Project 6. NoSQL을 이용한 대용량 *N*-Gram 언어모델 검색시스템

박운상 교수

Office: K 336

Tel: 705-8936

Email: unsangpark@sogang.ac.kr

프로젝트 목표

- 본 프로젝트 에서는 한 컴퓨터에 올릴 수 없는 대용량의 ***N*-Gram** 언어모델을 검색하기 위하여, 분산 데이터베이스 형식으로 구현된 비-관계형 **DB**인 **NoSQL** 시스템을 도입하여 해결한다.
- **N-Gram** 언어모델은 관계형 **DB**의 모습을 띄고 있지만, 대용량을 분산하여 보관하기 위해서 비-관계형 **DB**인 **NoSQL**을 사용합니다.
- 이를 위해 **Python**과 **AWS**의 **NoSQL**시스템인 **DynamoDB**를 사용한다.

프로젝트 요구사항 및 설명

- 프로젝트 목표 설정
  - 본 프로젝트에서는 대용량 **Bigram(*N*-Gram에서 *N*=2)** 언어모델을 **Python**과 **AWS DynamoDB** 시스템을 이용한 분산데이터베이스를 통해 빠르게 검색한다.
- 합성
  - 입력
    - **Amazon S3에 저장된 nGB의 Bigram 언어모델**
      - 각 줄에 단어 조합과 그 단어 조합이 나온 횟수를 Tab으로 구분하여 저장되어있음.

프로젝트 설명

- 합성
  - 처리
    - **Python에서 AWS S3의 Bigram언어모델을 읽어, AWS DynamoDB에 저장**
  - 결과
    - 대용량 언어모델이 저장된 자신의 **DynamoDB Table**
    - 이에 접근하여 특정한 단어 조합의 빈도를 **Command Line Interface**에서 조회할 수 있는 **Python Code**
- 제작
  - **AWS를 Python에서 사용할 수 있는 라이브러리인 "Boto" 사용**
  - **AWS EC2, S3 & DynamoDB를 이용**
    - **Boto와 AWS DynamoDB는 별도의 강의자료를 통하여 설명**

프로젝트 설명

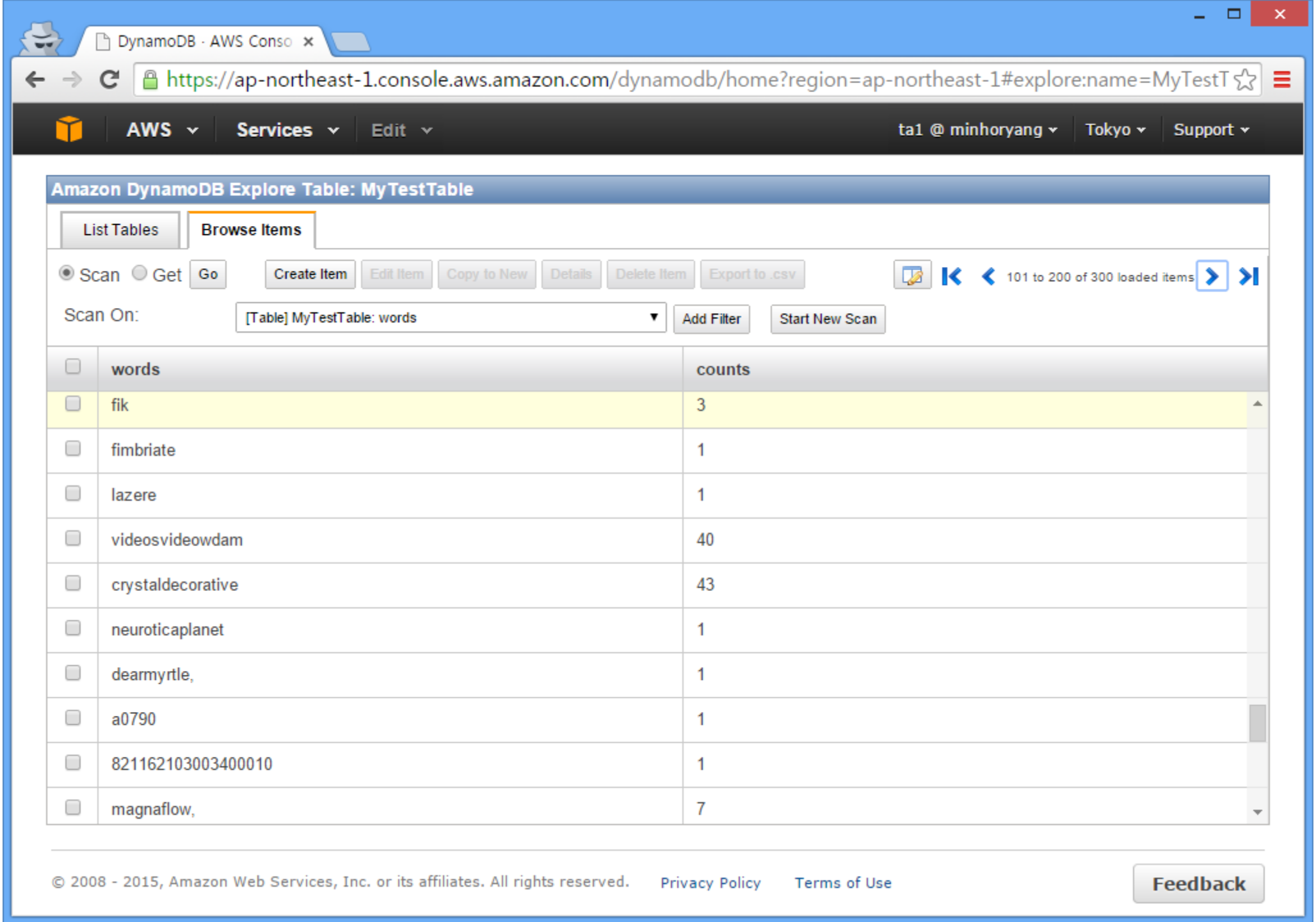
- 시험
  - 입력 파일은 조교의 S3 버킷에 있다.
  - AWS DynamoDB를 이용하여 Bigram 언어모델을 보관할 Table을 생성한다.
    - 속도와 요금을 결정하는 Read Capacity Units와 Write Capacity Units는 1로 고정한다.
    - Table의 Index는 String 타입의 “words” 이름으로 고정하고, 여기에는 Bigram의 단어조합을 저장한다.
  - 생성한 최종 출력은 자신의 DynamoDB Table에 저장되어야 함.

