Front page

Written exam in

IN4080 Natural Language Processing

2020 Fall

Disclosure of exam assignment: December 2 at 3:00 PM
Submission deadline: December 2 at 7:00 PM

It is important that you read this page carefully before you start.

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

Digital comforting round (trøsterunde)
If you have any questions regarding the exam questions, you may ask them in Zoom at 4:00 PM. There will be two Zoom channels

- Channel 1 (Jan Tore) Answering questions related to questions 1-3
  Zoom link:
  Meeting ID: 652 6323 7154
  Passcode: 088496
- Channel 2 (Pierre) Answering questions related to questions 4-5
  Zoom link:
  Meeting ID: 658 2513 5812
  Passcode: 736643

When you enter the Zoom channel, you will be put into a waiting room where you will await your turn. Any answers that may be important for more students will be posted on the semester web page:
https://www.uio.no/studier/emner/matnat/ifi/IN4080/h20/

General information:

- It is important that you check the course's semester page regularly. Important messages during the exam will be posted on the semester page.
- Remember that your submission need to be anonymous, do not write your name in your submission.
- All examination support materials are permitted. You need to gather information from available sources, assess the information quality, and put it together in a submission based on your own processing of the content. The submission must reflect your individual level of knowledge.

Collaboration
You are not allowed to collaborate or communicate with others, however. Note that after the home exam you can be randomly selected for a "control interview". The control interview will not affect the grade you have received.
https://www.mn.uio.no/english/about/hse/corona/kontrollsamtale.html

After the interview, if the teacher suspects that you have not written the assignment yourself, the department can issue a suspicion-of-cheating case.
https://www.uio.no/english/about/regulations/studies/studies-examinations/routines-cheating.html

Support
For technical questions during the exam contact: https://www.mn.uio.no/english/studies/exam/user-support.html

For more information regarding the exam at the MN faculty the fall 2020, see
https://www.mn.uio.no/english/about/hse/corona/examination-2020.html

Introduction

The goal is to build a sentiment classifier for sentences. We are given a corpus of sentences and each sentence is marked for sentiment. To make it possible to estimate a model manually, we have made a synthetic – and highly artificial – corpus consisting of 50 sentences, as you can see in the table.

Ref	Sentence	Label	Number of copies
a	It is bad.	NEG	20
b	It is good.	POS	15
c	It is not bad.	POS	5
d	It is not good.	NEG	10

1(a) Stop words

What do we mean by "stop words"? Give example of stop words. Why should we consider removing stop words before training the sentiment classifier?
Fill in your answer here

Maximum marks: 3

1(b) Naive Bayes

We will consider 'it' and 'is' as stop words and remove them. We will then build a simple multinomial naive Bayes bag of words text classifier, and train it on the corpus. The classifier will only use single words as features.

To see how well the model fits the training corpus, we will evaluate it on the training corpus. The classifier will classify the example sentence (c), "It is not bad.", wrongly. Show how the classifier achieves this result.

If you think any assumptions are missing, make your own and state them clearly.

Fill in your answer here

Maximum marks: 10

1(c) Accuracy

What is the accuracy of the classifier on the training corpus?

Fill in your answer here

Maximum marks: 4

1(d) Evaluation

- What are the precision, recall and f-score for the classifier for each of the two classes?
- What are the macroaverage precision, recall and F-score across the two classes?
- What is the microaverage precision across the two classes?

Fill in your answer here

Maximum marks: 6

1(e) Alternatives

Say, you are not content with the performance of the classifier on the training set, and you want to construct a classifier which fits the training set better. Discuss various strategies which can be applied for building better classifiers given the training data.

Fill in your answer here

Maximum marks: 7

2(a) Definition

Explain shortly (1-4 sentences) what is meant by a "language model (LM)".

Fill in your answer here

Maximum marks: 3

2(b) Training and testing

We train a bigram language model on the corpus of 50 sentences repeated below. Which probability will this model ascribe to sentence (c)? What is the perplexity of the sentence in the model? If you think any assumptions are missing, make your own and state them clearly.

Ref	Sentence	Label	Number of copies
a	It is bad.	NEG	20
b	It is good.	POS	15
c	It is not bad.	POS	5
d	It is not good.	NEG	10

Fill in your answer here

Maximum marks: 7

2(c) Neural networks

An alternative to an n-gram language model is to base the language model on (only) a feed-forward neural network and a "sliding window". In which respects is such a model different from an n-gram language model? Which advantages does it have compared to n-gram models?

Fill in your answer here

Maximum marks: 5

2(d) Recurrent neural networks

The "sliding window" feedforward neural net language models have some shortcomings which may be taken care of by a language model using a recurrent neural network (RNN). What kind of shortcomings? Illustrate with language examples. Explain how a RNN may overcome these problems.

Fill in your answer here

Maximum marks: 5

3(a) Approaches

In this course we have seen two ways of representing words as vectors based on their distribution in a corpus. The first is based on word-context matrices and the other is called "word embeddings". Describe shortly the main ideas of the two approaches. In particular, compare the two approaches with respect to the form of the vectors and how the vectors are derived.

