빅데이터 분석 솔루션

TEXTOM 매뉴얼

빅데이터 수집에서 시각화까지!

V5.0

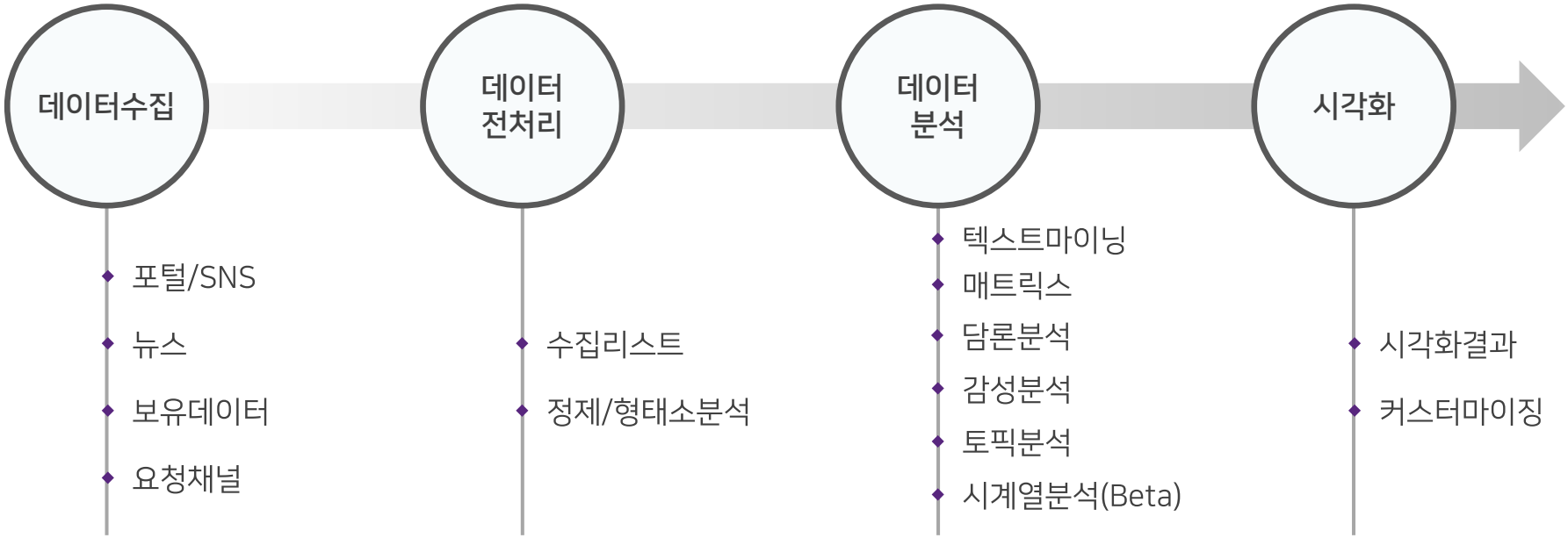
TEXTOM

한눈에 보는 TEXTOM 프로세스

방대하고 복잡한 텍스트 자료,

이제 **효율적으로 분석**하여

연구, 마케팅, 여론분석 등 다양한 곳에 **효과적으로 활용**해보세요!

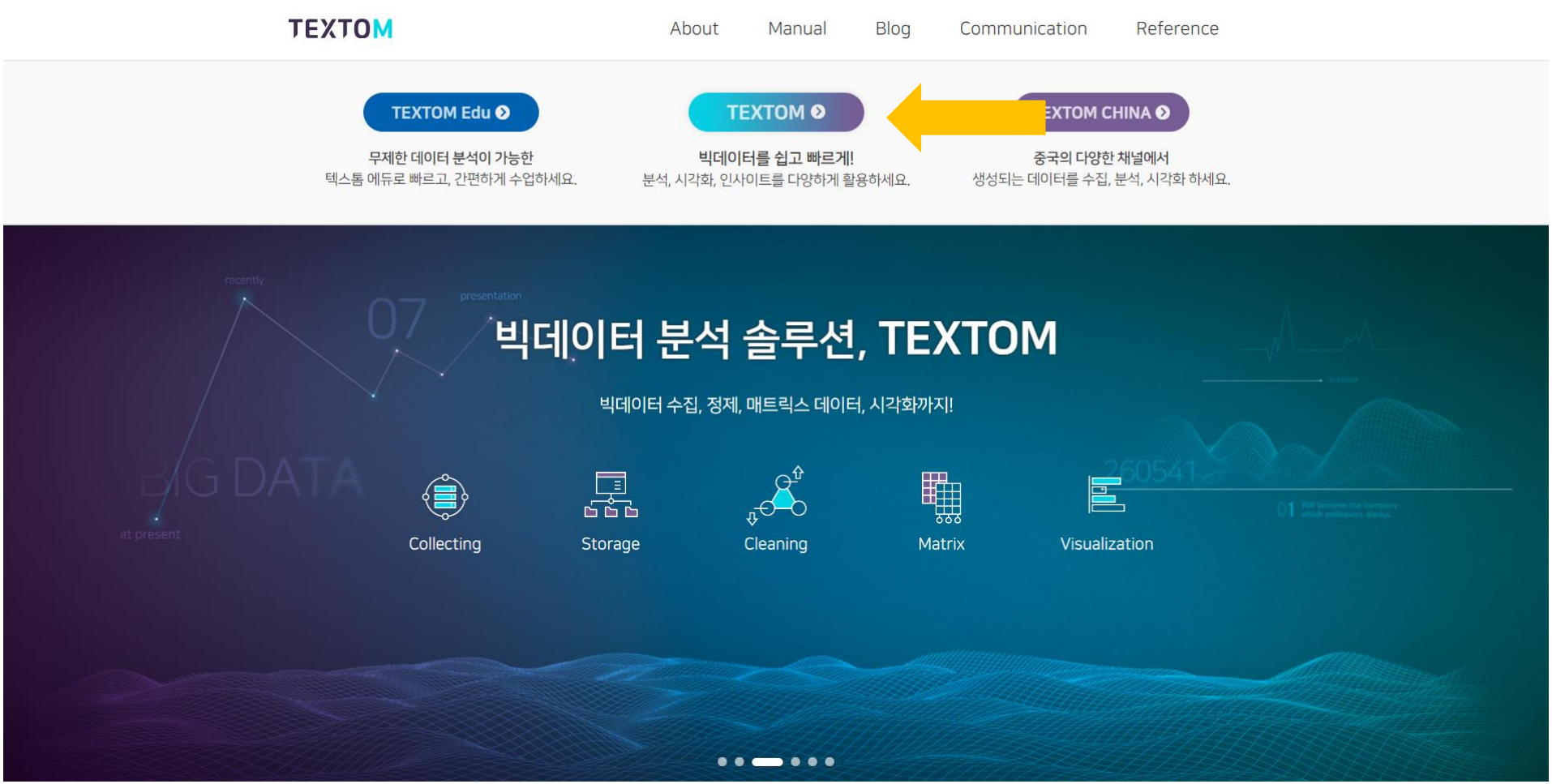


* 텍스톰은 Chrome 브라우저에 최적화되어 있습니다

Contents

※ 클릭 시 해당 단계 첫 페이지로 이동합니다

회원가입 / 로그인



TEXTOM 을 눌러 회원가입&로그인 페이지로 이동합니다

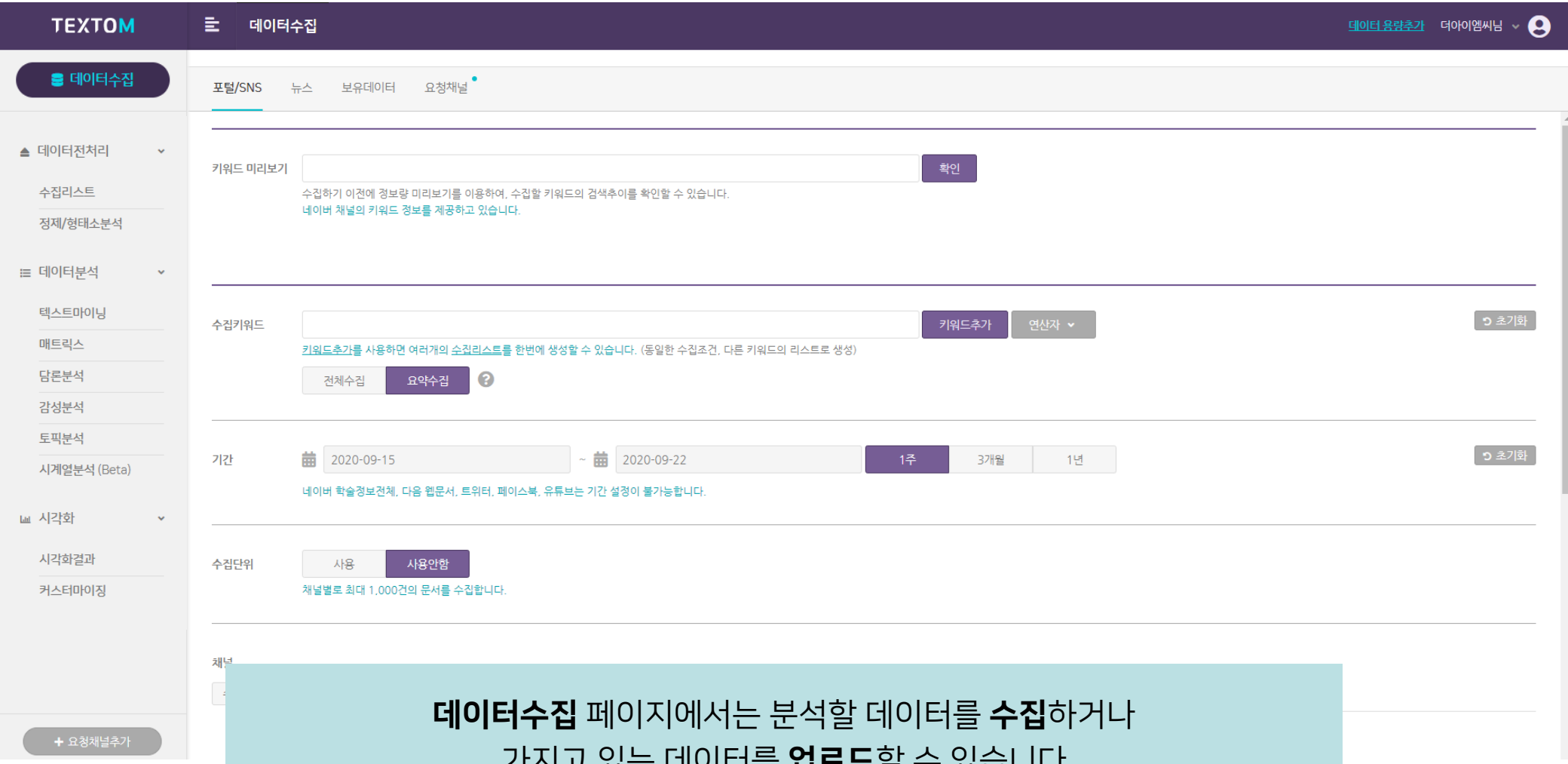
[회원가입 증명서류 및 무료데이터용량 지급 관련 안내 바로가기]

회원가입 / 로그인

가입 승인이 완료되면 로그인 후 솔루션 사용이 바로 가능합니다

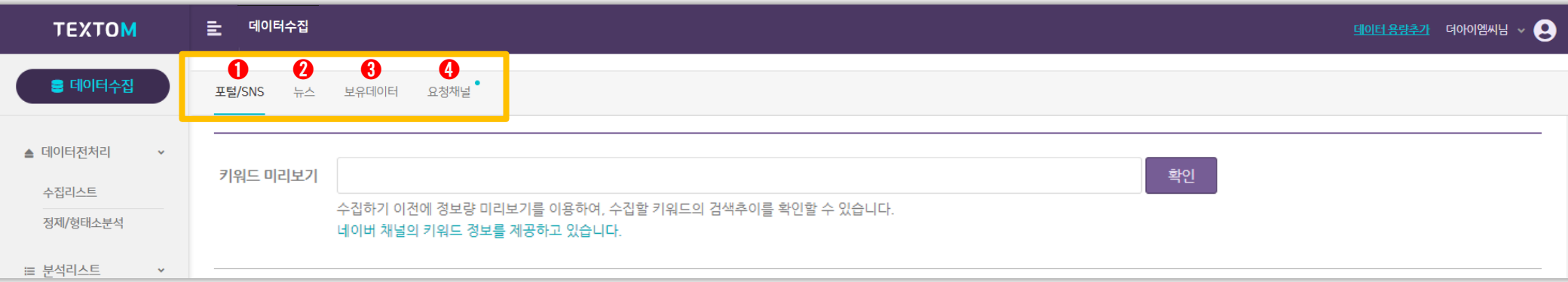
가입 승인은 영업일 기준 1~3일 이내에 완료됩니다

로그인 시 첫 화면



데이터수집 페이지에서는 분석할 데이터를 **수집**하거나
가지고 있는 데이터를 **업로드**할 수 있습니다

수집 탭



❶ **포털/SNS** 네이버, 다음, 구글, 바이두, 유튜브, 트위터, 페이스북 데이터 수집이 가능합니다

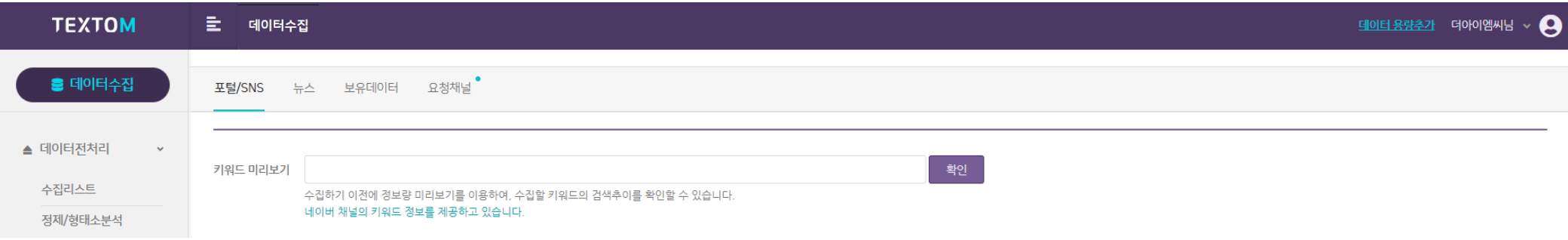
❷ **뉴스** KBS, MBC, SBS, YTN, 조선일보, 중앙일보, 동아일보, 한겨레, 경향신문 등 언론사 20곳의 기사 데이터 수집이 가능합니다

❸ **보유데이터** 텍스트로 작성된 pdf, txt, xlsx 형식의 파일 업로드가 가능합니다

❹ **요청채널** 텍스톰이 제공하는 채널 외 특정 사이트의 데이터 수집이 필요할 때, 별도의 추가 비용을 지불하고 이용 가능합니다

※ 데이터 수집은 데이터 용량이 차감되지 않습니다

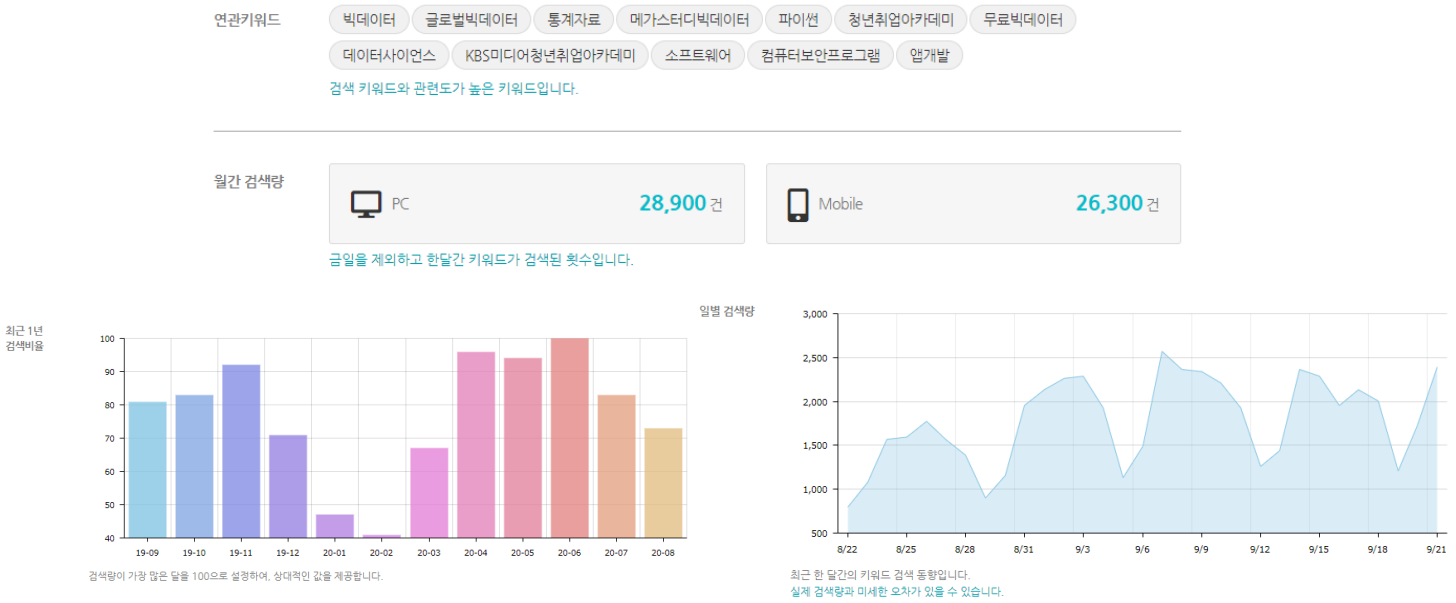
키워드 미리보기



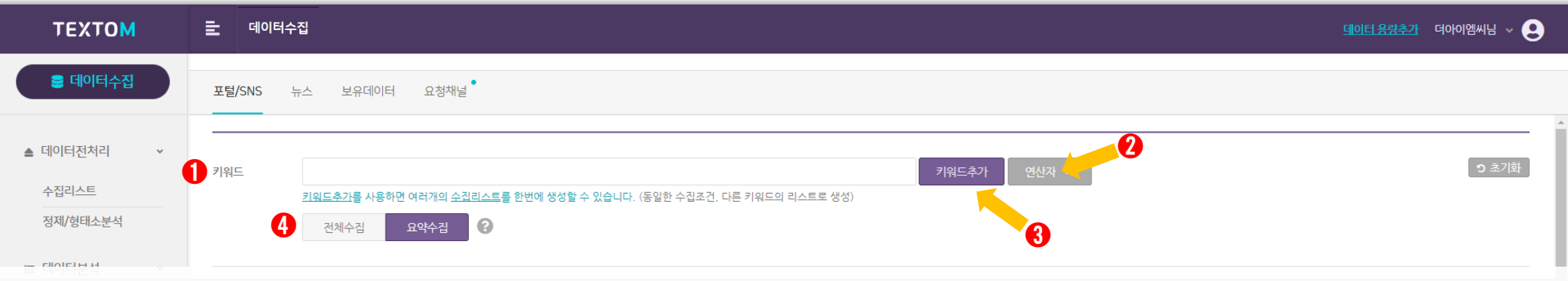
❶ 데이터수집 이전에 정보량 미리보기를 이용하여, 수집할 키워드의 검색추이를 확인할 수 있습니다.

