04 Data Preprocessing (2)전처리 및 URL 임베딩

전처리

- 노이즈 많은 Full_URL보다는 Domain 값 활용
- owership_1, ownership_2 결측치 수정
- owernship_1, ownership_2 결측 행 제거
- Domain .com, .co.kr 등 불용어 제거
- 11분 간 활동을 기준으로 Session_ID 부여

URL 임베딩

- Word2Vec Model
- 하나의 문헌 = 한 사람당 한 세션의 Domain들을 나열한 list

  Example. [‘auction’, ‘auction’, ‘kbstar’, ‘naver’]
- Dimension = 3, sg = 1, min_count = 1 로 학습
- 3차원 임베딩 값을 시각화에 이용할 RGB값으로 활용

6월 한 달, 매일, 매 분의 로그를 하나의 이미지로 나타내는
Log2Image 진행