Fill in your answer here

Maximum marks: 10

3(b) Negative sampling

One-method for deriving embeddings is called "skip-gram with negative sampling". Describe what is meant by "negative sampling" in this setting. How does it work? Why is it introduced?

Fill in your answer here

Maximum marks: 6

4(a) Grounding

Why is it important to think about (conversational) grounding when developing dialogue systems? Support your answer with a few examples.

Fill in your answer here

Maximum marks: 5

4(b) Speech processing

Calculate the Word Error Rate (WER) between this utterance:

could you go to my office and pick up my NLP book

And the hypothesis generated by a speech recogniser:

could you got you my office and pickup my NLB book

Show your calculations using an edit distance matrix. You can assume that insertions, deletions and substitutions all have a cost of 1.

Fill in your answer here

Maximum marks: 6

MDP

You wish to develop a (phone-based) spoken dialogue system that will call random U.S. citizens in order to collect opinion poll data for the next US election, with two candidates on the ballot: Kamala Harris and Ivanka Trump.

This system is framed as a *Markov Decision Process* (MDP) formalised as such:

- We have five possible states:
  - s_1 is the starting state
  - s_2 if the callee indicated their intention to vote for Kamala Harris
  - s_3 if the callee indicated their intention to vote for Ivanka Trump
  - s_4 if the callee expressed something else (that was not understood)
- The set of actions that can be taken by the dialogue system are as follows:
  - a_1 : Say "Hi, I'm a automated bot developed to collect polling data. May I ask you for whom you plan to vote in the next election?"
  - a_2 : Say "Sorry I did not understand. Who do you wish to vote for?"
  - a_3 : Say "Ok, thank you for your help, and have a nice day!"
- The transition model is as follows:
  - In state s_1, only action a_1 is possible, with three possible transitions:
    $P(s'=s_2|s=s_1,a=a_1)=0.48 P(s'=s_3|s=s_1,a=a_1)=0.40 P(s'=s_4|s=s_1,a=a_1)=0.12$
  - In states s_2 and s_3, only a_3 is possible and terminates the dialogue.
  - In state s_4, only a_2 is possible, with the following transitions:
    $P(s'=s_2|s=s_4,a=a_2)=0.36 P(s'=s_3|s=s_4,a=a_2)=0.32 P(s'=s_4|s=s_4,a=a_2)=0.32$
- Finally, the reward model is defined as such:
  - $R(s=s_2,a=a_3)=R(s=s_3,a=a_3)=10$ (if the system manages to register the callee's political preference)
  - $R(s=s_4,a=a_2)=-1$ (to capture the annoyance of asking the callee to repeat)
  - Other actions have a reward of zero.

4(c) MDP (a)

Based on this MDP model, calculate the expected cumulative reward of asking the callee to repeat when their answer was not properly understood, that is: $Q(s=s_4,a=a_2)$. You can assume a discount factor of 0.9.

Fill in your answer here

Maximum marks: 8

4(d) MDP (b)

One limitation of this MDP model is that is assumes that the dialogue system will always be 100 % certain it has correctly understood the political preference expressed by the callee. In practice, this will not always be the case, because of e.g. speech recognition or NLU errors, or because the callee may intentionally provide unclear or misleading information. How could this model be modified to capture those uncertainties?
Fill in your answer here

Maximum marks: 3

Fairness

You have developed an NLP model for automated essay scoring in Norwegian, and you wish to ensure your model is *fair*, in particular when it comes to whether the student is ethnically Norwegian or not.
To this end, you compare the essay scores with scores assigned by experienced teachers. To simplify our problem we will rely on binary pass/fail scores. In addition, we will assume that the human teachers themselves are free from social biases regarding the ethnicity of the students.
Here are the scores that are respectively produced by your NLP model and by the human teachers for a group of 21 students:

ID	Ethnical Norwegian?	Score from model: Pass (✓) or Fail (F)	Score from teachers: Pass (✓) or Fail (F)
1	*Yes*	✓	✓
2	**No**	✓	✓
3	*Yes*	✓	✓
4	*Yes*	**F**	**F**
5	*Yes*	✓	✓
6	*Yes*	**F**	**F**
7	*Yes*	✓	✓
8	**No**	**F**	**F**
9	**No**	✓	✓
10	*Yes*	✓	✓
11	*Yes*	✓	✓
12	**No**	✓	✓
13	*Yes*	✓	✓
14	**No**	**F**	**F**
15	*Yes*	✓	✓
16	*Yes*	✓	**F**
17	**No**	✓	✓
18	*Yes*	**F**	✓
19	**No**	✓	**F**
20	**No**	**F**	✓
21	*Yes*	✓	**F**

5(a) Fairness (a)

Based on this data, determine which fairness criteria covered during the course (demographic parity, predictive parity and equalised odds) are satisfied or not satisfied by your essay scoring model.

NB: We assume the essay scoring model does not have direct access to the ethnicity of the student, and the "unawareness" criteria is thus irrelevant here.

Fill in your answer here

Maximum marks: 10

5(b) Fairness (b)

Would you consider your NLP model as being fair to the students that are not ethnical Norwegian? Explain your answer.

Fill in your answer here

Maximum marks: 2