※ 네이버 채널의 키워드 정보를 제공하고 있습니다.

※ 연산자는 적용되지 않습니다.



수집 키워드



❶ 키워드 란에 **기입한 형태 그대로**, 각 채널 검색창에 검색을 하고 그 검색 결과를 수집하게 됩니다

❷ 분석의 목적과 주제에 맞는 정확한 데이터를 데이터수집 위해서는 세밀한 키워드 선정이 필요합니다

※ **네이버**와 **구글**은 연산자 기능이 사용 가능한 채널이므로, 상황에 맞게 적절한 **특수문자(연산자)**를 사용하면 더욱 정확한 데이터를 수집할 수 있습니다

❸ **키워드 추가** 기능은 여러 키워드를 동일한 설정(기간, 수집단위, 채널)으로 수집하고 싶을 때 사용하면 유용한 기능입니다

❹ **전체수집, 요약수집** 설정은 데이터의 본문 영역을 어떻게 수집할 지를 정하는 기능입니다

전체수집, 요약수집 추가 설명 바로가기

※ 전체수집 시 키워드와 관련 없는 스팸 데이터가 많습니다. 왜 그런가요?
각 사이트의 페이지마다 레이아웃이 모두 일관적이지 않아 **본문 근처의 리스트, 광고 등의 텍스트**가 모두 수집되는 경우가 많습니다.
본문 전체 내용이 꼭 필요한 경우가 아니라면, 요약수집을 사용해 주셔도 키워드에 대한 분석은 충분히 가능합니다.

수집 기간 / 단위

❶ 설정한 기간 내에 만들어진 데이터를 수집하게 됩니다

※ 포털/SNS의 **네이버 학술정보전체, 다음 웹문서, 유튜브, 트위터, 페이스북**은 기간 설정이 적용되지 않으며,
수집이 진행되는 시점의 해당 채널 검색 결과를 그대로 수집하게 됩니다

※ **뉴스의 언론사 전체 채널**은 최대 3개월까지 수집 가능하므로, 긴 기간을 수집하려면 **요청채널 서비스(유료)**를 이용해주시거나
3개월씩 나누어 여러 번 수집해 주셔야 합니다

❷ **수집단위**는 일/주/월/년의 시간 단위 중 선택된 단위로 데이터를 최대 1,000건까지 수집하는 기능입니다

※ 1년(365일)의 데이터를 일단위로 수집할 경우, 최대 365,000건의 데이터를 수집할 수 있습니다
동일한 기간을 수집단위 사용안함 으로 수집할 경우, 최대 1,000건 수집됩니다

※ 수집단위 기능을 사용할 수 있는 채널은 네이버 블로그, 카페, 지식IN, 뉴스 / 다음 블로그, 카페, 뉴스 / 구글 뉴스입니다

수집 채널

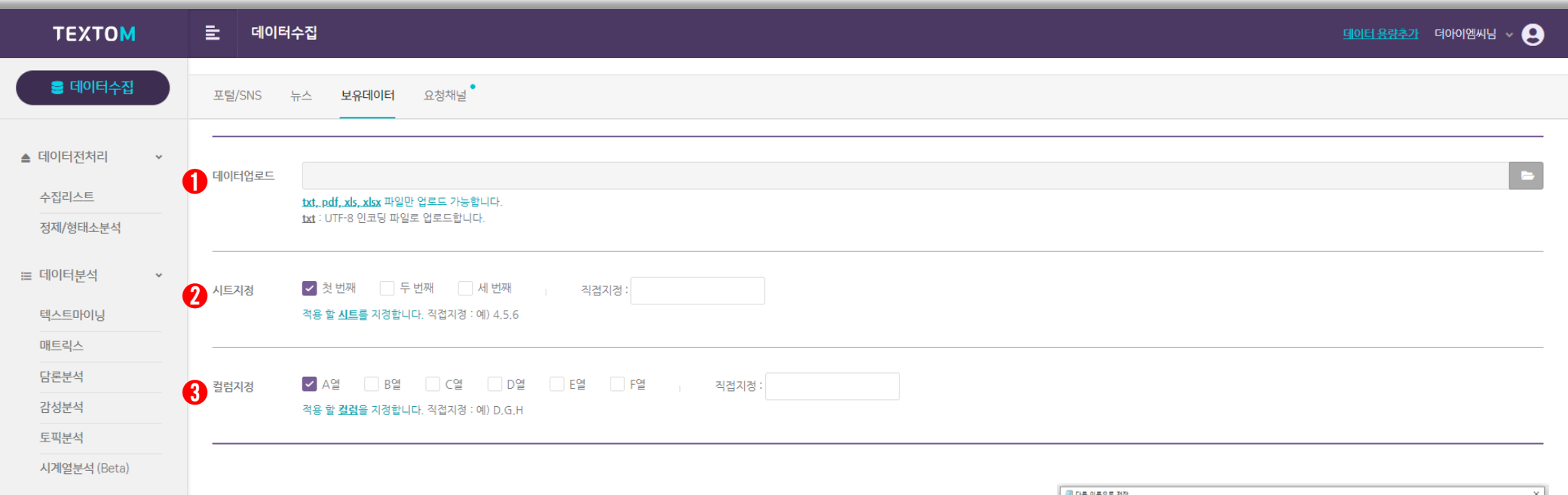
❶ 포털/SNS

❷ 뉴스

❶ **포털/SNS** 주요 포털 사이트 및 소셜 네트워크 서비스의 검색 결과를 수집할 수 있습니다

❷ **뉴스** 주요 방송사 및 언론사의 공식 사이트 검색 결과를 수집할 수 있습니다

보유데이터 업로드



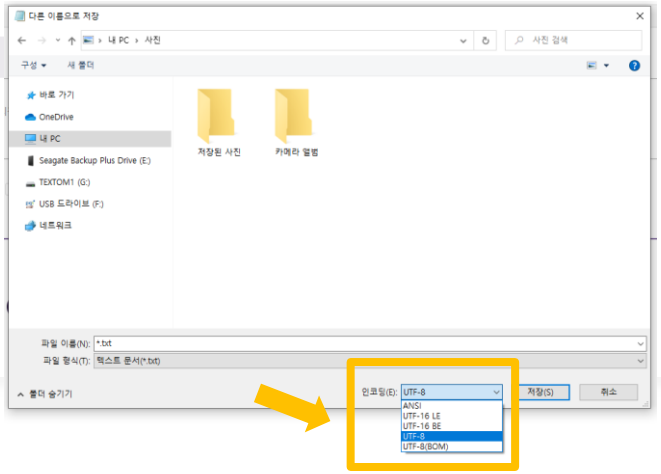
❶ 가지고 있는 데이터를 텍스톰에 업로드하여 정제, 분석, 시각화 할 수 있습니다

※ 파일 형식은 txt, pdf, xlsx, xls 파일만 가능합니다

※ txt 파일은 UTF-8 인코딩 파일로 업로드 해 주셔야 합니다 UTF-8 인코딩 설정방법 참고 ▶

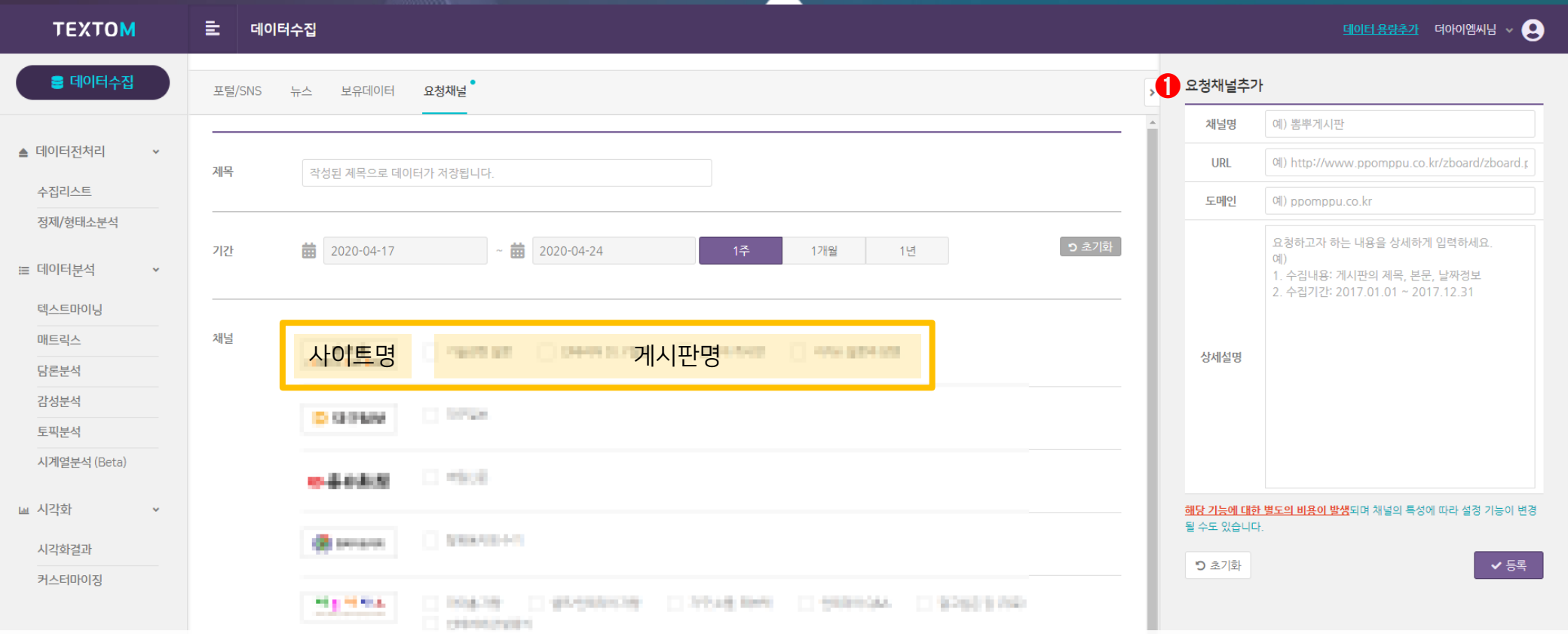
※ pdf 파일 내 이미지는 인식하지 않습니다

❷, ❸ **시트지정** 및 **컬럼지정** 기능은 **엑셀 파일** 업로드 시, 많은 시트와 열들 중에서 원하는 데이터만 지정할 수 있는 유용한 기능입니다



※ 컬럼 다중 선택 시, 컬럼들은 한 컬럼으로 합쳐지게 됩니다

요청채널



❶ 텍스톰이 제공하는 채널 외의 데이터 수집을 원하면 요청채널로 등록해주세요

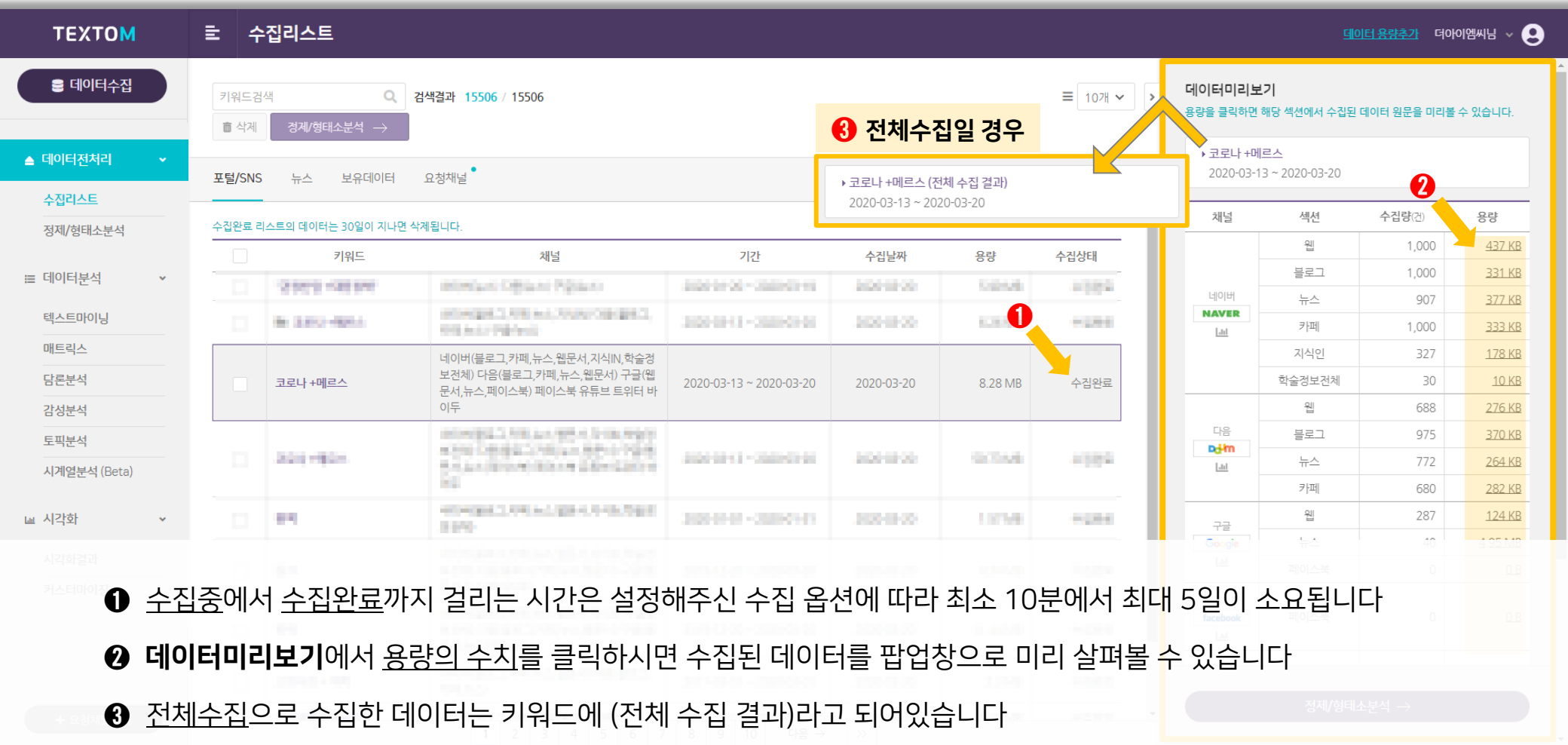
※ 적어 주신 내용을 바탕으로 요청 사이트의 데이터 수집가능여부 파악 후, 산출법에 따라 견적이 책정됩니다(추가 유료 서비스)

