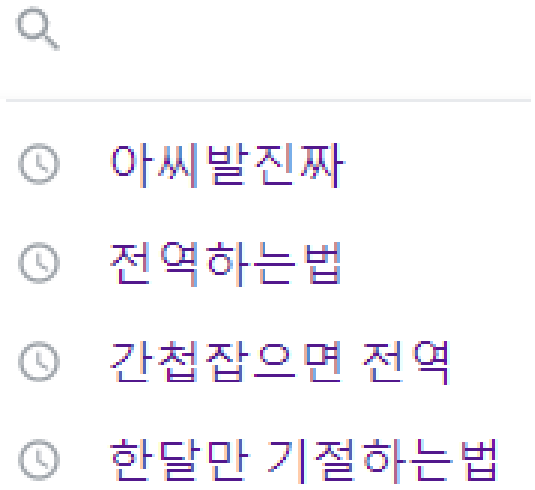
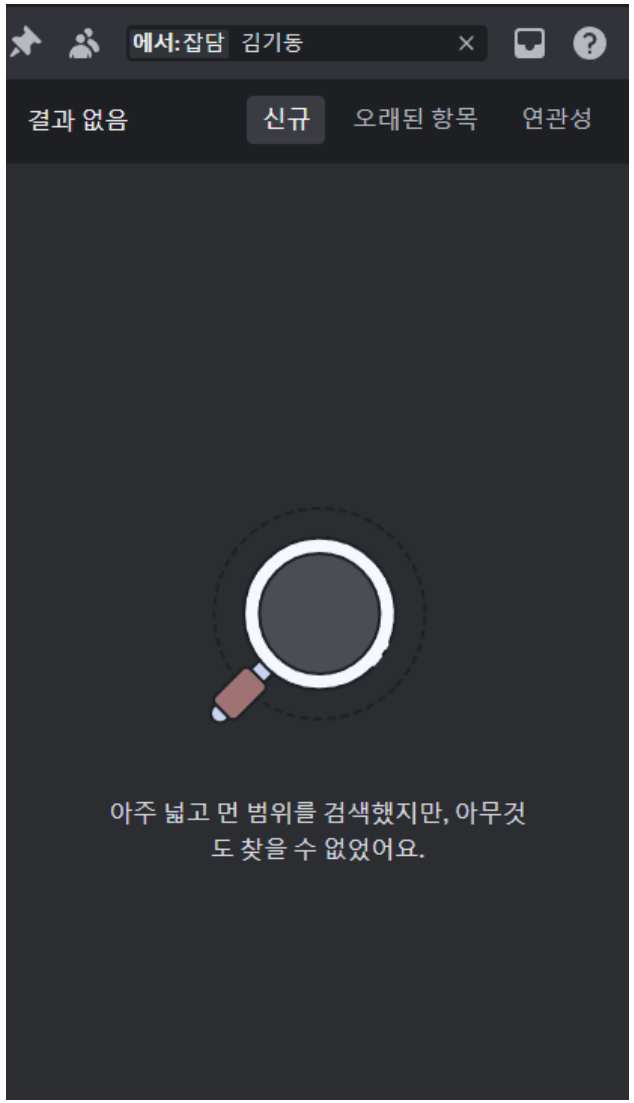
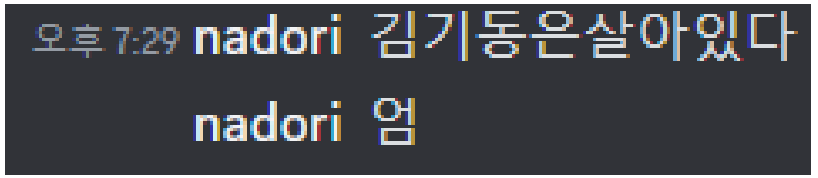
왜 Discord 검색 기능은 이따구인가??



