基于TF-IDF,DKN,CF混合优化的新闻推荐模型

姓名：肖文璇 学号：2018202163

姓名：雷雨寒 学号：2018202173

1 TF-IDF算法

1.1 基本原理与思路

基于内容相似度的推荐算法

基于内容，即把与**用户喜欢看**的新闻**内容相似**新闻推荐给用户。基于内容的推荐算法的主要优势在于无冷启动问题，只要用户产生了初始的历史数据，就可以开始进行推荐的计算。而且随着用户的浏览记录数据的增加，这种推荐一般也会越来越准确。

利用TF-IDF判断新闻相似性

找到两个新闻的共同关键词，分别计算该关键词在两篇文章的TF-IDF值（即该关键词在两篇文章的重要程度），然后将所得的两个TF-IDF值相乘，该乘积越高，则说明这个关键词在两篇文章中都很重要。最后求得所有共同关键词TF-IDF乘积的和，即得两篇文章的相似性。

公式如下：

$$Similarity(A,B)=\sum_{i\in m}TFIDF_{A_i}*TFIDF_{B_i}$$

其中m是两篇文章重合关键词的集合。

根据用户浏览历史构建用户喜好库

根据用户的浏览历史，我们可以得到一系列用户观看过的新闻，根据新闻的标签、标题、内容，我们可以提取出每个新闻的关键词，这些关键词构成了该用户的新闻喜好库，如果用户浏览历史中的关键词重复出现的次数越多，则该关键词的权重越高。

在对用户进行新闻推荐的时候，我们在总新闻库中计算每个新闻与用户的新闻喜好库的共同关键词的匹配程度，即

$$Similarity(news,user)=\sum_{i\in m}W_i*TFIDF_{user_i}*TFIDF_{news_i}$$

其中m是该新闻与用户喜好库中重合关键词的集合，W_i代表关键词i在用户喜好库中的权重。

算法基本框架

1. 对新闻数据进行清理
2. 利用`fit`函数，将所有新闻用TF-IDF矩阵的形式表示
3. 对新闻库中的每篇新闻，利用`recommender_top_k_items`在新闻库中找到与每篇最相似的前5篇新闻，以及它们的相似度分数。

4. 对于用户阅读的每篇历史新闻，提取它们在新闻库中最相似的前5篇新闻和分数，重复出现的新闻相似度分数会叠加。
5. 将这些得到的相似新闻根据重新计算的相似度分数排序，得到一个基于用户偏好的新闻推荐。

1.2 数据获取与处理

MIND数据集

新闻推荐的MIND数据集是从Microsoft新闻网站中收集的数据，我们从中使用了用户的行为数据和新闻数据。其中用户的行为数据 behaviors.tsv包含浏览日志和用户的新闻点击记录：Impression ID、User ID、Time、History、Impressions。新闻数据 news.tsv包含详细的新闻信息：News ID、Category、SubCategory、Title、Abstract、URL。

数据清理

利用`remove_duplicates`与`remove_nan`去除重复行与带有空值的行。

1.3 模型训练与优化

利用`TfidfRecommender`推荐器（微软实现的一个基于sklearn中TfidfVectorizer构建的推荐器）它用于文章推荐，即根据一篇文章，搜索与它最相似的文章，我们利用到该推荐器中的一些函数实现基于内容的新闻推荐。

其中的函数包括

tokenize_text：利用bert进行英文分词，并初始化文本向量化的工具，设置屏蔽一些停用词（TfidfVectorizer自带）。

fit：利用`tokenize_text`初始化的文本向量化工具，将文本转换成tf-idf矩阵，便于后续的分类检索。

recommend_top_k_items：根据前面得到的文本的tf-idf矩阵，计算文本库中所有文本与该文本的相似性，最后得到某个文本的前k个文本推荐。

将新闻内容导入`TfidfRecommender`推荐器进行训练：

```
def user_recommendation(recommender,news,reader):
    #data preparation
    news=clean_dataframe(news)
    news=news.reset_index()[0:1000]

    #fit in tfidf recommender
    tf,vectors_tokenized=recommender.tokenize_text(news,text_col='Abstract')
    recommender.fit(tf,vectors_tokenized)
    topk=recommender.recommend_top_k_items(news,k=5)
    cols_to_keep=['News ID','Title','Abstract']

    #generate result
    rec=pd.DataFrame()
    like1=reader1.like
    for i in range(len(like1)):
        rec=rec.append(recommender.get_top_k_recommendations(news,like1[i],cols_to_keep))
    rec=rec.reset_index()

    drop=[]
    for i in range(len(rec)):
        for j in range(i+1,len(rec)):
            if rec['News ID'][i]==rec['News ID'][j]:
                rec['similarity_score'][i]=rec['similarity_score'][j]+rec['similarity_score'][i]
                drop.append(j)
    rec=rec.drop(drop)

    rec=rec.sort_values(by=['similarity_score'],ascending=False)
    rec['rank']=range(len(rec))
    rec=rec.reset_index()
    rec=rec.drop(columns='index')
    rec=rec.drop(columns='level_0')
```

