자연어 처리 가이드

by. 김우빈

이 가이드는..

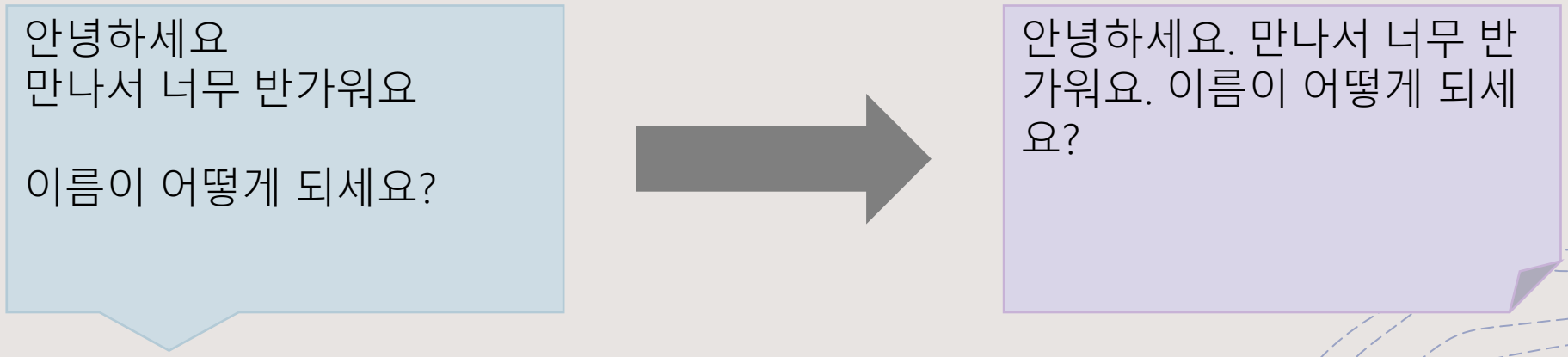
- 셀비스휴 프로젝트의 "감정 분석 인공지능 모델" 개발 과정 중 여러 사이트에서 수집한 익명성 고민 글을 전처리 하며 정리한 내용입니다.
- 자연어 처리에 관심있으신 분, 또는 validation(검사기) 개발을 고민하시는 분이 참고하시면 좋겠습니다.
- 자연어 처리 프로세스에 대한 간단한 정리입니다.

자연어 처리 프로세스

1. 공백문자 처리
2. URL 처리
3. 이메일 처리
4. 전화번호 처리
5. 반복 글자/단어 처리
6. 구두점 처리
7. 사이트별 시그니처 처리
8. 욕설 처리
9. 필요 문자 범위로 필터링
10. 맞춤법 교정기 활용

1. 공백문자 처리

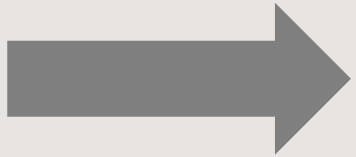
+ 데이터 전처리에 있어 제일 우선적으로 하는 처리입니다.
+ Newline(₩n), Tab(₩t)와 같은 공백문자(Whitespace Character)를 처리합니다.
+ 필요시 그 외의 공백문자를 제거합니다. 하지만 이는 9번 프로세스 "필요 캐릭터 범위로 필터링" 적용 중에 해결됩니다.
+ ₩n 처리는 완벽하게 처리하기에는 어렵습니다. 문장 사이에 있을 수도 있으며, 구두점(.) 대신에 사용될 경우도 있습니다. 일반적으로 ₩n 있는 경우 구두점(.)으로 대체해줍니다.