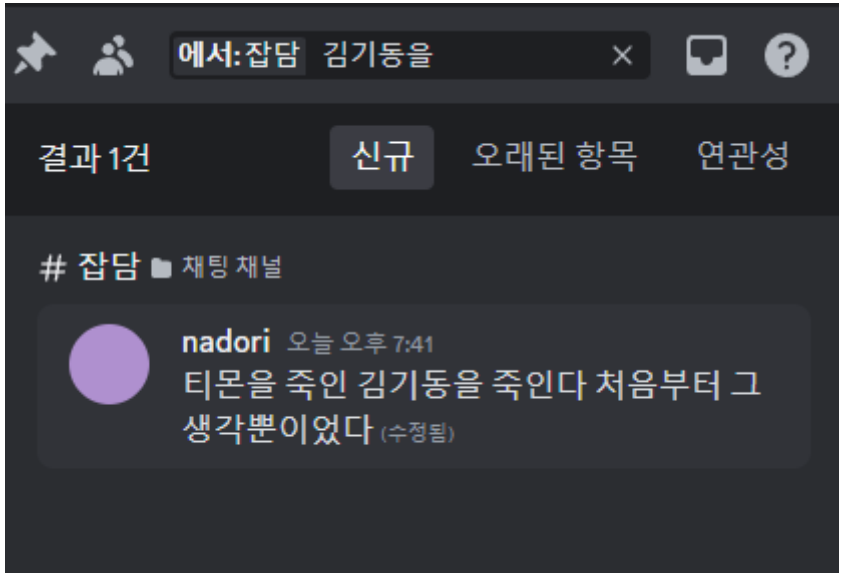
디코에서 한 대화의 키워드로 검색을 하고 싶을 때가 있다.
요즘 무슨 기술이 핫한지
김기동은 살아있는지
소다는 전역을 했는지 등등...



도키도키한 마음으로 검색했을 때
나를 반기는 것은
노력했다는 따뜻한 말 한마디뿐...



물론 김기동에 대한 채팅은 1e9개 넘게 존재한다.
그 중 하나도 식별하지 못한 디코에게
험한말을 참는 것은 쉽지 않다.



엥 근데 어절 단위로 검색하니까
제대로 찾아낸다.

격조사	주격	이/가, 께서
	보격	이/가
	관형격	의
	목적격	을/를/*ㄹ
	부사격	같이, 까지, 께, 만큼, 보다, 에, 에게, 에게로, 에게서, 에는, 에다가, 에도, 에서/*서, (으)로, (으)로부터, (으)로서, (으)로써, 처럼, 하고, 한테, 한테서 등
	인용격	고(간접인용), 라고(직접인용)
	서술격	이다
	호격	아/야
보조사		까지, (이)나, 은/는/*ㄴ, 대로, 도, (이)라도, 마다, 마저, 만, 밖에, 부터, 뿐, (이)나마, (이)든가, (이)든지/든, (이)야, (이)야말로, 요, 조차, (ㄴ/는/은)커녕
접속조사		고, 와/과, (이)며, (이)나, (이)랑, 하고
조사 결합형		고는, 고도, 까지나, 까지는/*까진, 까지의, 께서는/*께선, 께서도, (으)로는, (으)로도, (으)로서의, 만으로, 만으로는/*만으론, 만의, 보다는, 보다도, 뿐만, 에게까지, 에게나, 에게는/*에겐, 에게만, 에만, 에서는/*에선/*서는, 에서도/*서도, 와의/과의, 와는/과는, 하고도, 한테도, 한테는 등

Bilingual Research Volume 2014, Issue57, p223~255, Dec 2014

ㅇㅋ 그러면 우리는
궁금한 명사에 대해 검색하기 위해
이 표를 참고해서 다 검색을 하면 된다.