进一步优化模型：

为了优化我们的推荐结果，在计算TF-IDF值的时候引入权重，降低在同一大类而不在同一小类的新闻所得的TF-IDF权值大小，而在不在同一大类的新闻降低更多的TF-IDF权值，即：

$$Similarity(news,user)=W_{kind}*\sum_{i\in m}w_i*TFIDF_{newsi}*TFIDF_{useri}$$

其中m是新闻与用户喜好库重合关键词的集合，W_{kind}与用户的喜欢类别与该新闻的类别有关，w_i代表关键词i在用户喜好库中的权重。

具体的实现思路是根据用户喜欢的文章，获取用户喜欢的类别，构建用户喜好类型字典，字典的结构为`{Category:[SubCategory,],}`。

在计算新闻与用户喜好库的TFIDF值时，如果这篇新闻的大类别不在用户的喜好类型字典中，则降低它的权重，因为有可能两者的关键词比较相近但是不在用户喜欢的类别之中；如果新闻的大类别在用户的喜好类型字典，但小类别不在用户的喜好类型库中，则小幅降低它的权重；如果新闻的大小类别在用户的喜好类型字典，则增加它的权重。

构建用户喜欢类型字典并生成推荐字典：

```
# 根据用户喜欢的文章，获取用户喜欢的类别，Category结构为{Category:[SubCategory,],}
def Get_like_Category(like,news):
    Category = {}
    # 遍历喜欢的文章
    for like_item in like:
        # 找到喜欢的类别
        key = news[news['News ID'] == like_item]['Category'].values[0]
        # 将小类存放到大类对应的列表中
        if key not in Category.keys():
            Category[key] = []
```

```
            Category[key].append(news[news['News ID'] == like_item]
['SubCategory'].values[0])

    return Category

# 根据用户喜欢的新闻类别，重新生成推荐结果，news为原始的新闻数据，rec为tfidf生成的推荐结果
def re_rec(Category, news, rec):
    # 遍历推荐的文章
    for i in range(len(rec)):
        rec_item = rec.iloc[i]
        news_item = news[news['News ID'] == rec_item['News ID']]
        if news_item['Category'].values[0] not in Category.keys():
            #如果不在喜欢的大类里面，相似值权重降低为0.8
            rec.loc[i, 'similarity_score'] = 0.8*rec_item['similarity_score']
        else:
            # 符合喜欢的大类，不符合喜欢的小类，权重值降低为0.9
            if news_item['SubCategory'].values[0] not in
Category[news_item['Category'].values[0]]:
                rec.loc[i, 'similarity_score'] = 0.9 *
rec_item['similarity_score']
            # 大类小类都符合，权重值提高为1.25
            else:
                rec.loc[i, 'similarity_score'] = 1.25 *
rec_item['similarity_score']
    # 根据新的相似值重新排序
    re_rec = rec.sort_values(by=['similarity_score'], ascending=False)
    print(re_rec[0:10])
    return re_rec
```

1.4 应用实例

模拟一个读者，他的浏览历史都是与皇室相关的新闻：

	rank	similarity_score	Title	Abstract	Category	SubCategory	URL
0	1	0.220413	Prince George's Royal Life in Photos	Photos of the future king of England who is third in line for the throne, behind his father and Prince Charles.	lifestyle	lifestyleroyals	https://assets.msn.com/labs/mind/AABDkOp.html
1	2	0.199708	Queen Elizabeth's Cousin Says Royal Family 'Don't Communicate Very Well'	Queen Elizabeth's Cousin: Royal Family 'Don't Communicate Very Well'	lifestyle	lifestyleroyals	https://assets.msn.com/labs/mind/AAGCsB5.html
2	3	0.161782	It's Not All About the Corgis - Here Are the Royal Family's Other Beloved Pets	We all know how much Queen Elizabeth loved her Corgis.	lifestyle	lifestyleroyals	https://assets.msn.com/labs/mind/AADmbCD.html
3	4	0.158303	Cutest photos of the royal Cambridge kids	See the best photos of Prince George, Princess Charlotte and Prince Louis!	lifestyle	lifestyleroyals	https://assets.msn.com/labs/mind/AAAM6VA.html
4	5	0.157612	What Do Prince George & Princess Charlotte Know About Their Royal Roles?	Do Prince William and Kate Middleton's kids know about their royal roles?	lifestyle	lifestyleroyals	https://assets.msn.com/labs/mind/AAGeZKp.html