※ 요청채널은 수집된 데이터만 전달받는 방법(일회성)과 텍스톰 요청채널 페이지 내 구축하는 방법이 있습니다

※ 수집된 데이터는 xlsx 파일 또는 txt 파일로 전달됩니다

※ 구축이 완료되면 데이터수집에서 제공하는 기존 채널과 동일한 방법으로 이용하시면 됩니다

수집리스트 살펴보기



❶ 수집중에서 수집완료까지 걸리는 시간은 설정해주신 수집 옵션에 따라 최소 10분에서 최대 5일이 소요됩니다

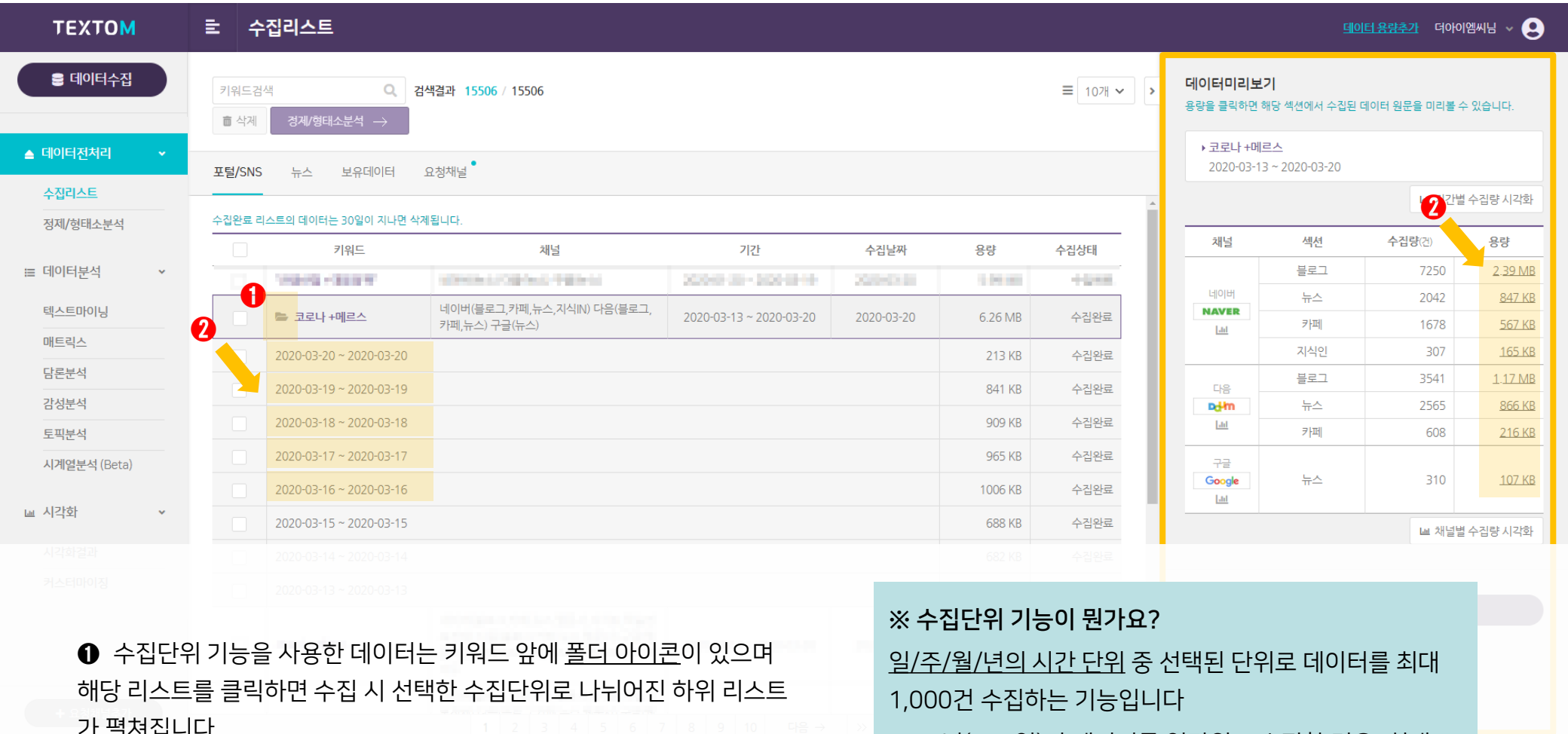
❷ **데이터미리보기**에서 용량의 수치를 클릭하시면 수집된 데이터를 팝업창으로 미리 살펴볼 수 있습니다

❸ 전체수집으로 수집한 데이터는 키워드에 (전체 수집 결과)라고 되어있습니다

※ 수집량이 1,000건인 데이터가 많습니다. 왜 그런가요?

오픈API를 통해서 데이터를 수집하는 경우, 섹션별로 한번에 최대 1,000건까지 수집할 수 있습니다
1,000건 이상의 데이터를 수집 해야 하는 경우 수집단위 기능을 사용해주세요

수집단위 사용한 수집리스트 살펴보기



❶ 수집단위 기능을 사용한 데이터는 키워드 앞에 폴더 아이콘이 있으며 해당 리스트를 클릭하면 수집 시 선택한 수집단위로 나뉘어진 하위 리스트가 펼쳐집니다

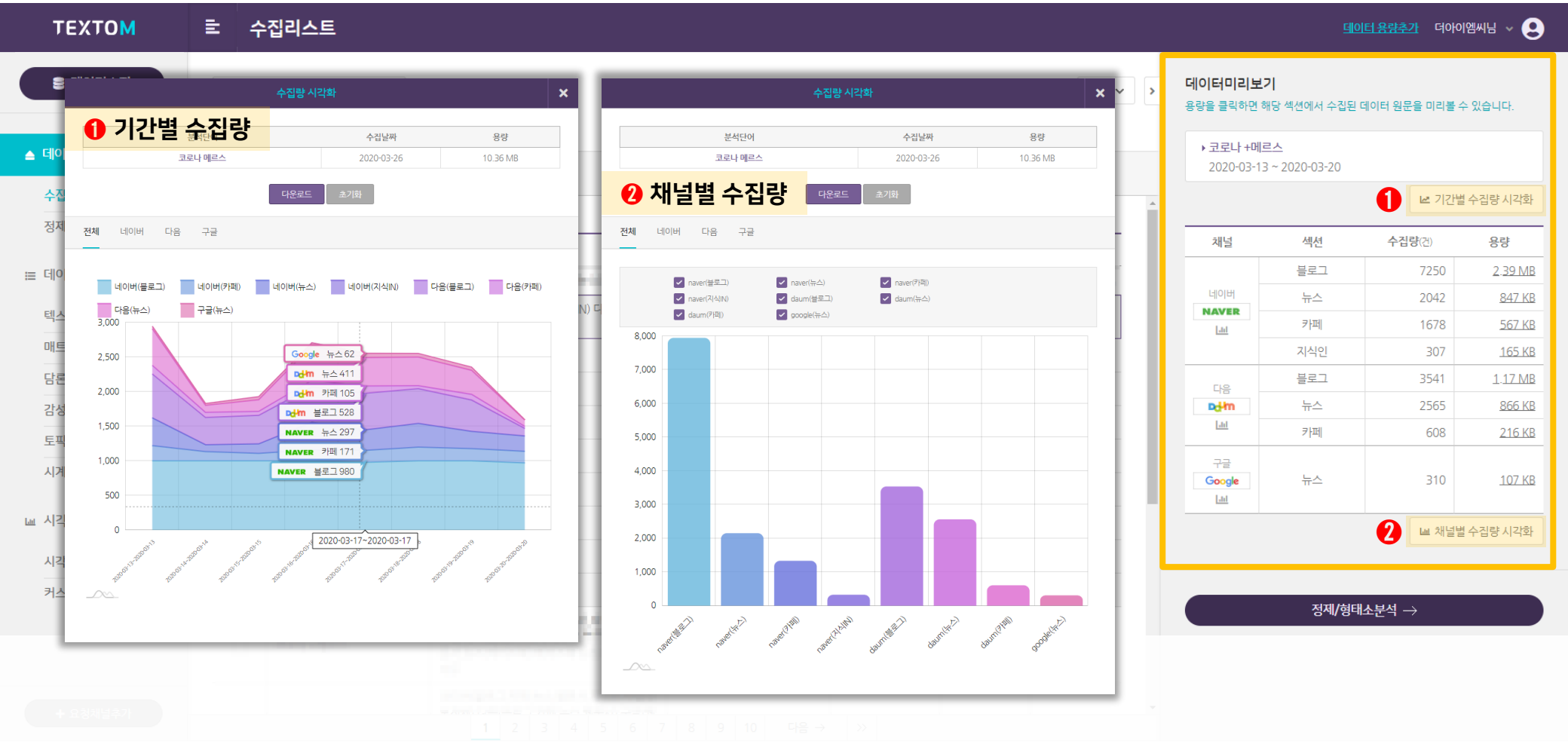
❷ 나뉘어진 데이터리스트를 개별적으로 선택하시어 **데이터미리보기**에서 용량의 수치를 클릭하시면 수집된 데이터를 팝업창으로 미리 살펴볼 수 있습니다

※ 수집단위 기능이 뭔가요?

일/주/월/년의 시간 단위 중 선택된 단위로 데이터를 최대 1,000건 수집하는 기능입니다

Ex. 1년(365일)의 데이터를 일단위로 수집할 경우, 최대 365,000건의 데이터를 수집할 수 있습니다 동일한 기간을 수집단위 사용안함 으로 수집할 경우, 최대 1,000건 수집됩니다

수집단위 사용한 수집리스트 살펴보기



수집단위 기능을 사용한 데이터는 데이터미리보기에서 기간별, 채널별 수집량 시각화를 확인할 수 있습니다

수집된 데이터 미리보기

- 데이터미리보기 팝업창 활용법 -

데이터미리보기 영역 각 섹션의 용량을 클릭하시면
왼쪽 이미지처럼 수집된 데이터를 팝업창으로
미리 살펴볼 수 있습니다

❶ 수집한 데이터는 **채널**과 그 채널의 **섹션**별로 분류되어 나타납니다

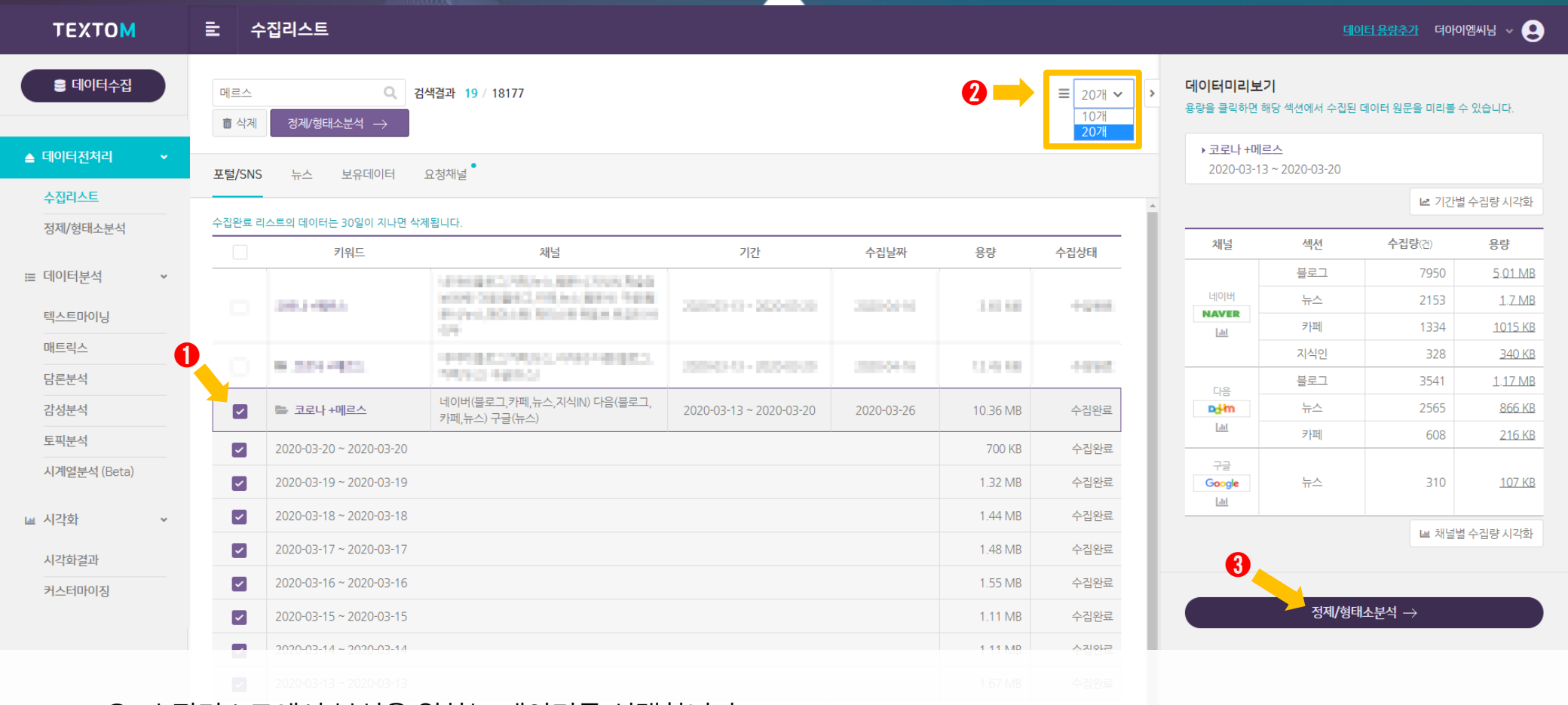
❷ 데이터미리보기 팝업창에서는 **제목, 본문 요약, URL**을 제공하며, 각 데이터의 **제목**이나 **URL**을 클릭하시면 해당 데이터의 실제 페이지로 이동할 수 있습니다

❸ 데이터미리보기 팝업창으로, 수집된 데이터를 전체적으로 미리 살펴보며 Garbage Data(광고성 글, 원하는 주제와 무관한 글)가 있는지, 수집이 정상적으로 잘 되었는지 체크해줍니다

※ Tip

Garbage Data 는 정제/형태소분석 페이지에서 **키워드필터링** 기능으로 제거 가능하므로, 해당 문서들의 공통적인 키워드들을 따로 메모해두시길 바랍니다

