

Text Analysis and Retrieval – Final Exam (AY 2015/2016)

The exam has **18 questions** for a total of **30 points**. The 16 multi-choice questions carry 1 point each (1/2 point subtracted for incorrect answer), while the two essay questions carry 7 points each. The page limit per essay question is **two A4 pages**. The exam duration is **135 minutes**. You must turn in the exam questions with your solutions.

Part A: Multi-choice questions (16 points)

1. (1 pt) Which of the following is *not* an information extraction task?
(a) named entity recognition and classification (c) parsing
(b) relation extraction (d) coreference resolution
2. (1 pt) We wish to determine the semantic similarity between three pairs of texts: $T_1 =$ “I called to the large dog that sat on the bank.”, $T_2 =$ A huge dog was by the river, and I yelled its name.”, and $T_3 =$ Police were called to the bank after large amount of money had been stolen.”. What are the likely predictions of a tf-idf vector space model on this task?
(a) (T_2, T_3) are more similar than (T_1, T_3)
(b) (T_1, T_2) are less similar than (T_1, T_3)
(c) (T_1, T_2) are more similar than (T_1, T_3)
(d) (T_1, T_2) are equally similar as (T_1, T_3)
3. (1 pt) Let the *capitalized-word* denote any word with the upper-cased first letter. We are extracting named entities from text using the following regular expression: `[capitalized-word]+`. What is the precision of this extraction pattern on the following text: “The 45-year-old has been transferred to [Lausanne university hospital] in [Switzerland], officials there say. [Schumacher] was placed in a medically induced coma after suffering a severe head injury in a skiing accident in the [French Alps].”? Gold annotations are shown in brackets. Assume strict evaluation.
(a) $\frac{3}{4}$ (b) $\frac{1}{2}$ (c) 1 (d) $\frac{3}{5}$
4. (1 pt) The sentence “I love Justin Timberlake the musician but his acting should stop so he can focus on music.” expresses some sentiment. Which approach to sentiment analysis will give us the most precise analysis of this sentence?
(a) sarcasm detection (c) document-level sentiment analysis
(b) aspect-based sentiment analysis (d) dictionary-based sentiment analysis
5. (1 pt) When constructing a sentiment lexicon, both dictionary-based and corpus-based start from an initial set of words, which is iteratively improved. What does the initial set of words contain?
(a) manually selected positive or negative seed words (c) all words from the vocabulary
(b) random words from the vocabulary (d) it is empty
6. (1 pt) Consider the following text snippet: “On Feb 19 in Brussels, David Cameron will prevail with all of his most important demands. The British prime minister will be standing alone at the summit, faced with opposition from his 27 EU counterparts. This summer, Cameron is planning to hold a referendum on Britain’s future in the EU.” How many named entity mentions and how many coreference chains are there in this text?
(a) 9 mentions and 6 chains (c) 3 mentions and 2 chains
(b) 6 mentions and 4 chains (d) 5 mentions and 2 chains

7. (1 pt) What type of question poses a challenge for QA systems, more than the other types of factoid questions?
 (a) *Who* questions (b) *Where* questions (c) *Why* questions (d) *Which* questions
8. (1 pt) What is the correct definition of *coreference*?
 (a) Mentions A and B corefer if they agree in gender and number
 (b) Two entity mentions corefer if they denote the same real-world entity
 (c) Mentions A and B corefer if one is a substring of another
 (d) Mention A corefers with mention B if it is mentioned in text before mention B
9. (1 pt) Which of the following word sequences is a good example of a keyphrase extracted from a document?
 (a) *whole nine yards* (b) *Chappaquiddick incident* (c) *last straw* (d) *the only state*
10. (1 pt) Which of the following is *not* true for aspect-oriented sentiment analysis?
 (a) Aspect-oriented sentiments can be used to improve overall sentiment classification of the review
 (b) Aspect-oriented sentiment analysis uses models that assign sentiment clues to aspects
 (c) Aspect-oriented sentiment analysis does not require an acquisition of a domain-specific sentiment lexicon
 (d) Aspect-oriented sentiment analysis aims to determine the sentiment expressed towards each mentioned aspect of each of the mentioned entities of interest (e.g., the lens of the camera)
11. (1 pt) What is the name of the summary that is generated by c/p sentences from the original text?
 (a) abstractive summary (c) executive summary
 (b) informative summary (d) indicative summary
12. (1 pt) We are using the BIO scheme to annotate Person, Organization, and Location named entities in text. We will then train a sequence labeler to do the NER task. On how many classes will the model effectively be trained?
 (a) 3 (b) 7 (c) 4 (d) 6
13. (1 pt) The objective of the document processing step in QA is to:
 (a) generate the textual gist of the document
 (b) determine the type of the question
 (c) extract the named entities and coreference chains
 (d) obtain a set of paraphrases likely to contain the answer
14. (1 pt) Which of the following word sequences is a good example of a collocation in English?
 (a) *three small advantages* (b) *Syria and Iraq* (c) *heavy rain* (d) *quite radical*
15. (1 pt) A two-step sentence-level sentiment analysis consists of:
 (a) a three-way classification followed by regression
 (b) a four-way classification
 (c) a four-way classification followed by a binary classification
 (d) a binary classification followed by a three-way classification
16. (1 pt) We wish to design a system that enables the user to select a person from a knowledge base and retrieve all texts in which this person is mentioned. For this the system will need to do:
 (a) anaphora resolution (b) entity linking (c) keyword extraction (d) relation extraction

