Master Degree in Computer Engineering

Natural Language Processing
Final Exam

September 18th, 2025

1. **[2 points]** Present Herdan/Heaps law and Zipf/Mandelbrot law, and discuss their relevance in the context of natural language processing.

2. **[6 points]** Consider the following term-context matrix, providing co-occurrence counts for the target words phone, drink, book, computer, and the context words shop, school, pub, job.

	shop	school	pub	job
phone	11	05	02	17
drink	12	02	32	00
book	13	41	04	08
computer	16	17	02	32

   (a) For all entries in the main diagonal, indicate how to compute the positive pointwise mutual information (PPMI). Use fractions and logarithms in your answers **without** computing these operators.

   (b) The rows of the PPMI term-context matrix can be used as sparse word embeddings. Discuss a method to obtain dense word embeddings from the PPMI term-context matrix.

3. **[5 points]** In the context of static word embeddings, answer the following questions.

   (a) Specify the notation and the algorithm called skip-gram with negative sampling (SGNS).

   (b) Introduce and discuss the objective function used by the SGNS algorithm.

4. **[2 points]** In the context of text tokenization, introduce the three approaches based on word, character, and subword tokenization. Compare these three methods, outlining their advantages and shortcomings.

(see next page)

5. **[4 points]** Introduce the neural network architecture called sentence-BERT (SBERT). Discuss the training and the inference phases for this architecture.

6. **[6 points]** In the context of decoder-only architectures for large language models (LLM), answer the following questions.

   (a) Introduce the notion of language modeling head and provide the full stack view of the decoder and the language modeling head.

   (b) Explain the technique of weight tying between the language modeling head and the embedding matrix.

   (c) Discuss the objective function for decoder-only LLM and explain the use of the technique called teacher forcing.

   (d) Introduce the task of text completion, and provide a few examples of how to cast complex NLP tasks as text completion.

7. **[6 points]** In the context of the task of part-of-speech (PoS) tagging, answer the following questions.

   (a) Introduce the probabilistic model called Hidden Markov model (HMM), and the related notions of emission and transition probabilities.

   (b) Introduce the decoding problem (also called inference problem) for HMMs, and discuss the Viterbi algorithm for its solution.

8. **[2 points]** Discuss the life cycle of a ChatBot, as presented in the course lectures.