Similarity Functions for user – user Collaborative Filtering
https://grouplens.org/blog/similarity-functions-for-user-user-collaborative-filtering/

일반적으로 user – user 협업 필터링은 Pearson 상관계수를 이용해서 사용자들을 비교한다.
(Spearman correlation, Cosine 유사도가 이용되었지만 Pearson 상관계수가 가장 좋아서,
한동안 이슈화 되지 않았다.)

그러나 평균중심화 백터의 Cosine 유사도(Adjusted Cosine)를 offine evaluation metrics에서 사용하였을 때 Pearson 상관계수보다 좋았다.

Pearson 계수는 유사도를 계산할 때 두 가지 문제를 가지고 있었다

첫 번째는 아이템에 대해서 평점을 매긴 사람과 그렇지 않은 사람과의 비교이다. 이 문제를 통계적으로 해결하기 위해선 평점을 매긴 사람들끼리만 고려하는 것이다. 그러나 실제에선 매우 유의한 차이를 만들어내지 못한다.

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v}(r_{ui}-\mu_u)(r_{vi}-\mu_v)}{\sqrt{\sum_{i \in I_u \cap I_v}(r_{ui}-\mu_u)^2}\sqrt{\sum_{i \in I_u \cap I_v}(r_{vi}-\mu_v)^2}}$$ (Pearson)

두 번째는 적은 아이템에 대해서 공통적으로 평점을 준 사용자들은 매우 높은 유사성을 가진다는 것이다. 단일 쌍에 대한 Pearson correlation은 정의 되지 않는다. (최소 2개의 쌍)
또한 한 사용자가 여러 아이템에 대해서 평점을 준 뒤, 두 아이템에 대해서 Pearson correlation을 계산하여 비슷하다고 말을 하는데 이것은 좋은 방법이 아니다.
최소 평점을 준 아이템의 개수가 50개를 공통적으로 가질 때까지 유사성이 선형적으로 감소한다.

● Adjusted Cosine

$$\mathrm{sim}(u,v) = \frac{\hat{\mathbf{u}} \cdot \hat{\mathbf{v}}}{\|\hat{\mathbf{u}}\|\|\hat{\mathbf{v}}\|} = \frac{\sum_i \hat{r_{ui}} \hat{r_{vi}}}{\sqrt{\sum_i \hat{r_{ui}}^2}\sqrt{\sum_i \hat{r_{vi}}^2}}$$ ($\hat{r_{ui}}$ is the normalized rating $r_{ui}-\mu_u$)

Adjust Cosine similarity는 Pearson correlation과 유사한 형태이다.
이 수식에 대해서 아직 i 값에 대해서 명시되지 않았을 때 Iu∩Iv의 합은 Pearson correlation과 같다.
그러나 Iu∪Iv의 값을 전부 더 했을 때는 rui^=0이다

따라서 사용자들이 동일한 item을 평가하였을 때는 Pearson correlation이 존재한다. 그러나 각각 다른 아이템을 평가하였을 때는 Pearson correlation이 존재하지 않는다. (그렇지만 나중에 전체값을 계산할 때는 사용) – 즉 엄격한 선형 스케일이 아니라 등급 값 계산에 영향을 미치고 basic idea를 전달하는데 유용하다. 이러한 유사도 함수를 self damping의 특성을 가진다고 한다.
오프라인 실험에서 self-damping은 유의한 가중치를 주는 것 보다 효과가 있고, 앞에서 본 50개의 컷오프에 영향을 받지 않는 것을 확인했다.

몇몇의 연구에서 Cosine similarity가 잘 작용을 하지 않았던 이유는 self-damping을 가지고 있지 않았기 때문이다. 앞서 연구에서 보았을 때 Adjusted Cosine이 유사도를 측정하는 가장 일반적인 모형임을 알아야 한다.