输出的新闻推荐结果如下图所示：

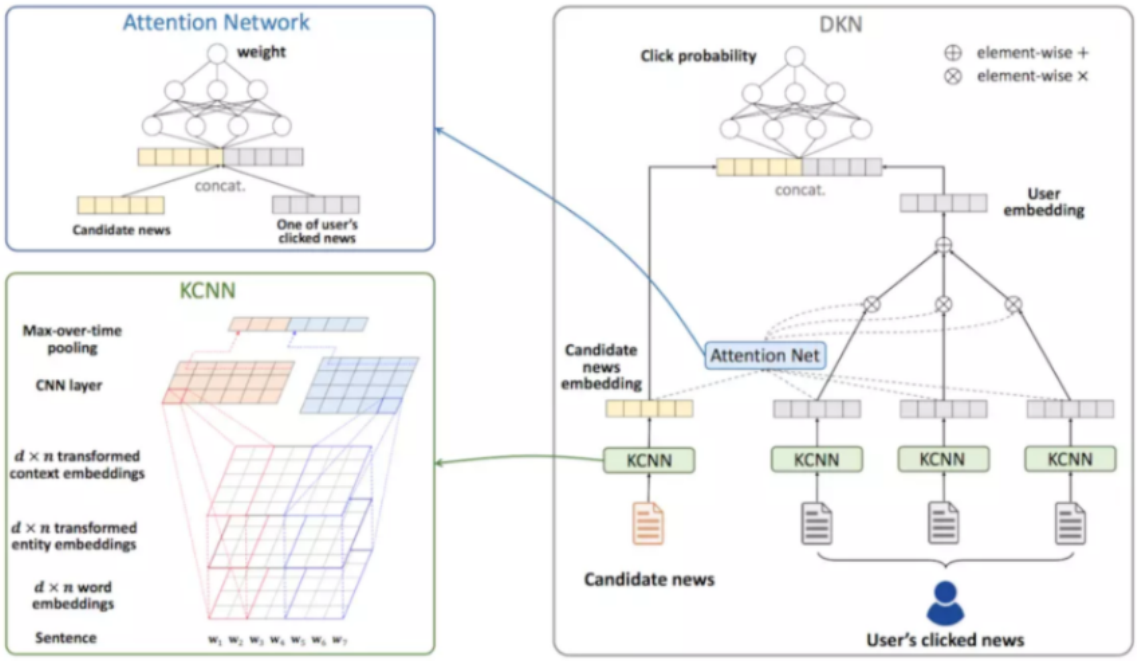
	rank	similarity_score	News ID	Title	Abstract
0	0	0.442929	N23937	6 gorgeous royal family heirlooms that Kate Mi...	Kate Middleton wears priceless heirlooms inclu...
1	1	0.372986	N60434	It's Not All About the Corgis - Here Are the R...	We all know how much Queen Elizabeth loved her...
2	2	0.372986	N51340	Queen Elizabeth's Cousin Says Royal Family 'Do...	Queen Elizabeth's Cousin: Royal Family 'Don't ...
3	3	0.270929	N1810	Where Do the Royals Reside? This Handy Guide W...	Even serious British royal family buffs might ...
4	4	0.239435	N34169	When royals lose their tempers, from the Queen...	See all the times the royal family have lost t...
5	5	0.153879	N40494	Cutest photos of the royal Cambridge kids	See the best photos of Prince George, Princess...
6	6	0.150135	N51736	2020 Ford Mustang Shelby GT500: 8 More GT500 E...	The more you know...
7	7	0.146020	N50187	Grover Norquist: Elizabeth Warren wants to rai...	Elizabeth Warren fashions herself a serious po...
8	8	0.129699	N8071	25 Photos of the Royal Family at Balmoral Cast...	The royal family has been visiting the Scottis...
9	9	0.093596	N14219	The rewriting of Democratic presidential campa...	Moderate Democrats are worried about one thing...