수집리스트 선택하기



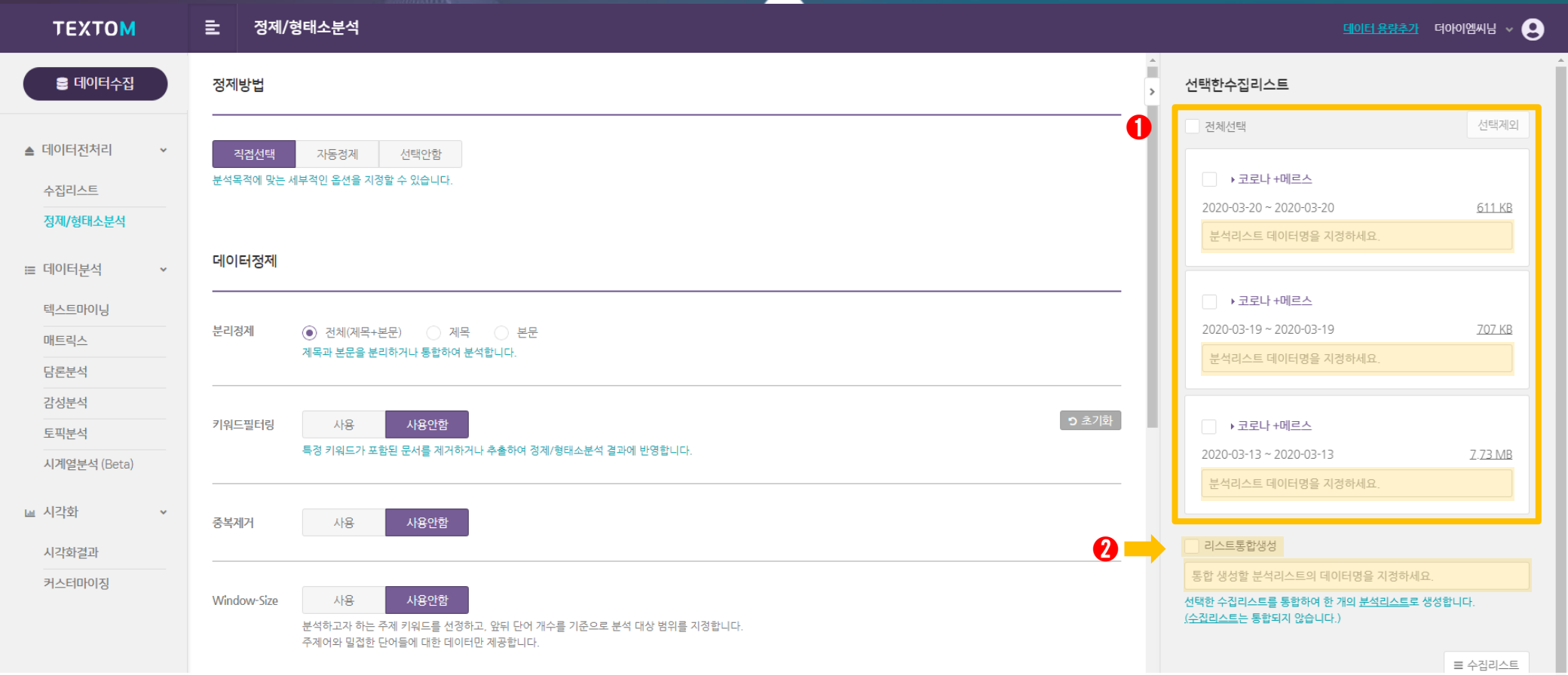
❶ 수집리스트에서 분석을 원하는 데이터를 선택합니다

※ 다중 선택이 가능하며, 다중 선택한 데이터들은 정제/형태소분석 단계에서 통합하거나 동일한 설정으로 분석리스트를 생성할 수 있습니다

❷ 한 페이지에서 보여주는 데이터 수를 더 늘릴 수 있습니다

❸ 선택한 데이터가 분석할 데이터가 맞는지 재확인한 뒤, **정제/형태소분석** 버튼을 눌러줍니다

수집리스트 설정하기

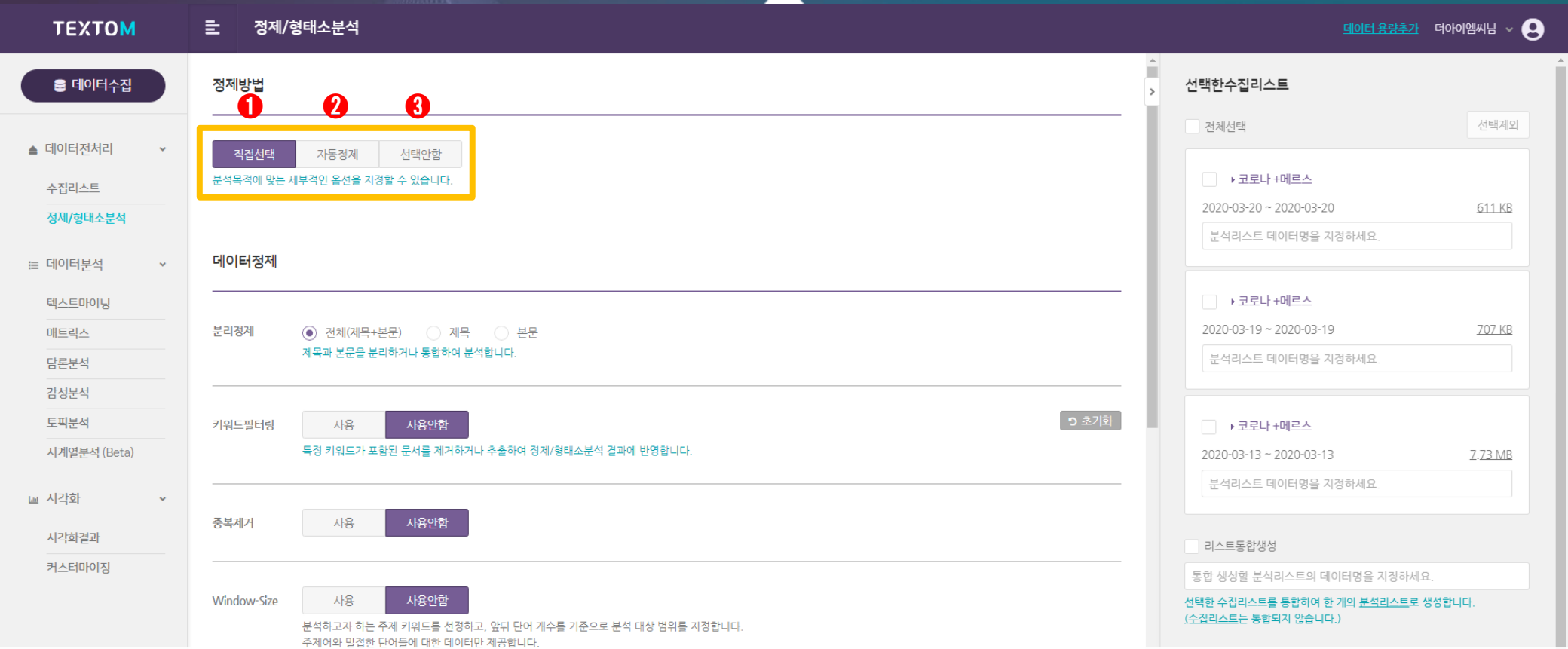


❶ 선택한 리스트에 대한 **데이터명**을 지정하고 체크박스를 선택합니다

※ 데이터명은 생성 시 설정한 분리정제, 분석기, 분석품사 정보를 넣어 지정하시면 추후 여러 데이터 사이에서 구별하기 좋습니다

❷ 리스트를 다중 선택한 경우, 해당 리스트들을 하나의 데이터로 통합하여 생성할 수 있습니다

정제방법 설정하기



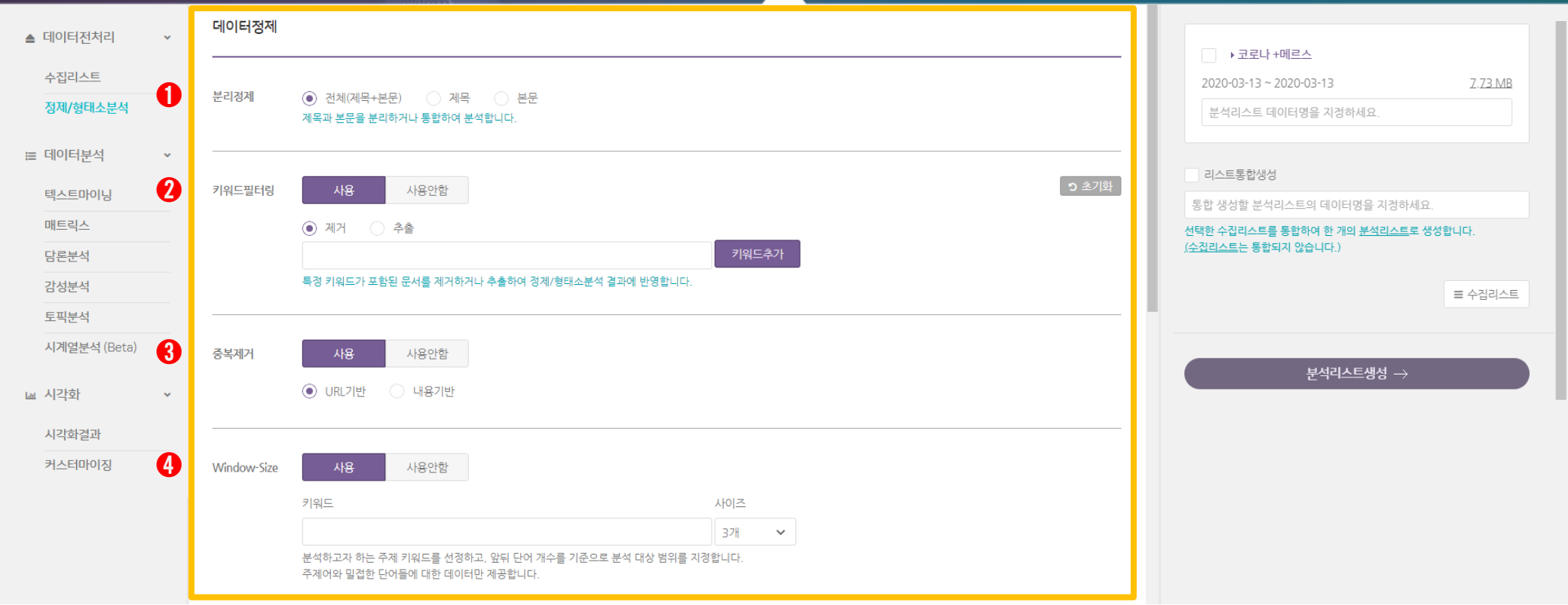
수집한 데이터 또는 업로드한 보유데이터의 정제, 형태소분석 방법을 지정할 수 있습니다

❶ **직접선택** 모든 설정을 이용자가 직접 선택할 수 있어 정교한 정제가 가능합니다

❷ **자동정제** 텍스톰이 지정해놓은 설정값으로 자동 선택되어, 어느 정도 다듬어진 정제데이터를 볼 수 있습니다

❸ **선택안함** 이미 정제한 데이터를 업로드하여 분석리스트를 생성하고자 하는 경우 선택안함을 해주시면 됩니다

데이터정제 설정하기



❶ **분리정제** 수집된 데이터에서 분석에 사용할 텍스트 영역을 지정할 수 있으며, 제목, 본문, 제목+본문 중에서 선택 가능합니다

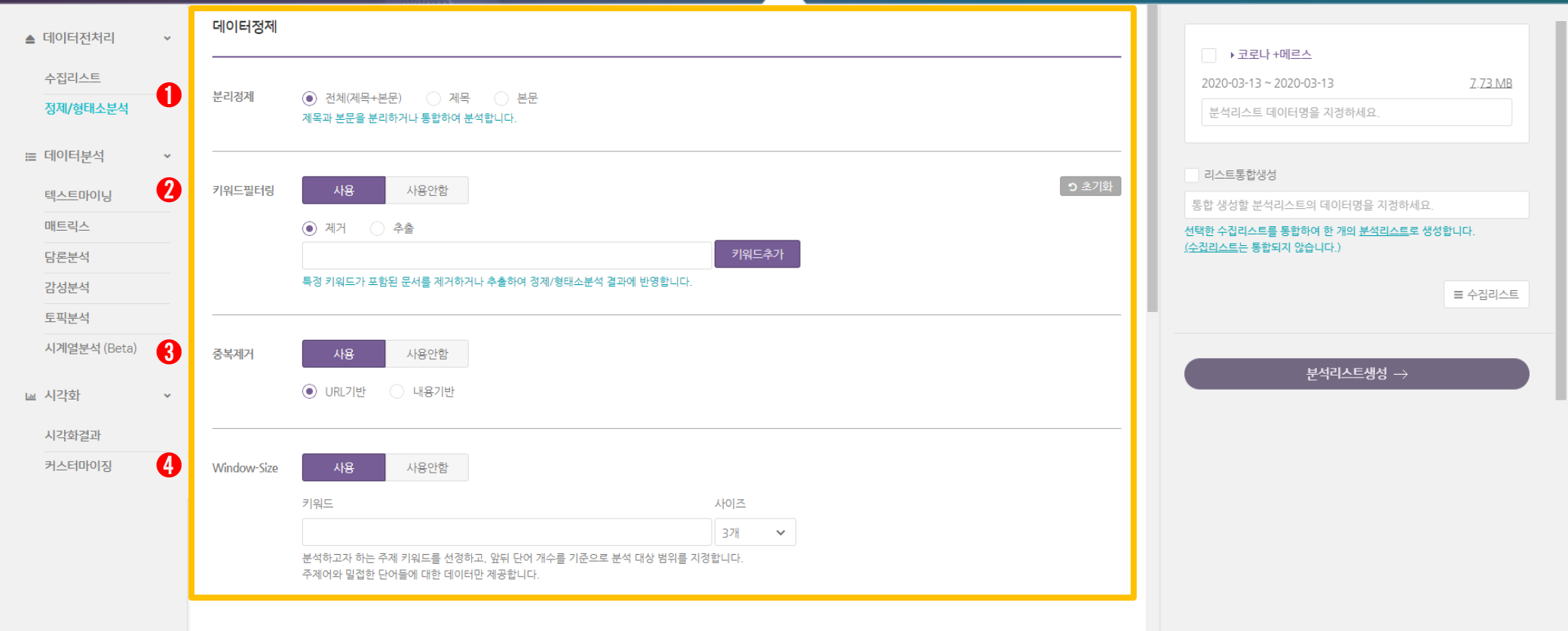
❷ **키워드필터링** 특정 키워드가 포함된 **문서**들을 제거하거나 추출할 수 있습니다

※ 한 필드에 한 단어 씩 기입해 주셔야 하며, 기입해주신 단어들 중 문서 내에 한 단어라도 있으면 적용이 됩니다

※ **제거 :** Garbage Data 제거에 유용한 기능으로, 해당 키워드가 속한 문서들을 제거한 정제데이터를 확인할 수 있습니다

※ **추출 :** 수집데이터에서 특정 키워드가 포함된 문서를 추출하여 정제할 수 있습니다

데이터정제 설정하기



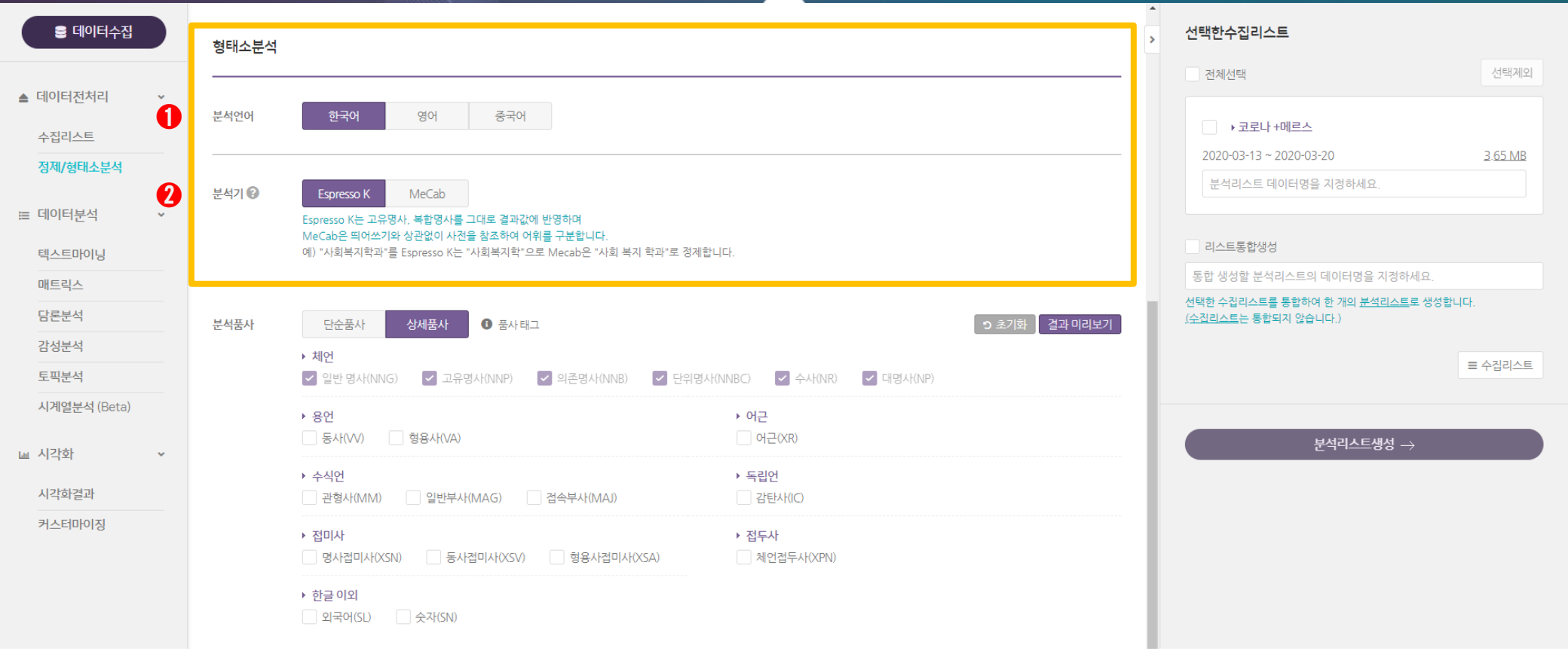
❸ **중복제거** 중복된 데이터를 제거할 수 있습니다

※ **URL기반**은 링크 전체가 완전히 동일할 때 제거됩니다

※ **내용기반**은 특수문자, 띄어쓰기까지 완전히 동일할 때 제거됩니다

❹ **Window-Size** 한 문서 내에서 특정 키워드 앞뒤로 단어 개수 범위를 지정할 수 있습니다

형태소분석 설정하기

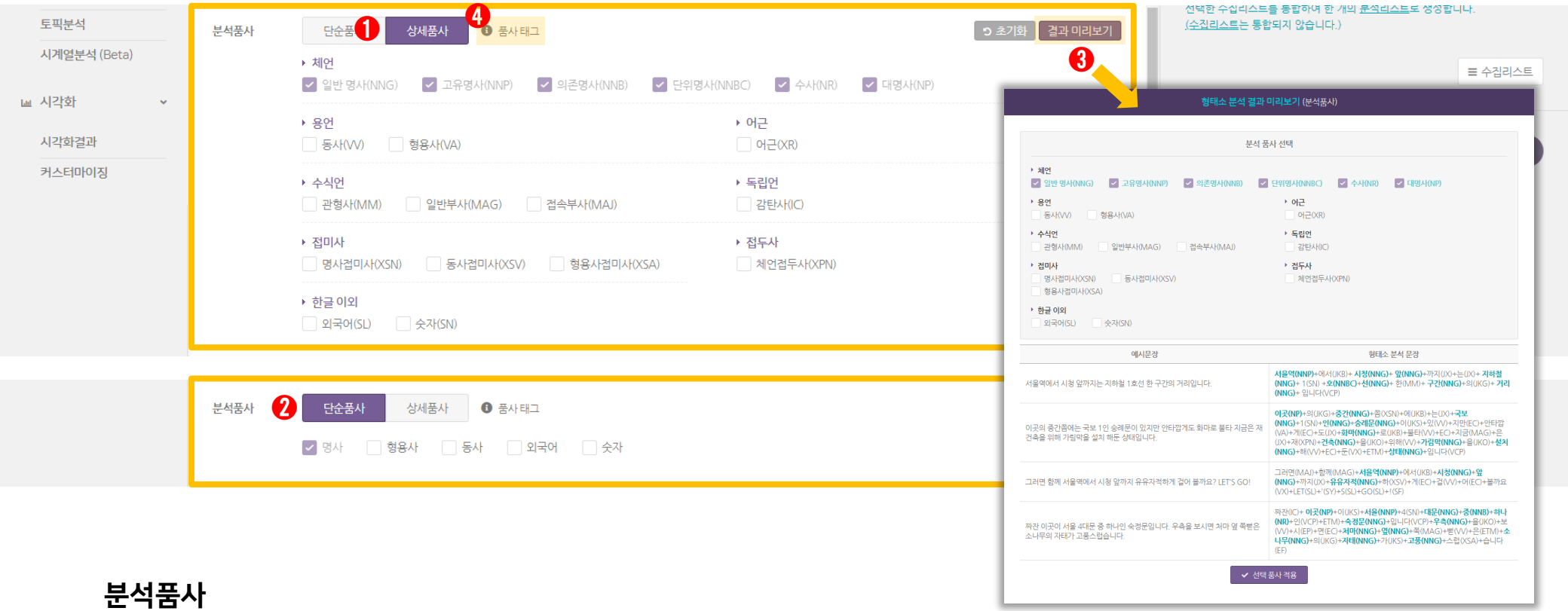


형태소 분석은 단어를 구성하는 각각의 형태소들을 인식하고 용언의 활용, 불규칙 활용이나 축약, 탈락현상이 일어난 형태소를 원형으로 복원하는 과정을 의미하며, **형태소 분석기**는 텍스트를 형태소 단위로 분석하고 품사를 함께 출력해주거나 특정 품사에 해당하는 형태소만 선별해주는 패키지를 의미합니다

