NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer Science (IDI)

Examination paper for TDT4310:
Intelligent Text Analytics and Language Understanding

Examination date:	May 27th, 2021
Examination time (from-to):	09:00 – 13:00
Permitted examination support material:	A / All support material is allowed

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

If you experience technical problems during the exam, contact Orakel support services as soon as possible before the examination time expires. If you don't get through immediately, hold the line until your call is answered.

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. 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.

About submissions:

- All questions should be answered in a pdf file and uploaded to Inspera.
- All files must be uploaded before the examination time expires.
- 30 minutes are added to the examination time to manage the files. The additional time is included in the remaining examination time shown in the top left-hand corner.
- NB! You are responsible for ensuring that the file(s) are correct and not corrupt/damaged. Check the file(s) you have uploaded by clicking "Download" when viewing the question. All files can be removed or replaced as long as the test is open.
- *How to digitize your sketches/calculations*
- *How to create PDF documents*
- *Remove personal information from the file(s) you want to upload*

Automatic submission: Your answer 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. This is considered as "did not attend the exam".

Withdrawing from the exam: If you become ill or wish to submit a blank test/withdraw from the exam for another reason, 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.

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

Cheating/Plagiarism: The exam is an individual, independent work. Examination aids are permitted, but it is not permitted to communicate with others about the exam questions or to distribute draft solutions during the exam. Such communication is regarded as cheating. 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**

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.

TDT4310: Exam Questions 2021 [maximum points: 50]

1) There are quite a few differences between human languages (such as English, Norwegian and Tok Pisin) and artificial languages (e.g., Python, ALGOL 58 and First Order Logic).
   Describe five such differences and discuss for each of them why that particular aspect of human language makes a difference (be it negative or positive) when trying to build language processing tools and systems. **[10pt]**

2) Recent years have seen plenty of use of terms such as 'fake news', 'conspiracy theories', 'big lie', 'false flag operations' and even 'InfoWars'. However, there is nothing new about trying to take the initiative and changing the *political narrative* in order to impose one's own version of the truth. History is indeed (mostly) written by the winners, but the ones who manage to write the story often become winners. When Trump refuses to accept losing the 2020 US presidential election, he shows that he understands that what matters is not the actual factual state, but whose view of it gets to dominate the story. On a global scale, actors such as China and the US work hard to ensure that their version of the current story gets to dominate, while often trying to disguise where that version of the story originates from.
   Suppose you want to build a system to track such political narratives, in order to figure out the actual driving force behind a (version of a) story; both who is trying to impose it and why.

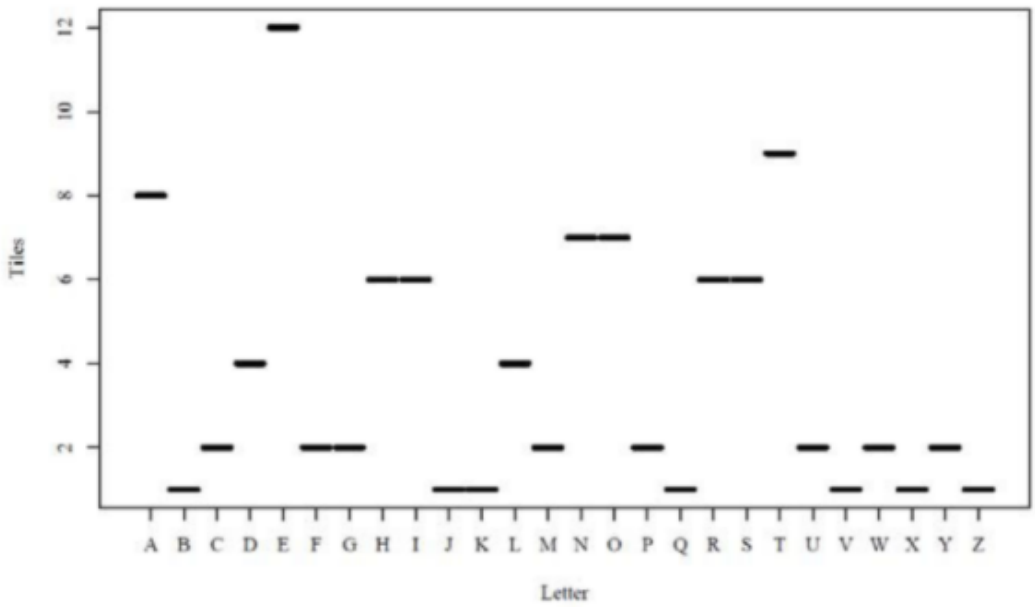
   a) What kind of data would you need? **[5 pt]**