2. URL 처리

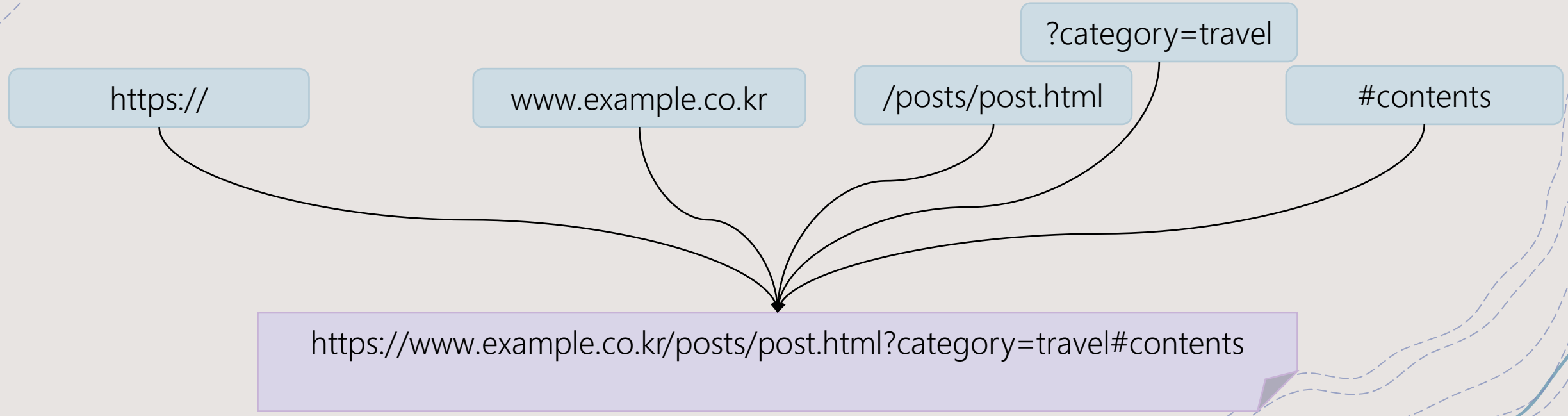
+ http://, https://, www. 로 시작하는 URI(URL)을 처리합니다.
+ 특정 문자/단어/기호 제거하기 전, 데이터 전처리 초기 단계에 우선적으로 하는 것이 좋습니다.
+ 필요시 그 외의 스킴(scheme)을 처리합니다 (예: ftp://, socket://).
+ http://, https://, www 가 없는 URL (예: google.com) 은 domain extension(예: .com, .co.kr, .gov)을 모두 체크하지 않는 이상 처리하지 않습니다. 잘 못 처리되는 경우, 띄어쓰기가 잘못된 글들이 제거됩니다 (not.a.link ← 제거됨).
+ 필요시 IPv4 / IPv6 형식의 주소를 처리하되, 금액(1,000, 1.000, 12.34), 번호(010.1111.2222) 등 숫자 관련 처리와 겹치지 않게 주의해야 합니다.
+ URL 구조는 다음 슬라이드 참고
+ URL 처리 시 고려사항 목록은 다다음 슬라이드 참고

http://example.com에 접속해서 쿠폰번호 #123 입력하시면 되요.



link에 접속해서 쿠폰번호 #123 입력하시면 되요.

2.1 URL 처리 : URL 구조



2.2 URL 처리 : 고려사항

- URL에 사용 가능/불가능한 특수 기호들이 있음으로 정규표현식(Regex) 사용 시 적절히 적용해야 합니다 (RFC 3986 표준 참고).
- URL은 ASCII 문자열로만 전송이 되기 때문에, 대부분의 특수기호 (또는 영어 외 다른 언어 문자)는 % 와 2개의 16진수로 표현이 됩니다. 예를 들어, 플러스(+)는 URL에서 %20 으로 표현됩니다 (한국어== %ED%95%9C%EA%B5%AD%EC%96%B4).
- Query에서 공백( )은 자동으로 플러스(+)로 대체됩니다. Path에서는 공백( )이 사용될 수 없으며, 플러스(+)로도 대체될 수가 없습니다.
- Path에서 . 또는 .. 은 redirect로 사용될 수가 있습니다 (스택 참고).

Unsafe:

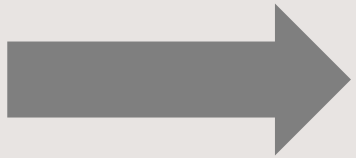
Characters can be unsafe for a number of reasons. The space character is unsafe because significant spaces may disappear and insignificant spaces may be introduced when URLs are transcribed or typeset or subjected to the treatment of word-processing programs. The characters "<" and ">" are unsafe because they are used as the delimiters around URLs in free text; the quote mark (""") is used to delimit URLs in some systems. The character "#" is unsafe and should always be encoded because it is used in World Wide Web and in other systems to delimit a URL from a fragment/anchor identifier that might follow it. The character "%" is unsafe because it is used for encodings of other characters. Other characters are unsafe because gateways and other transport agents are known to sometimes modify such characters. These characters are "{", "}", "|", "₩", "^", "~", "[", "]", and "`".

