Dataset Descriptions

1 General Data Description

The MIND dataset for news recommendation was collected from the real user behavior logs of Microsoft News. We randomly sampled 1 million users who had at least 5 news click records during 6 weeks from October 12 to November 22, 2019. In order to protect user privacy, each user is de-linked from the production system when securely hashed into an anonymized ID.

2 Training and Development Data

The data for training and development is a zip-compressed folder, which contains three different files:

- train.tsv
- docs.tsv
- valid.tsv

2.1 train.tsv

The train.tsv file records the user news impression logs. It contains 3 columns divided by tab ('\t'):

- Userid.
- History. This column contains the historical news click behaivors. The data format is "[news id 1]#TAB#[time1]#N#[news id 2]#TAB#[time2] ...". You can use these histories to predict whether the news samples in the third column are clicked.
- Impressions. This column is organized in the format below: "[positive news ids]#TAB#[negative news ids]#TAB#[impression time]". For positive and negative news, their gold labels are 1 and 0, respectively. Different news ids are separated by whitespaces.

2.2 valid.tsv

The valid.tsv file contains 3 columns, which are divided by tab:

- Userid.
- History (This column has the same format with that in the train.tsv file)
- Impressions. This column is structured by a format like "[news ids]#TAB#[impression time]". You need to predict the click scores (i.e., 1 for clicked news and 0 for non-clicked news) of these news articles.

2.3 docs.tsv

The docs.tsv file contains 7 columns, which are divided by tab:

- Newsid
- Vertical (the vertical category of a news article)
- Subvertical (a finer-grained subvertical category of a news article)
- Title
- Abstract
- URL
- Body

3 Test Data

Test data will be released soon. After it is released, participants can make predictions on the test set and submit their results. Note that only the last submission of each team will be regarded as the official submission.