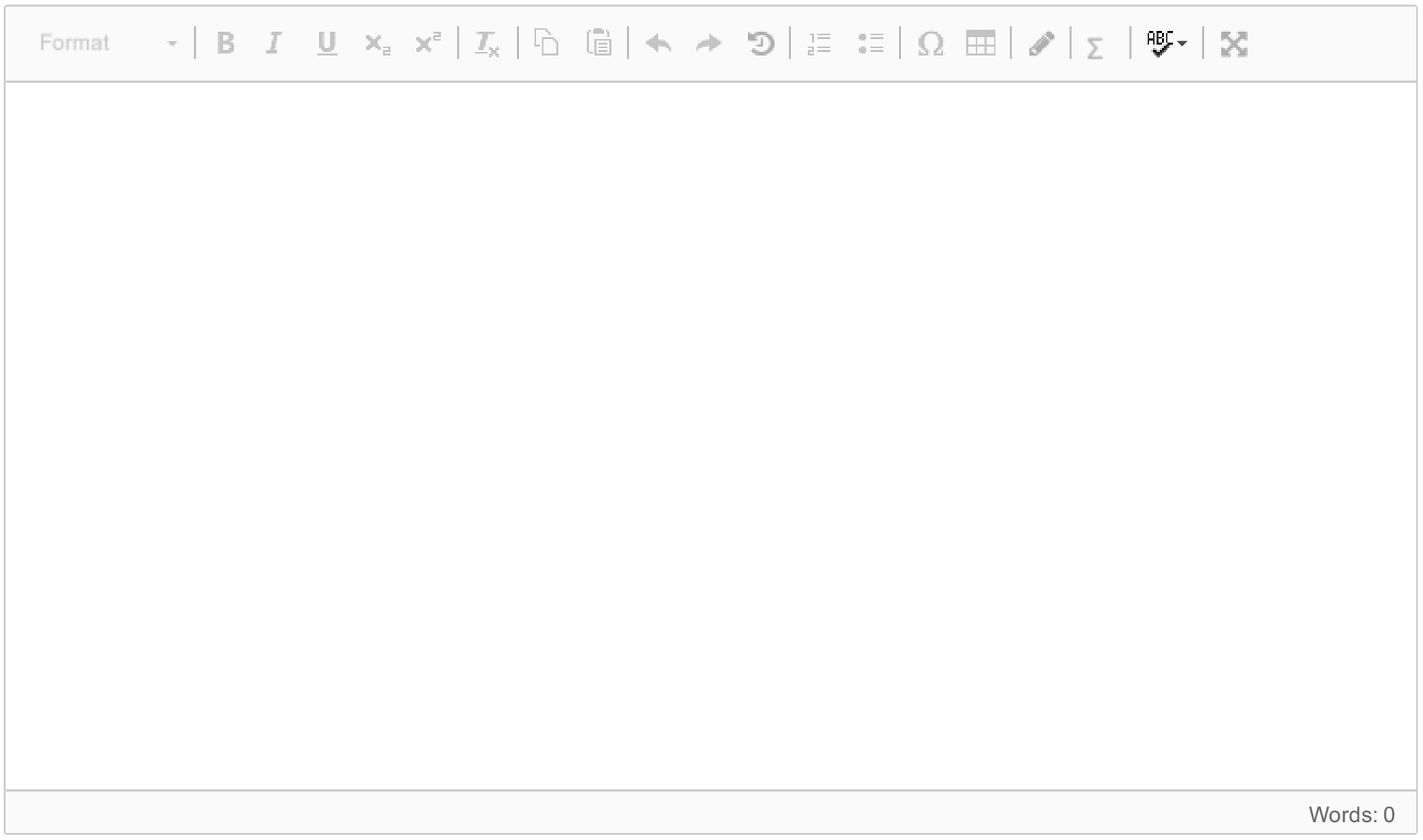
You may answer in English, Norwegian, Danish or Swedish.

(a)

Use

Name at least two practical tasks where language models may be useful.

Fill in your answer here



Maximum marks: 4

(b)

Model

Explain how a bigram language model will assign a probability to a sequence of words. You are advised to use (simplified) equations in your explanation. Which assumption is made by the model?

Fill in your answer here

Format | B I U x_2 x^2 I_x | Ω | ABC

Words: 0

Maximum marks: 6

i

Corpus 1

Corpus 1

<s> Sam likes Pam <\s>
<s> Pam likes Sam <\s>
<s> Sam likes egg and ham<\s>

(c)

Bigram probability

Consider a language model trained on corpus 1. Which conditional bigram probability will it ascribe to P(likes | Sam), assuming a straightforard unsmoothed maximum likelikehood estimation?

