

## 1 Этап evaluation

Пусть  $I_x$ ,  $I_y$  – наборы элементов, для которых есть оценки у пользователей  $x$  и  $y$  соответственно;  $I_{xy}$  – набор элементов, для которого есть оценки у обоих пользователей  $x$  и  $y$ ;  $r_{xi}$  и  $r_{yi}$  – значения оценок для объекта  $i$ .

Можем использовать меру схожести объектов.

### 1.1 Cosine similarity

Вычисляет сходство между векторами на основе угла между ними. Значения близкие к 1 обозначают сильную корреляцию, близкие к 0 – слабую.

$$\text{Cosine}(x, y) = \frac{\sum_{i \in I_{xy}} r_{xi} r_{yi}}{\sqrt{\sum_{x \in I_x} r_{xi}^2 \sum_{y \in I_y} r_{yi}^2}}$$

### 1.2 Pearson Correlation Coefficient (PCC)

Оценивает линейную связь. Значения близкие к 1 обозначают сильную положительную корреляцию, близкие к 0 – слабую, близкие к -1 – сильную отрицательную.

$$\text{PCC}(x, y) = \frac{\sum_{i \in I_{xy}} (r_{xi} - \bar{r}_x)(r_{yi} - \bar{r}_y)}{\sqrt{\sum_{x \in I_x} (r_{xi} - \bar{r}_x)^2 \sum_{y \in I_y} (r_{yi} - \bar{r}_y)^2}}$$

### 1.3 Manhattan similarity

Основана на манхэттенском расстоянии между векторами. Значения близкие к 1 обозначают сильную корреляцию, близкие к 0 – слабую.

$$d_1(x, y) = \sum_{i \in I_{xy}} \|r_{xi} - r_{yi}\|$$
$$ES(x, y) = \frac{1}{1 + d_1(x, y)}$$

### 1.4 Euclidean similarity

Основана на расстоянии между векторами в евклидовом пространстве. Значения близкие к 1 обозначают сильную корреляцию, близкие к 0 – слабую.

$$d(x, y) = \sqrt{\sum_{i \in I_{xy}} (r_{xi} - r_{yi})^2}$$
$$ES(x, y) = \frac{1}{1 + d(x, y)}$$

## 1.5 Jaccard similarity

Вычисляет сходство и разнообразие двух наборов. Измеряет отношение количества элементов, совместно используемых двумя наборами, к общему количеству элементов в обоих наборах. Значения близкие к 1 обозначают сильную корреляцию, близкие к 0 – слабую.

$$J(x, y) = \frac{\|x \cap y\|}{\|x \cup y\|}$$

## 1.6 Анализ

Поскольку работаем с бинарными оценками, можем использовать Jaccard similarity. Тогда оценка того, понравится ли пользователю  $x$  статья  $a$  будет выглядеть так:

$$like(x, a) = \frac{\sum_{y \in Y_a} J(x, y)}{|Y_a|}$$

где  $Y_a$  – множество пользователей, которым понравилась статья  $a$ .

Для рекомендации авторов можно поступить аналогично.

## 2 Этап проверки качества

### 2.1 Recall at K

Считаем долю релевантных элементов из отобранных  $k$  элементов от общего количества релевантных элементов.

$$Recall@k = \frac{\sum_{i=1}^k Relevance@i}{\sum_{i=1}^N Relevance@i}$$

### 2.2 Precision at K

Считаем долю релевантных элементов из отобранных  $k$  элементов. Недостаток – не учитывает порядок элементов.

$$Precision@k = \frac{1}{k} \sum_{i=1}^k Relevance@i$$

### 2.3 Average precision at K

Считаем сумму  $Precision@i$  по индексам  $i$  от 1 до  $k$  только для релевантных элементов и делим на количество релевантных элементов.

$$AP@k = \frac{\sum_{i=1}^k Relevance@i \cdot Precision@i}{\sum_{i=1}^k Relevance@i}$$

## 2.4 Mean average precision at K

Считаем  $AP@K$  для каждого объекта и усредняем.

$$MAP@k = \frac{1}{N} \sum_{j=1}^N AP@k_j$$

## 2.5 Coverage

Доля элементов, которую выдает рекомендательная система.

$$coverage = \frac{I_s}{I}$$

## 2.6 Diversity

Показывает насколько разнообразные элементы выдает рекомендательная система.

$$diversity = 1 - similarity$$

## 2.7 Анализ

Можем использовать  $MAP@k$  и дополнительно учитывать coverage и diversity.