❶ **분석언어** 데이터를 인식할 언어를 선택해줍니다

❷ **분석기** 각 분석기의 특성을 참고하여 형태소분석기를 선택해줍니다 Espresso K 와 MeCab 분석기 차이 자세히보기

형태소분석 설정하기



분석품사

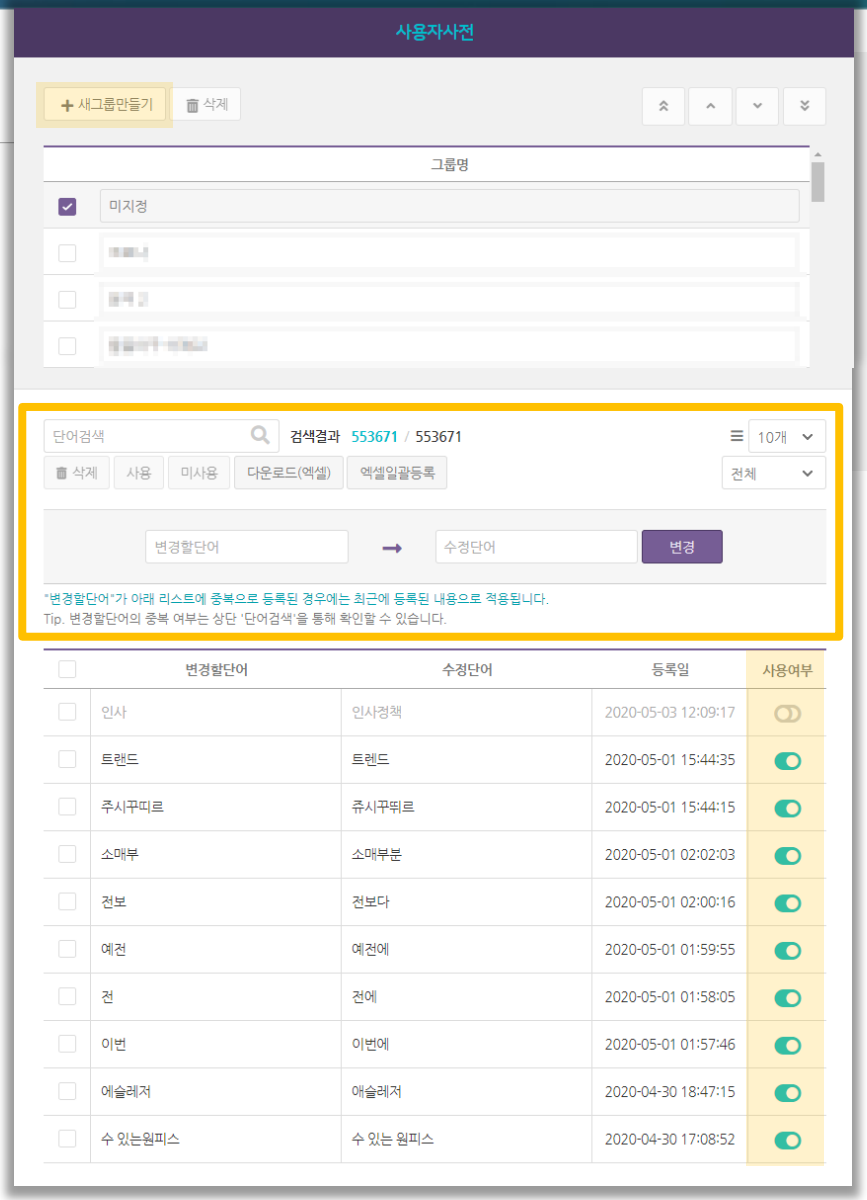
❶ **상세품사** 체언, 용언, 어근, 수식언, 독립언, 접미사, 접두사, 외국어, 숫자를 선택할 수 있습니다

❷ **단순품사** 명사, 형용사, 동사, 외국어, 숫자를 선택할 수 있습니다

❸ **결과 미리보기** 선택한 품사가 예시문장에 바로 적용되어 정제데이터에 선별될 단어를 미리 확인할 수 있습니다

❹ **품사 태그** 분석기별 품사 태그를 확인할 수 있습니다

사용자사전 설정하기

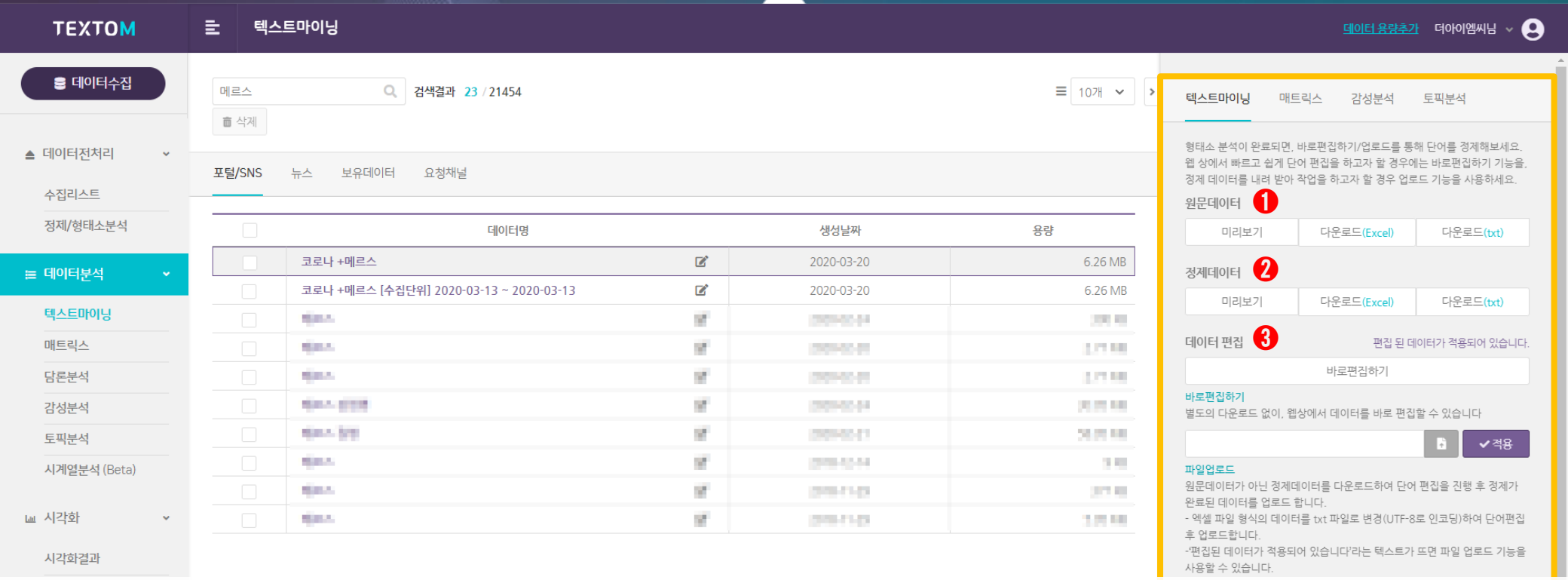


❶ **사용자사전** 유사한 주제의 데이터 또는 동일한 주제의 기간 및 채널만 다른 데이터를 반복 정제해야 하는 경우, 변경할 단어들을 사전으로 구축해두면, 형태소 분석 과정에서 지정해둔 수정 단어들로 일괄 변경되는 기능입니다

❷ **사용자사전설정** 그룹별로 단어를 등록할 수 있으며, 등록된 단어 사용여부도 매번 설정 가능합니다

❸ **그룹지정** 그룹지정 란에서 사용할 사전의 그룹명을 선택해 주시면 됩니다

텍스트마이닝하기



❶ **원문데이터** 수집된 데이터의 원문을 미리보기 혹은 xlsx 파일, txt 파일로 다운로드 할 수 있습니다.

※ **원문데이터의 파일은 데이터 정제 후에도 내용이 변하지 않으며** 언제든지 다운로드 받을 수 있습니다.

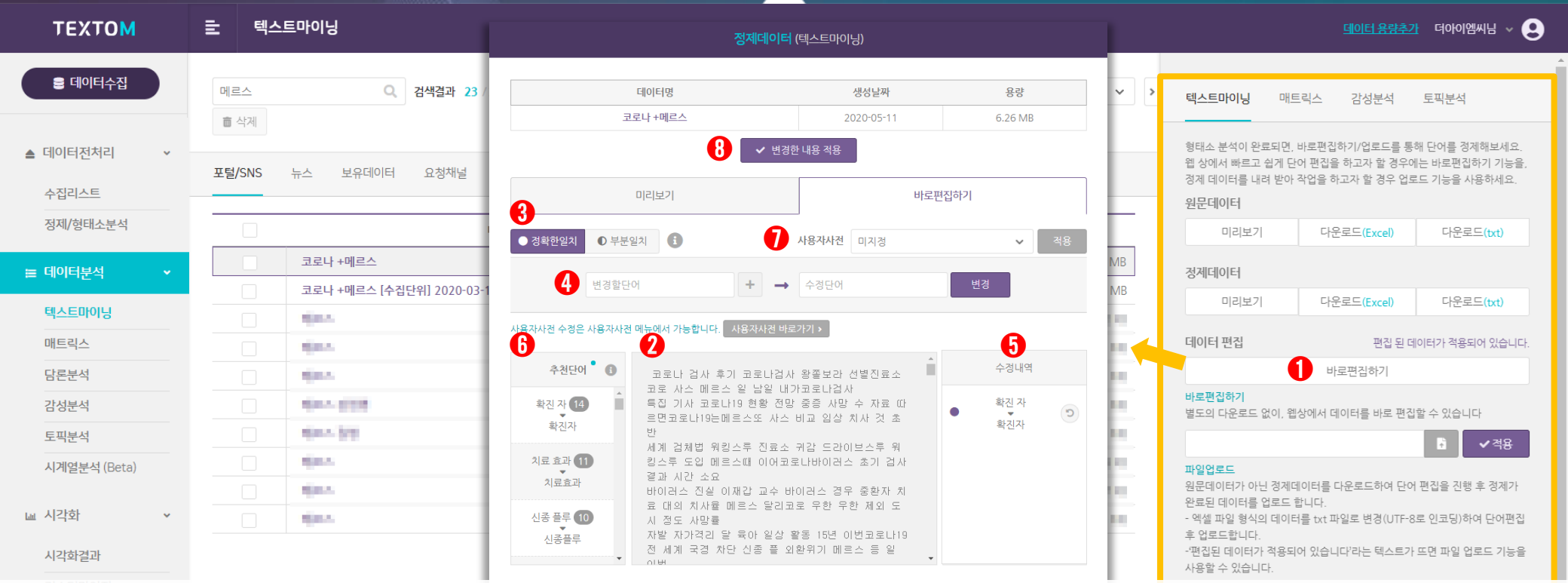
❷ **정제데이터** 정제/형태소 분석 결과 데이터를 미리보기 혹은 xlsx 파일, txt 파일로 다운로드 할 수 있습니다.

※ **정제데이터의 파일은 데이터를 정제 할 때마다 내용이 변경**됩니다.

※ 한 번 정제된 데이터는 이전상태로 되돌릴 수 없기 때문에 **최초의 정제 데이터를 내려 받아 두고 작업하는 것을 추천**해드립니다.

❸ **데이터 편집** 고유명사 및 복합명사, 동의어, 불용어 제거 등 데이터 수정을 할 수 있습니다.

텍스트마이닝하기

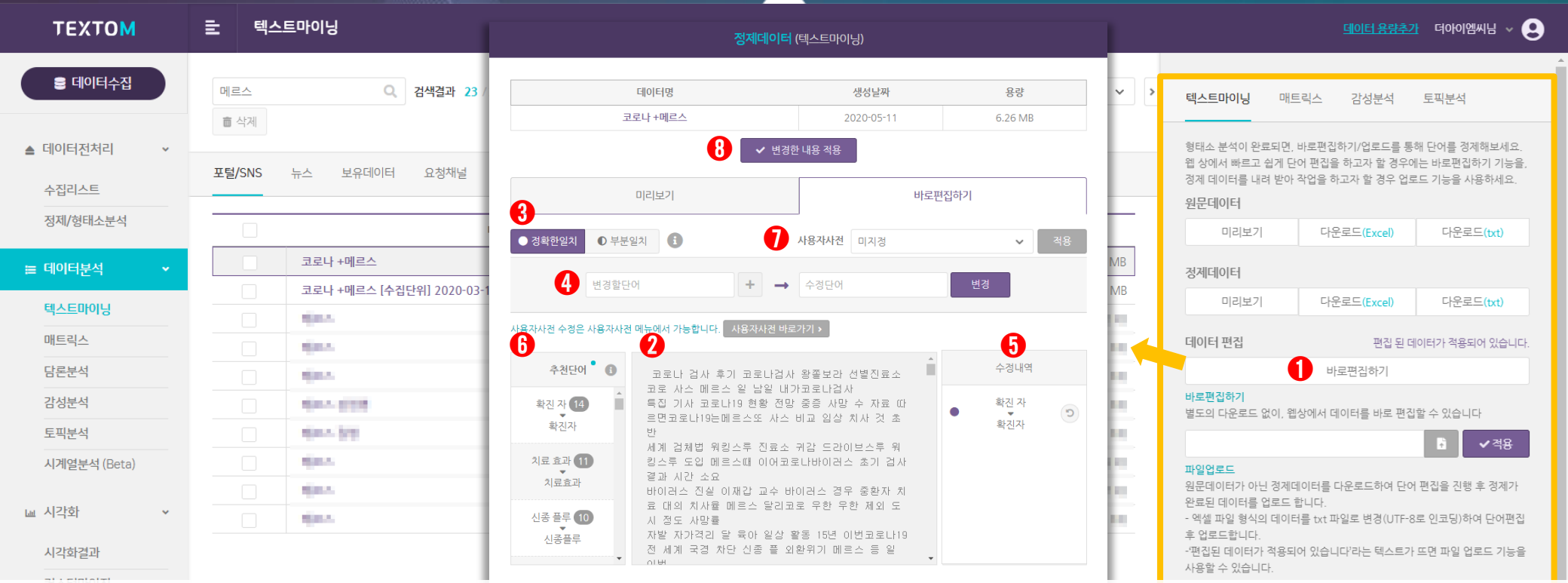


❶ **바로편집하기** 데이터편집의 한 방법으로, 파일 업로드 없이 웹상에서 바로 데이터를 정제할 수 있는 기능입니다.

※ 바로편집하기에서 변경한 내용 적용이 되면 되돌릴 수 없으므로 정제데이터를 내려 받아 두고 작업하는 것을 추천드립니다.

❷ **정제데이터 미리보기 화면**으로 변경할 데이터 찾기 또는 수정내역이 적용된 데이터를 미리 확인 할 수 있습니다.

텍스트마이닝하기



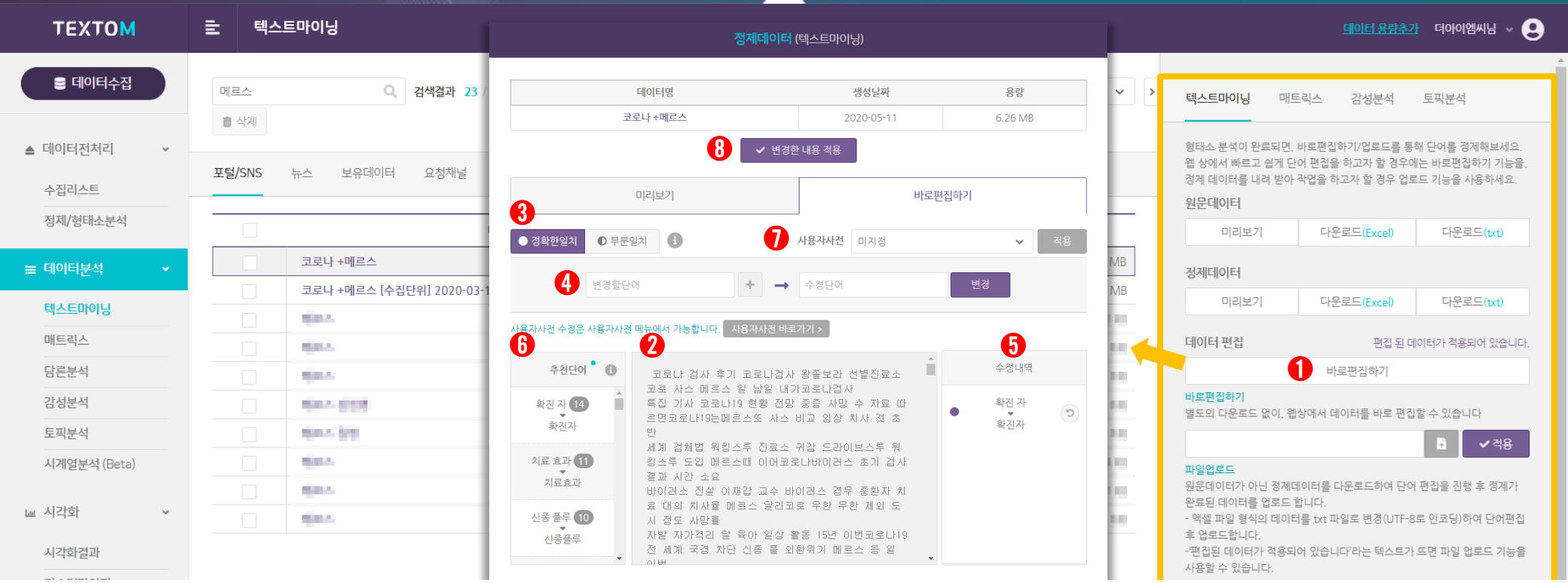
❸ 정확한일치/부분일치 **정확한일치**는 정확하게 일치하는 문자열만 변경합니다.