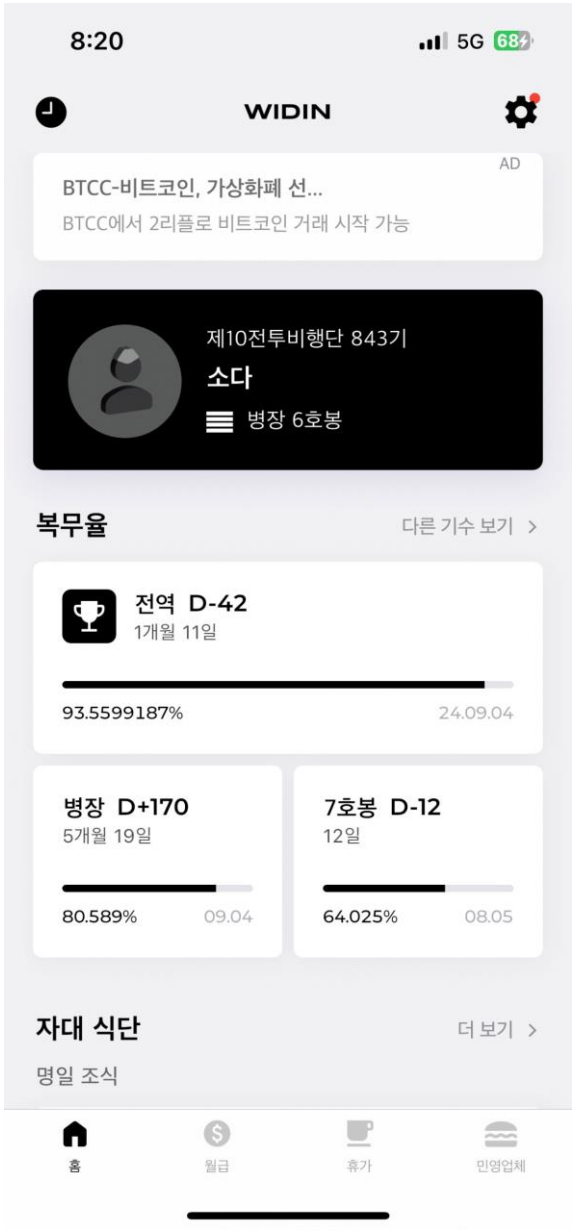
물론 명사는 선녀다.

나무위키

목차

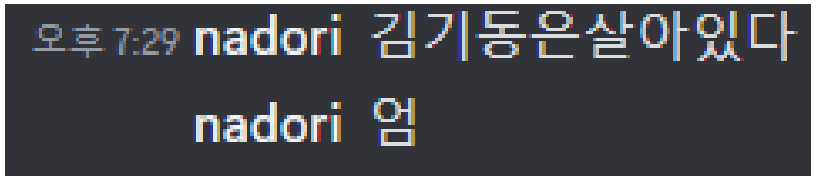
모든 분야를 전문가인 것 처럼 알려주는
나쌤의 목차를 봐보자.

한국어는 우리팀 탑이 부러워 할 정도로
어미가 많다.



물론 금부자인 소다 the 말년병장같은
사람은 할 게 없어서
다 검색해 볼 수 있지만,

일반인의 신분으론 쉽지 않다.



그러면 대체 왜 디코는 어절 단위로 검색을 하게 만드는걸까..
대체 무슨 이유일까..
메시지를 어떤 방식으로 검색하는지 알아보자.

디코는 기본적으로 fuzzy string search를 사용한다.
이게 뭔지 위키피디아님께 여쭤보면

컴퓨터 과학에서 근사 문자열 매칭(흔히 퍼지 문자열 검색이라고 도 함)은 패턴과 대략적으로 일치하는 문자열을 찾는 기술입니다 (정확하게 일치하는 것이 아님).

한줄요약: 네입버 -> 네이버를 찾으셨나요?

최소 편집

Gold III

Time Limit	Memory Limit	Submissions	Accepted	Solved	Ratio
2 seconds	512 MB	3089	1659	1339	56.426%

Description

두 문자열 A와 B가 주어졌을 때, A에 연산을 최소 횟수로 수행해 B로 만드는 문제를 "최소 편집" 문제라고 한다.

A에 적용할 수 있는 연산은 총 3가지가 있으며 아래와 같다.

