벤포드 법칙의 실제 성립 여부 검증
<진법(3~10)별, 분야(정치, 경제, 사회, 과학)별>

과 목 명 컴퓨팅 기초 | 팀 명 **인지** 팀

지도교수 변해선 교수님 | 학 과 / 이 름 첨단융합학부 박**인**수, 문**지**오

연구배경 및 필요성

벤포드의 법칙: 광범위한 분포를 보이는 수치 데이터들의 가장 큰 자리 숫자는 작은 숫자인 경향을 보이는 것 .

Ex) 100원 -> 200원: 2배 증가
200원 -> 300원: 1.5배 증가
900원 -> 1000원: 10/9배 증가
"1의 수명이 가장 길고, 9의 수명이 가장 짧다."

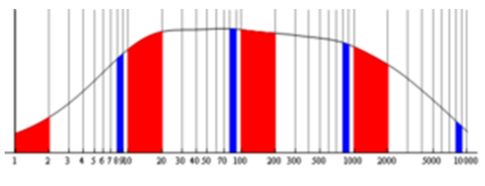
벤포드의 법칙 이용 사례: 재무회계 수치의 신뢰성 탐지/ 전기요금 고지서/ SNS, TV 프로그램 등의 데이터 조작 검증

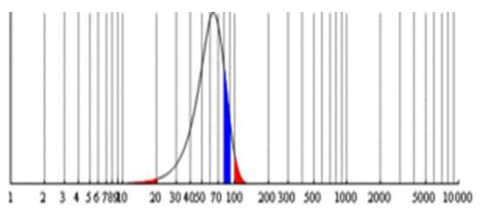
목표: 3진법 이상의 임의의 진법 체계에 대해 벤포드의 법칙이 항상 잘 성립할까?
정치, 경제, 사회, 과학 또한 벤포드의 법칙이 잘 성립할까?

방법: 1. 인터넷 상의 뉴스 기사 본문들을 수집해 3-10진법으로 정제 후 첫 자리 숫자의 분포 분석, 시각화, 적합도 검정하여 벤포드의 법칙이 잘 성립하는지를 검증한다.
2. 인터넷 상의 뉴스 기사들을 정치,경제,사회,과학 4분야로 나눈 후 각각의 벤포드의 법칙 성립 여부 또한 검증한다.

$$P(d) = \log_b(d+1) - \log_b(d) = \log_b\left(1 + \frac{1}{d}\right).$$

d	$P(d)$	Relative size of $P(d)$
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	





프로젝트 진행과정

1. 함수들 정의하기

- **extract_numbers(text) 함수**

: 입력받은 텍스트에서 숫자를 추출해 리스트 형태로 반환하는 함수
re 모듈을 사용해 숫자 문자열을 추출하고 float 또는 int로 변환한다.
예) extract_numbers('3년새 30억 뛰었다...') -> [3, 30]

- **convert_to_base(a,b) 함수**

: 10진법인 실수형 데이터 a를 b진법으로 변환해서 앞 한자리만 반환하는 함수
a를 정수와 소수 부분으로 분리하고 각각을 b로 계속 나누거나 곱해서 변환한다.
예) convert_to_base(6.12,3) -> 2

- **convert_list_to_bases(decimal_list) 함수**

: 10진법인 실수형 데이터 리스트 각각의 원소들을 3~10진법으로 변환해 2차원 리스트를 반환하는 함수로, 앞서 만든 convert_to_base 함수를 활용한다.
예) convert_list_to_bases([3,30]) -> [[1,1],[3,1],[3,1],[3,5],[3,4],[3,3],[3,3],[3,3]]

2. 뉴스 기사 본문 텍스트 크롤링 및 숫자 추출

- 네이버 뉴스 탭에서 정치, 경제, 사회, 과학 분야를 포함한 기사 url들을 복사해 하나의 리스트(URL_tot)에 담는다.
- request와 BeautifulSoup 모듈을 사용하고 'article' 태그를 이용해서 본문 텍스트를 가져온다.
- 앞서 정의한 extract_numbers 함수로 본문에서 숫자를 추출하고 이를 반복해, 여러 url들에서 가져온 숫자 데이터를 하나의 리스트(Tot)에 누적해 저장한다.
- 4가지 분야에 해당하는 기사는 각 분야별 리스트(URL_pol, URL_eco, URL_soc, URL_sci)에 추가로 담고, 각각에서 추출한 숫자 리스트도 개별적으로 저장한다.

