**Rate Limits**

서블릿 공부 중에 서블릿에서 클라이언트의 요청이 너무 많이 증가하면 서버 컴퓨터에 과부화가 발생할텐데 이를 어떻게 관리하는지 궁금해서 찾아봤다

Rate Limits API 제한에 제한을 둠으로써 하나의 클라이언트가 서버의 리소스를 너무 많이 사용하지 않도록 하는 방법이다.

예를 들어, 개발자가 한 기업의 API를 사용해서 애플리케이션을 만든다고 생각하자. 이 때, 사용자가 이 애플리케이션을 통해서 기업의 API에 요청을 보내게 되는데 이를 한번에 너무 많이 보내면 서버 측에서 사용자 한 명에 대한 리소스가 많아진다.

Rate Limits는 여러 가지 방법이 있는데 그 중에서 RPM(requests per minute – 시간당 제한), RPD(request per day – 일당 제한), TPM(tokens per minute – 시간당 토큰), TPD(tokens per day – 시간당 토큰), IPM(images per minute – 분당 이미지)같이 여러 가지 방법이 있다.

.

참고

OpenAI API 사이트 - https://platform.openai.com/docs/guides/rate-limits?context=tier-free