C++

문자 인식 알고리즘과 이를 적용한 오버워치 핵 사용자 국적(문화권) 구분 관련 개발

최종 보고서

제출일자: 2023 – 12 - 24
제출자명: 한동호
제출자학번:203734

1. 프로젝트 목표

1) 배경 및 필요성

세계화 시대인 것을 감안했을 때, 사용자의 국적은 다양한 목적에 사용될 수 있는 중요한 정보입니다. 예를 들어, 사용자의 언어 선호도, 문화적 배경, 구매 성향 등을 파악하는 데 사용할 수 있습니다. 그러나 특정한 데이터에는 사용자의 국적이 포함되어 있지 않을 수도 있습니다. 국적이 포함되어 있지 않은 경우, 사용자의 이름, 이메일 주소, IP 주소 등의 정보를 기반으로 국적을 추정할 수 있습니다. 이중 사용자의 이름을 이용해 국적을 분석하는 알고리즘을 개발하고, 이를 통해 해당 이용자의 국적을 추정을 자동화하는 프로그램을 만들면 추후 한국인과 외국인들이 섞인 명단을 다룰 때 업무자동화에 큰 역할을 할 것을 예상합니다.

이를 적용하여, 10000개 이상의 (아시아 서버 기준) 오버워치 핵 사용자 명단 분석을 자동화하고 국적(문화권)을 조사하여 가장 핵을 많이 쓰는 국가(문화권)가 어딘지 추정해보는 개인적인 호기심 또한 해결해보고자 합니다.

2) 프로젝트 목표

한국어와 외국어를 분류하여 국적을 추정하는 알고리즘을 개발합니다.

알고리즘을 활용해 한국인과 외국인이 뒤섞인 특정 명단(오버워치 핵 사용자 명단)을 분석하는 작업을 자동화 시킵니다.

얻어낸 국적 및 문자 관련 데이터를 토대로 어떤 나라(문화권)에서 가장 많이 핵을 사용했는지 도식화 시킵니다.

3) 차별점

기존 프로그램들은 이름이 일부만 공개되고 나머지 부분이 특수문자로 처리된 경우엔 언어(문자) 분석이 원활하게 이뤄지지 않는 경우가 있었습니다. 이를 극복하고 분석해내는 알고리즘을 개발할 것입니다. 또한 이렇게 얻어낸 데이터를 도식화 시켜 나타내는 기능을 추가해 직관적인 수치 분석이 가능하게 만들 것입니다

2. 기능 계획

1) 기능 1

- 문자를 통한 언어 인식기능

(1) 세부 기능 1

- 중국어는 한자를 쓰지만 일본어는 한자와 가나를 혼용하므로 가나가 하나라도 포함된 유저 데이터는 일본어(가나)로 구분하는 기능 추가,

2) 기능 2

- 언어 인식기능을 이용한 데이터 분석 및 수치화 기능

(1) 세부 기능 1

- 유저 이름 데이터의 일부가 특수문자로 가려진 경우를 감안하여 데이터를 가공하는 기능 추가

2) 기능 3

- 얻어낸 데이터를 도식화해주는 기능(콘솔창으로 구현)

3. 진척사항

1) 기능 구현

1) 기능 1

- 문자를 통한 언어 인식기능

2) 기능 2

- 언어 인식기능을 이용한 데이터 분석 및 수치화 기능

(1) 세부 기능 1

- 유저 이름 데이터의 일부가 특수문자로 가려진 경우를 감안하여 데이터를 가공

```
// 언어 분석(기능1+2)
     if (std::regex_match(std::string(1, person.name.front()), std::regex("[가-힣]"))) {
         person.language = "한국인";
         koreanCount++;
     }
     else if (std::regex_match(std::string(1, person.name.front()), std::regex("[₩₩x4e00-
₩₩x9fa5]"))) {
         person.language = "중국인 + 일본인";
         chineseCount++;
     }
     else {
         person.language = "미국인";
         englishCount++;
     }
```

} 하는 기능 추가

2) 기능 3

- 얻어낸 데이터를 도식화해주는 기능(콘솔창으로 구현)

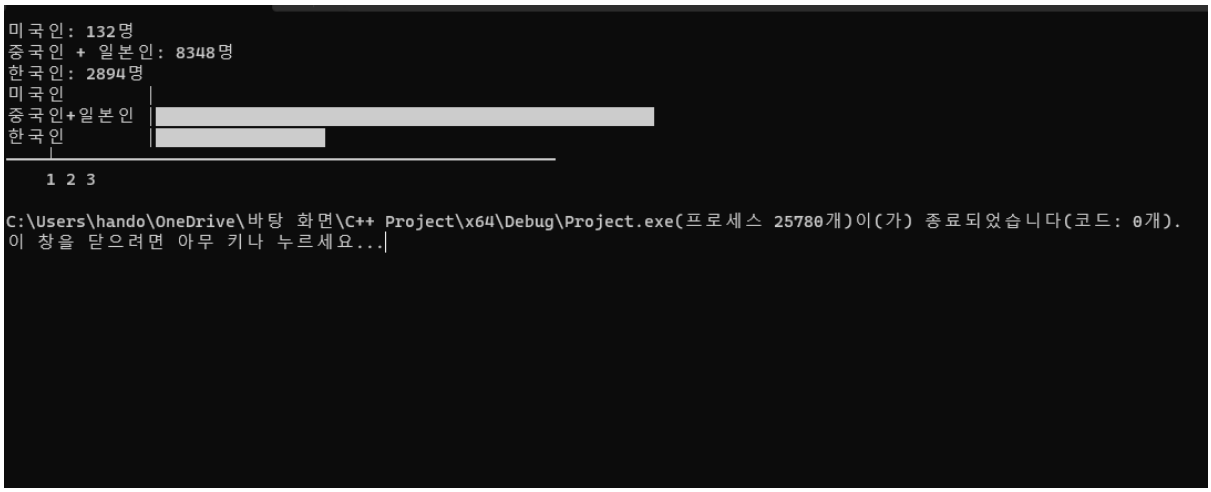
```
void drawGraph(const std::vector<int>& values, const std::vector<std::string>& labels) {
    int max_value = *std::max_element(values.begin(), values.end());

    for (int i = 0; i < values.size(); i++) {
        std::cout << labels[i] << "  |";
        int bar_length = values[i] * 50 / max_value; // 막대의 길이를 조정하기 위해 비율 계산

        for (int j = 0; j < bar_length; j++) {
            std::cout << "";
        }
        std::cout << '₩n';
    }
```

- 적용된 배운 내용 => 반복문, 조건문, 배열, 파일 입출력, 구조체, 문자열

2) 테스트 결과



중국인이 가장 많은 비율을 차지 할꺼라고 예상했고 실제로도 그렇긴 하지만, 한국인도 은근히 핵을 많이 쓴다는 사실을 알게됨

4. 계획 대비 변경 사항

일본어와 중국어는 공용으로 사용하는 한자들이 많아 통합하여 통계를 내기로 함

3. 프로젝트 일정