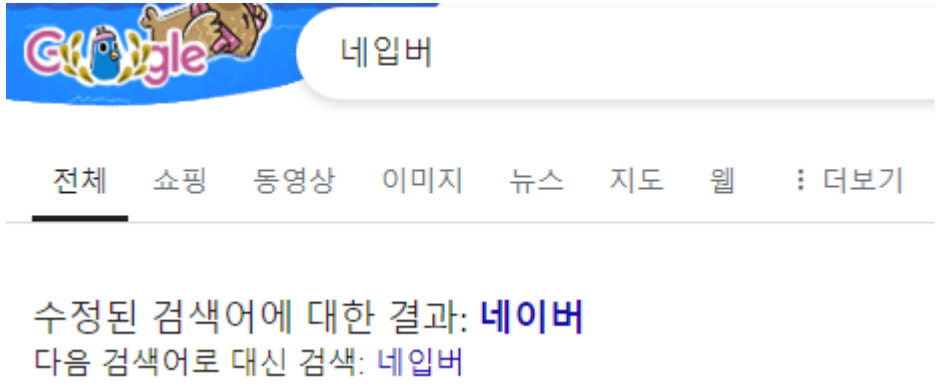
1. 삽입: A의 한 위치에 문자 하나를 삽입한다.
2. 삭제: A의 문자 하나를 삭제한다.
3. 교체: A의 문자 하나를 다른 문자로 교체한다.

두 문자열이 주어졌을 때, 최소 편집 횟수를 구하는 프로그램을 작성하시오.

Fuzzy search의 기초적인 알고리즘으로는
삽입, 삭제, 교체 연산으로
문자열 S -> T로의 거리가
일정 이하면 유사하다고 판정하는 것이 있다.

물론 가중치를 주고 이리 저리 비빔밥을
하기 때문에 이리 간단하지는 않다.

그러면 Fuzzy search의 장점이 뭘까?

