***CAREER*: Human-Machine Sensemaking of Text via Dimension Reductions**

Sensemaking for large collections of data is often a balance between automation and human control. Dimension reductions (DR) often serve as a first step in sensemaking to get an overview of the dataset and identify similar items. However, they notoriously suffer from a lack of interpretability and offer limited human control and feedback to adjust based on human knowledge. This proposal aims to enable **human-in-the-loop (H-I-L) DR** methods, specifically for text, and study their impacts on analysis workflows. This five-year project advances H-I-L DR for text by considering three questions:

1. Given recent advances in expalainable DR for tabular data via DR gradients, how can gradients by leveraged for explaianble DRs of text? We aim to design **gradient-based explanations** for DRs of text that capture the effects of low-level text features (e.g. words) on the DR space.

2. Explainable DR's provide a starting point for sensemaking tasks. However, in the case where DR's do not capture meaningful information or reflect the analysts prior knowledge, how can we enable human feedback to teach DRs based on domain knowledge without requiring labeled data or the inspection of individual documents? We seek to design interaction methods that operate on the dual text feature space to enable meaningful feedback in the early stages of sensemaking.

3. Finally, how do H-I-L DR's impact trust and bias in DR results? With explanations and interaction in place, we seek to study the impact of human interaction on trust in DR results and H-I-L DR as a means to both mitigate machine bias and inject human bias in DR results through carefully designed studies.

Intellectual Merit: This project will investigate methods to bring humans into the loop in DRs of text for sensemaking tasks. The plan will design new methods of explainability and interaction in DRs to facilitate bi-directional learning between the human and the machine. Explainability methods will leverage DR gradients to capture the effects of low-level text features on the DR results, enabling explanations situated in the context of document features. Interactions with low-level text features will be enabled in a dual feature space to eliminate the need for individual document inspection and facilitate exploratory interactions without deep prior knowledge of the corpus.

Finally, it plans to investigate the impacts of these methods on analytics, in particular on trust and bias in results generated by interactions. H-I-L analytics has the power to reduce machine bias from skewed training data while at the same time opens the door for human bias from interactions. Studying the impacts of H-I-L methods on bias and trust is critical for enabling responsible and fair analytics.

Broader Impacts: The proposed research has broad applications in a range of fields, such as intelligence analysis. We will create open source toolkits to enable the use of our proposed methods by analysts outside of visualization research. Additionally, we will propose/organize Human-AI workshops at conferences in both human-centered computing venues and data science and AI focused venues, to help bridge the gap between visualization research and data science/AI research, and disseminate our approaches to a broader audience that may need them. Additionally, this proposed plan will support undergraduate researchers on this project and engage with the Tulane-Newcomb Scholars to mentor young women in research activities.