

C++

문자 인식 알고리즘 과 이를 적용한 오 버워치 핵 사용자 국적(문화권) 구분 관련 개발

진척 보고서 #1

제출일자: 2023 - 11 - 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 극복하

고 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) 구현한 기능 이름

- 한글, 알파벳, 중국어, 일본어 구별 알고리즘
- 설명 - 구현은 했으나 인코딩 문제로 추후 수정할 계획
- 적용된 배운 내용 => 반복문, 조건문, 문자열, 배열

```
텍스트를 입력하세요: hi hello  
The input text is in English.
```

```
텍스트를 입력하세요: 한국어  
입력한 텍스트는 한국어입니다.
```

중국어, 일본어 구별 알고리즘을 만들고 출력을 따로 해봤더니 글자가 깨짐, 해당 문자는 한국어로 인식되는 문제 발생

```
텍스트를 입력하세요: されたテキストは日  
입력한 텍스트는 한국어입니다.
```

```
텍스트를 입력하세요: 本是中文  
입력한 텍스트는 한국어입니다.
```

2) 테스트 결과

(1) 한국어, 중국어, 일본어, 영어 구별 알고리즘

=> 인코딩 문제로 인한 결과물 출력시 글자 깨짐 및 인식장애

-(한국어와 영어는 구별이 가능, 일본어와 중국어는 불가능 및 글자깨짐)

=>알고리즘에는 문제가 없음을 확인, 다음 일정 전까지 수정 계획

4. 계획 대비 변경 사항

1) 도식화 관련 기능

- 도식화 기능에 대해 어떻게 구현할 것인지 명시하지 않았는데, 이는 수업에서 배우지 않는 내용임을 뒤늦게 인지하고 해당 기능을 콘솔로 구현예정

3. 프로젝트 일정

TASKS	11/3	11/15	11/22	11/29	12/6	12/13	12/22
제안서							
기능 1							
세부기능(기능1)							
기능 2							
세부기능 (기능 2)							
기능 3							
코드 정리 및 마무리							