University of Ljubljana, Faculty of Computer and Information Science

Natural language processing, written exam, 03 June 2022

Each student may have one sheet of notes in A4 format; other literature is not allowed. The use of any electronic device is considered cheating. All five questions count equally. Duration: 90 minutes.

Students who wish to look into the written exam results can do so on Tuesday, 07 June 2022, at 8:30 in the room of Prof Robnik Šikonja.

1. From a given corpus with extracted named entities (NEs), take the list of persons. From them you want to extract the politicians and assign them their political wing: left, right, or center. Table on the right gives an example output.
   From the table, calculate the precision, recall and F_1-score for the right-wing politicians. A correct instance is one in which the *System output* and *Ground truth* agree on a label. The *Ground truth* column only contains entries for politicians.

Persons in NEs	**System output**	**Ground truth**
Barack Obama	center	left
Bill Clinton	center	left
Bugs Bunny	left	
Donald Trump		right
George W. Bush	right	right
Hillary Clinton	right	right
Howard Stern		center
Jason Brown	center	
John Mayor	right	center
Jonathan Swift	left	
Karl Marx	left	
Mitt Romney	right	center
Noam Chomsky	left	left
Pippi Longstocking	center	
Ralph Nader	left	left
Richard Cheney	right	right
Sarah Palin	right	right
Shaquille O'Neal	center	

2. Explain the positional encoding used in transformer models.

3. The English language has a relatively fixed word order, i.e. Subject + Verb + Object (+ Adverb Of Place + Adverb Of Time), e.g., *Milley meets George at the park every day*. The sentences that do not obey this form, look ungrammatical.
   Propose how would you use the BERT model to determine if in a given sentence there is a problem with the word order. Suggest a design of a complete system, describe the inputs, outputs, used dataset and algorithms, and describe the training and evaluation of the model.

4. Describe the three phases of pre-neural summarization pipeline.

5. Explain how to address question answering with large pretrained encoder-decoder language models. Give an example of pretraining and fine-tuning.