부분일치는 문서에서 부분적으로 일치하는 문자열을 포함하여 모두 변경합니다.

※ 예시) '이번v메르스v사태'와 '이번메르스사태'에서 '메르스'를 '코로나'로 변경하고자 할 때,

- **정확한일치**는 **띄어쓰기**가 있는 '이번v메르스v사태'를 '이번v코로나v사태'로 변경하고, '이번메르스사태'는 변경하지 않습니다.
- **부분일치**는 '이번v메르스v사태', '이번메르스사태'를 '이번v코로나v사태', '이번코로나사태'로 '메르스'를 포함한 **모든 문자열을 변경**합니다.

텍스트마이닝하기



❹ 단어변경 좌측 칸에 변경하고자 하는 단어를 입력하고 우측 칸에는 수정된 결과단어를 입력합니다.

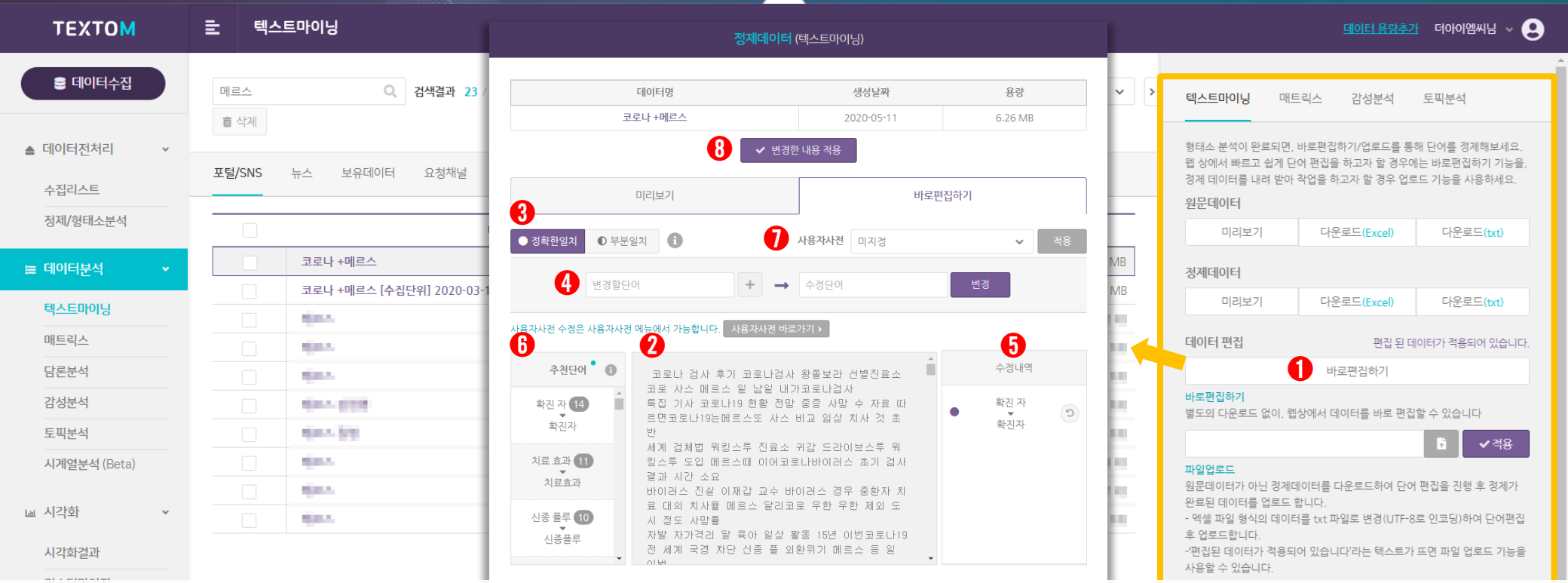
※ 변경할 단어 여러 개를 하나의 수정단어로 변경하고자 할 때, +(플러스)버튼을 클릭하면 변경할 단어를 추가할 수 있습니다.

※ 단어빈도와 N-gram 분석결과를 참고하면 수정할 유의어, 제거할 불용어, 결합할 단어 찾기에 효과적입니다.

❺ 수정내역 되돌리기 아이콘을 클릭하면 해당 단어의 변경을 취소할 수 있습니다.

※ 변경한 내용 적용 버튼을 누른 후에는 되돌릴 수 없습니다.

텍스트마이닝하기



❻ 추천단어 N-gram기반의 확률 모델을 통해 단어쌍을 선별하여 추천합니다.

※ 추천단어 리스트에서 변경하고자 하는 **추천단어를 클릭하면 입력을 하지 않아도 바로 수정**이 됩니다.

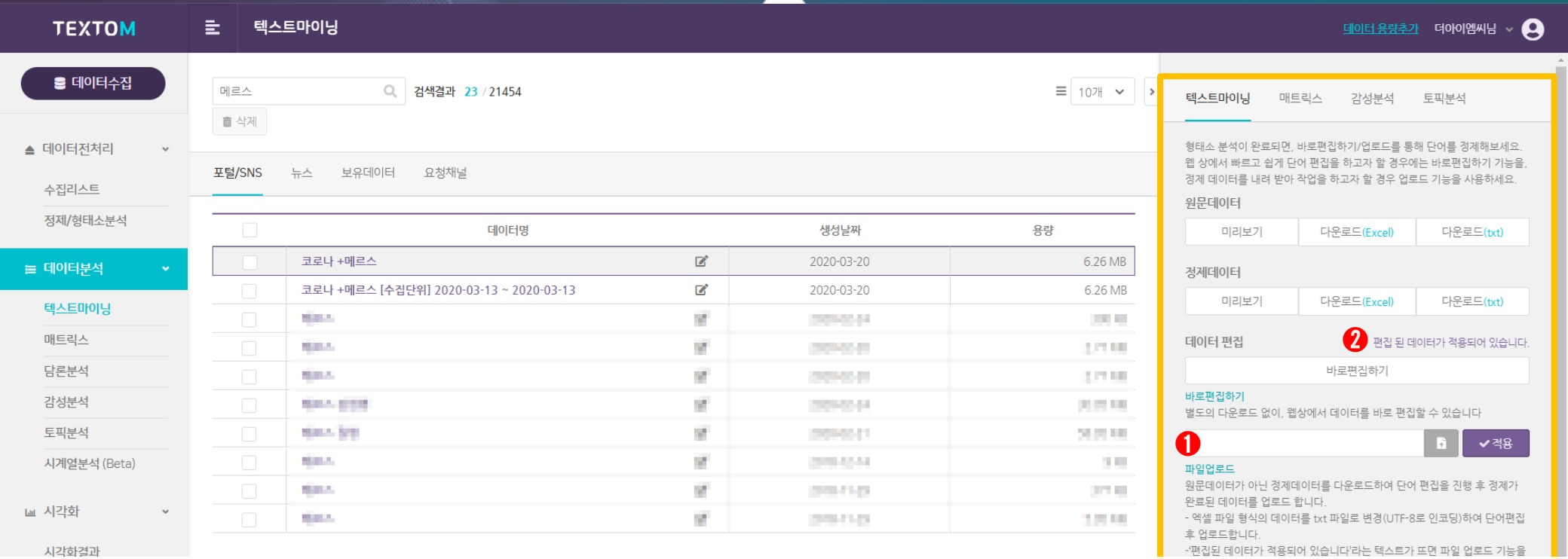
※ N-gram : 데이터에서 2개 단어가 연쇄적으로 표현된 횟수를 기록한 값

❼ 사용자사전 적용하기 사용자사전 설정 후 이용할 수 있는 기능으로, 적용하기를 누르면 데이터가 수정됩니다.

❽ 변경한 내용 적용하기 수정내역의 내용을 데이터에 적용하고자 한다면 '변경한 내용 적용'버튼을 클릭합니다.

※ 변경 내용을 적용한 후에는 수정 전으로 되돌릴 수 없으니, 수정내역 검토 및 정제데이터를 미리 내려 받아 두는 것을 추천합니다.

텍스트마이닝하기



❶ **파일업로드** 정제데이터를 다운받아 단어 편집 후, 편집이 완료 된 파일을 업로드 하는 데이터 편집 방법입니다.

※ 원문데이터가 아닌 정제데이터를 다운로드 받아 편집 해야 합니다.

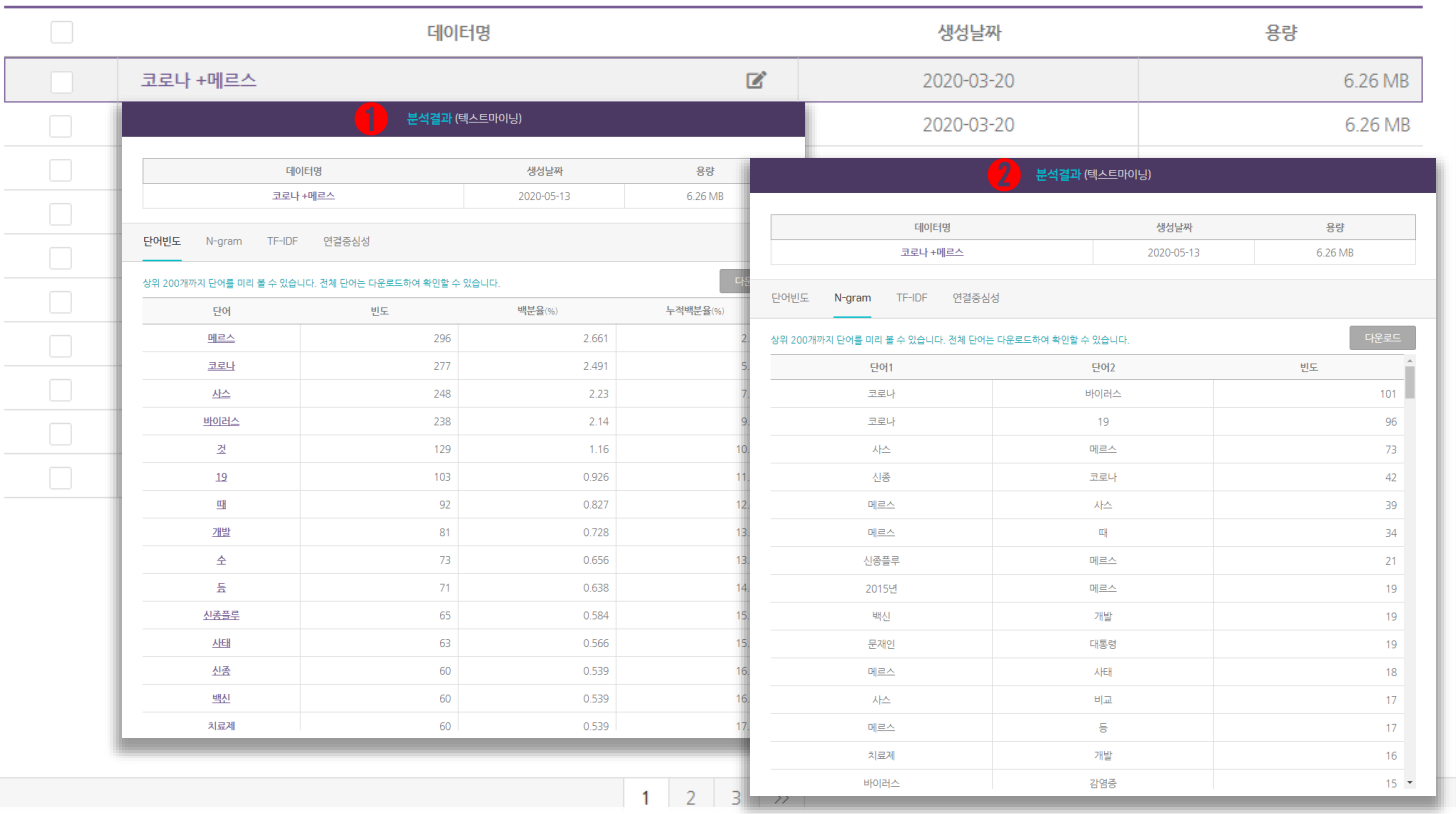
※ **엑셀파일 형식의 데이터를 txt 파일로 저장해야하며, txt 파일은 저장 시 인코딩을 UTF-8로 설정하여 저장 후 업로드해야 합니다.**

❷ 편집한 데이터의 **적용 진행 상황**을 확인할 수 있습니다.

※ 데이터 편집이 완료되면 '편집된 데이터가 적용되었습니다.'라는 문구가 등장하고 데이터 확인 후 다시 편집을 진행할 수 있습니다.

※ 단, **'편집한 내용을 적용하고 있습니다.'라는 문구가 등장할 경우엔 편집이 진행되고 있는 상황으로 도중에 편집을 진행하면 오류가 생길 수 있습니다.**

텍스트마이닝하기



❶ **단어빈도** 추출된 단어와 데이터 내 해당 단어의 빈도수를 나타냅니다.

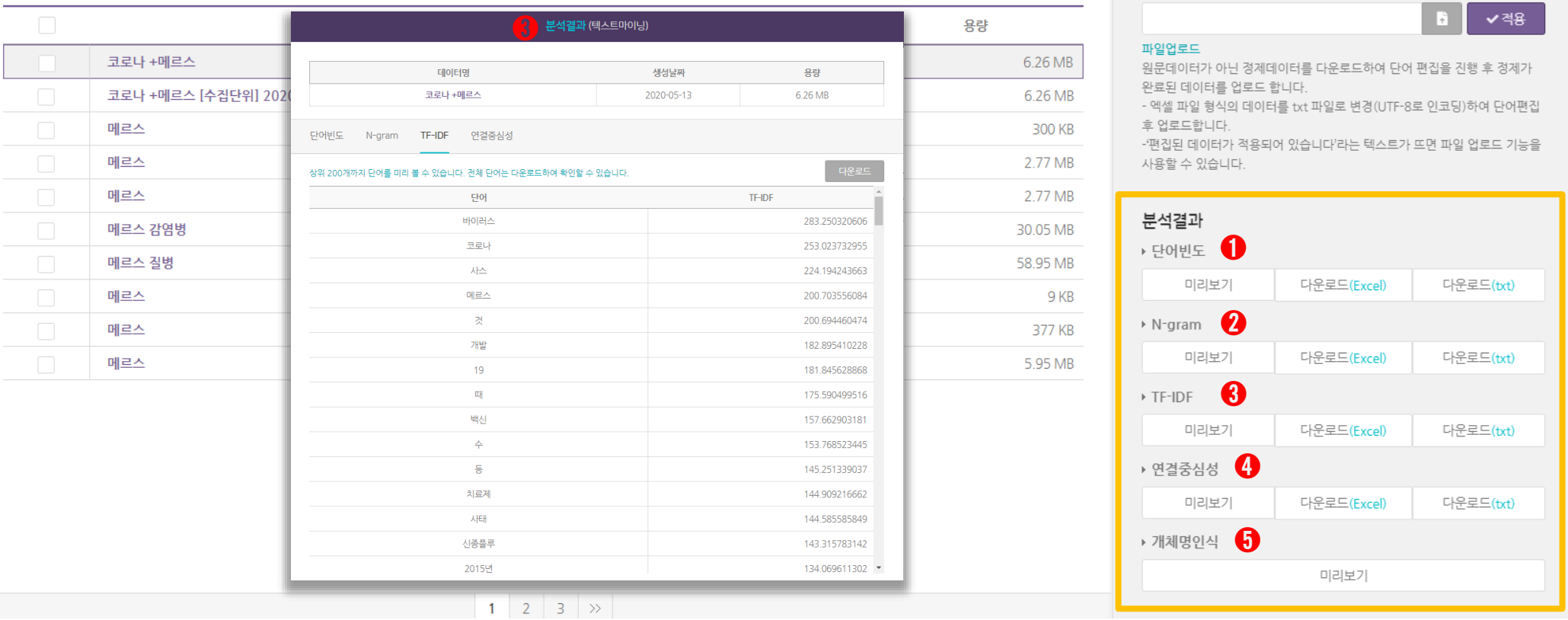
※ 빈도수가 높다는 것은 정제데이터 내에 해당 단어가 등장하는 빈도가 높다는 것을 의미합니다.

※ 백분율은 전체 언급량을 100으로 했을 때 언급된 양을 의미합니다.

❷ **N-gram** n개 단어의 연쇄를 의미합니다.

※ 단어1과 단어2의 빈도가 높다는 것은 두 단어가 나란히 등장하는 빈도가 높다는 것을 의미합니다.

