数据预处理流程与输出格式说明

本说明总结了 `preprocess.ipynb` 脚本的数据处理方法，并演示了各类输出数据的格式。

---

1. 数据处理方法总结

1.1 加载与初始化

- 导入文本处理、数据分析、NLP、文件操作等库。
- 设置数据文件路径和全局参数（如最大长度、词频阈值等）。
- 下载并加载英文停用词。

1.2 文本分词与词典构建

- 使用自定义 `word_tokenize` 函数对文本进行分词。
- `read_news` 函数读取新闻数据，对标题和正文分词，统计词频，过滤低频词，生成：
  - 新闻内容字典 `news`
  - 新闻索引 `news_index`
  - 类别字典 `category_dict`
  - 词典 `word_dict`
- 保存上述字典到本地文件。

1.3 用户编码器输入生成

- `get_rep_for_userencoder` 函数将每条新闻的类别、标题、正文转为索引序列（定长，超长截断，不足补零）。
- 输出并保存：
  - 新闻类别数组 `news_vert`
  - 新闻标题数组 `news_title`
  - 新闻正文数组 `news_body`

1.4 Seq2Seq模型输入输出生成

- `get_rep_for_seq2seq` 函数将正文和标题分别转为定长索引序列，生成：
  - 编码器输入 `sources`
  - 解码器输入 `target_inputs`（以 `<sos>` 开头）
  - 解码器输出 `target_outputs`（以 `<eos>` 结尾）
- 保存上述数组。

1.5 词向量矩阵生成

- `load_matrix` 函数加载预训练的 GloVe 词向量，为词典中的词生成嵌入矩阵 `embedding_matrix`，未命中词用正态分布初始化。
- 保存嵌入矩阵。

1.6 用户日志样本处理

- 通过 `parse_train_user`、`parse_valid_user`、`parse_test_user` 分别处理训练、验证、测试用户日志，生成：
  - 用户点击历史（定长索引序列）
  - 正负样本对（训练/验证），正样本及重写标题（测试）
- 保存所有用户和样本数据。

---

2. 输出数据格式举例

2.1 新闻内容字典（news）

```
{
  'N12345': ['Sports', ['man', 'wins', 'race'], ['the', 'man', 'won', 'the',
'race', '.']],
  ...
}
```

2.2 新闻索引（news_index）

```
{
  'N12345': 1,
  'N12346': 2,
  ...
}
```

2.3 类别字典（category_dict）

```
{
  'Sports': 1,
  'Politics': 2,
  ...
}
```

2.4 词典（word_dict）

```
{
  'unk': 0,
  '<sos>': 1,
  '<eos>': 2,
  'man': 3,
  'wins': 4,
  ...
}
```

2.5 用户编码器输入（news_vert, news_title, news_body）

```
news_vert.shape  # (新闻数+1, )
news_title.shape # (新闻数+1, MAX_TITLE_LEN)
news_body.shape  # (新闻数+1, MAX_BODY_LEN)
# 例如 news_title[1] = [3, 4, 5, 0, 0, ...]  # 3,4,5为词索引，后面补零
```

2.6 Seq2Seq输入输出（sources, target_inputs, target_outputs）

```
sources.shape        # (新闻数+1, MAX_CONTENT_LEN)
target_inputs.shape  # (新闻数+1, MAX_TITLE_LEN)
target_outputs.shape # (新闻数+1, MAX_TITLE_LEN)
# 例如 target_inputs[1] = [1, 3, 4, 0, ...]  # 1为<sos>，后面是词索引
#      target_outputs[1] = [3, 4, 5, 2, 0, ...]  # 2为<eos>
```

2.7 用户日志样本（以训练集为例）

```
TrainUsers[0]  # [0, 0, 0, ..., 12, 34, 56]  # 用户点击历史，长度MAX_CLICK_LEN
TrainSamples[0]  # [userindex, [pos_id, neg_id], [1, 0]]
# 例如 [5, [123, 456], [1, 0]]  # 第5个用户，正样本新闻123，负样本456
```

2.8 测试样本

```
TestSamples[0]  # [userindex, pos_id, rewrite_title]
# 例如 [5, 123, 'the man wins the race']
```

如需查看某个具体变量的内容或格式，可以在notebook中直接 `print()` 或显示变量。