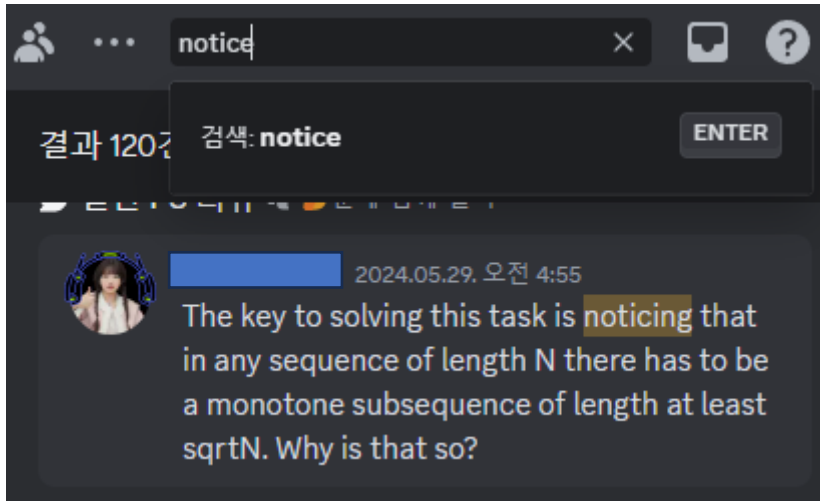


검색어의 오타 정정

https://www.lgcns.com/blog/cns-tech/aws-ambassador/52009

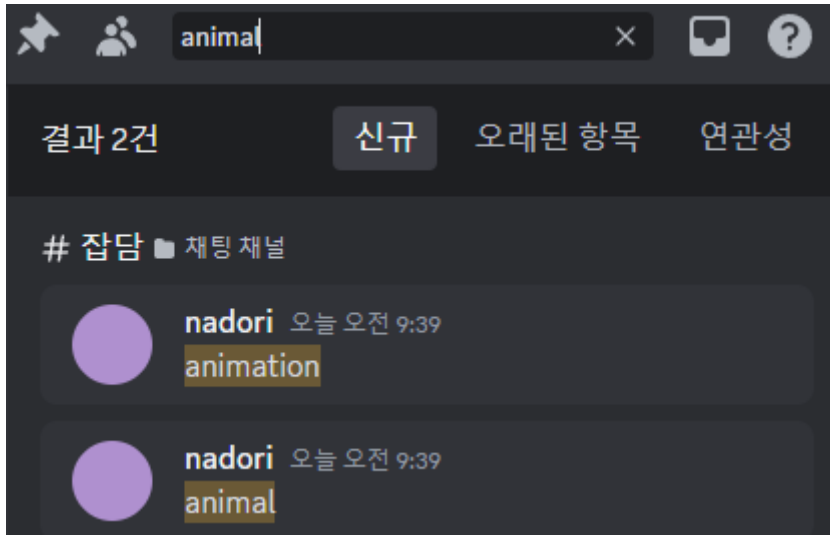
표기의 흔들림 정정

그러면 Fuzzy search의 장점이 뭘까?

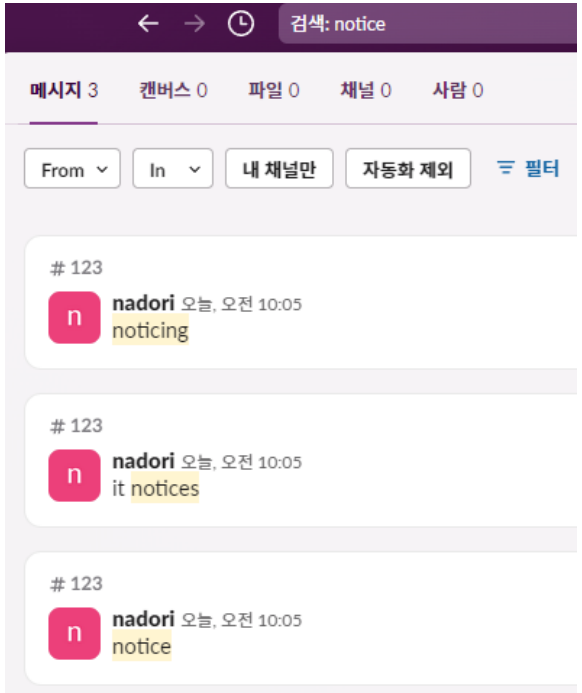


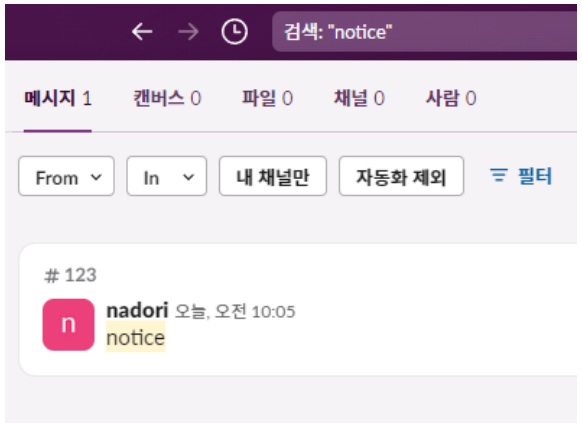
영어권 goat

그러면 Fuzzy search의 단점이 뭘까?



?





사실 exact search 기능 하나만 넣어줘도
편의성 측면에서 압도한다.

그리고 텔레그램 빼고 다른 메신저들은 다 밥먹듯이
하고있다.

그럼 exact search를 구현하면 되잖아.
왜 안함??

얘네도 안하고싶은 건 아닐거고
단어 단위로 관리하는 것이 훨씬 가볍기 때문이다.

예를 들어
내가 '서면'을 검색했는데
서면을 단어로 처리를 했으면
이런 검색 결과만 나오겠지만

'서면'을 그냥 무언가의 substring으로 본다면
멈춰서면, 다가서면 이런 결과도 같이 뽑아낼 것이다.

