Bachelor thesis, Institute of Computer Science, Freie Universität Berlin

Human-Centered Computing (HCC), AG NBI

Comparing Interpretability Techniques for Unsupervised Topic Modeling

– Exposé –

Tim Korjakow

tim.korjakow@campus.tu-berlin.de

Supervisor: Jesse Josua Benjamin

First examiner: Prof. Claudia Müller-Birn

Second examiner: Prof. Klaus-Robert Müller

Berlin, July 20, 2019

1 Motivation of the thesis

Research on Explainable Artificial Intelligence, often called XAI, is currently wildly distributed and characterized by several competing ideas and approaches from a vast number of fields including computer science, mathematics, social sciences and philosophy. Additionally to that development stands the fact that most of humanity's knowledge is encoded in texts. Therefore there is an urgent need to research these methods in the context of NLP.

Project IKON, which aims to develop a data-driven application for the discovery of knowledge transfer potentials between research projects and society at a major natural history research institution, serves as an use case to analyze and apply these methods. The core of this application is a cluster visualization which links research projects by their semantic similarity and is supplemented by links to infrastructures (e.g., collections or labs) and knowledge transfer activities. This cluster view is driven by a Topic Modelling Pipeline with a Singular Value Decomposition at its heart. Arras et al. showed that this method shares a limited amount of expressiveness with different other linear methods due to its purely linear nature [AHM+17]. Furthermore building the pipeline brought up serious concerns and uncertainty concerning the meaningfulness of approaches such as parameter manipulation or choosing between algorithms for dimensionality reduction [BMG18] in our use case. Interviews with future users of the application were conducted and the results reflected our uncertainty and revealed problems with the interpretation of output produced by the pipeline.

Therefore the integral problem in Project IKON is: Which parts of the pipeline need explainability, what kind of models can be used for these parts and what kind of artifacts can be extracted in order to support the process of identifying potentials for knowledge transfer at the research institution?

2 Related work

With the surge of the application of machine learning (ML) systems in our daily life there is an increasing demand to make operation and results of these systems interpretable for people with different backgrounds (ML experts, non-technical experts etc.). Contrary to these efforts, interpretability as a term is an ill-defined objective [Lip16] for research and development in ML algorithms since there is no widely agreed upon definition of it. This leads to a research landscape where every paper defines interpretability by itself, which, in turn, makes comparisons between techniques complicated.

Miller et al. [MHS17] support this point by conducting a literature study and uncovering that interpretability research is rarely influenced by insights from the humanities, especially connected fields as explainability or causality research.

This thesis builds upon these findings and tries to transfer their critical insights into the sub field of natural language processing - an often overlooked discipline in the context of interpretability research.

3 Goal setting and procedure

As formulated in the introduction, the main problem for project IKON is missing knowledge about the interpretability of the existing model. This thesis should therefore examine existing interpretability techniques, their applicability to the existing model or a contending one and subsequently a decision for one of them by systematically and iteratively sourcing feedback from the principal researchers of project IKON. This goal setting directly leads to the following list of steps:

1. In the beginning a workshop with the researchers from project IKON should result in a list of questions that scientists at the museum try to answer while interacting with the visualization. This is based on previous interviews and workshops which were conducted at the museum itself. Since we as developers did not possess an exhaustive list of techniques to enhance explainability for unsupervised NLP models exist, a thorough and reproducible literature analysis on the status of XAI research in the field of NLP according to Petersen et al. [PFMM08] is going to be conducted. This should result in a number of papers which are, according to the process, good representatives of the literature base and therefore also of current research efforts.

   A quantitative analysis of these papers should summarize occurring XAI methods and categorize them according to the proposed categories of Hohman et. al. [HHC+19]. These categories are not a perfect fit for a thesis dealing with explainability for non-technical experts since it also categorizes techniques according to their mathematical inner workings, but Hohman et al. extended the categories proposed by Lipton [Lip16], which formulated the starting hypothesis for this thesis and is the closest to a nontechnical assessment of interpretability research I could find.

2. The currently existing topic extraction pipeline can be generalized into the following four components: document embedding, dimensionality reduction into a topic space, clustering and another dimensionality reduction into 2D. Based on the results of the previous step for each component either a directly applicable method (e.g. a clustering algorithm) from a paper or a model which supports most collected methods (e.g. a neural network for document embedding) is chosen and implemented. Since the new pipeline should capture atleast as much information as the old one, each component will be quantitatively accessed according to applicable measures e.g. ([RBH15]). This is necessary to ensure that one is actually interpreting existing and captured semantic relations and not random artifacts generated by the various methods.