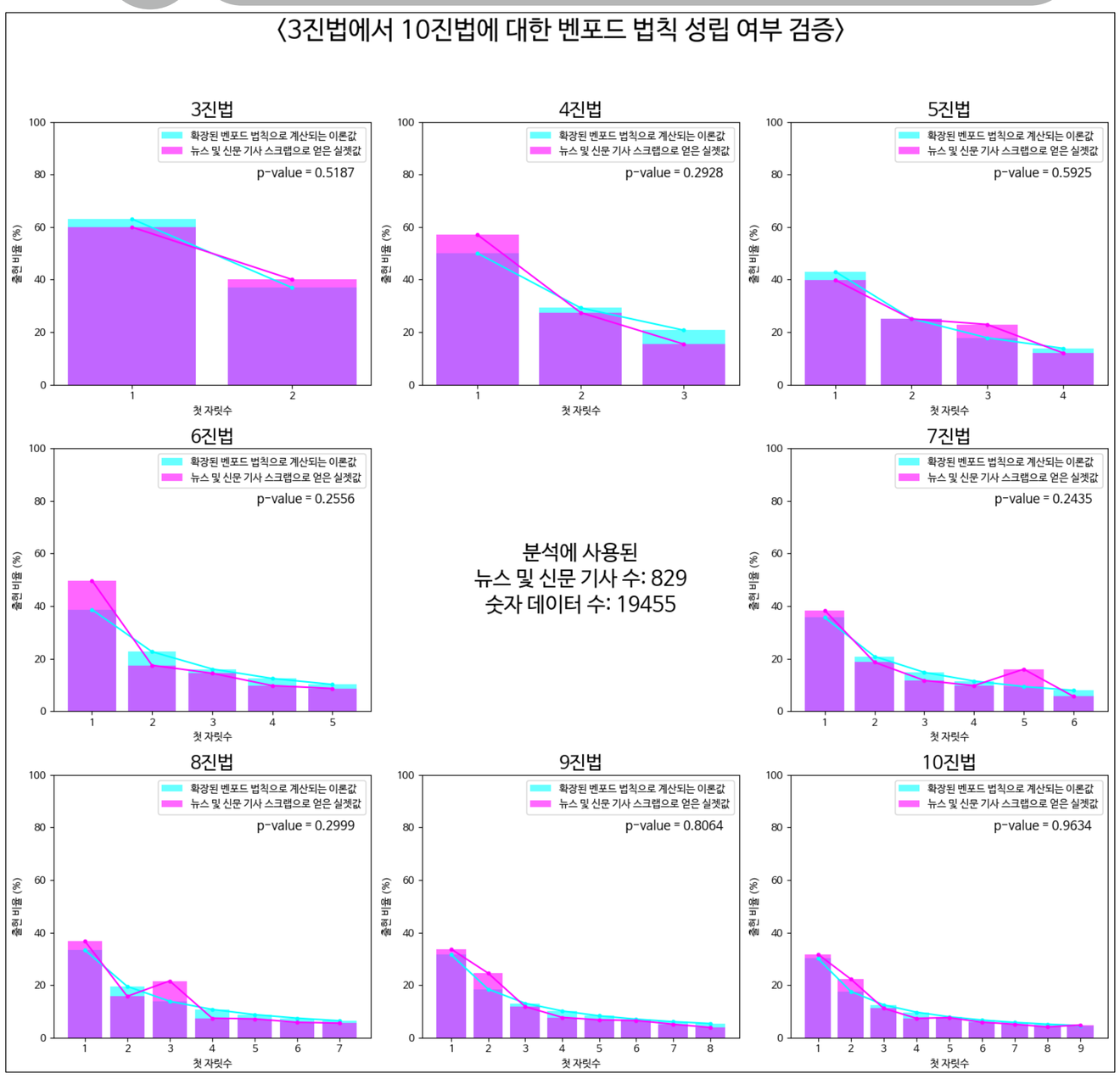
3. 8가지 진법(3~10)별 벤포드 법칙에 대한 적합도 검정 및 시각화