Part B: Essay questions (14 points)

This part of the exam contains a set of real-world TAR usage scenarios. You are asked to describe how you would build a system that addresses these scenarios. When presenting your solution, always give a motivation behind your choices and discuss the benefits of your choices compared to the alternatives. In case your answer includes a machine learning model, make sure to comment on what model and features you would use, as well as how you would tune it. You are also required to describe the inputs and outputs of each subsystem in your pipeline.

17. (7 pts) PR sensations.

As some guys in TakeLab are quite into PR, we decided to show off at a Croatian version of TV show *The Voice*. Besides participating in the show as singers, we have also implemented a system that enriches the viewer experience. It extracts the most commonly used positive and negative phrases that describe each of the participants from Facebook comments. Knowing that this use-case is trivial for an average TAR student, we ask you to explain how would you go about tackling it.

You are given a list of The Voice contestants, and a bunch (over 9,000) of Facebook comments (translated into English for your convenience):

“Đurđa, man, dat gurl, dat astonishing voice. Amazing! :)”
“Even though his performance was only average, Domagoj was one helluva view.”
“he is so energetic and positive. gooooo, Tutek, go! xoxoxo #coffeeisthething”
“Mladen, weird and wonderful!!one Go #TeamŠnajder”
“arrogant, but nails each and every pop song. < 3 Goran!!”
“Ljudmila is a woman of great power #ljudmilarocks”

For sake of simplicity, assume that each tweet concerns only a single contestant, and that all names and surnames are unique.

In at most two pages, describe the design of a system that solves the above task. In your description, be sure to include answers to all of the following questions:

- Would you opt for a supervised or an unsupervised approach? Explain.
- What components (TAR tasks) would constitute your end-to-end system? Explain their roles.
- Your tweets are not labeled with the participant name it concerns. How would you go about this (without resorting to regular expressions or basic string matching)?
- How would you detect positive and negative phrases in a tweet? What approach and features would you use? Be specific.
- How to decide how many phrases should system produce in the end?
- How would you evaluate your system?

18. (7 pts) Too long; didn't read.

Tina is an industrious student studying for the TAR exam. After she had went through all the available lecture slides, she decided to take a look at the reading materials provided by prof. Tailor. Unfortunately, she soon became utterly frustrated, as the reading materials are practically endless. Thinking about the generations to come, Tina decided to make their lives easier. For every learning outcome, she (1) read all the materials and highlighted paragraphs that are concerned with it, and (2) went through all the highlighted paragraphs, writing a summary that covers all the materials about that learning outcome. After devoting her entire summer to this, she managed to complete the first three chapters of TAR (©). Luckily, this effort resulted in a dataset that, for each of the learning outcomes, contains a set of related paragraphs, along with the summary of this set. As a top-tier TAR student, she quickly realized she could make a supervised model that would do all the dirty work for her. Being highly confident in your TAR abilities, she outsources this task to you (and goes to the beach to relax)!

Your task is to build a system that, given a learning outcome and a complete set of reading materials for the entire course (divided into paragraphs), produces an extractive summary (20–30 sentences) of texts that deal with the topic of a given learning outcome. In at most two pages, describe its design. In your description, be sure to include answers to all of the following questions:

- What components (TAR tasks) would constitute your end-to-end system? Explain their roles.
- How would the system identify related paragraphs for a learning outcome?
- How would the system judge which sentences should be included in the summary?
- Can your system produce redundant information? If not, explain, otherwise elaborate on how you would remove the redundancy. Be specific.
- Do you think that using an abstractive summary would benefit your system? What makes you stick with an extractive one in practice? Elaborate.
- How would you evaluate your system's performance?