可以看到新闻推荐结果也都与皇室相关。

2 DKN算法

2.1 基本原理与思路

给定一个用户$user_i$，他的点击历史$\{t_1, t_2, \dots t_n\}$是该用户过去一段时间内曾点击过的新闻标题，n代表用户点击过新闻的总数，DKN要解决的问题就是给定用户的点击历史，以及标题单词和知识图谱中实体的关联，预测一个用户i是否会点击一个特定新闻t_j。



DKN的网络输入有两个：候选新闻集合，用户点击过的新闻标题序列。输入数据通过KCNN来提取特征，之上是一个attention层，计算候选新闻向量与用户点击历史向量之间的attention权重，在顶层拼接两部分向量之后，用DNN计算用户点击此新闻的概率。

新闻特征提取 KCNN

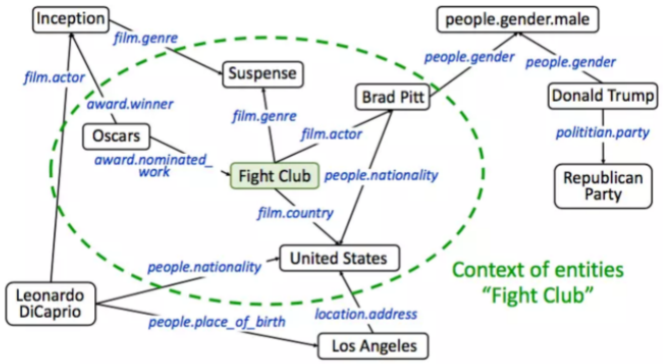
特征提取需要获得三个嵌入，标题中每个单词的Embedding、获取标题中每个单词对应的实体的Embedding 、得到每个单词的上下文Embedding。

1. 标题单词的Embedding可以通过预训练的word2vec模型得到。

2. 标题单词对应实体的Embedding，通过下面四个步骤得到：
    - 识别出标题中的实体并利用实体链接技术消除歧义
    - 根据已有知识图谱，得到与标题中涉及的实体链接在一个step之内的所有实体所形成的子图。
    - 构建好知识子图以后，利用基于距离的翻译模型得到子图中每个实体embedding。
    - 得到标题中每个单词对应的实体embedding。
3. 上下文Embedding

    为了更好地利用一个实体在原知识图谱的位置信息，可以利用一个实体的上下文来进一步的刻画每个实体，即用每个实体相连的实体embedding的平均值来进一步刻画每个实体，计算公式如下：

$$\bar{e} = \frac{1}{|context(e)|} \sum_{e_i \in conext(e)} e_i$$

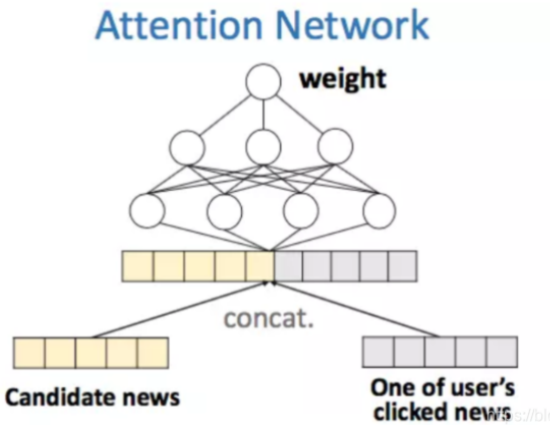


使用mutlti-channel和word-entity-aligned KCNN进行embedding的拼接，先把实体的embedding和实体上下文embedding映射到一个空间里，映射的方式可以选择线性方式$g(e) = Me$，也可以选择非线性方式$g(e) = tanh(Me + b)$，这样我们就可以拼接三部分作为KCNN的输入：

$$w = [[w_1 g(e_1) g((\bar{e_1}))][w_2 g(e_2) g((\bar{e_2}))] \dots [w_n g(e_n) g((\bar{e_n}))]] \in R^{d*n*3}$$

基于注意力机制的用户兴趣预测

获取到用户点击过的每篇新闻的向量表示以后，计算候选文档对于用户每篇点击文档的attention，再做加权求和，计算attention：



2.2 模型准备与训练

环境配置与数据处理

dkn需要安装tensorflow，本次实验使用的是`tensorflow1.15.0-CPU`环境。从Mind数据集中的`behavior.tsv`中得到用户历史文件`user_history_file`。从新闻标题获得标题单词的embedding`word_embeddings_5w_100.npy`与对应实体的embedding`entity_embeddings_5w_100.npy`，并从新闻文档信息中得到新闻的特征`doc_feature.txt`。

模型训练与预测

利用我们设定的参数`hparams`初始化DKN模型，然后使用我们的训练数据`train_file`与验证数据`valid_file`对模型进行反复训练

```
model = DKN(hparams, DKNTextIterator)
model.fit(train_file, valid_file)
```

