NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer Science (IDI)

Examination paper for TDT4310:
Intelligent Text Analytics and Language Understanding

Academic contact during examination:	Prof. Björn Gambäck
Phone:	+46 70 568 15 35
E-mail:	gamback@ntnu.no
Technical support during examination:	Orakel support services
Phone:	+47 73 59 16 00
Examination date:	June 2nd, 2020
Examination time (from-to):	09:00 – 13:00
Permitted examination support material:	All support material is allowed

Other information: Since all the course material and the lectures were in English, the exam is too. However, answers may of course be given in either English or Norwegian. Read the questions carefully. If a question is unclear/vague, make your own assumptions and specify in your answer the premises you have made. Only contact academic contact in case of errors or insufficiencies in the question set.

Maximum number of points on the exam: 50. The exam counts for 25% of the overall course grade. You should answer 5 questions, counting for 10 points each. Note that which questions you should answer will depend on factors included in the specification at the beginning of each question. That all questions have the same maximum points does neither imply that the questions are at an equal level of difficulty, nor that the answers to those questions should be of about equal length. The opposite is rather the case, so some questions are most certainly easier than others and require shorter answers. It will thus make sense to first reply to those. Further, be aware that the questions are *not* necessarily ordered by level of difficulty, so read all questions carefully before starting to answer the ones you consider "low-hanging fruit".

Notifications: If there is a need to send a message to the candidates during the exam (e.g., if there is an error in the question set), this will be done by sending a notification in Inspera. A dialogue box will appear. You can re-read the notification by clicking the bell icon in the top right-hand corner of the screen. All candidates will also receive an SMS to ensure that nobody misses out on important information. Please keep your phone available during the exam.

Students will find the examination results in Studentweb. Please contact the department if you have questions about your results. The Examinations Office will not be able to answer this.

File uploading: All files must be uploaded *before* the examination time expires:

1. A PDF document with the answers to the exam questions.
2. A PDF document containing your project report.
3. A PDF document with the slides for your project presentation.
4. (optionally either a, b or both:)
    a. A zip-file with your project code
    b. A link to a page (e.g., on GitHub) with the code
    (include the link in the project report and/or on your exam answering sheet)

How to create PDF documents
Remove personal information from the file(s) you want to upload
How to digitize your sketches/calculations

Saving: Answers written in Inspera are automatically saved every 15 seconds. If you are working in another program remember to save your answer regularly.

Submission: Answers written in Inspera will be submitted automatically when the examination time expires and the test closes, if you have answered at least one question. This will happen even if you do not click "Submit and return to dashboard" on the last page of the question set. You can reopen and edit your answer as long as the test is open. If no questions are answered by the time the examination time expires, your answer will not be submitted.

Accessing your answer post-submission: You will find your answer in Archive when the examination time has expired.

Withdrawing from the exam: If you wish to submit a blank test/withdraw from the exam, go to the menu in the top right-hand corner and click "Submit blank". This *cannot* be undone, even if the test is still open.

Cheating/Plagiarism: The exam is an individual, independent work. Examination aids are permitted. All submitted answers will be subject to plagiarism control. *Read more about cheating and plagiarism here.*

Language:	**English**
Number of pages (front pages excluded):	**3**
Number of pages enclosed:	**5**

TDT4310: Exam Questions 2020 [maximum points: 50]

1) Your plan to sail solo around the world sadly ended badly when you were hit by a severe storm in the middle of the Pacific. You lost control of the steering and of all communication channels to the outside world. Your boat drifted for days with the currents and the winds. You thought all was lost, but fortune smiled upon you when it finally stranded on the shores of an isolated island, which is off most charts and hitherto unknown to be inhabited. However, you quickly find out that there is a population on the island. Friendly people who help you, feed you and shelter you. Still your boat is a total wreck, you fail to re-establish communication with the outside, and the locals speak a language you don't understand (you can safely assume that they speak a human language, but one so far not analysed by any linguist).

