N-gram 결과 보고

목차

1. N-gram 이란?

N-gram 이란?

100개의 주소 데이터에서 각 단어의 출현 빈도

1 : 서울특별시 서초구 25개
2 : 서울특별시 서촉구 2개

위와 같은 경우에서
1번의 개수가 많으므로 ‘서울특별시’ 뒤에는
‘서초구’가 나올 확률이 더 높으므로 ‘서촉구’는 오타로 판별

정상인 두 데이터

A : 서울시 강서구 등촌동
B : 부산시 강서구 명지동

위 경우 데이터에서 ‘서울시 강서구’ 가 많은 경우
B의 경우에서도 ‘서울시 강서구’ 로 변환할 수 있음

정상인 두 데이터 (역순)

A : 등촌동 강서구 서울시
B : 명지동 강서구 부산시

위의 경우에는 등촌동 뒤에는 강서구, 명지동 뒤에는 강서구가
나올 확률이 높다고 판단

(단점 : ~동 은 올바른 데이터라고 가정해야함
-> 번지수로 동을 판별할 수 없으므로)

2. 목적

N-gram을 주소에 적용하는 목적

- 주소의 오탈자 검출 및 수정 가능
  - ‘서초구 서초대로56길 40 디제이빌딩 3층’ 의 데이터를
    ‘서울특별시 서초구 서초대로56길 40 디제이빌딩 4층’ 으로 수정 가능