训练的运行过程如下图所示：

```
at epoch 1
train info: logloss loss:0.6707635026867107
eval info: auc:0.5601, group_auc:0.5114, mean_mrr:0.172, ndcg@10:0.2394, ndcg@5:0.1708
at epoch 1 , train time: 76.3 eval time: 11.4
at epoch 2
train info: logloss loss:0.6216596258898913
eval info: auc:0.5653, group_auc:0.5307, mean_mrr:0.1798, ndcg@10:0.2427, ndcg@5:0.1852
at epoch 2 , train time: 74.7 eval time: 11.6
at epoch 3
train info: logloss loss:0.5895718016614349
eval info: auc:0.5854, group_auc:0.5445, mean_mrr:0.1818, ndcg@10:0.2448, ndcg@5:0.1909
at epoch 3 , train time: 75.0 eval time: 11.5
```

使用`predict`函数即可对用户与新闻的匹配程度打分，分数越高则代表用户可能对这则新闻越感兴趣。

```
model.predict(test_file,output_file)
```

2.3 应用实例

我们设定一个用户历史阅读的新闻内容为关于皇室的新闻，部分新闻描述如下：

	News ID	Category	SubCategory	Title	Abstract	URL	Title Entities	Abstract Entities
0	N3112	lifestyle	lifestyleroyals	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Philip, Duke of Edinburgh",...	[]
17	N25192	lifestyle	lifestyleroyals	Every outfit Duchess Kate has worn in 2019	See Kate Middleton's style choices this year f...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[{"Label": "Catherine, Duchess of Cambridge", ...
66	N1172	lifestyle	lifestyleroyals	The surprising age differences between your fa...	Here are the age differences between Meghan Ma...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[{"Label": "Meghan, Duchess of Sussex", "Type"...
197	N16959	lifestyle	lifestyleroyals	Queen Elizabeth's Favorite Beauty Products Hav...	Here, all the brands the British monarch swear...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[]

DKN模型的结果如下：

	News ID	Category	SubCategory	Title	Abstract	URL	Title Entities	Abstract Entities
26414	N12985	music	music-celebrity	Broadway Star Laurel Griggs Suffered Asthma At...	Teen star Laurel Griggs, who passed away on No...	https://www.msn.com/en-us/music/music-celebrit...	[{"Label": "Broadway theatre", "Type": "F", "W...	[{"Label": "Once (musical)", "Type": "W", "Wik...
17978	N5323	lifestyle	lifestyleroyals	Prince Harry Talked to Another Royal About His...	Prince Albert of Monaco shares the personal ad...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Harry, Duke of Sussex", "Ty...	[{"Label": "Albert II, Prince of Monaco", "Typ...
20019	N13667	lifestyle	lifestyleroyals	Prince Harry and Meghan Markle just shared a n...	The Duke and Duchess of Sussex shared a new ph...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Charles, Prince of Wales", "Type":...	[{"Label": "Charles, Prince of Wales", "Type":...
19953	N11302	tv	tv-celebrity	Counting On's Josiah, Lauren Welcome 1st Child...	Counting On's Josiah, Lauren Welcome 1st Child...	https://www.msn.com/en-us/tv/tv-celebrity/coun...	[{"Label": "19 Kids and Counting", "Type": "W"...	[{"Label": "19 Kids and Counting", "Type": "W"...
306	N10331	lifestyle	lifestyleroyals	Meghan and Harry to take 'family time' off, sa...	The Duke and Duchess of Sussex will take a bre...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Harry, Duke of Sussex", "Ty...	[{"Label": "Duke of Sussex", "Type": "U", "Wik...
3103	N16759	lifestyle	lifestyleroyals	How Kate Middleton and Prince William's royal ...	Harry and Meghan had a much more relaxed appro...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince William, Duke of Cambridge"...	[{"Label": "Prince William, Duke of Cambridge"...
26685	N579	lifestyle	lifestyleroyals	Why Kate & Meghan Were on Different Balconies ...	There's no scandal here. It's all about the or...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Meghan, Duchess of Sussex", "Type"...	[]
24837	N19695	lifestyle	lifestyleroyals	Why Prince Harry Wore His Remembrance Poppy Di...	Why Prince Harry's Poppy Was Worn on His Cap	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Harry, Duke of Sussex", "Ty...	[{"Label": "Prince Harry, Duke of Sussex", "Ty...
22434	N22351	lifestyle	lifestyleroyals	Meghan Markle and Prince Harry Won't Spend Chr...	They'll hang out with baby Archie and Meghan's...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Meghan, Duchess of Sussex", "Type"...	[{"Label": "Meghan, Duchess of Sussex", "Type"...