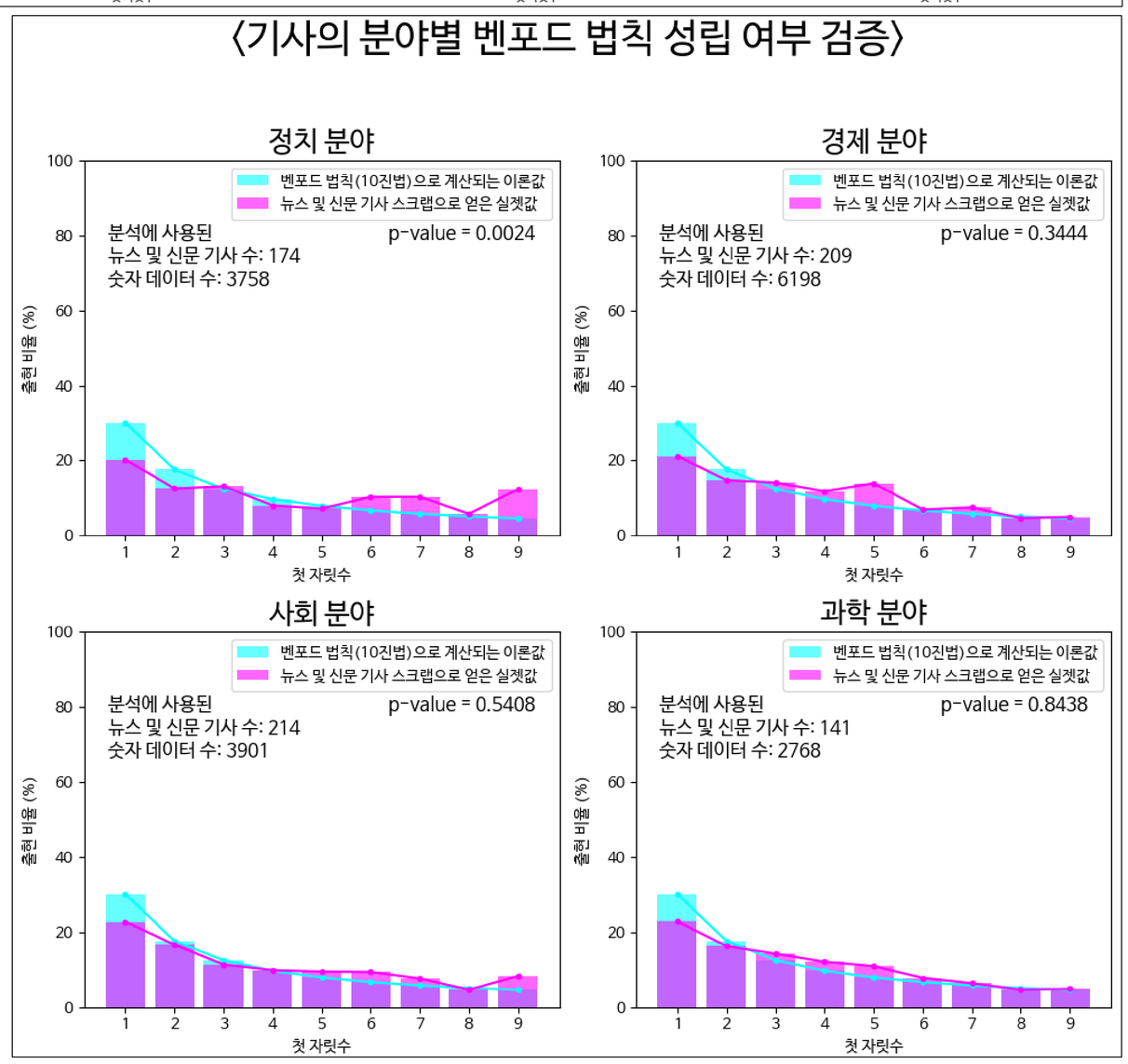
- convert_list_to_bases 함수와 count 함수를 사용해서 각 진법별 숫자의 분포 비율을 계산한 리스트를 만든다.
- math 모듈에서 log 함수를 사용해 각 진법별로 벤포드 법칙을 따르는 이론적 분포 비율을 계산한 리스트도 만든다.
- scipy.stats의 chisquare 모듈로, 실제 분포 비율이 이론적 분포 비율과 얼마나 유사한지 각 진법별로 적합도 검정하여 p-value를 구한다.
- matplotlib.pyplot 모듈을 이용해서, 각 진법별 분포 비율의 실젯값과 이론값을 막대 그래프와 꺾은선 그래프로 나타내고 p-value도 표시해 시각화한다.

4. 4가지 분야(정치, 경제, 사회, 과학)별 벤포드 법칙에 대한 적합도 검정 및 시각화

- 기사의 분야별로 벤포드 법칙의 성립 여부를 검증하는 과정에서는 10진법만 사용한다.
- 3번과 동일한 방법으로 각 분야에 대해 적합도 검정하여 p-value를 구하고, 막대 그래프와 꺾은선 그래프로 시각화한다.

프로젝트 성과





프로젝트 기대효과

가짜뉴스 판별: 벤포드의 법칙 이론값과, 정상적인 뉴스의 실젯값의 차이와, 이론값과 가짜뉴스의 실젯값의 차이를 비교하면, 가짜뉴스의 차이값이 더 큼을 알 수 있는데, 이를 통해 뉴스가 가짜 뉴스인지 아닌지를 판단할 수 있다. 인위적인 데이터를 삽입한다면 벤포드의 법칙이 깨질 것이기 때문이다.

회계부정, 투표 조작, 스포츠 데이터 조작 판별