Introduction to MIND and MIND-small datasets

Overall Introduction

The MIND dataset for news recommendation was collected from anonymized behavior logs of Microsoft News website. We randomly sampled 1 million users who had at least 5 news clicks during 6 weeks from October 12 to November 22, 2019. To protect user privacy, each user is de-linked from the production system when securely hashed into an anonymized ID. We collected the news click behaviors of these users in this period, which are formatted into impression logs. We used the impression logs in the last week for test, and the logs in the fifth week for training. For samples in training set, we used the click behaviors in the first four weeks to construct the news click history for user modeling. Among the training data, we used the samples in the last day of the fifth week as validation set. In addition, we release a small version of MIND (**MIND-small**), by randomly sampling 50,000 users and their behavior logs. Only training and validation sets are contained in the MIND-small dataset.

The datasets are intended for non-commercial research purposes only to promote advancement in the field of artificial intelligence and related areas, and is made available free of charge without extending any license or other intellectual property rights. The dataset is provided “as is” without warranty and usage of the data has risks since we may not own the underlying rights in the documents. We are not be liable for any damages related to use of the dataset. Feedback is voluntarily given and can be used as we see fit. Upon violation of any of these terms, your rights to use the dataset will end automatically. If you have questions about use of the dataset or any research outputs, we encourage you to undertake your own independent legal review. For other questions, please feel free to contact us at mind@microsoft.com.

Dataset Format

Both the training and validation data are a zip-compressed folder, which contains four different files:

File Name	Description
behaviors.tsv	The click histories and impression logs of users
news.tsv	The information of news articles
entity_embedding.vec	The embeddings of entities in news extracted from knowledge graph
relation_embedding.vec	The embeddings of relations between entities extracted from knowledge graph

behaviors.tsv

The behaviors.tsv file contains the impression logs and users' news click hostories. It has 5 columns divided by the tab symbol:

- Impression ID. The ID of an impression.
- User ID. The anonymous ID of a user.
- Time. The impression time with format "MM/DD/YYYY HH:MM:SS AM/PM".
- History. The news click history (ID list of clicked news) of this user before this impression. The clicked news articles are ordered by time.
- Impressions. List of news displayed in this impression and user's click behaviors on them (1 for click and 0 for non-click). The orders of news in a impressions have been shuffled.

An example is shown in the table below:

Column	Content
Impression ID	91
User ID	U397059
Time	11/15/2019 10:22:32 AM
History	N106403 N71977 N97080 N102132 N97212 N121652
Impressions	N129416-0 N26703-1 N120089-1 N53018-0 N89764-0 N91737-0 N29160-0

news.tsv

The docs.tsv file contains the detailed information of news articles involved in the behaviors.tsv file. It has 7 columns, which are divided by the tab symbol:

- News ID
- Category
- SubCategory
- Title
- Abstract
- URL
- Title Entities (entities contained in the title of this news)
- Abstract Entities (entites contained in the abstract of this news)

The full content body of MSN news articles are not made available for download, due to licensing structure. However, for your convenience, we have provided a utility script to help parse news webpage from the MSN URLs in the dataset. Due to time limitation, some URLs are expired and cannot be accessed successfully. Currently, we are tring our best to solve this problem.

An example is shown in the following table:

Column	Content
News ID	N37378
Category	sports
SubCategory	golf
Title	PGA Tour winners
Abstract	A gallery of recent winners on the PGA Tour.

Column	Content
URL	https://www.msn.com/en-us/sports/golf/pga-tour-winners/ss-AAjnQjj?ocid=chopendata
Title Entities	[{"Label": "PGA Tour", "Type": "O", "WikidataId": "Q910409", "Confidence": 1.0, "OccurrenceOffsets": [0], "SurfaceForms": ["PGA Tour"]}]
Abstract Entites	[{"Label": "PGA Tour", "Type": "O", "WikidataId": "Q910409", "Confidence": 1.0, "OccurrenceOffsets": [35], "SurfaceForms": ["PGA Tour"]}]

The descriptions of the dictionary keys in the "Entities" column are listed as follows:

Keys	Description
Label	The entity name in the Wikidata knwoledge graph
Type	The type of this entity in Wikidata
WikidataId	The entity ID in Wikidata
Confidence	The confidence of entity linking
OccurrenceOffsets	The character-level entity offset in the text of title or abstract
SurfaceForms	The raw entity names in the original text

entity_embedding.vec & relation_embedding.vec

The entity_embedding.vec and relation_embedding.vec files contain the 100-dimensional embeddings of the entities and relations learned from the subgraph (from WikiData knowledge graph) by TransE method. In both files, the first column is the ID of entity/relation, and the other columns are the embedding vector values. We hope this data can facilitate the research of knowledge-aware news recommendation. An example is shown as follows:

ID	Embedding Values
Q42306013	0.014516 -0.106958 0.024590 ... -0.080382

Due to some reasons in learning embedding from the subgraph, a few entities may not have embeddings in the entity_embedding.vec file.