텍스트마이닝하기



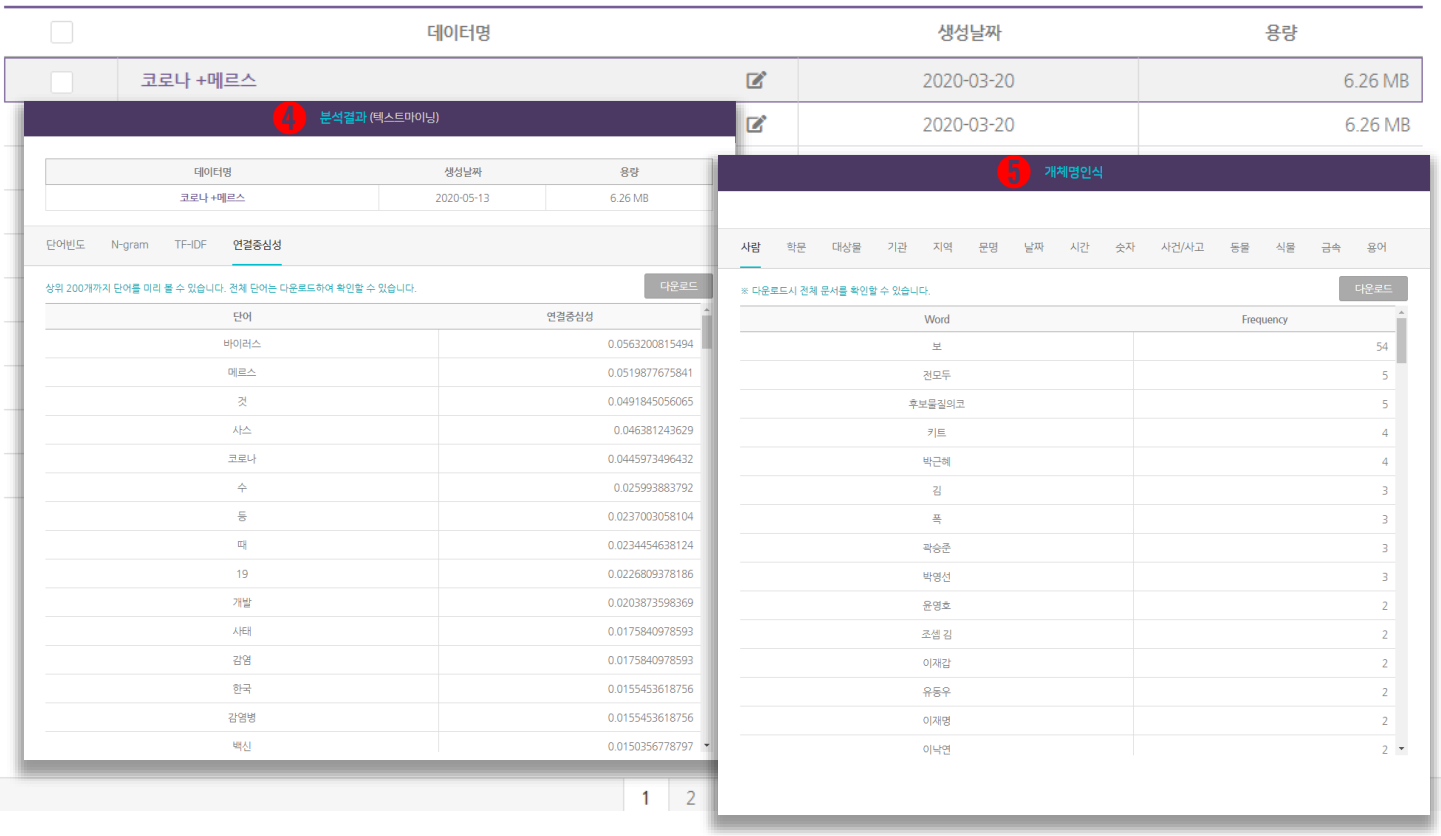
❸ TF-IDF TF(단어빈도)와 IDF(문서빈도의 역수)를 곱한 값으로 단어가 특정 문서에서 얼마나 중요한 지를 나타냅니다.

※ TF : 문서 내 특정단어의 빈도수 / DF : 여러 문서 내의 특정단어 빈도수 / IDF : DF의 역수

※ TF - IDF = TF X 1/DF

※ 특정 범위 내에서 모든 단어들의 빈도수와 단어가 포함된 문서(정보)들의 빈도수를 구한 후 역수를 취해 곱하여 문서의 중요도를 찾는 방법

텍스트마이닝하기



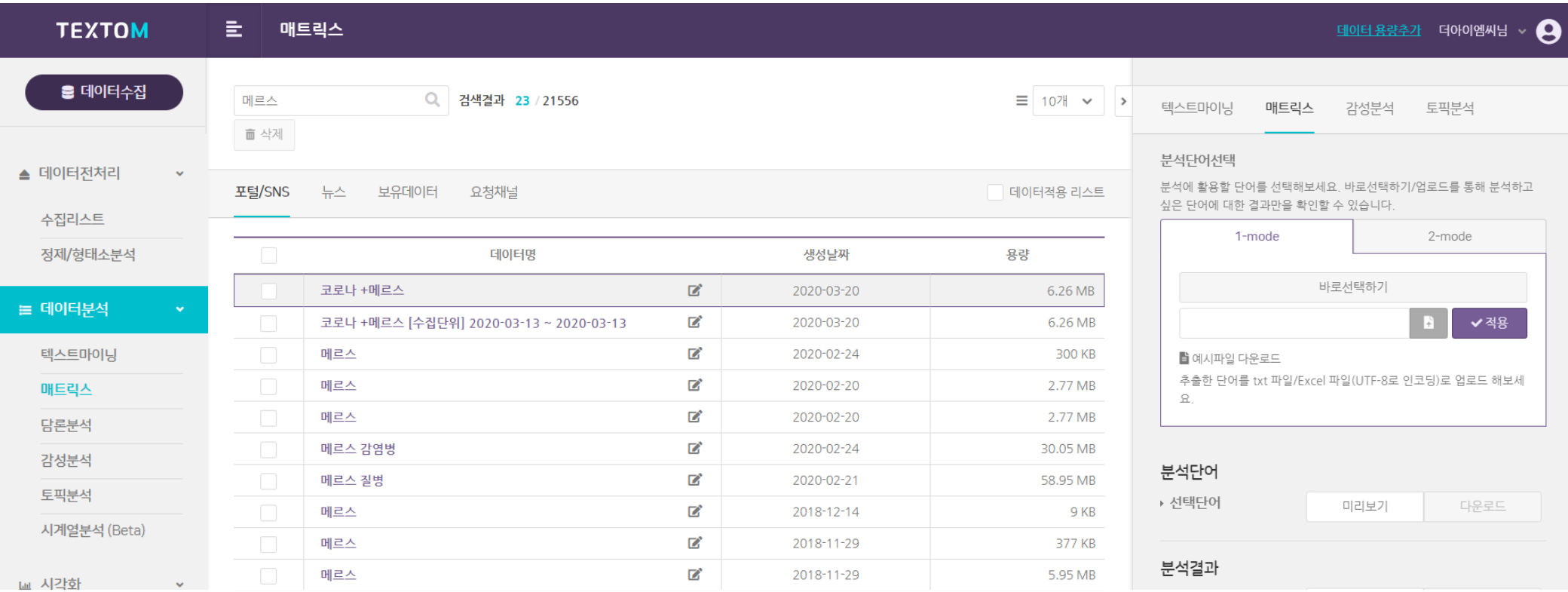
❹ 연결중심성 A라는 단어가 B라는 단어와 얼마나 많이 연결되어 있는지를 나타냅니다.

❺ 개체명인식 단어를 13개의 개체명 범주에 따라 분류하고 그 빈도수를 나타냅니다.

※ 13개의 개체명 : 사람, 학문, 대상물, 기관, 지역, 문명, 날짜, 시간, 숫자, 사건/사고, 동물, 식물, 금속, 용어

※ 단어빈도, N-gram, TF-IDF, 연결중심성, 개채명인식 모두 데이터를 정제를 거칠 때마다 순위 및 내용이 변동됩니다.

매트릭스 만들기

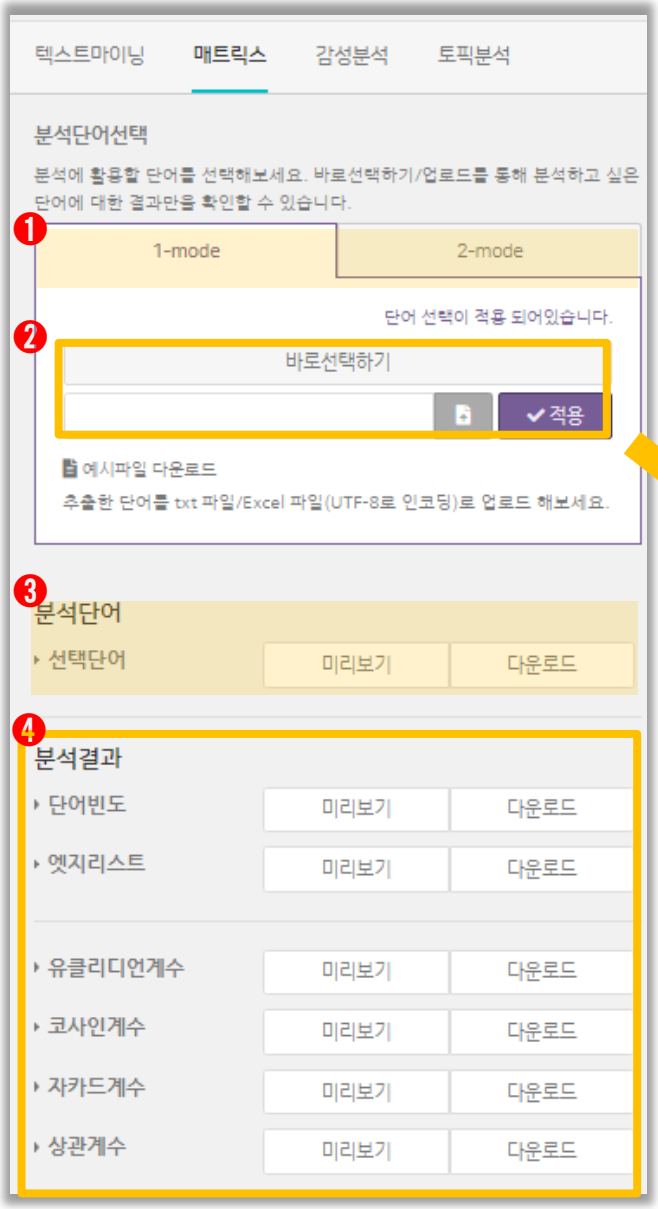


정제된 단어들 간의 동시출현빈도 (공출현빈도: co-occurrence)로 작성된 **매트릭스 데이터**를 제공합니다.
수학적 분석을 위한 소셜네트워크 데이터의 기본 형태는 **테이블 형태의 매트릭스**입니다.

※ 공출현 : 전체 텍스트 내(한 행) 특정 범위에서 노드들이 같이 출현하였을 때 이 범위 내에 있는 모든 노드들 간에 의미론적으로 상호 연관되는 관계가 있다고 가정

동시출현 빈도 행렬을 산출하여 분석하고자 하는 해당 데이터의 공출현 빈도, 근접관계 등을 통해 네트워크로 시각화가 가능합니다.
※ 텍스톰에서 제공하는 매트릭스를 이용하여 UCINET, NodeXL, Netminer, Gephi 등의 프로그램에서 시각화가 가능합니다.

매트릭스 만들기

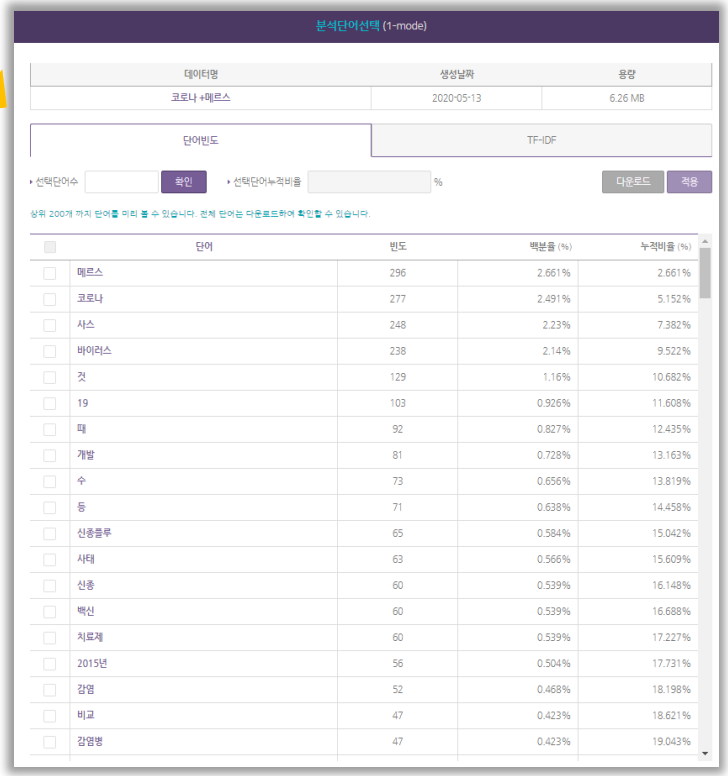


❶ **분석단어 선택** 1-mode 또는 2-mode 분석 방법 선택

※ 1-mode는 단어(메인노드)간의 관계를 분석하고 싶을 때, 2-mode는 단어(메인노드)와 범주(서브노드)간의 관계를 분석하고 싶을 때 선택합니다.

❷ **바로선택하기**

분석하고 싶은 단어를 웹상에서 바로 선택하거나, 단어리스트를 직접 업로드 할 수 있습니다.



❸ **분석단어**

선택한 단어와 빈도를 미리보기 또는 다운로드 받을 수 있습니다.

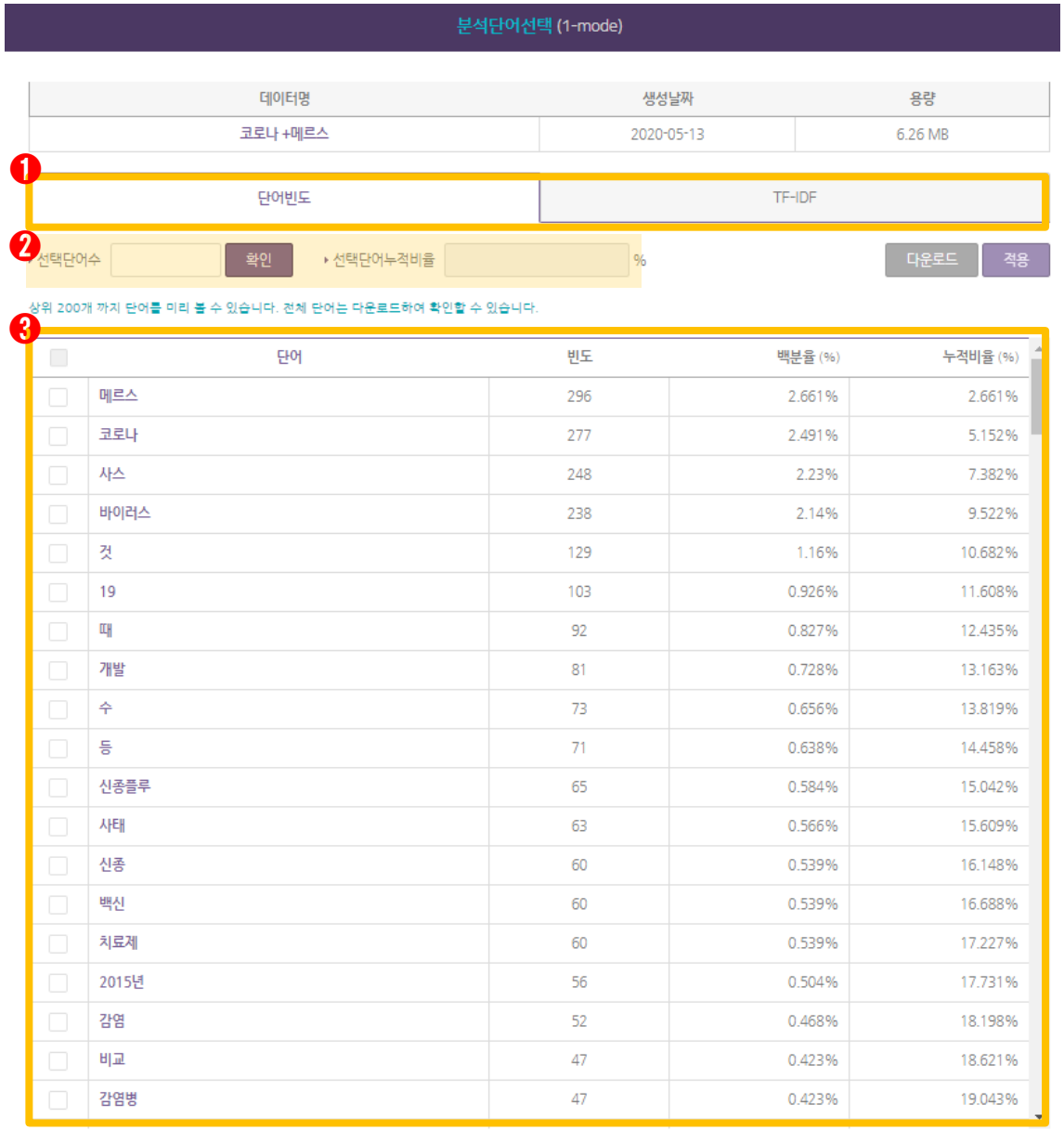
❹ **분석결과**

단어빈도와 엣지리스트 그리고 단어 간 공출현을 통한 유사도 계수를 계산한 방식에 따라 4가지 결과값으로 제공합니다.

※ 작성한 매트릭스로 UCINET를 활용하여 시각화를 하실 수 있습니다.

UCINET 활용법 알아보기

매트릭스 만들기



❶ **단어빈도** 또는 **TF-IDF** 순으로 단어를 볼 수 있습니다.

❷ **선택 단어 수**

선택 할 단어의 개수를 입력 후 확인을 누르면 선택 단어누적 비율이 자동으로 계산되어 나오며 빈도 또는 TF-IDF순으로 선택이 됩니다.

※ 가시성이 좋은 시각화를 만들기 위하여 선택 단어 수를 50~70개 사이로 선택 할 것을 추천 드립니다.

