# Politecnico di Torino
# Dipartimento di Automatica e Informatica
# Deep NLP (01VIXSM)
# Written Exam

January 27th, 2022

---

**Name**: _____
**Surname**: _____
**Student ID**: _____

---

**Exam rules:**

- The present exam consists of 6 pages (including this cover page) and 7 questions overall. Any inconsistencies/printing errors in the written exam content must be reported to the teacher *at the beginning* of the exam.
- Exam duration: *60 minutes*.
- Withdraw is allowed only *at the end* of the exam.
- The exam is *closed-book*. Electronic devices, mobile phones, smart watches, and extra papers (even blank papers) are *not allowed*.
- Closed-ended questions: cross the right answer (just one) at pag. 2. Wrong or missing answers to closed-ended questions will receive *no penalty*.
- Open questions: write your answers below the text of the question. If you need more space please use the last page (i.e., pag. 6) and/or the back side of the paper.

**Evaluation grid**

| Question: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total |
|---|---|---|---|---|---|---|---|---|
| Points: | 1 | 1 | 1 | 1 | 5 | 5 | 6 | 20 |
| Score: | | | | | | | | |

1. (1 point) Which of the following is a valid definition of n-gram?
   - A set of n sentences that are part of a document.
   - A bag of words describing a document topic.
   - A contiguous sequence of n characters that are part of a word.
   - A set of n phonemes that are part of a phrase.
   - None of the above.
2. (1 point) Which of the following is a word normalization technique in NLP?
   - Cosine distance.
   - Lemmatization.
   - Stopword removal.
   - TF-IDF.
   - None of the above.
3. (1 point) What of the following metrics can be used to evaluate the performance of a summarization system?
   - F1-score.
   - BLEU score.
   - CER: Character Error Rate.
   - Accuracy.
   - None of the above.
4. (1 point) When does a *cold start problem* occur?
   - While looking for the initial learning rate for training a RNN.
   - While recommending item to new users.
   - While retrieving content from recently indexed documents.
   - While starting a new conversation between a human and AI agent.
   - None of the above.

5. (5 points) Explain the **PageRank** algorithm:

   1. Enumerate and explain the main algorithm steps.
   2. Exemplify at least one context of usage in NLP.
   3. Specify the conditions under which PageRank can be applied to a given graph. Explain how to proceed when the conditions are not satisfied.

**Draft solution 5.1:**

- Rule 1. Links from a graph node to itself are ignored.
- Rule 2. Multiple outgoing edges from one node to another are treated as a single edge.
- Rule 3. Set the same initial value for all nodes (e.g., $P(x) = \frac{1}{N}$).
- Apply the Pagerank formula to all vertices at each step $P(x) = \frac{1-\lambda}{N} + \lambda \sum_{y \to x} \frac{P(y)}{out(y)}$ ($\lambda$: damping factor, $N$: number of vertices, $y \to x$: node $x$'s incoming edge).

**Draft solution 5.2:**

It is used in information retrieval to get an estimation of the overall importance of a node in the graph. It can be used as additional relevance score for building search engines.

**Draft solution 5.3:**

The graph must not have sinks (i.e., vertices with only incoming edges). To solve this problem it is possible to add an outgoing edge from the sink vertex to every other node in the graph.

6. (5 points) Elaborate on the **Named Entity Recognition** task.

    1. Formulate the task and clarify the main goals.
    2. Illustrate at least two example rules that can be manually defined to identify people's names in text.
    3. Provide a high-level description of the architecture proposed by **LUKE** (*LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention by I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto*) to address NER.

**Draft solution 6.1:** The NER task consists in identifying portion of text referring to named entities (people, locations, known-objects). The two-fold aims are to detect word n-grams that explicitly refer to particular entities and to classify them as one of the predefined category of named entities.

**Draft solution 6.2:**

- A sequence of words starting with a capital letter, followed by lowercase letters (E.g., Barack Obama)
- Never include special characters: % / $ / =

**Draft solution 6.3:**

Key points:

- The model is derived from pretrained BERT.
- It shares with BERT the architecture.
- It also considers specific token embeddings to model the entities
- The model uses the attention mechanism to attend both word and entity relationships
- During the training phase the entities within the annotated text are randomly masked. The trained model predicts the original masked entities.

7. (6 points) The *Politecnico di Torino* wants to automate the process of teaching content curation. The catalogue of the courses offered by the university consists of 10000 courses, each one described by a 500-word document. Course descriptions can be written either in Italian or in English (a mix of the above is not allowed).

    1. Design a NLP pipeline aimed at summarizing the description of each course in a 100-word abstract to be visualized in the Polito Webpage. At this stage, let us assume that no humanly-generated annotations are provided. Enumerate the main steps, the names of the models and algorithms used, and any further settings/relevant assumptions made in the design process.
    2. Some volunteers provided manually generated summaries of 1000 course descriptions. How can we exploit the annotated data in order to improve the quality of the generated summaries?
    3. What are the procedures and metrics used to evaluate the performance of the summarization pipeline at Step 2?

    **Draft Solution 7:**

    Key points:

    - Language identification.
    - Separately for each document, apply an unsupervised summarization method to select the most relevant sentences (e.g., TextRank) $\rightarrow$ explain the selected method.
    - Exploit the annotated data to finetune a pretrained model (e.g., BERTSUM) $\rightarrow$ motivate the choice.
    - Rouge or BLEU: explain the selected metric.
    - Which Rouge score? Fixed summary length $\rightarrow$ Recall.
    - Describe how to compare the generated summaries with the golden summaries.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.