   You realise that you'll probably be stuck on the island for a long time, so want to be able to chat with the inhabitants. On the positive side, your solar-powered PC is fully equipped with all the resources you used in a course on text analytics and language understanding that you took at NTNU before your trip.

   a. **Outline roughly the main issues that you would need to address in order to build a system that would allow you to communicate with the island's local population. [2pt]**

      (Hint: if you're unable to think of at least 5-6 steps you need to take to solve this problem, you're probably trying to make life on the island too simple for your own good.)

   b. **Select four of the issues you described in (1a) and for each of those issues in turn discuss in more detail how you'd address it under the given circumstances. [8pt]**

2) Recent language processing research is characterized by the combination of traditional AI (often called "GOFAI")/symbolic/rule-based and corpus-based/statistical/deep learning approaches.

   a. Outline briefly the philosophy of the two approaches:
      How does each approach model language processing, what view(s) does it take of natural language and the properties of languages, and what does it assume (perhaps unrealistically in the extreme case) is feasible? **[5pt]**

   b. The GOFAI/symbolic/rule-based approach to natural language understanding determines the syntactic structure of sentences before carrying out semantic interpretation.
      Give some arguments (with examples) for why syntactic structure is relevant to determining semantics. **[2pt]**

   c. Discuss which of the two approaches (or combination of them) that you would adopt in building a natural language interface to a specific database, and why.
      (You don't need to describe the actual system, only your choice of overall approach.) **[3pt]**

3) Note: In this question you need to answer (a) and then either (b) or (c):

Health informatics include many topics that could be addressed using language technology tools. Hence, for example, electronic patient records contain a lot of interesting information, but the access to the information is in general quite restricted and the hospitals rarely utilise the patient records to their full potential. At a psychiatric clinic, a patient record would commonly contain a description (in words) of the patient's overall mental health condition, followed either by a diagnosis (or the lack of diagnosis, i.e., that the patient is healthy) as assigned by the doctor/psychiatrist or the conclusion that the patient does show some symptoms, but that no specific diagnosis can be assigned at this time.

How would you go about helping such a clinic create applications that could use the information stored in the patient records? Which sets of data would you create, and how? What types of representations would you use?

Depending on your answer to (a) below, consider the use cases described in either (b) or (c):

a. How many characters are there in your surname? **[0pt]**

b. If your answer to (a) is an odd number:

Suppose a psychiatrist fails to diagnose a patient and hence wants to see other patient records with condition descriptions similar to those of the current patient. How would you go about **retrieving patient records with similar mental health condition descriptions? [10p]**

c. If your answer to (a) is an even number:

Suppose the clinic wants to completely replace their psychiatrists with a system, which would be **suggesting diagnoses for new patients only based on their mental health condition descriptions.** How would you go about doing that? Would you be able to claim that your system was completely reliable? (Why or why not?) **[10p]**

4) Note: In this question you need to answer (a) and then either (b) or (c):

a. Which day of the month is your birthday? **[0pt]**

b. If your answer to (a) is an odd number:

List the (in your view) **five most important currently used techniques** in text analytics and language understanding, for each of those techniques motivating in detail why you think it is important enough to be included on your top-5 list. **[10pt]**

c. If your answer to (a) is an even number:

List the (in your view) **five most important issues** in text analytics and language understanding that were addressed in TDT4310 (either in the lectures, in the labs and/or in the TAP or NLTK books), for each of those issues motivating why you think it is important enough to be included on your top-5 list. **[10pt]**

5) Entailment is a common issue addressed in natural language inferencing, with the main problem being whether one sentence entails another sentence. It is defined as that Sentence 1 *entails* Sentence 2 if and only if Sentence 2 is true whenever Sentence 1 is true, as in the following examples:

- The corona virus has spread to all continents. There is corona virus in Antartica.
  `Entailment`
- Håland scored 2 goals and Zlatan scored 1. Håland scored more goals than Zlatan.
  `Entailment`
- Håland was born in England. Håland was born in Leeds.
  `Non-entailment`
- The corona virus has spread to all continents. Håland scored one more goal than Zlatan.
  `Non-entailment`
- The corona virus has spread to all continents. The corona virus has not spread to all continents.
  `Non-entailment`

Suppose that you trained a machine learning model on examples such as those above and the model then deduced that entailment is a property which depends on the second sentence being 6 words long.

a) Explain why this could this happen. **[2pt]**

b) Why would it be wrong? **[2pt]**

c) What could you do to address the problem?
(Suggest more than one way in which the non-desired effects could be mitigated.) **[6pt]**