3.12 *Lab: 도트곱을 사용하여 투표기록 비교하기*

미국 상원의원의 투표기록을 $\mathbb{R}$상의 벡터로서 나타내고 도트곱을 사용하여 투표기록을 비교할 것이다. 이 lab에서는 리스트만 사용하여 벡터를 나타낸다.

여기서는 벡터를 사용하여 상원 의원들의 정치적 성향을 객관적으로 평가할 것이다. 각 의원의 투표기록은 벡터로서 나타낼 수 있으며 이 벡터의 각 원소는 상원 의원이 주어진 입법 현안에 어떻게 투표했는지 나타낸다. 두 의원의 "투표 벡터"의 차이를 비교함으로써 의원들이 어떤 성향을 보이는지 알 수 있다.

사용하는 데이터는 다소 오래된 것이다. 하지만 오바마가 상원 의원으로서 어떤 성향을 보였는지 알아볼 수 있다. 좀 더 최근 데이터를 사용해 보고자 하는 경우 `resources.codingthematrix.com`에 올려놓을 더 많은 데이터 파일을 사용할 수 있을 것이다.

3.12.1 *파일 읽어 들이기*

마지막 lab에서와 같이 lab에서 사용할 정보는 여백으로 구분되는 텍스트 파일에 저장되어 있다. 109대 상원의 투표기록은 `voting_record_dump109.txt` 파일에 들어 있다.

파일의 각 라인은 다른 상원 의원의 투표기록을 나타낸다. 파일에서 데이터를 읽어 들이는 방법을 잊어버린 경우 다음과 같이 하면 된다.

```
>>> f = open('voting_record_dump109.txt')
>>> mylist = list(f)
```

프로시저 `split(·)`을 사용하여 파일의 각 라인을 리스트로 분리할 수 있다. 리스트의 첫 번째 원소는 상원 의원의 이름, 두 번째는 소속 정당, 세 번째는 소속 주, 나머지는 법안들에 대한 의원의 투표기록이다. "1"은 찬성, "−1"은 반대, 그리고 "0"은 기권을 나타낸다.

Task 3.12.1: 프로시저, `create_voting_dict(strlist)`을 작성해 보자. 이 프로시저는 주어진 문자열의 리스트(소스 파일에서 읽어 들인 투표기록)에 대해 상원 의원의 성을 그 의원의 투표기록을 나타내는 리스트에 매핑하는 딕셔너리를 리턴한다. 정수의 문자열 표현(예를 들어, '1')을 실제 정수(예를 들어, 1)로 변경해 주는 내장 프로시저, `int(·)`을 사용할 필요가 있을 것이다.

3.12.2 *도트곱을 사용하여 투표를 비교하기 위한 두 가지 방법*

$\boldsymbol{u}$ 와 $\boldsymbol{v}$는 두 개의 벡터라고 하자. 엔트리들이 모두 1, 0, 또는 -1인 간단한 경우를 생각해 보자. $\boldsymbol{u}$와 $\boldsymbol{v}$의 도트곱은 다음과 같이 정의된다.

$$\boldsymbol{u} \cdot \boldsymbol{v} = \sum_k \boldsymbol{u}[k]\boldsymbol{v}[k]$$

k번째 엔트리를 고려해 보자. 만약 $\boldsymbol{u}[k]$와 $\boldsymbol{v}[k]$ 둘 다 1이면, 위 식의 대응하는 항은 1이다. 만약 $\boldsymbol{u}[k]$와 $\boldsymbol{v}[k]$ 둘 다 -1이면, 대응하는 항은 또한 1이다. 따라서, 위 식의 항이 1이라는 것은 의견의 일치(동의)를 나타낸다. 한편, 만약 $\boldsymbol{u}[k]$와 $\boldsymbol{v}[k]$가 서로 다른 부호이면, 대응하는 항은 -1이다. 따라서, 위 식의 항이 -1인 것은 의견의 불일치를 나타낸다. 만약 $\boldsymbol{u}[k]$와 $\boldsymbol{v}[k]$ 중 어느 하나 또는 둘 다 영이면, 대응하는 항은 0이 되고, 이것은 해당 엔트리가 의견 일치 또는 불일치에 대한 정보를 제공하지 않는다는 것을 의미한다. 그러므로 $\boldsymbol{u}$와 $\boldsymbol{v}$의 도트곱은 $\boldsymbol{u}$와 $\boldsymbol{v}$가 얼마나 일치하는지를 나타내는 척도이다.

3.12.3 *정책 비교*

주어진 두 명의 상원 의원의 마음이 얼마나 일치하는지를 알아보고자 한다. 벡터 $\boldsymbol{u}$와 $\boldsymbol{v}$의 도트곱을 사용하여 얼마나 자주 두 의원의 의견이 일치하는지를 판단할 것이다.