물론 그 과정에서 여태 있던 모든 메시지의 substring
에 '서면'이 존재하는지 검사하기도 할 테고.

메시지의 포함된 단어들을 적절히 색인하고,

사전에 존재하는 단어들의 변화형을 적절히 전처리한
다면 효율적인 관리가 가능한 것이다.

그러면 이제 다른 서비스와의 갈드컵을 해보자.

카톡은 fuzzy search는 안하지만 exact search를 제대로 해주잖아
둘다 똑같이 메신저인데 왜 얘넨 해주고 디코는 못함?

카톡은 메시지 저장을 유저한테 짬때리니까 더 쉽죠
애초에 30일 지나면 메시지 불러오기가 안되잖앙

디코는 새로운 기기로 접속해도 이전 메시지 정보가 다 남아있어요
그리고 백베에서 검색한번하면 ㅈㄴ오래걸림

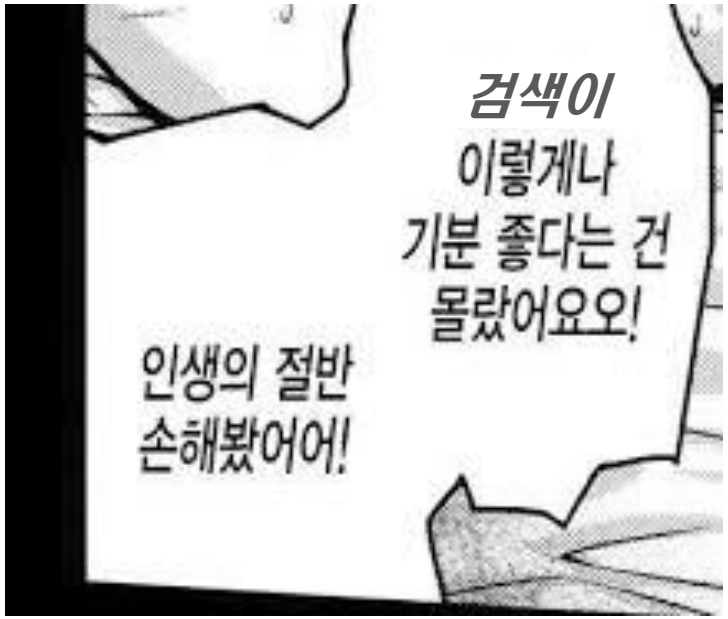
구글 검색은 둘 다 해주는데
왜 얘넨 해주고 디코는 못함?

뇌피셜인데
구글은 채팅방이 여러 개가 아니잖아요.
Ps적으로 substring을 위해 트라이 하나로 관리한다고 가정해보면

채팅방 1에서 asdsgfs이라고 치고
채팅방 2에서 asdsg라고 검색하면 검색이 안나와야하잖삼
그냥 관리가 빡셈

그래서 결론이 뭐냐?

이거 뭐 어쩔수없네...
각 메신저의 장단점으로 봐야하지 않을까?

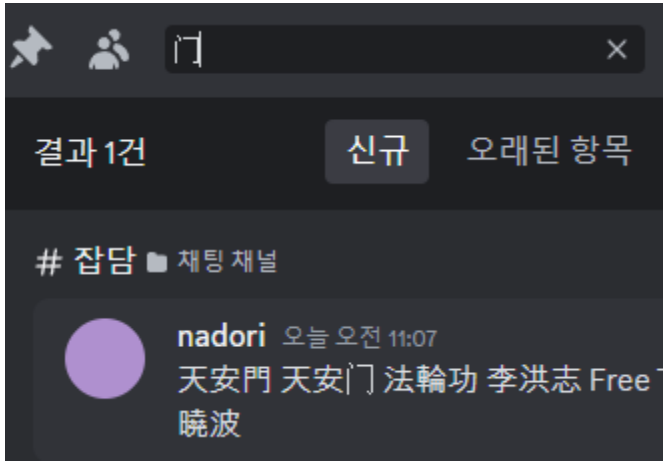


이렇게 끝내면 아쉬우니 어떤 언어권이 제일 손해봤는지 알아보자.