   b) How would you go about processing and representing that data? **[5 pt]**

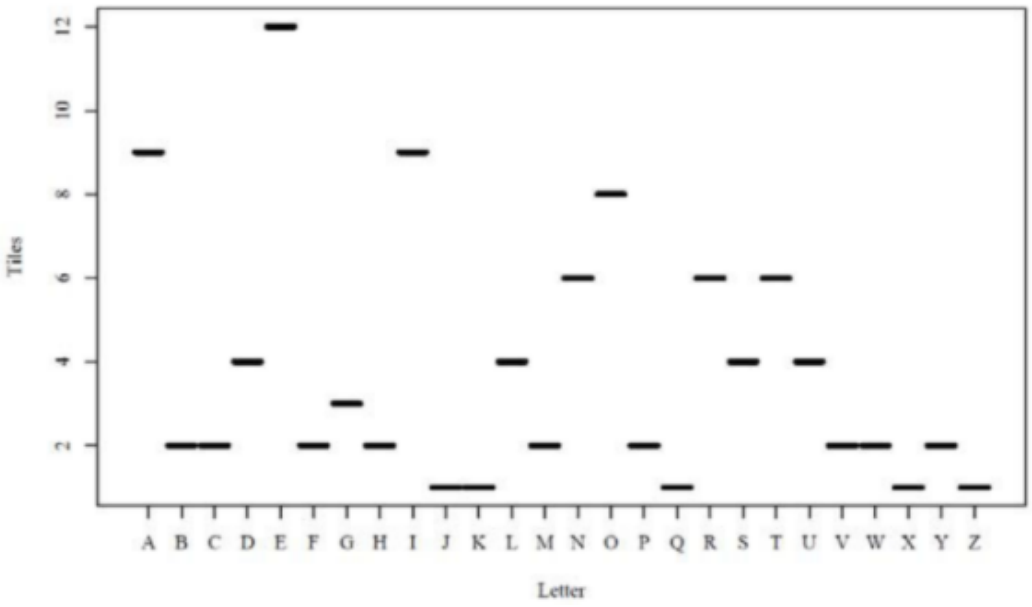
3) Of course it would be very nice to build a general system to track all types of political narratives and all possible sources, in all possible languages, but maybe you want to simplify the problem somewhat and decide to build a classifier which only accounts for three possible sources of a narrative: China, the US, and "someone else", while covering news outlets in China and the US only.
   Describe the key issues you would need to address to create such a system. **[10 pt]**

4) List the (in your view) five most important *currently* used techniques and methods in text analytics and language understanding that were addressed in TDT4310 (that is, either in the lectures, in the student projects, in the lab exercises, or in course books and articles). For each of those techniques motivate in detail why you think it is important enough to be included on your top-5 list. **[10pt]**

5) The gaming company 'The Blizzard King Valve' has come up with a product which will change gaming forever. Instead of using a gaming computer, they want to build their system using a peculiar hardware platform consisting of 100 small wooden tiles that will be put on something they call a "board". Two of those tiles will be empty, but the remaining 98 tiles will have a letter of the English alphabet (A-Z) and a score written on them. The score associated with each letter should be inversely proportional to how common the letter is in the English language.
The company asked you to come up with the scores. Being a bit lazy, you simply took the frontpage of the local newspaper and counted the characters. The distribution you found looked like this:



However, the sneaky company didn't trust your expertise, but hired another consultant who suggested that the distribution rather should be along the following lines:



Upset as you are over this flagrant lack of trust, you retort to the course in text analytics that you once took at NTNU. You remember having had access to the million-word Brown corpus, so you decide to

check the 20 most frequent words in the corpus, which together actually account for about 31% of the total tokens in the corpus. Those words are (in frequency order):

the, of, and, to, a, in, that, is, was, he, for, it, with, as, his, on, be, at, by and *I*.

a) In what way and for which letters would this list help explain the discrepancy between the two suggested distributions above? **[2 pt]**

However, those 20 words are of course not equally frequent, but rather follow Zipf's law, as follows.

	Word	Frequency (%)
1.	the	6.8872
2.	of	3.5839
3.	and	2.8401
4.	to	2.5744
5.	a	2.2996
6.	in	2.1010
7.	that	1.0428
8.	is	0.9943
9.	was	0.9661
10.	he	0.9392
11.	for	0.9340
12.	it	0.8623
13.	with	0.7176
14.	as	0.7137
15.	his	0.6886
16.	on	0.6636
17.	be	0.6276
18.	at	0.5293
19.	by	0.5224
20.	I	0.5099

b) How would that information influence your explanation of the difference? **[2 pt]**

c) You might notice a gender bias among the most frequent terms. The Brown Corpus consists of American English texts published in 1961. How can that explain the gender bias? **[2 pt]**

d) Describe how you could go about debiasing the Brown corpus. **[2 pt]**

e) Suppose you were to collect a social media version of the Brown corpus today, how would you try to ensure that it wasn't gender biased? **[2 pt]**