前10条新闻推荐结果大多数是关于皇室的新闻。

3 CF算法

3.1 基本原理与思路

User similarity-based CF

给定一个“活跃用户”Alice和其并未见过的新闻i，找到一个用户集合（最近邻），集合中的用户之前与Alice喜欢同样的一些新闻，并且已经对新闻i打分，预测Alice对该新闻的打分，比如这些用户对新闻i的打分的均值，对Alice 未见过的所有新闻重复以上过程，推荐其中得分最高的新闻。

- 给定一个用户如Alice，计算和Alice相似的用户，得到一个近邻集合N
- 给定一个新的新闻，根据N中用户的打分来预测Alice的打分，如求平均

皮尔逊相关系数

在基于用户最近邻的协同过滤中广泛使用的相似性度量：皮尔逊相关系数

$$sim(a,b)=\frac{\sum_{p\in P}(r_{a,p}-\overline{r_a})(r_{b,p}-\overline{r_b})}{\sqrt{\sum_{p\in P}(r_{a,p}-\overline{r_a})^2}\sqrt{\sum_{p\in P}(r_{b,p}-\overline{r_b})^2}}$$

其中，a,b:表示用户，$r_{a,p}$表示用户a对新闻p的评分，P表示被a和b打分的新闻集合，sim介于-1和1之间。

预测函数

$$pred(a,p)=\overline{r_a}+\frac{\sum_{b\in N}sim(a,b)*(r_{b,p}-\overline{r_b})}{\sum_{b\in N}sim(a,b)}$$

计算邻居用户对新闻i的打分是否高于或低于平均值，以与用户a的相似度为权重结合这种评分差异，然后在活跃用户的平均值上加/减邻居的偏见值，以此作为预测值。

为了优化预测结果，使用预测函数时进行了一些改变，所有邻居的评分并非具有同等价值，给相似度高（>0.8）的用户更高的权值，同时设置相似度阈值，只看那些相似度绝对值>0.5的邻居。

基于DKN预测分数的思路

我们采用了DKN模型的预测结果来获得每个用户对新闻的评分，然后基于这个分数计算皮尔逊相关系数，并利用预测函数计算推荐结果。

3.2 算法实现

计算皮尔逊相关系数

根据上述公式计算目标用户与选取的1000个邻居用户的皮尔逊相关系数。

```
def construct_sim(df):
    left = 0
    ra = np.mean(df[0])
    for p in range(0,1000):# 新闻
            left += (df[0][p] - ra)*(df[0][p] - ra)

    left = left ** 0.5

    sim = {}
    for b in range(0,1000): # 用户
        rb = np.mean(df[b])
        up = 0
        right = 0
        for p in range(0,1000):# 新闻
            up += (df[0][p] - ra)*(df[b][p] - rb)
            right += (df[b][p] - rb)*(df[b][p] - rb)

        right = right ** 0.5
        down = left*right
        sim[b] = up / down

     return sim
```

构建预测结果

```
def predict(sim,df)
    pred = {}
    for p in range(0,1000): # 新闻
        up = 0
        down = 0
        pred[p] = ra
        for b in range(0,1000): # 用户
            if (sim[b] > 0.5 and sim[b] < 0.8) or (sim[b] < -0.5 and sim[b] >
-0.8):
                rb = np.mean(df[b])
                up += sim[b] * (df[b][p] - rb)
                down += sim[b]
            if(sim[b] >= 0.8) or (sim[b] <= -0.8):
                rb = np.mean(df[b])
                up += sim[b] * (df[b][p] - rb)*2
                down += sim[b]
        pred[p] = ra + up/down
    return pred
```

3.3 应用实例

同例2.3，设定一个用户历史阅读的新闻内容为关于皇室的新闻，部分新闻描述如下：

	News ID	Category	SubCategory	Title	Abstract	URL	Title Entities	Abstract Entities
0	N3112	lifestyle	lifestyleroyals	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Philip, Duke of Edinburgh",...	[]
17	N25192	lifestyle	lifestyleroyals	Every outfit Duchess Kate has worn in 2019	See Kate Middleton's style choices this year f...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[{"Label": "Catherine, Duchess of Cambridge", ...
66	N1172	lifestyle	lifestyleroyals	The surprising age differences between your fa...	Here are the age differences between Meghan Ma...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[{"Label": "Meghan, Duchess of Sussex", "Type"...
197	N16959	lifestyle	lifestyleroyals	Queen Elizabeth's Favorite Beauty Products Hav...	Here, all the brands the British monarch swear...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[]