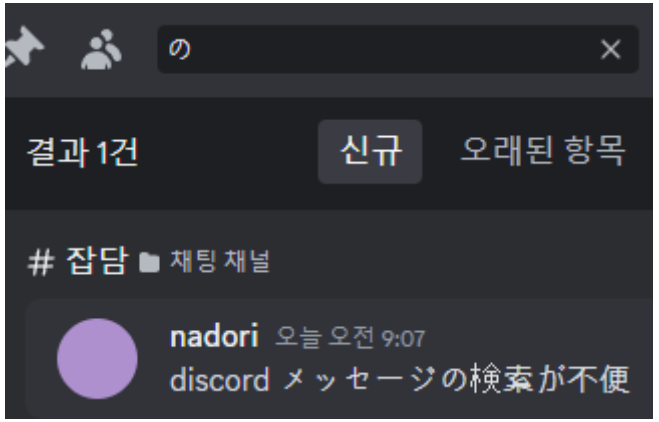
일단 띄어쓰기 안하는 일본, 중국은 손해좀 봐야지?


不便
결과 1건
신규
오래된 항목
잡담
채팅 채널
nadori 오늘 오전 9:07
discord メッセージの検索が不便

?


门
결과 1건
신규
오래된 항목
잡담
채팅 채널
nadori
오늘 오전 11:07
天安門 天安门 法輪功 李洪志 Free
曉波

??


の
결과 1건
신규
오래된 항목
잡담
채팅 채널
nadori 오늘 오전 9:07
discord メッセージの検索が不便

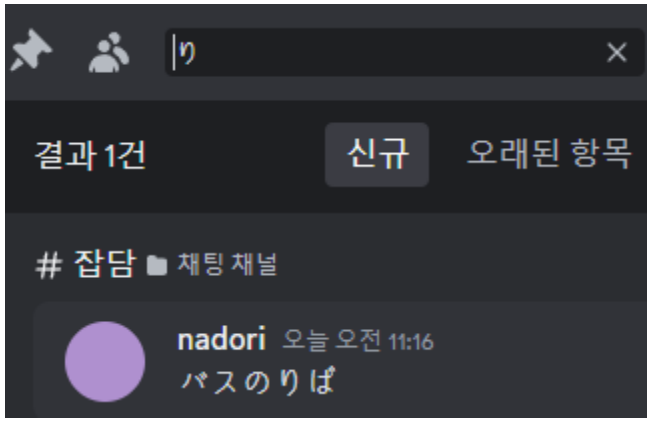
???

그렇다...

한자는 표의문자이기 때문에 1한자 = 1단어로 취급한다.
통용되는 한자가 약 7000개라 관리측에서도 별 문제가 없다.

이미 사전에 있으니까
힝스티비 이런걸 안쳐대니까
글자 하나에 의미가 있으니까
색인하기 훨씬 좋은 것 아닐까

아니 순수 한자인 중국어까진 ㅇㅈ하는데
왜 일본어는 히라가나까지 단독 검색이 되는건가요?



일단 히라가나는 보통 단독으로 잘 쓰이지 않고
보통 조사나 어미로만 사용되니 그런 것 같기도 하고...
히라가나(48)가 한글(11172)보다 훨씬 적은 것도 있는 것 같고...

한글도 이렇게 하면 안되나요?

일단 표음문자다 보니까 저렇게 하는게 훨씬 불편하고
위 사진을 보면 노바리라고 검색했는데 노리바가 나온다.
표의문자면 괜찮지만 표음문자라면
입시미술을 검색했는데 미시입술이 나올수도 있기에
득보단 실이 크다.

결론

또 해내셨습니까... GOAT

미국이 부러워하고 중국이 벌벌떨며 일본이 시기하는 한국
디코 편의성 뒤에서 1위 달성!