2023학년도 2학기 언어데이터과학

제8강 데이터과학 (4)

박수민

서울대학교 인문대학 언어학과

2023년 9월 27일 수요일

지난 시간에 배운 것

1. 데이터프레임의 정의와 용도
2. `pandas`를 이용한 데이터프레임 처리
3. `seaborn`을 이용한 데이터 시각화

오늘의 목표

1. Token frequency, type frequency, document frequency의 개념 차이를 설명할 수 있다.
2. 세 가지 빈도를 각기 해석하는 방법을 설명할 수 있다.

Frequency의 두 가지 의미

- 주파수
- 빈도 ← 오늘의 개념

언어학자는 셈

- 코퍼스에서 '삼국지'와 관련된 문서가 몇 건 있는가?
- 이 책에서 가장 자주 사용된 형태소가 무엇이고 몇 번 사용되었는가?
- 국어사전에 2음절 한자어 명사 표제어가 몇 개 있는가?
- 청취 실험/문법성 테스트에서 a라고 답한 참여자가 몇 명인가?
- ...

여러 가지 빈도 개념

- Document frequency
  - 코퍼스에서 '삼국지'와 관련된 **문서가 몇 건** 있는가?
- Token frequency
  - 이 책에서 가장 자주 사용된 형태소가 무엇이고 **몇 번 사용**되었는가?
- Type frequency
  - 국어사전에 2음절 한자어 명사 **표제어가 몇 개** 있는가?

예시

'달아 달아 밝은 달아'는 **몇 개 단어**로 이루어져 있는가?

Type freq. 두 개 단어 ['달아', '밝은']

Token freq. 네 개 단어 {'달아': 3, '밝은': 1}

빈도 개념 활용 예시 (1)

오늘의 논문

이정복. (2010). 〈인터넷 통신 공간의 여성 비하적 지시 표현〉. 《사회언어학》 18권 2호. 215–247.

질문

‘김여사’와 ‘김기사’는 의미상 대등한 한 쌍인가?

빈도 개념 활용 예시 (1)

‘김여사’ vs. ‘김기사’ 현상 (이정복 2010, 226면)

‘김여사’에 대응하는 남성형도 다수 나타났다. (4가)의 ‘김기사’는 […] 인터넷에서도 ‘김여사’와 함께 쓰이며 비슷한 뜻을 갖고 있다. […] 그런데 이러한 남녀 대응형은 **사용 빈도** 및 용법 면에서 큰 차이가 있다. […] 2010년 9월 중순 현재 ‘김여사’는 677개의 게시글에서 쓰인 반면 ‘김기사’는 50개에서 쓰였다.

두 가지 빈도 개념으로 해석하기

Type freq. 여성형 1개 $=$ 남성형 1개

Doc freq. ‘김여사’ 677개 $>$ ‘김기사’ 50개

Token freq. 알 수 없음

빈도 개념 활용 예시 (1)

구분	김여사	김기사	합
이야기 게시판	677(93.1)	50(6.9)	727개(100%)
토론 게시판	599(87.4)	86(12.6)	685개(100%)

<표 2> '김여사'와 '김기사'의 쓰임 차이 [게시글 수(백분율)]

문제: 백분율

이야기 게시판에서 '김여사'의 비율이 93.1%라는 수치가 무엇을 의미하는가?

빈도 개념 활용 예시 (1)

표에 백분율(%)을 넣고 싶을 때 주의할 점

분자에 빈도 값이 들어가고, 분모에는 …?

- 분모가 무엇인지를 명확히 해야 한다.
- 분모가 '전체'를 나타내야 한다.
- 셀의 값의 총합이 분모와 같아야 한다.

빈도 개념 활용 예시 (1)

구분	김여사	김기사	합
이야기 게시판	677(93.1)	50(6.9)	727개(100%)
토론 게시판	599(87.4)	86(12.6)	685개(100%)

<표 2> '김여사'와 '김기사'의 쓰임 차이 [게시글 수(백분율)]

백분율 사용의 문제점

1. 분모의 문제: 727개는 게시판 전체의 게시글 개수가 아니다.
   - 100%가 실은 '전체'가 아니다!
2. 분자의 문제: 677개 게시물과 50개 게시물이 겹칠 수 있다.
   - '김여사'와 '김기사'를 둘 다 포함하는 게시물이 존재하면 중복해서 집계된다!

빈도 개념 활용 예시 (2)

오늘의 발표문

박수지. (2021). 〈뉴스 기사 제목에 나타난 '○○女' 와 '○○男' 의 사용 양상 비교: 1990-2021년 사회면 기사를 중심으로〉. 《2021년 한국사회언어학회 가을 학술대회》. https://github.com/suparklingmin/news-title-gender

[숙제05] Word Cloud

제출 기한 변경: 2023-10-06 → 2023-10-10 13:00까지

[숙제02] AWK 연습

2023-10-06 13:00까지

다음 시간에 배울 것

한국어 텍스트 데이터를 다룰 때 고려해야 할 요소

1. 문자 인코딩
2. 유니코드