Fill in your answer here

Format

Words: 0

Maximum marks: 2

(d) Sentence probability

Which probability will the model ascribe to the following sequence?
<s> Sam likes Sam <\s>
Show how the number is calculated.

Fill in your answer here

Format | B I U x_2 x^2 | I_x | | | | | Ω | | | Σ | ABC | |

Words: 0

Maximum marks: 4

(e)

Problem

Which problems will the bigram model face in assigning a probability to the following sentence?
<s> Sam likes ham <\s>

Fill in your answer here

Format

Words: 0

Maximum marks: 2

(f)

Interpolation

One way of fixing these problems is to use interpolation. Explain how interpolation works for a bigram model.

Fill in your answer here

Format

Words: 0

Maximum marks: 4

(g)

Applying interpolation

Which probability will an interpolated bigram model ascribe to the sentence?
<s> Sam likes ham <\s>
You may assume a 0.5-0.5 weighting.
Show how you get the result.

Fill in your answer here

Format

Words: 0

Maximum marks: 4

(h)

Add one

In general, it is not a good idea to use Laplace smoothing ("add-one") to smooth bigram probabilities. Explain why!

Fill in your answer here

Format

Words: 0

Maximum marks: 3

2

Text classification

(a) Naive Bayes

$$\arg\max_{c \in C} P(c \mid \mathbf{f}) = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(f_i = v_i \mid c)$$

The formula shows the model for Naive Bayes classification.

- Give a short description of the formula:
  - What is C and c?
  - What is **f**, f_i, v_i and n?
  - What is meant by *argmax*?
- Which simplifying assumptions are made?

Fill in your answer here

Format

Words: 0

Maximum marks: 5

i

Data set

Data set:

Document	Class
fun, fun, good, good, bad	pos
exciting, good, exciting	pos
fun, exciting, exciting, bad	pos
bad	neg
terrible, terrible, terrible, terrible, terrible	neg

(b)

Multinomial

There are several ways the Naive Bayes model can be used for text classification. One of them is called *Multinomial Naive Bayes*. Given the two classes *pos* and *neg*, and the data set, what are the values of the $P(f_i = v_i \mid c)$ for $i = 1, 2, \ldots n$?

Fill in your answer here

Format | B I U x_2 x^2 | I_x | | | | | Ω | | | Σ | ABC | |

Words: 0

Maximum marks: 5

(c)

Binarized NB

A variant of Multinomial Naive Bayes is called Binarized Multinomial Naive Bayes. We saw in mandatory assignment 2 how this binarized variant was potentially better for sentiment classfication. How does the binarized variant differ from the standard Multinomial NB? What are the values of the $P(f_i = v_i \mid c)$ for $i = 1, 2, \ldots n$ in the the binarized model?

Fill in your answer here

Format

Words: 0

Maximum marks: 5

3(a)

Word vectors and cosine similarity

When calculating the similarity of two vectors, we may use cosine as a similarity metric. Cosine similarity between two vectors A and B is defined as:

$$sim(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

The nominator is the *dot product* of vectors A and B: $\sum_{i=1}^{n} A_i B_i$

The denominator is the product of the *magnitudes* or *lengths* of vectors A and B: $\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}$

Below is a term-context matrix, containing TF-IDF values for a subset of the **target** and *context* words. The values are artificial, in order to make the calculations easier.

	pizza	**soup**	**lunch**
a	0	2	1
anchovies	2	0	0
ate	3	2	1
enjoys	1	2	2
lunch	1	3	0
pizza	0	0	1
soup	1	0	1
spoon	0	2	1

Let's call the vectors for the words **pizza**, **soup** and **lunch** for V_{pizza}, V_{soup} and V_{lunch}, respectively.

First, let's calculate the *magnitude* or *length* of the term-context vectors.

What is $\|V_{pizza}\|$, i.e. the *magnitude* of V_{pizza}? Give your answer as a numeric value, nothing else.

What is $\|V_{soup}\|$, i.e. the *magnitude* of V_{soup}?

What is $\|V_{lunch}\|$, i.e. the *magnitude* of V_{lunch}?