Task 3.12.2: 프로시저, `policy_compare(sen_a, sen_b, voting_dict)`을 작성해 보자. 이 프로시저는 주어진 두 상원 의원의 이름과 투표기록을 나타내는 리스트에 이름을 매핑하는 딕셔너리에 대해 두 의원의 투표정책 사이의 유사도를 나타내는 도트곱을 리턴한다.

Task 3.12.3: 프로시저, `most_similar(sen, voting_dict)`을 작성해 보자. 이 프로시저는 주어진 상원 의원의 이름과 투표기록을 나타내는 리스트에 이름을 매핑하는 딕셔너리에 대해 정치적 성향이 입력한 의원과 가장 유사한 의원의 이름을 리턴한다.

Task 3.12.4: 프로시저, `least_similar(sen, voting_dict)`을 작성해 보자. 이 프로시저는 이름이 `sen`인 의원과 투표기록이 가장 일치하지 않는 의원의 이름을 리턴한다.

Task 3.12.5: 이 프로시저들을 사용하여 어느 상원 의원이 로드 아일랜드의 전설적인 의원 링컨 채피(Lincoln Chafee)와 가장 비슷한지 알아 보자. 그다음에, 이 프로시저들을 사용하여 어느 의원이 펜실베니아의 릭 샌토럼(Rick Santorum)과 가장 일치하지 않는지 찾아보자. 이들의 이름을 리턴한다.

Task 3.12.6: 가장 좋아하는 주 출신의 두 상원 의원들의 투표기록이 얼마나 유사한가?

3.12.4 *평균적 민주당원과의 비교*

Task 3.12.7: 프로시저, `find_average_similarity(sen, sen_set, voting_dict)`을 작성해 보자. 이 프로시저는 주어진 상원 의원의 이름 `sen`에 대해 이 의원의 투표기록을 `sen_set`에 있는 모든 의원들의 투표기록과 비교하여 각각에 대해 도트곱을 계산하여 평균 도트곱을 리턴한다.

작성한 프로시저를 사용하여 어느 의원이 민주당(Democrats)의 집합(입력 파일에서 이 집합을 추출할 수 있음)과 평균 유사도가 가장 높은지 계산해 보자.

마지막 Task에서는 각 의원의 기록을 민주당 상원 의원 각각의 투표기록과 비교해야 한다. 만약 동일한 계산을 모든 넷플릭스(Netflix) 가입자들의 영화에 대한 선호도에 대해 수행한다면 너무 시간이 오래 걸려 실용적이지 않을 것이다.

다음으로, 도트곱의 대수적 성질, 즉 *분배성*을 사용하면 계산을 간편화할 수 있음을 알아볼 것이다.

$$(v_1 + v_2) \cdot x = v_1 \cdot x + v_2 \cdot x$$

이다.

$$(\boldsymbol{v}_1 + \boldsymbol{v}_2) \cdot \boldsymbol{x} = \boldsymbol{v}_1 \cdot \boldsymbol{x} + \boldsymbol{v}_2 \cdot \boldsymbol{x}$$

Task 3.12.8: 프로시저, `find_average_record(sen_set, voting_dict)`을 작성해 보자. 이 프로시저는 상원 의원들의 이름으로 구성된 주어진 집합에 대해 평균 투표기록을 찾는다. 즉, 의원들의 투표기록을 나타내는 리스트에 대해 벡터 덧셈을 수행하고, 그다음에 벡터 덧셈의 결과인 합을 벡터들의 개수로 나눈다. 결과는 벡터이다.

이 프로시저를 사용하여 민주당의 집합에 대한 평균 투표기록을 계산하고 그 결과를 변수 `average_Democrat_record`에 할당하자. 다음에, 어느 의원의 투표기록이 평균 민주당의 투표기록과 가장 유사한지 찾아보자. Task 3.12.7와 동일한 결과를 얻었는가? 이유를 설명할 수 있는가?

3.12.5 *최대 경쟁자*

Task 3.12.9: 의견 일치가 가장 안 되는 두 상원 의원이 누구인지 찾는 프로시저 `bitter_rivals(voting_dict)`을 작성해 보자.

이 일을 위해서는 투표기록의 각 쌍을 비교해야 한다. 이것을 누구나 아는 뻔한 방식보다 더 빠르게 할 수는 없을까? *빠른 행렬곱*을 사용하는 조금 더 효율적인 알고리즘이 있다. 행렬곱에 대해서는 나중에 살펴볼 것이다. 그렇지만 이론적으로 빠른 알고리즘에 대해서는 다루지 않을 것이다.

3.12.6 *개방형 연구*

현대 정치의 보지적한 [illegible]