출처: RFC 1738

3. 이메일 처리

+ account@domain.com 형식의 이메일을 처리합니다.
+ URL 처리 후 바로 적용 하는게 좋습니다.
+ 일반적으로 이메일 아이디는 알파벳, _, ., 그리고 숫자의 조합으로만 사용하되, 도메인별로 다를 수도 있습니다 (RFC 5322 표준 참고).

hong.gildong88@example.com로 연락주세요~

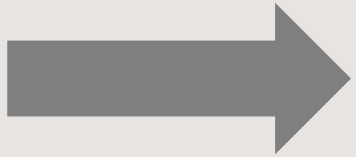


email로 연락주세요~

4. 전화번호 처리

- 000-0000-0000 형식의 국내 전화번호를 처리합니다.
- 대표번호는 제외합니다. 일반 전화번호는 개인정보 누출 우려가 있지만, 대표번호는 그렇지 않습니다. 더불어, 대표번호를 통해서 정보를 추출할 수가 있습니다.
- 상황에 따라서 하이픈(-) 대신에 점(.) 또는 공백( )이 사용되는 경우가 있으며, 어떤 기호도 없는 경우도 있습니다. 안 좋은 케이스는 대시(—)가 사용되는 경우인데, 대시도 여러 종류가 있기 때문에, 모든 기호를 잡거나 아니면 특수 케이스들은 예외로 처리합니다.
- URL & 이메일 처리와 마찬가지로 데이터 전처리 초기 단계에 적용 하는게 좋습니다.

010-123-5678, 또는 032 7777 8888으로 연락주세요. FAX는 070.5555.6666입니다.



phone, 또는 phone으로 연락주세요. FAX는 phone입니다.

5. 반복 글자/단어 처리

- 반복 글자(ㅋㅋㅋㅋ, 하하하하하)와 반복 단어(와우와우와우와우)를 처리합니다.
- 반복 글자(1개의 문자)를 우선 처리하며, 3번 이상 반복되는 글자를 2번만 반복하게 처리합니다 (예: ㅋㅋㅋ → ㅋㅋ, 하하하하 → 하하).
- 반복 글자/단어 처리시 숫자는 제외합니다.
- 반복 단어(2개 이상 문자의 조합) 처리시 2번 이상 반복되는 문자를 반복되지 않게 처리합니다 (예: 정말 정말 정말 → 정말, 무한도전무한도전무한도전 → 무한도전).
- URL/이메일 처리 이후에만 처리해야 합니다.