Using cosine similarity as a similarity metric, what is the similarity between the following word pairs?

pizza and **soup**

pizza and **lunch**

Maximum marks: 8

3(b) Distributional semantics

The key concept in distributional semantics is to derive vector representations of words, given their distribution in text. Two popular alternatives are:

1. **tf-idf vectors**, sometimes also also called *term-context vectors* or *co-occurence vectors*
2. **word embeddings**, where *Word2Vec* is one of the most popular algorithms

Briefly describe *two* central differences between tf-idf vectors and word embeddings. Try to both explain what the difference consist of, and what the consequences for applying the approach in NLP applications are.

Fill in your answer here

Format

Words: 0

Maximum marks: 7

3(c) # TFIDF

TFIDF is a function that computes a score (or weight) for term t given:

- a corpus of documents, D
- a specific document d, where $d \in D$
- a term t, where $t \in d$

TFIDF consist of two parts, TF and IDF.

1) Which of the parameters t, d, D are necessary to calculate TF?

- t
- d
- D
- t, d
- t, D

2) Which of the parameters t, d, D are necessary to calculate IDF?

- t
- d
- D
- t, d
- t, D

3) What does the TFIDF formula try to capture about the relationship between the term, the document and the corpus?

- The meaning of a term in a document, given a corpus.
- The importance of a term in a document, given a corpus.
- The importance of a document in a corpus, given a term.

Maximum marks: 6

4

Dependency structure and dependency parsing

i

Structure 1

Structure 1

	word	lemma	upos	xpos	head	type
1	What	what	PRON	WP	6	obj
2	do	do	AUX	VBP	4	aux
3	you	you	PRON	PRP	4	nsubj
4	like	like	VERB	VB	0	root
5	to	to	PART	TO	6	mark
6	do	do	VERB	VB	4	xcomp
7	?	?	PUNCT	.	4	punct

(a)

Projectivity

This (simplified) dependency structure is taken from the UD_English treebank.

Is the structure projective or not?
State reasons for your answer.

Fill in your answer here

Format

Words: 0

Maximum marks: 5

(b)

Evaluation metrics

In evaluation of dependency structures, two metrics are often used: unlabeled attachment score (**UAS**) and labeled attachment score (**LAS**). Explain briefly these metrics and in particular the differences between them.

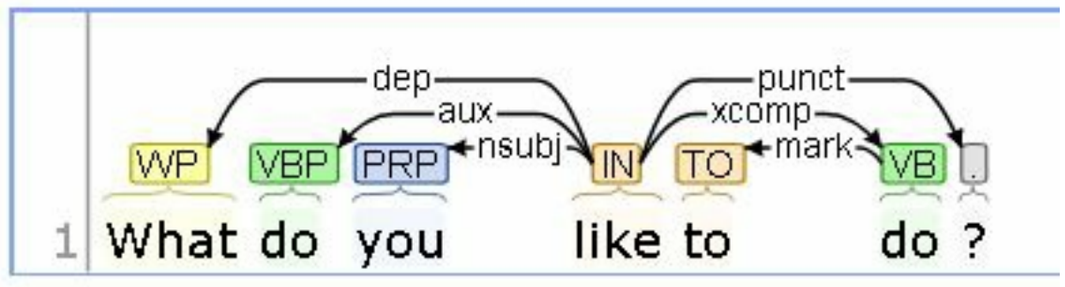
Fill in your answer here

Format

Words: 0

Maximum marks: 4

(c) LAS and UAS



When the CoreNLP system parses the sentence, it produces the structure above, call it structure (2).

Taking structure(1) as the gold standard, what is the LAS and UAS of structure (2)?

Fill in your answer here

Format

Words: 0

Maximum marks: 3

(d) Transition-based dependency parsing

Show step-by-step how a transition-based parser can produce structure (2). Be explicit with respect to the data structures used and the effect of each step.

Fill in your answer here

Format

Words: 0

Maximum marks: 8

5

Information extraction

(a)

Steps

What are the typical steps of an information extraction system? Explain what the goals are for each step. You do not have to explain how the actual steps are carried out.

Fill in your answer here

Format

Words: 0

Maximum marks: 5

(b)

Manual method

One of the steps in an information extraction system is relation extraction.
There are several different methods for relation extraction. One method
is to use hand-written patterns. Explain shortly the main principles of this approach.
What are the bottlenecks of this approach?

Fill in your answer here

Format

Words: 0

Maximum marks: 5

(c) Supervised

An alternative method is to use supervised classification. Explain shortly the main principles of this method. What are the bottlenecks of this method?

Fill in your answer here

Format

Words: 0

Maximum marks: 5