**Final Exam guide**

The final exam will take place on August 10th at 9:00 am. Please note that the exam location is the International Hall, not the Lee hall. The duration of the exam will be 2 hours, and it will is a closed-book exam. You will not see any numerical questions, hence no computational tools will be necessary.

The exam questions may come from both lecture materials and discussions. It's important to grasp the concepts from all homework assignments except homework 5. Now, let’s review the subjects we’ve covered. Keep in mind that the exam will cover the topics below, but do not expect having the same questions on the actual exam; they are provided to assist you in your preparation.

1. Information retrieval

In this section, we learned how to approach the NLP task conceptually. The core processes involve: identifying information needs, formulating queries, and evaluating outcomes. Imagine a compilation of Shakespeare's plays. Let’s create a hypothetical situation: I want to locate a play where the central characters have connections to regions outside Europe. How should I structure the query? Which search techniques would be suitable?

1. Vector Space Model and text representation

Here, we learned the mathematical representation of the text. Starting from the one-hot vectors, we delved into the significance of word frequency within the text. Despite its infrequent usage in deep learning models, the bag-of-words and TF-IDF approaches remain prevalent in various contexts. It is crucial to understand the assumptions that underlie each of these representations, which are intricately linked to vector-based logic like cosine similarity. You might be asked:

Elaborate on the notion of term frequency-inverse document frequency (TF-IDF) as an expansion upon the bag-of-words concept. Illustrate how TF-IDF enhances both document representation and search ranking.

Discuss the role of stop words in a general document and how can they impact the representation of documents? When do we need (not) to keep them?

Outline the procedure for computing cosine similarity between two documents utilizing their vectorized forms. Why is cosine similarity preferred over other distance metrics?

In the homework, we represented text in different ways. Please check them out and make sure you understand how it was done. We also discussed about the ‘Lazy price’ paper in both class and homework. Reviewing this process based on the paper would be helpful for you.

1. Classification task

Once more, I won’t ask any calculations or precise equations. If you’ve completed the homework correctly, you should comprehend the merits and drawbacks of each model. Typically, your task guides your modeling decisions. Imagine you have access to a dataset of Yelp user reviews. What sort of questions would you wish to address? On the flip side, if your goal is to recognize high-quality restaurants, what data would be essential? What feasible models could you employ? It’s essential to acknowledge the constraints of each model and demonstrate how you address them. Assess your understanding through verbal evaluation using the problem set questions.