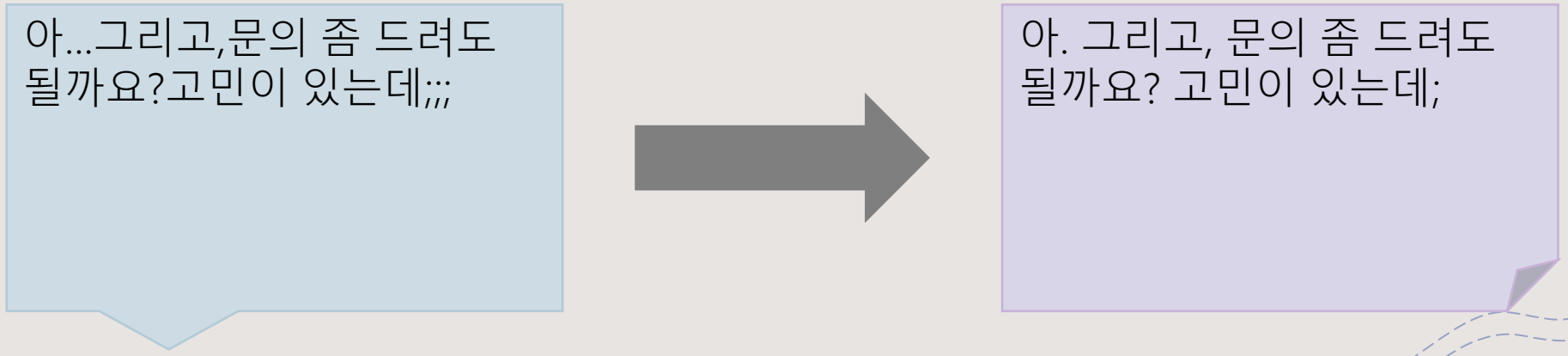
ㅋㅋㅋㅋㅋㅋ 이거 너무 너무 웃긴데요. 황ㅎ하

→

ㅋㅋ 이거 너무 웃긴데요. 황ㅎ하

6. 구두점 처리

- 구두점(. , ? ! ;)을 처리합니다 - 반복되는 구두점 제거, 그리고 띄어쓰기가 적용됩니다.
- 점(.) 처리시 숫자 소수점인지 체크가 필요합니다 (예: 10.11, €10.000, $15.99).
- 단어 위주로만 분석을 하는 경우 이 단계를 넘어가시고 10번 "필요 문자 범위로 필터링" 단계에서 구두점을 포함한 모든 기호를 제거합니다.
- URL/이메일/전화번호 처리 이후에만 처리해야 합니다.

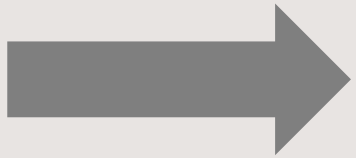


7. 사이트별 시그니처 처리

- 특정 사이트들은 글 끝에 사인/주석/워터마크가 달립니다 (예: "- from dc", "- uploaded from Android"). 이러한 경우 처리해줍니다.

제 고민을 들어 주셔서 감사합니다~

- dc official App



제 고민을 들어 주셔서 감사합니다~

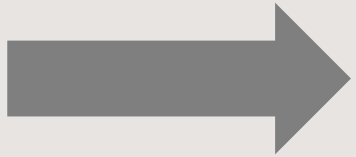
8. 욕설 처리

- 일반적인 욕설 처리는 단어 사전을 기반으로 처리합니다.
- 단어 사전 기반 욕설 처리는 명확한 한계가 있습니다 (예: ㅅㅣㅂㅂㅂ바ㄹ, shibal).
- 욕설 처리는 통계적인 방안으로도 가능하며, 머신러닝 모델로도 처리가 가능합니다. 이러한 기법들을 사용하여 단어 사전의 한계를 극복할 수 있습니다 (예: ㅅㅣㅂㅂㅂ바ㄹ → 90%확률로 욕설 판정).
- 욕설 처리 과정에서 일반 단어들이 처리되는 경우가 있어 주의해야 합니다(예: 해OO 않고는 몰라요, 잠 OO 못 해서 피곤, 떡볶이 소스는 OO장으로). 상황에 따라 이러한 욕설들을 예외를 하거나, 머신러닝 기법을 사용하여 욕설인지 아닌지 판단하게 할 수도 있습니다 (하지만 아쉽게도 한국어 욕설 감지 모델은 오픈소스가 없습니다).
- 맞춤법 교정 후 다시 한번 처리해주면 더 좋은 결과가 나올 수 있습니다 (예: 십발→교정 후 욕설 처리).
- 욕설을 다른 문자로 대체하는 경우, 특수 기호(■), 알파벳(O), 또는 마크({curse})로 대체할 수 있습니다. 이는 다른 정제 작업과 겹치지 않도록 해야 합니다 - 예를 들어, 네이버 맞춤법 교정기 같은 경우 특수기호를 단어로 여기지 않아서, 특수기호와 붙어있는 문자들에 대하여 제대로 된 교정을 하지 않습니다.
- 상황에 따라 욕설 처리를 안 할 수 있습니다. 예를 들어, 글에 대하여 긍정/부정 분류를 하는 경우, 욕설을 유지 하는게 더 좋은 정확도를 낸다면 욕설을 유지해야 합니다.