프로젝트 설명

- 평가
  - S3에 저장된 Hadoop Parted Output에서 자신의 DynamoDB Table로 데이터를 전송을 할 때 빠지는 단어 조합이 있는지, Count수가 같은지를 검사
  - CLI를 통하여 특정 단어 조합을 검색할 때 Count를 잘 출력하는지, 그리고 그 Count가 맞는지 검사
    - 수행 시 Read Capacity Units와 Write Capacity Units를 초과하여 계산하였을 경우 경고
  - 성능평가를 위해 3번을 돌려 평균을 내어 차등 점수 부여

프로젝트 설명

- 환경구성
  - 자신의 AWS EC2 Instance
  - AWS S3
  - AWS DynamoDB
- 제출물
  - DynamoDB_Import_S3.py
  - DynamoDB_Query_CLI.py
  - AWS DynamoDB의 Table 구성과 내부가 담긴 캡쳐화면 (.png)
  - 학번.conf
    - 자신의 AWS 계정이 기록된 설정파일
    - 형식은 실습시간에 제공하며, 확인을 위해서만 사용됩니다.


DynamoDB · AWS Conso
https://ap-northeast-1.console.aws.amazon.com/dynamodb/home?region=ap-northeast-1#explore:name=MyTestT
AWS
Services
Edit
ta1 @ minhoryang
Tokyo
Support
Amazon DynamoDB Explore Table: MyTestTable
List Tables
Browse Items
Scan
Get
Go
Create Item
Edit Item
Copy to New
Details
Delete Item
Export to .csv
101 to 200 of 300 loaded items
Scan On:
[Table] MyTestTable: words
Add Filter
Start New Scan
words
counts
fik 3
fimbriate 1
lazere 1
videosvideowdam 40
crystaldecorative 43
neuroticaplanet 1
dearmyrtle, 1
a0790 1
821162103003400010 1
magnaflow, 7
© 2008 - 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Feedback

제출 방법

- **sp학번_proj6라는 이름의 디렉터리를 만들고, 이 디렉터리에 제출파일, Document, readme 파일을 넣어서 디렉터리를 tar로 압축하여 한 파일로 만들어 메일로 보내시기 바랍니다.**
- **제출 파일은 sp<학번>_proj6.tar 입니다.**
- **제출 주소 : sp2016proj@gmail.com**

  메일제목 형식 : [SP숙제 #6] 학번 이름

주의 사항은 이전 프로젝트와 같습니다.