Literature review

\*\*Vectorization Methods\*\*

The transformation of text data into a format understandable by machine learning algorithms, known as vectorization, is crucial in Natural Language Processing (NLP). Different techniques have been proposed and explored in the literature, including CountVectorizer (Salton & McGill, 1983), Word2Vec (Mikolov et al., 2013), Doc2Vec (Le & Mikolov, 2014), and TF-IDF (Ramos, 2003). CountVectorizer, a frequency counter for text documents, offers a simple way to both tokenize a collection of text documents and build a vocabulary of known words (xxx). Word2Vec and Doc2Vec are prediction-based methods which learn word representations using shallow neural networks, while TF-IDF, or Term Frequency-Inverse Document Frequency, is a statistical method that reflects how important a word is to a document in a collection (Ramos, 2003).

\*\*Recommendation Systems\*\*

Recommendation systems have been extensively researched in the field of information retrieval and personalized user services. The principle behind recommendation systems is to provide users with recommendations based on their preferences and behavior. Cosine similarity and Jaccard similarity are among the common techniques used in recommendation systems (Adomavicius & Tuzhilin, 2005).

\*\*Text Generation Models\*\*

In recent years, text generation models and automatic summarization methods have gained considerable attention in NLP research. TextRank is an extractive summarization technique based on the concept of PageRank, which involves ranking sentences in a document based on their relevance and significance (Mihalcea & Tarau, 2004). Meanwhile, transformer-based models like T5 (Raffel et al., 2019) and Bart (Lewis et al., 2019) have been proposed and achieved state-of-the-art performance on multiple NLP tasks. These models adopt the transformer architecture (Vaswani et al., 2017) which is based on self-attention mechanisms. Despite the rising dominance of transformer-based models, older techniques like TextRank still find utility in specific contexts due to their simplicity and interpretability."

\*\*Topic Modeling\*\*

Topic modeling techniques such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999) have been widely used for discovering abstract topics from document collections. More recently, BERTopic (de Vries et al., 2020), which combines BERT embeddings (Devlin et al., 2018) and c-TF-IDF, has been proposed as an approach to generate more coherent and interpretable topics.

\*\*OpenAI\*\*

The use of language models has been dramatically transformed by the rise of large transformer-based architectures, and APIs have been developed to allow easy access to these models. OpenAI's text-davinci-003 and text-embedding-ada-002 are two such APIs which can be leveraged for a variety of natural language tasks (OpenAI, 2021). The former generates human-like text based on a prompt, while the latter produces embeddings for the given input text, which can be beneficial for downstream tasks like clustering or similarity estimation.

\*\*Clustering Recommendation \*\*

Clustering algorithms, such as KMeans, have been employed in various applications for grouping similar instances together. KMeans works well with high dimensional data, especially when dimension reduction techniques like Principal Component Analysis (PCA) are used to capture the essential structure of the data (Jolliffe & Cadima, 2016). Before applying KMeans, it's often advisable to standardize the data to ensure all features have the same scale (Hastie, Tibshirani & Friedman, 2001). To determine the optimal number of clusters, silhouette analysis can be used, which provides a succinct graphical representation of how well each object lies within its cluster (Rousseeuw, 1987)."

Reference

Blei, D., Ng, A. and Jordan, M. (2003) 'Latent dirichlet allocation', Journal of Machine Learning Research, 3, pp. 993-1022.

Mihalcea, R. and Tarau, P. (2004) 'TextRank: Bringing order into text', in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, pp. 404-411.

Adomavicius, G. and Tuzhilin, A. (2005) 'Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions', IEEE Transactions on Knowledge and Data Engineering, 17(6), pp. 734-749.

de Vries, M., Pryzant, R., Adel, H. and Jurafsky, D. (2020) 'BERTopic: Leveraging BERT and c-TF-IDF to create coherent and interpretable topics', arXiv preprint arXiv:2008.10931.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', arXiv preprint arXiv:1810.04805.

Le, Q. and Mikolov, T. (2014) 'Distributed Representations of Sentences and Documents', in Proceedings of the 31st International Conference on Machine Learning, Beijing, China, pp. 1188-1196.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. (2019) 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', arXiv preprint arXiv:1910.13461.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient Estimation of Word Representations in Vector Space', arXiv preprint arXiv:1301.3781.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J. (2019) 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer', arXiv preprint arXiv:1910.10683.

Ramos, J. (2003) 'Using tf-idf to determine word relevance in document queries', in Proceedings of the First Instructional Conference on Machine Learning, Piscataway, NJ, pp. 133-142.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017) 'Attention is All You Need', in Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, pp. 5998-6008.

OpenAI (2021) 'OpenAI API', OpenAI.

Jolliffe, I.T. & Cadima, J. (2016) 'Principal component analysis: a review and recent developments', Philosophical Transactions of the Royal Society A, 374(2065), 20150202.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) 'The Elements of Statistical Learning', New York: Springer series in statistics.

Rousseeuw, P.J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', Journal of Computational and Applied Mathematics, 20, pp. 53-65.