3. A full user study would normally be necessary to assess how the implemented methods may support a non-technical expert in interpreting the results of the pipeline, but in order to keep the volume of this thesis in a feasible frame I will resort to a cognitive walkthrough from the point of view of a researcher from the

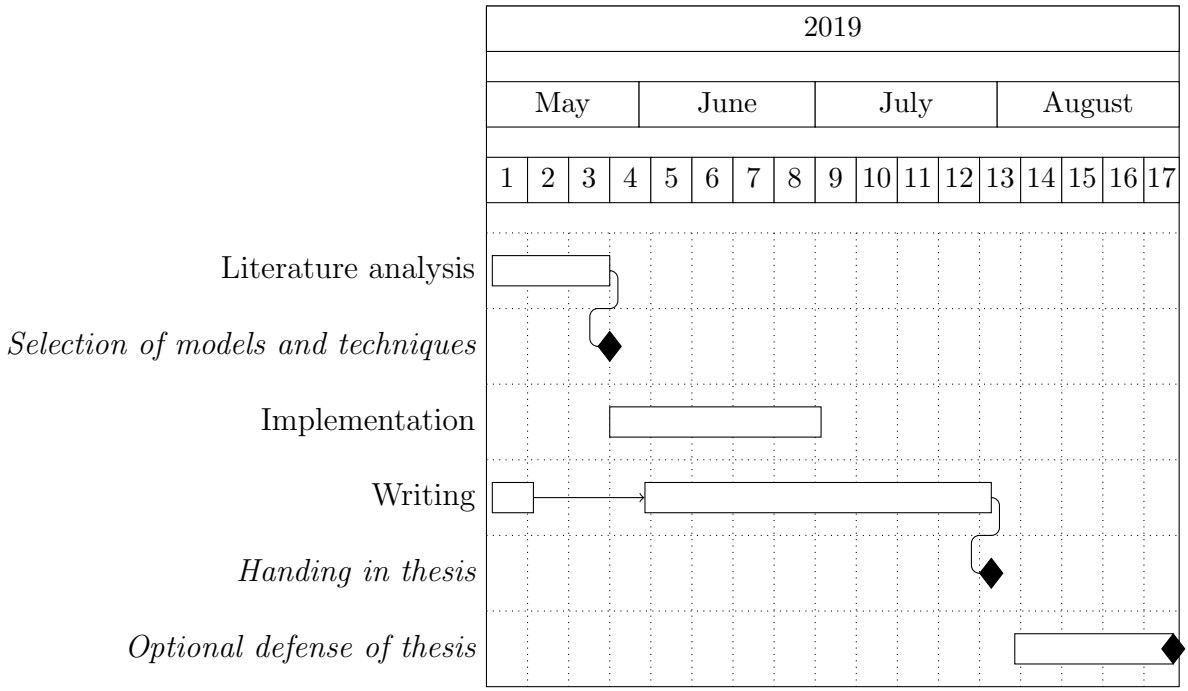
national natural history museum. Since ensuring robustness in such qualitative tests is always a concern, information from previous interviews with domain experts from the museum will be used to derive meaningful tasks. The walkthrough should show how the implemented techniques may help with answering the initially sourced questions from step 1.

4 Implementation

One of the main technical challenges and parts of the implementational work will be the augmentation of the current topic modelling pipeline by a document embedding technique. Since the performance of the model greatly depends on this step, it is crucial to have well learned vector representations of the document base. Currently there is a corpus of circa 114000 scientific documents available in order to train the model. If that is not enough to gain expressive document embeddings, one may include pretrained word embeddings via e.g. BERT [DCLT18] to introduce external information into the model and enhance the semantic coherence of the learned embeddings. This path should be taken with caution since it is connected to an hardly determinable amount of complexity. Research in the field of transfer learning for document embeddings is still in its infancy.

In order to adhere to the current research and industry standards, the implementation of this thesis is going to be done in Python. Specifically, all the data preprocessing will be done with spaCy and all models will be implemented either in Gensim or Keras. Based on the chosen model which is going to be augmented with explainability mechanisms further packages and technologies may be selected.

5 Time calculation



References

[AHM+17] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *PLOS ONE*, 12(8):e0181142, August 2017.

[BMG18] Jesse Benjamin, Claudia Müller-Birn, and Rony Ginosar. Transparency and the Mediation of Meaning in Algorithmic Systems. October 2018.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, October 2018.

[HHC+19] Fred Hohman, Andrew Head, Rich Caruana, Rob DeLine, and Steven Drucker. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. January 2019.

[Lip16] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]*, June 2016.

[MHS17] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]*, December 2017.

[PFMM08] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic Mapping Studies in Software Engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Swindon, UK, 2008. BCS Learning & Development Ltd.

[RBH15] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, pages 399–408, Shanghai, China, 2015. ACM Press.