9. 필요 문자 범위로 필터링

- 데이터 분석 범위에 따라서 필요한 문자 범위로만 필터링을 합니다. 한국어 위주로 분석을 하는 경우, 알파벳(a-z, A-Z 범위), 한국어(ㄱ-ㅎ, ㅏ-ㅣ, 가-힣 범위)와 기호 몇 개만 예외해서 나머지는 제거하면 됩니다.
- 필요시 한자도 통과 범위 내에 포함하거나, 또는 한자→한글 변환을 해줍니다.
- 필요시 외국어를 감지하여 제거하는 대신 번역을 할 수도 있습니다.

꿀 팁 주셔서 감사합니다. ありがとう ಥ_ಥ

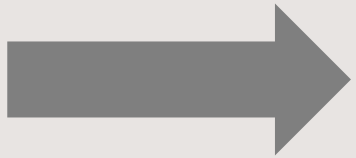


꿀 팁 주셔서 감사합니다.

10. 맞춤법 교정기 활용

- 일반적으로 맞춤법 교정은 텍스트 데이터 정제 마지막 단계로 진행되며 오타, 띄어쓰기 등을 처리해줍니다.
- 상황에 따라서 맞춤법 교정이 불필요할 수도 있습니다. 예를 들어, "일반인이 자주 틀리는 맞춤법 파악"이 데이터 분석의 목적이라면 맞춤법 교정은 적절하지 않습니다.
- 맞춤법 교정은 일반적으로 좋은 결과를 내지만, 주기적으로 업데이트되지 않는 교정기는 반대로 나쁜 결과를 냅니다 - 신조어/외래어/속어 등 신규 단어를 처리하지 못할 경우가 있습니다.
- 파이썬(Python)에서 사용할 수 있는 한국어 맞춤법 교정 라이브러리는 hanspell 입니다. 이 라이브러리는 네이버 맞춤법 검사기를 사용합니다.
- 파이썬 한국어 품사(POS, Part-Of-Speech) 태깅 라이브러리 KoNLPy는 품사 태깅 과정에 normalization(정규화) 옵션을 사용하면 맞춤법 교정 효과를 낼 수 있습니다.

안녕 하세용~ 만나 서 반가어요!



안녕하세요~ 만나서 반가워요.

자연어 처리 예시 (실 데이터)

아 종강했는데 뭐하냐? 알바는 12일뒤부터 하는데 하고싶은것도 할것도 없는데 뭐하는게 좋을까?₩n시발 시험끝나자마자 뇌풀려서 멍함;

→

아 종강했는데 뭐 하냐? 알바는 12일 뒤부터 하는데 하고 싶은 것도 할 것도 없는데 뭐 하는 게 좋을까? OO 시험 끝나자마자 뇌풀려서 멍함;

나 십대 후반 이십대 초 까지 인디밴드 내 귀에 도청장치 되게 좋아했는데. https://youtu.be/2RK6seztvV8₩n₩nhttps://youtu.be/lj2ZC_TSpp0₩n₩nhttps://youtu.be/VfEVO6yv6sA₩n₩nhttps://youtu.be/0BoLYOSAsfs₩n₩n곡 몇개 추천하고 감₩n₩n나도 간만에 들으니 좋네

→

나 십 대 후반 이십 대 초까지 인디밴드 내 귀에 도청장치 되게 좋아했는데. link. link. link. link. 곡 몇 개 추천하고 감. 나도 오래간만에 들으니 좋네.

요즘. https://youtu.be/hkt-AWerpU8₩n₩n이거 들으면서 독서 하는데 집중력 미침₩n- dc official App

→

요즘. link. 이거 들으면서 독서하는데 집중력 미침.

마음잘맞는사람, 친구구해여! 여잔데 동성친구면 좋구요..₩n꼭 아니어도 되는데 말이 잘? 통하는 사람 찾아여ㅠㅠ₩n₩n₩n카톡 hiruka₩n번호 010 2891 4165

→

마음 잘 맞는 사람, 친구 구해요! 여잔데 동성친구면 좋고요. 꼭 아니어도 되는데 말이 잘? 통하는 사람 찾아요ㅠㅠ. 카톡 hiruka. 번호 phone_number.

자연어 처리 추가 개선 방향

- 광고/스팸 글 감지
- 외국어 감지 및 번역
- 이모지 감지 (っ °Д °;)っ ಥ_ಥ
- 주소 감지
- 은행 계좌번호 감지
- 혐오글 감지