Reproducing:
Topic Modeling on Podcast Short-Text Metadata

YAHYA JABARY*, TU Wien, Austria
BIBERGER JOACHIM, TU Wien, Austria
REINHARD SELINA, TU Wien, Austria
CHALISSERY JULIA, TU Wien, Austria
AHMED NUSAIBA, TU Wien, Austria

...

Additional Key Words and Phrases: Podcasts, Short-text, Topic modeling, Named entities

ACM Reference Format:
Yahya Jabary, Biberger Joachim, Reinhard Selina, Chalissery Julia, and Ahmed Nusaiba. 2024. Reproducing: Topic Modeling on Podcast Short-Text Metadata. 1, 1 (December 2024), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Named Entity informed Corpus Embedding (NEiCE), is a topic modeling algorithm that uses named entities (NE) to improve the quality of topics extracted from short-text content. The algorithm was introduced in the paper "Topic Modeling on Podcast Short-Text Metadata" by Valero et al. [2] from Deezer Research. It is based on the CluWords [3] algorithm, which clusters words based on their nearest neighbors. The authors claim that NeICE outperforms other topic modeling algorithms of the same class, such as Non-negative matrix factorization (NMF), Short-text topic modeling via non-negative matrix factorization (SeaNMF) [1] and Clustering words (CluWords) [3], on podcast metadata from Deezer, Spotify and iTunes.

We put the claims of the authors, regarding the performance of NEiCE on the 3 datasets, to the test and report our findings in this paper. We were able to successfully reproduce the results of the original paper with a few exceptions, which we will discuss in the following sections.

1 Introduction

References

[1] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 world wide web conference*. 1105–1114.

[2] Francisco B Valero, Marion Baranes, and Elena V Epure. 2022. Topic modeling on podcast short-text metadata. In *European Conference on Information Retrieval*. Springer, 472–486.

Authors' Contact Information: Yahya Jabary, jabaryyahya@gmail.com, TU Wien, Austria; Biberger Joachim, TU Wien, Austria; Reinhard Selina, TU Wien, Austria; Chalissery Julia, TU Wien, Austria; Ahmed Nusaiba, TU Wien, Austria.

Manuscript submitted to ACM

Dataset	Deezer				Spotify				iTunes			
	20	50	100	200	20	50	100	200	20	50	100	200
NEiCE (0.2, 0.3)	50.2	48.9	51.4	48.4	51.7	49.0	45.2	46.5	49.3	43.3	49.5	47.0
NEiCE (0.2, 0.4)	53.1	49.2	50.8	50.6	48.7	48.7	43.5	41.7	47.2	49.5	50.7	51.3
NEiCE (0.3, 0.3)	48.5	52.1	51.5	49.8	52.2	49.0	47.5	47.6	50.3	52.5	49.0	48.2
NEiCE (0.3, 0.4)	53.3	50.9	55.3	51.6	50.1	48.5	51.1	49.8	52.5	49.5	49.2	49.8
NEiCE (0.4, 0.3)	53.2	51.5	52.2	50.0	53.2	49.5	50.5	45.9	52.8	50.1	50.6	51.1
NEiCE (0.4, 0.4)	56.4	52.6	48.1	49.0	51.0	48.2	47.3	47.8	52.4	51.9	49.9	47.4
NEiCE (0.5, 0.3)	52.5	56.3	50.8	55.4	51.3	47.7	45.6	45.4	50.6	46.5	46.7	49.0
NEiCE (0.5, 0.4)	56.3	60.6	54.9	53.3	55.0	49.9	46.7	45.0	50.5	52.0	48.7	46.1

Table 1. NEiCE dataset performance on Deezer, Spotify, and iTunes

[3] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 753–761. https://doi.org/10.1145/3289600.3291032

A System Specifications

All experiments were conducted on a consumer-grade laptop with the following specifications:

```
$ system_profiler SPSoftwareDataType SPHardwareDataType
Software:

    System Software Overview:

      System Version: macOS 14.6.1 (23G93)
      Kernel Version: Darwin 23.6.0
      Boot Volume: Macintosh HD
      Boot Mode: Normal
      Computer Name: Yahya's MacBook Pro
      User Name: Yahya Jabary (sueszli)
      Secure Virtual Memory: Enabled
      System Integrity Protection: Enabled
      Time since boot: 103 days, 2 hours, 59 minutes

Hardware:

    Hardware Overview:

      Model Name: MacBook Pro
      Model Identifier: Mac14,10
      Model Number: <redacted>
      Chip: Apple M2 Pro
      Total Number of Cores: 12 (8 performance and 4 efficiency)
      Memory: 16 GB
      System Firmware Version: 10151.140.19
      OS Loader Version: <redacted>
      Serial Number (system): <redacted>
      Hardware UUID: <redacted>
      Provisioning UDID: <redacted>
      Activation Lock Status: Disabled
```