CF算法的结果如下：

	NewsID	Category	SubCategory	Title	Abstract	URL	Entity1	Entity2
0	N17473	finance	finance-real-estate	The penthouse of NYC's Woolworth Building just...	The condo comes as a "white-box" unit, meaning...	https://www.msn.com/en-us/finance/finance-real...	[{"Label": "Woolworth Building", "Type": "F", ...	[]
1	N5387	lifestyle	lifestyleroyals	The Most Adorable Photos of Archie Harrison Mo...	The royal baby boy is growing up so fast!	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[]
2	N10887	lifestyle	lifestyleroyals	19 Rarely Seen Photos of Royal Siblings	We've pulled together some rarely seen photos ...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[]
3	N3627	news	newscrime	Motorcyclist dies after crash with car in Quee...	A man who was riding a motorcycle died of his ...	https://www.msn.com/en-us/news/newscrime/motor...	[{"Label": "Queen Creek, Arizona", "Type": "G"...	[{"Label": "Queen Creek, Arizona", "Type": "G"...
4	N13667	lifestyle	lifestyleroyals	Prince Harry and Meghan Markle just shared a n...	The Duke and Duchess of Sussex shared a new ph...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Charles, Prince of Wales", "Type":...	[{"Label": "Charles, Prince of Wales", "Type":...
5	N10063	tv	tv-celebrity	Ellie Kemper announces birth of second child	Ellie Kemper reveals her second child, Matthew...	https://www.msn.com/en-us/tv/tv-celebrity/elli...	[{"Label": "Ellie Kemper", "Type": "P", "Wikid...	[{"Label": "Ellie Kemper", "Type": "P", "Wikid...
6	N19904	lifestyle	lifestyleroyals	Prince Harry and Prince William's Rift Is "One...	Santa is shook.	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Harry, Duke of Sussex", "Ty...	[{"Label": "Santa Claus", "Type": "R", "Wikida...
7	N22351	lifestyle	lifestyleroyals	Meghan Markle and Prince Harry Won't Spend Chr...	They'll hang out with baby Archie and Meghan's...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Meghan, Duchess of Sussex", "Type"...	[{"Label": "Meghan, Duchess of Sussex", "Type"...
8	N9951	lifestyle	lifestyleroyals	Best Looks: Queen Maxima of the Netherlands	We're so very fortunate to live in a world fil...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Queen M\u00e1xima of the Netherlan...	[]
9	N2210	sports	football_ncaa	Tennessee QB Maurer suffered concussion agains...	The freshman QB for the Vols has been replaced...	https://www.msn.com/en-us/sports/football_ncaa...	[]	[{"Label": "Tennessee Volunteers football", "T...

可以看到大部分新闻关于皇室，同时还推荐了金融，犯罪，电视，运动领域的文章，为新闻推荐提供了多样性。

4 混合模型

4.1 基本原理与思路

得到上诉三种不同模型的推荐结果之后，我们要做的就是将他们所给出的推荐结果混合到一起，我们先将每种推荐算法结果的分值都归一化处理，再将每种模型推荐结果的权重分别设置为DKN-0.4，CF-0.5，TFIDF-0.1，合并时，同时处于多个推荐模型结果中的新闻的分值将叠加到一起,相反，如果一条新闻不在其他模型的推荐结果中，那么他在对应推荐模型下的得分直接视为0。

公式如下:

$$score_f = score_d/3 + score_c/3 + score_t/3$$

4.2 算法实现

```
ns={}
for i in range(20):#默认对每种模型的前20条结果进行处理
    if ns1.loc[i,'newslist'] not in ns.keys():
        ns[ns1.loc[i,'newslist']]=0
    if ns2.loc[i,'newslist'] not in ns.keys():
        ns[ns2.loc[i,'newslist']]=0
    if ns3.loc[i,'newslist'] not in ns.keys():
        ns[ns3.loc[i,'newslist']]=0
    ns[ns1.loc[i,'newslist']]+=float(ns1.loc[i,'scorelist'])/3
    ns[ns2.loc[i,'newslist']]+=float(ns2.loc[i,'scorelist'])/3
    ns[ns3.loc[i,'newslist']]+=float(ns3.loc[i,'scorelist'])/3

#将上述结果按分值重排序
result_dict=dict(sorted(ns.items(), key=lambda d:d[1], reverse = True))
```

4.3 应用实例

同前，设定一个用户历史阅读的新闻内容为关于皇室的新闻，部分新闻描述如下：

	News ID	Category	SubCategory	Title	Abstract	URL	Title Entities	Abstract Entities
0	N3112	lifestyle	lifestyleroyals	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Philip, Duke of Edinburgh",...	[]
17	N25192	lifestyle	lifestyleroyals	Every outfit Duchess Kate has worn in 2019	See Kate Middleton's style choices this year f...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[{"Label": "Catherine, Duchess of Cambridge", ...
66	N1172	lifestyle	lifestyleroyals	The surprising age differences between your fa...	Here are the age differences between Meghan Ma...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[{"Label": "Meghan, Duchess of Sussex", "Type"...
197	N16959	lifestyle	lifestyleroyals	Queen Elizabeth's Favorite Beauty Products Hav...	Here, all the brands the British monarch swear...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[]