3. 실험 결과

창규님의 알고리즘을 통해 1차로 주소 정규화 한 후

N-gram 으로 주소 정규화 작업

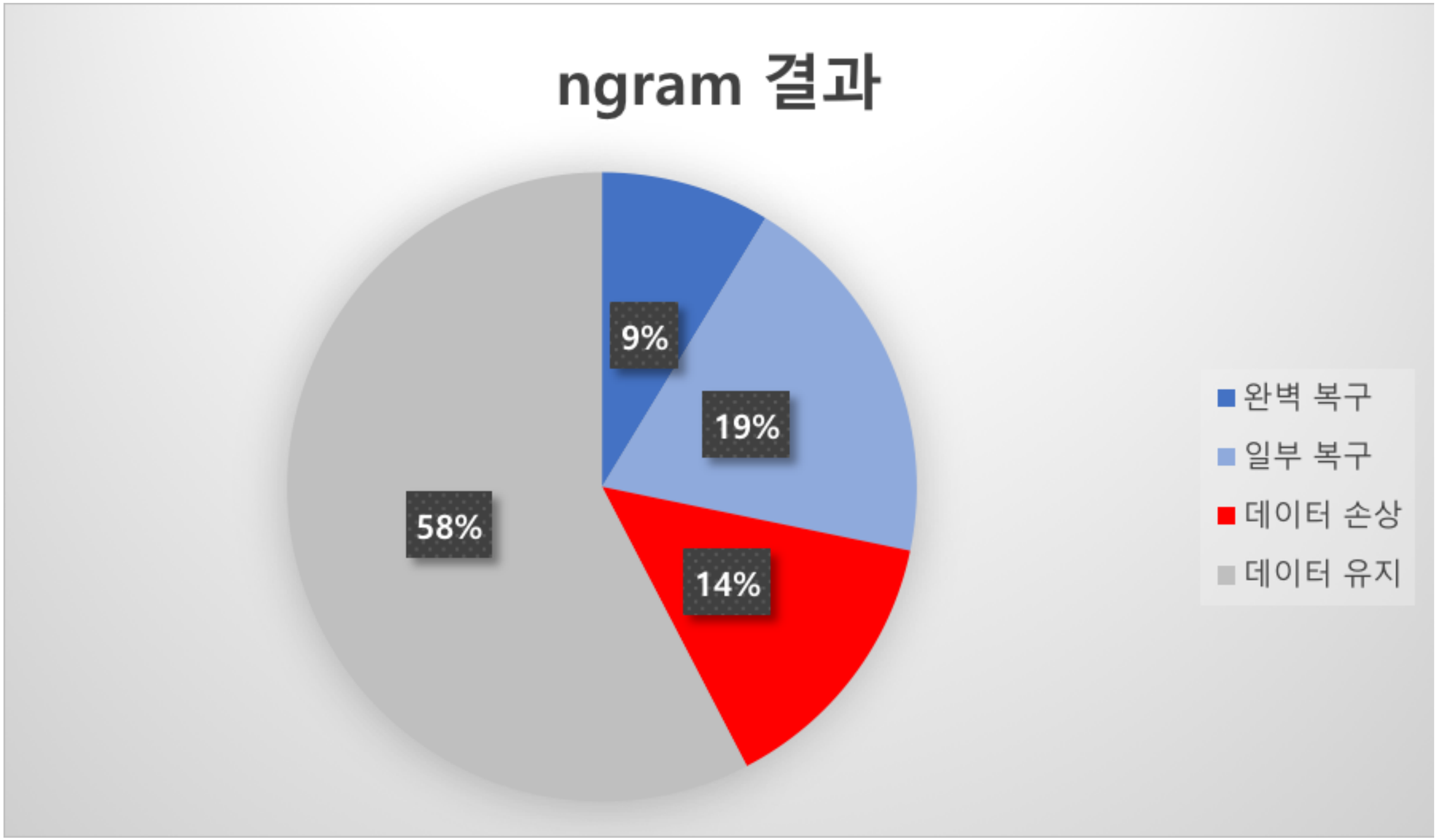
창규님의 주소 정규화 알고리즘

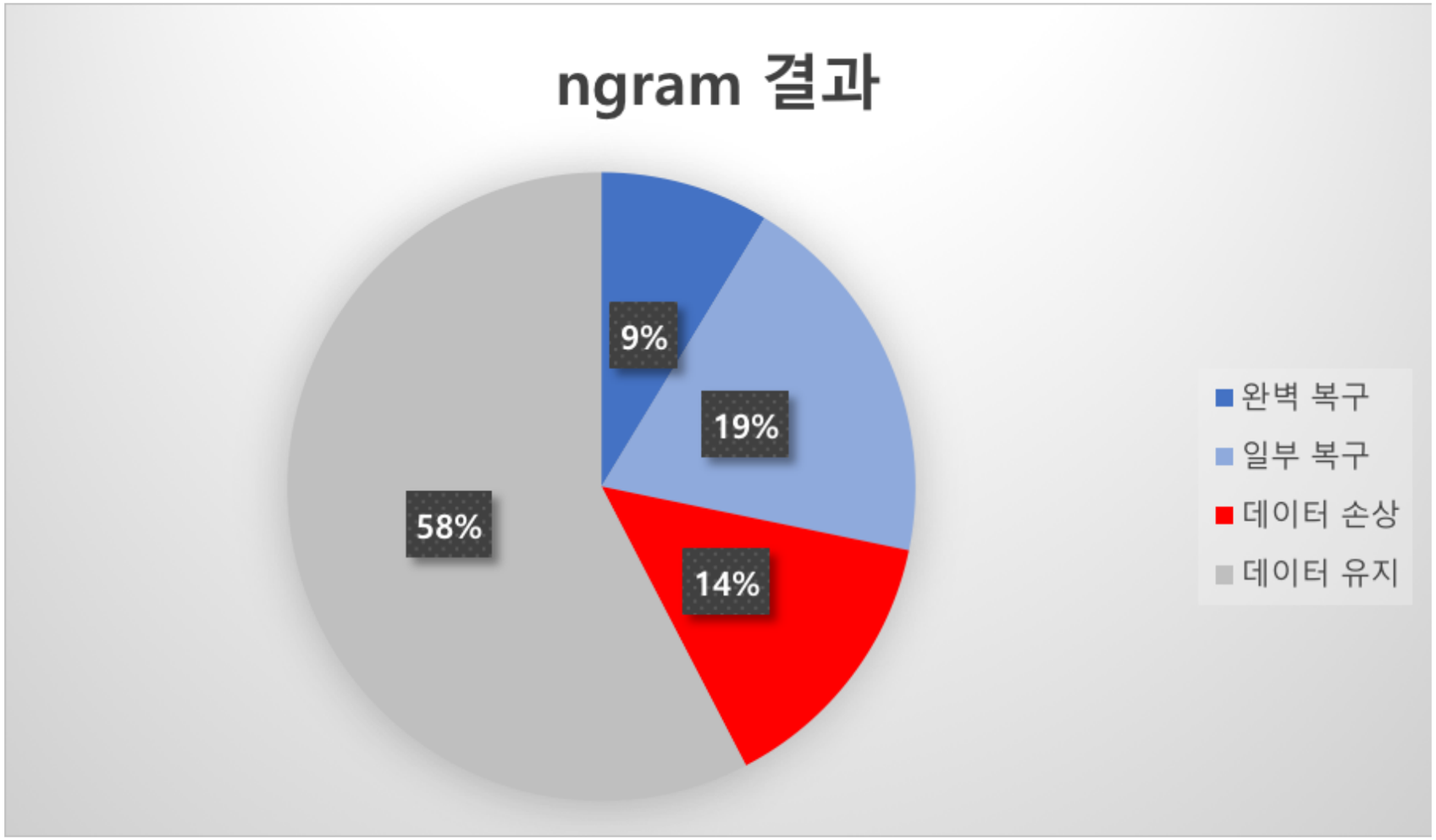
- **Split (**띄어쓰기가 되어있지 않은 주소 정보 띄어쓰기**)**

- 주소 통일

addr_list['충청북도']=('충청북도', '충북')
sub_addr_list['충청북도']=('제천시', '충주시', '단양군', '음성군', '진천군', '증평군', '괴산군',
'청주시', '보은군', '옥천군', '영동군')

- 크롤링 (위도 경도값을 이용하여 주소값 반환 -
http://mygeoposition.com/)


ngram 결과
9%
19%
58%
14%
완벽 복구
일부 복구
데이터 손상
데이터 유지



총 데이터 : 92개 (100%)

완벽 복구 : 8개 (8.70%)

일부 복구 : 18개 (19.57%)

데이터 손상 : 13개 (14.13%)

데이터 유지 : 53개 (57.61%)

복구(완벽 복구 + 일부 복구) : 28.27%

데이터 손상 : 14.13%

데이터 어느정도 복구 예 :

- 서울특별시를 복원하지만 지번 숫자 제거

- 대구달성군 -> 달성군 으로만 복구

데이터 손상의 예 :

- 아예 다른 주소로 변환

- 인천광역시 -> 울산광역시

- 띄어쓰기 없는 경우 데이터 손상

데이터 미변화의 예 :

- 쓸 수 없는 데이터 -> 쓸 수 없는 데이터

- 구주소 -> 신주소 (정확하게 변환되었음)

4. 결론

- 결국 N-gram 은 확률이다.

- 정확할 수도 있고 정확하지 않을 수도 있다.

- N-gram 의 기반이 되는 단어 출현 빈도 문서를 정확한 문서를
  사용하면 N-gram 의 정확도는 더 높아질 것이다.
  (현재는 창규님이 수정한 COMPANY table 의 주소값으로
  단어 출현 빈도 문서 생성후 이용)
  (DB 공공데이터 API 는 사파리, 크롬에서 사용 불가)

- 오탈자의 데이터 수가 적어서 손으로 직접 하는 것이 정확하고
  효율적이다.