**Events Sequence Prediction**

Max. Marks: 1

This problem is no longer available for practice. Apology for any inconvenience!

**Problem Statement**

Medaplexis, a premium payer wants to improve the medical care using Machine Learning. They want to predict next events of Diagnosis, Procedure or Treatment that is going to be happen to patients. They've provided with patient journey information coded using ICD9. The detail reference for ICD9 can be found [here](http://www.icd10codesearch.com/) .

In this challenge, you have to predict the next 10 events reported by patient in order of occurrence in 2014.

[**Download Dataset**](https://he-s3.s3.amazonaws.com/media/hackathon/zs-young-data-scientist-challenge-2017/events-sequence-prediction/2f97065a-6-YDS_dataset.zip)

**Data Description**

There are three files given to download: train.csv, test.csv and sample\_submission.csv   
The train data consists of patients information from Jan 2011 to Dec 2013. The test data consists of Patient IDs for the year 2014.

| **Variable** | **Description** |
| --- | --- |
| ID | Patient ID |
| Date | Period of Diagnosis |
| Event | Event ID (ICD9 Format) - Target Variable |

**Submission**

A participant has to submit a .csv file with PID and predicted events in order of recency. Check the sample submission file for reference.

| **PID** | **Event1** | **Event2** | **Event3** | **Event4** | **Event5** | **Event6** | **Event7** | **Event8** | **Event9** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1028890 | 2605 | 2382 | 3222 | 2761 | 1809 | 3620 | 2781 | 2178 | 7132 |
| 1021443 | 3786 | 3553 | 2706 | 3553 | 3786 | 2706 | 3553 | 3786 | 2706 |
| 1016194 | 2605 | 2382 | 3222 | 2761 | 1809 | 3620 | 2781 | 2178 | 7132 |
| 1017298 | 2706 | 3553 | 2706 | 3553 | 3786 | 2706 | 3553 | 3786 | 2706 |

**Evaluation Metric**

The evaluation matrix (Final Leaderboard score) for this competition is mean NDCG (Normalized discounted cumulative gain)@K where K=10. NDGC is calculated as:

nDCG(k)=DCG(k)/(IDCG(k)nDCG(k)=DCG(k)/(IDCG(k)

where, DCG(k)DCG(k) is Discounted Cumulative Gain@K and is computed as:

DCG(k)=∑ki=12reli−1log2(i+1)DCG(k)=∑i=1k2reli−1log2(i+1)

IDCG(k)IDCG(k) is Ideal or maximum possible value attainable. The NDCG calculations are normalized between 0.0 to 1.0. The relireli is relevance of prediction and is defined as:

1 if predicted event is correct @k

min(act\_pos/pred\_pos, pred\_pos/act\_pos) if predicted event is in top k

else 0

where, act\_pos is actual position of predicted event and pred\_pos is position of predicted event. For more information on this metric, [read here](https://en.wikipedia.org/wiki/Discounted_cumulative_gain#Normalized_DCG).

**Submission Guidelines**

Please make your final submission based on [guidelines mentioned here](https://s3-ap-southeast-1.amazonaws.com/he-public-data/2017%20YDS%20Submission%20Content%20Detailse0fbbc1.pdf). Participants failing to follow these guidelines will not be shortlisted for the next round.