

[C++] 뉴스 클러스터링(난이도: 중)

2018년 3월 20일에 게시됨

5. 뉴스 클러스터링(난이도: 중)

여러 언론사에서 쏟아지는 뉴스, 특히 속보성 뉴스를 보면 비슷비슷한 제목의 기사가 많아 정작 필요한 기사를 찾기가 어렵다. Daum 뉴스의 개발 업무를 맡게 된 신입사원 튜브는 사용자들이 편리하게 다양한 뉴스를 찾아볼 수 있도록 문제점을 개선하는 업무를 맡게 되었다.

개발의 방향을 잡기 위해 튜브는 우선 최근 화제가 되고 있는 "카카오 신입 개발자 공채" 관련 기사를 검색해보았다.

- 카카오 첫 공채.. '블라인드' 방식 채용
- 카카오, 합병 후 첫 공채.. 블라인드 전형으로 개발자 채용
- 카카오, 블라인드 전형으로 신입 개발자 공채
- 카카오 공채, 신입 개발자 코딩 능력만 본다
- 카카오, 신입 공채.. "코딩 실력만 본다"
- 카카오 "코딩 능력만으로 2018 신입 개발자 뽑는다"

기사의 제목을 기준으로 "블라인드 전형"에 주목하는 기사와 "코딩 테스트"에 주목하는 기사로 나뉘는 걸 발견했다. 튜브는 이들을 각각 묶어서 보여주면 카카오 공채 관련 기사를 찾아보는 사용자에게 유용할 듯싶었다.

유사한 기사를 묶는 기준을 정하기 위해서 논문과 자료를 조사하던 튜브는 "자카드 유사도"라는 방법을 찾아냈다.

자카드 유사도는 집합 간의 유사도를 검사하는 여러 방법 중의 하나로 알려져 있다. 두 집합 A, B 사이의 자카드 유사도 $J(A, B)$ 는 두 집합의 교집합 크기를 두 집합의 합집합 크기로 나눈 값으로 정의된다.

예를 들어 집합 $A = \{1, 2, 3\}$, 집합 $B = \{2, 3, 4\}$ 라고 할 때, 교집합 $A \cap B = \{2, 3\}$, 합집합 $A \cup B = \{1, 2, 3, 4\}$ 이 되므로, 집합 A, B 사이의 자카드 유사도 $J(A, B) = 2/4 = 0.5$ 가 된다. 집합 A 와 집합 B 가 모두 공집합일 경우에는 나눗셈이 정의되지 않으니 따로 $J(A, B) = 1$ 로 정의한다.

자카드 유사도는 원소의 중복을 허용하는 다중집합에 대해서 확장할 수 있다. 다중집합 A 는 원소 "1"을 3개 가지고 있고, 다중집합 B 는 원소 "1"을 5개 가지고 있다고 하자. 이 다중집합의 교집합 $A \cap B$ 는 원소 "1"을 $\min(3, 5)$ 인 3개, 합집합 $A \cup B$ 는 원소 "1"을 $\max(3, 5)$ 인 5개 가지게 된다. 다중집합 $A = \{1, 1, 2, 2, 3\}$, 다중집합 $B = \{1, 2, 2, 4, 5\}$ 라고 하면, 교집합 $A \cap B = \{1, 2, 2\}$, 합집합 $A \cup B = \{1, 1, 2, 2, 3, 4, 5\}$ 가 되므로, 자카드 유사도 $J(A, B) = 3/7$, 약 0.42가 된다.

이를 이용하여 문자열 사이의 유사도를 계산하는데 이용할 수 있다. 문자열 "FRANCE"와 "FRENCH"가 주어졌을 때, 이를 두 글자씩 끊어서 다중집합을 만들 수 있다. 각각 {FR, RA, AN, NC, CE}, {FR, RE, EN, NC, CH}가 되며, 교집합은 {FR, NC}, 합집합은 {FR, RA, AN, NC, CE, RE, EN, CH}가 되므로, 두 문자열 사이의 자카드 유사도 $J(\text{"FRANCE"}, \text{"FRENCH"}) = 2/8 = 0.25$ 가 된다.

입력 형식

- 입력으로는 `str1`과 `str2`의 두 문자열이 들어온다. 각 문자열의 길이는 2 이상, 1,000 이하이다.

- 입력으로 들어온 문자열은 두 글자씩 끊어서 다중집합의 원소로 만든다. 이때 영문자로 된 글자 쌍만 유효하고, 기타 공백이나 숫자, 특수 문자가 들어있는 경우는 그 글자 쌍을 버린다. 예를 들어 "ab+"가 입력으로 들어오면, "ab"만 다중집합의 원소로 삼고, "b+"는 버린다.
- 다중집합 원소 사이를 비교할 때, 대문자와 소문자의 차이는 무시한다. "AB"와 "Ab", "ab"는 같은 원소로 취급한다.

출력 형식

입력으로 들어온 두 문자열의 자카드 유사도를 출력한다. 유사도 값은 0에서 1 사이의 실수이므로, 이를 다루기 쉽도록 65536을 곱한 후에 소수점 아래를 버리고 정수부만 출력한다.

예제 입출력

str1	str2	answer
FRANCE	french	16384
handshake	shake hands	65536
aa1+aa2	AAAA12	43690
E=M*C^2	e=m*c^2	65536

■ C++ code

```
#include <iostream>
#include <map>
#include <algorithm>

using namespace std;

int main()
{
    int size1, size2, ans = 65536;
    double all = 0, dup = 0;
    string str1, str2;
    map<string, int> map1;

    getline(cin, str1);
    getline(cin, str2);

    transform(str1.begin(), str1.end(), str1.begin(), ::tolower);
    transform(str2.begin(), str2.end(), str2.begin(), ::tolower);

    size1 = str1.size();
    size2 = str2.size();

    for (int i = 0; i < size1-1; i++)
    {
        if (str1[i] < 97 || str1[i] > 122) continue;
        if (str1[i+1] < 97 || str1[i+1] > 122) continue;

        string str;
        str = str1[i];
        str += str1[i+1];
        all++;

        if (map1.count(str))    map1[str]++;
        else    map1[str] = 1;
    }

    for (int i = 0; i < size2-1; i++)
    {
        if (str2[i] < 97 || str2[i] > 122) continue;
        if (str2[i+1] < 97 || str2[i+1] > 122) continue;
```

```
string str;
str = str2[i];
str += str2[i+1];

if (map1.count(str))
{
    if (map1[str] > 0)
    {
        map1[str]--;
        dup++;
    }
    else    all++;
}
else
    all++;

if (all != 0)
    ans = (int)((double)ans * (dup / all));
cout << ans;

return 0;
}
```

광고

REPORT THIS AD

REPORT THIS AD

이 글 공유하기:

Twitter

Facebook

관련

[C++] 프렌즈4블록(난이도: 상)
"기초 개발실력 다지기"에서

[C++] 문제3. 파일명 정렬
"기초 개발실력 다지기"에서

[C++] 문제2. 압축
"기초 개발실력 다지기"에서

이 글은 [기초 개발실력 다지기](#) 카테고리에 분류되었고 [C++](#), [카카오](#), [Kakao](#) 태그가 있습니다. [고유주소](#) 북마크.