混合模型的推荐结果如下：

	News ID	Category	SubCategory	Title	Abstract	URL	Title Entities	Abstract Entities
20019	N13667	lifestyle	lifestyleroyals	Prince Harry and Meghan Markle just shared a n...	The Duke and Duchess of Sussex shared a new ph...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Charles, Prince of Wales", "Type":...	[{"Label": "Charles, Prince of Wales", "Type":...
20187	N19904	lifestyle	lifestyleroyals	Prince Harry and Prince William's Rift Is "One...	Santa is shook.	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Harry, Duke of Sussex", "Ty...	[{"Label": "Santa Claus", "Type": "R", "Wikida...
22434	N22351	lifestyle	lifestyleroyals	Meghan Markle and Prince Harry Won't Spend Chr...	They'll hang out with baby Archie and Meghan's...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Meghan, Duchess of Sussex", "Type"...	[{"Label": "Meghan, Duchess of Sussex", "Type"...
3103	N16759	lifestyle	lifestyleroyals	How Kate Middleton and Prince William's royal ...	Harry and Meghan had a much more relaxed appro...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince William, Duke of Cambridge"...	[{"Label": "Prince William, Duke of Cambridge"...
21225	N15539	lifestyle	lifestyleroyals	Harry and Meghan Revive Feud Rumors By Staying...	So, what are you doing for Christmas?	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Meghan, Duchess of Sussex", "Type"...	[{"Label": "Christmas", "Type": "H", "Wikidata...
97	N17473	finance	finance-real-estate	The penthouse of NYC's Woolworth Building just...	The condo comes as a "white-box" unit, meaning...	https://www.msn.com/en-us/finance/finance-real...	[{"Label": "Woolworth Building", "Type": "F", ...	[]
1271	N5387	lifestyle	lifestyleroyals	The Most Adorable Photos of Archie Harrison Mo...	The royal baby boy is growing up so fast!	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[]
15207	N10887	lifestyle	lifestyleroyals	19 Rarely Seen Photos of Royal Siblings	We've pulled together some rarely seen photos ...	https://www.msn.com/en-us/lifestyle/lifestyler...	[]	[]
551	N3627	news	newscrime	Motorcyclist dies after crash with car in Quee...	A man who was riding a motorcycle died of his ...	https://www.msn.com/en-us/news/newscrime/motor...	[{"Label": "Queen Creek, Arizona", "Type": "G"...	[{"Label": "Queen Creek, Arizona", "Type": "G"...
1267	N837	lifestyle	lifestyleroyals	Already 6 Months! Prince Harry, Duchess Meghan...	Already 6 Months! Prince Harry, Duchess Meghan...	https://www.msn.com/en-us/lifestyle/lifestyler...	[{"Label": "Prince Harry, Duke of Sussex", "Ty...	[{"Label": "Prince Harry, Duke of Sussex", "Ty...

可以看到在大部分新闻类别在保证用户喜欢的情形下，也保持了一些其他类别新闻的混合。

5 总结与反思

5.1 分工协作

在整个过程中，我们双方相互协作，学习了一系列不同的推荐算法，收获颇丰。

肖文璇主要负责文献查找与数据处理分析，研究讨论了TF-IDF与DKN算法与优化思路，完成了CF新闻推荐算法的代码实现。

雷雨寒主要完成了模型准备，TD-IDF模型与DKN模型的代码实现与模型优化，研究讨论了CF新闻推荐的算法与改进思路。

5.2 模型优势

本新闻推荐模型采用了TF-IDF算法,DKN算法,User-based-CF算法的混合优化，TF-IDF模型能通过用户的新闻浏览历史库得到与库中新闻十分类似的新闻推荐结果，保证了新闻推荐结果是用户较为感兴趣的内容或接近用户感兴趣的新闻，但推荐结果较为单一。而DKN模型能通过知识图谱发现不同新闻主题与内容的联系，进行更为发散的新闻推荐。基于用户的协同过滤方法能够发现了用户与用户之间的潜在联系，利用用户之间相似或相反的关系，推荐出更加多样化。三者相互混合补充，最终取得了较好的新闻推荐结果。

5.3 模型进一步优化

本模型还有一些地方需要继续改进，在TF-IDF算法中，因为算法需要构建较大的TF-IDF矩阵，因此时间和空间开销都较大。同时本模型基于静态的数据，未来还可以基于动态的新闻数据进行实时推荐。