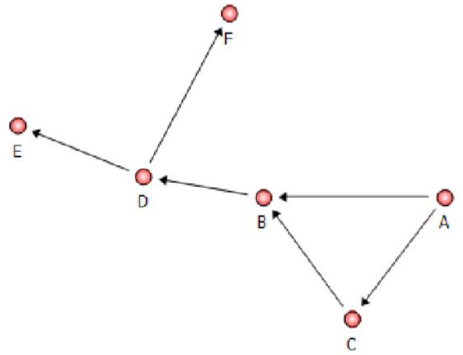
❸ **단어선택**

상위 200개의 단어 중 분석하고 싶은 단어를 직접 선택할 수 있습니다.

※ 더 많은 데이터를 보고 싶은 경우, 텍스트마이닝 – 분석결과 – 단어빈도 데이터를 다운로드 받아 상위빈도 순으로 전체 단어 리스트에서 원하는 단어를 추출하여 업로드합니다.

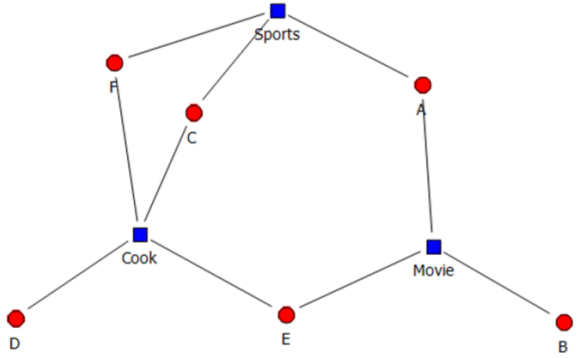
매트릭스 만들기

❶ 1-mode 분석 방법



	A	B	C	D	E	F
A	0	1	1	0	0	0
B	0	0	0	2	0	0
C	0	3	0	0	0	0
D	0	0	0	0	1	2
E	0	0	0	0	0	0
F	0	0	0	0	0	0

❷ 2-mode 분석 방법



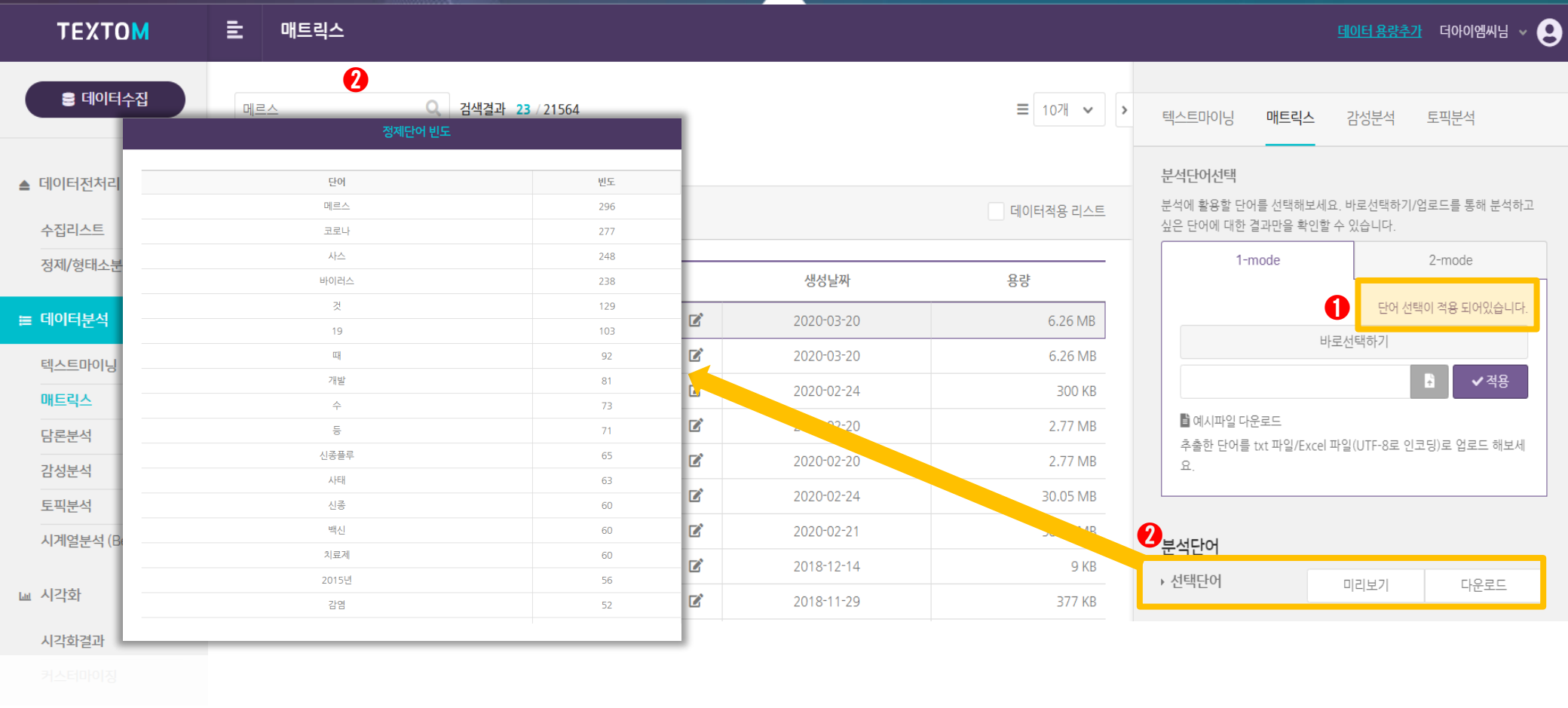
	Sports	Movie	Cook
A	1	1	0
B	0	1	0
C	1	0	1
D	0	0	1
E	0	1	1
F	1	0	1

※ 1-mode와 2-mode의 차이점은 무엇인가요?

- 1-mode 의 경우, 매트릭스와 같이 행과 열이 같은 단어로 이루어진 매트릭스 (모든 네트워크의 가장 일반적인 유형)
- 2-mode 의 경우, 매트릭스의 행과 열이 다른 단어로 이루어진 매트릭스
  (서로 다른 유형의 조직 간의 관계, 조직과 조직에 속한 조직구성원 간의 관계, 이벤트와 그 이벤트에 참여하는 참석자 간의 관계 등. 예: 기업과 문화재단, 동호회와 회원, 구직자와 취업박람회, 영화와 영화배우 간의 관계)

2-mode로 분석한 사회 이슈 자세히 보기

매트릭스 만들기



❶ **단어 선택이 적용 되어있습니다** 라는 문구가 나타나면 아래의 분석단어와 분석결과들을 확인 할 수 있습니다

❷ 선택한 단어와 빈도값을 미리보기 또는 다운로드 할 수 있습니다. 다운로드 버튼 클릭 시 엑셀 파일로 다운로드 됩니다.

매트릭스 만들기

분석결과

❶ 단어빈도 — 미리보기 | 다운로드

❷ 엣지리스트 — 미리보기 | 다운로드

유클리디언계수 — 미리보기 | 다운로드

코사인계수 — 미리보기 | 다운로드

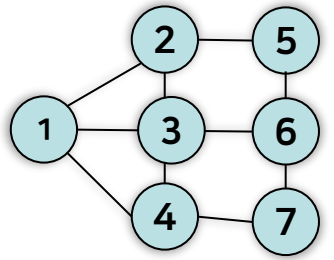
자카드계수 — 미리보기 | 다운로드

상관계수 — 미리보기 | 다운로드

❶ 단어빈도의 매트릭스

분석결과 (1-mode) - 빈도

	메르스	코로나	사스	바이러스	것	19	때	개발	수	등	신종플루	사태	신종	백신	치료제	2015년	감염	비교	감염병	한국	확산
메르스	0	207	229	168	83	65	74	36	36	50	51	52	38	40	37	44	25	40	27	22	25
코로나	207	0	176	230	77	147	42	36	40	22	55	44	64	39	10	52	29	26	12	30	43
사스	229	176	0	183	87	45	58	20	23	39	36	17	50	32	20	42	23	34	26	13	31
바이러스	168	230	183	0	71	34	19	35	32	24	18	12	77	21	35	28	49	12	9	14	26
것	83	77	87	71	0	27	30	6	15	8	12	13	22	12	11	25	8	9	5	6	17
19	65	147	45	34	27	0	22	11	25	7	31	16	7	9	3	17	7	9	13	7	10
때	74	42	58	19	30	22	0	0	19	7	20	24	6	2	3	10	7	16	10	9	8
개발	36	36	20	35	6	11	0	0	15	9	16	8	10	56	45	2	9	0	14	6	7
수	36	40	23	32	15	25	19	15	0	7	17	6	6	10	12	8	5	5	6	22	2
등	50	22	39	24	8	7	7	9	7	0	15	11	7	7	4	5	5	6	22	2	3
신종플루	51	55	36	18	12	31	20	16	17	15	0	7	2	10	3						
사태	52	44	17	12	13	16	24	8	6	11	7	0	6	1	6						
신종	38	64	50	77	22	7	6	10	6	7	2	6	0	1	4						
백신	40	39	32	21	12	9	2	56	10	7	10	1	1	0	31						
치료제																					



<Mini Social graph>

❷ 엣지리스트의 매트릭스

분석결과 (1-mode) - 엣지리스트

word1	word2	weight
메르스	코로나	207
메르스	사스	229
메르스	바이러스	168
메르스	것	83
메르스	19	65
메르스	때	74
메르스	개발	36
메르스	수	36
메르스	등	50
메르스	신종플루	51
메르스	사태	52
메르스	신종	38
메르스	백신	40
메르스	치료제	37
메르스	2015년	44
메르스	감염	25
메르스	비교	40

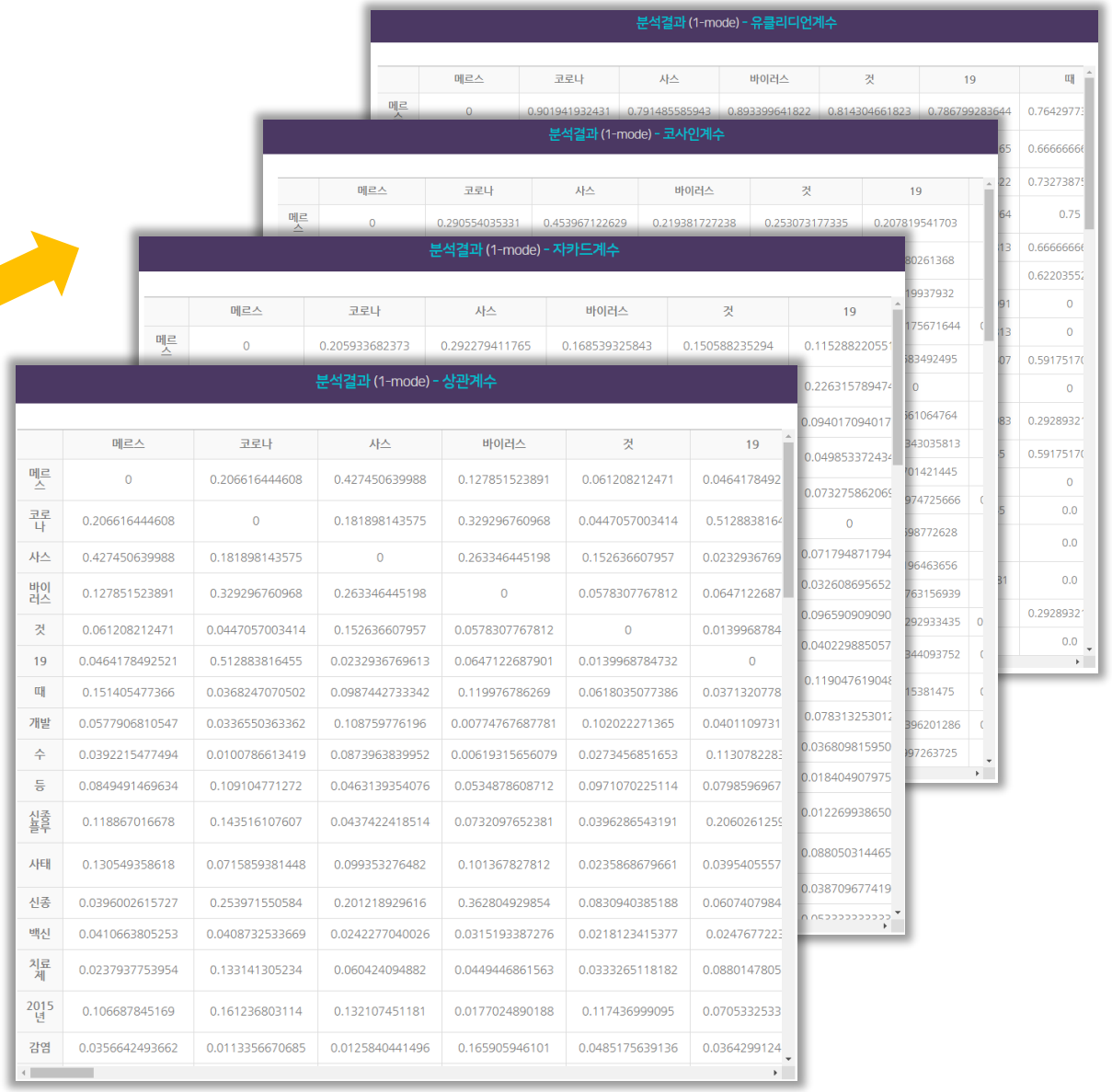
❶ 단어빈도

전체 문서 내에서 단어의 출현빈도가 높은 순서대로 단어와 빈도 수를 표시합니다.

❷ 엣지리스트(Edge List)

단어와 단어, 노드와 노드를 짝지어 목록으로 보여줍니다.

매트릭스 만들기



❶ 유클리디언 계수

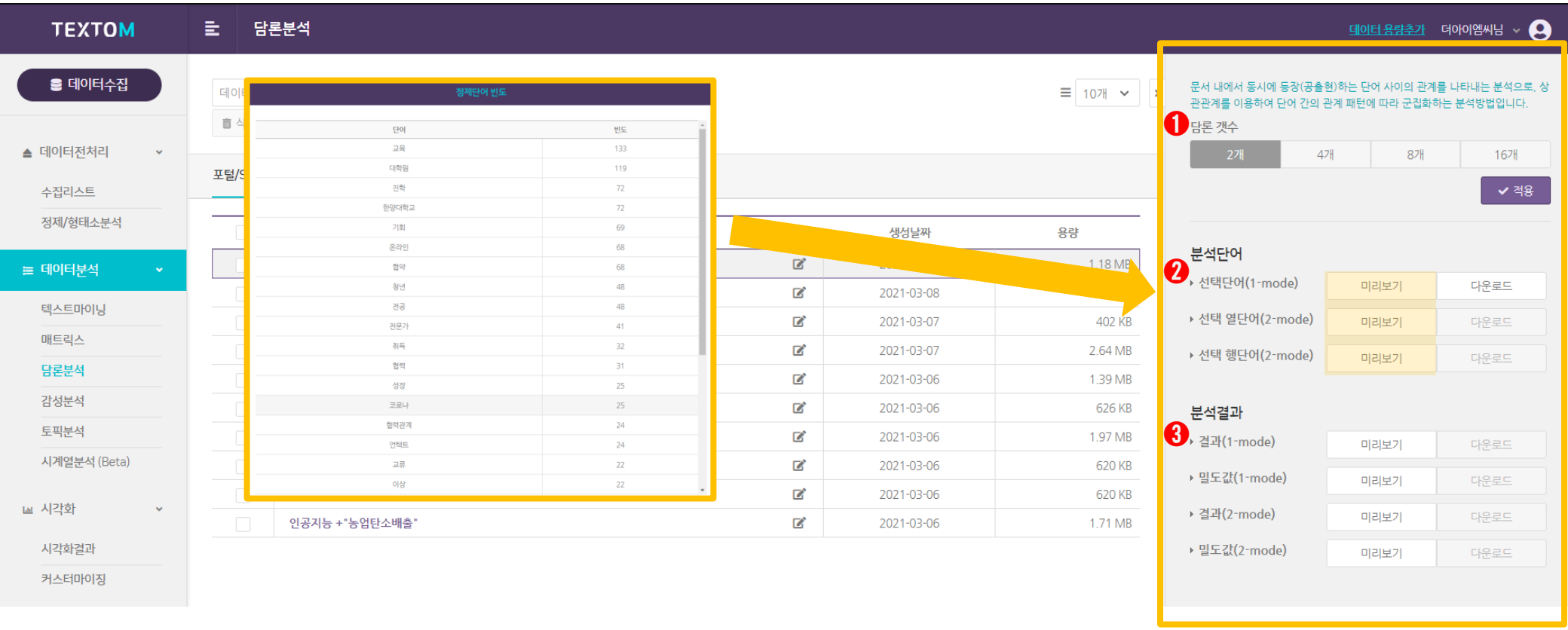
❷ 코사인 계수

❸ 자카드 계수

❹ 상관 계수

단어 간의 유사도 측정방법 자세히 알아보기

담론분석 하기



❶ **블록 개수** 담론분석은 각 키워드의 관계적 유사성을 측정하기 위해 상관분석을 반복적으로 실시해 상관계수 값에 따라 그룹화되는 것을 보며 블록개수를 설정 및 반복하여 최적의 블록개수를 찾습니다.

❷ **분석단어** 매트릭스 단계에서 추출한 단어 미리보기 입니다.

❸ **분석결과** 상관분석에서 도출되는 상관계수의 밀도값은 -1과 1 사이의 값을 가지게 됩니다.
일반적으로, 상관분석에서 -1, 1의 값으로 결과가 도출 됩니다.

담론분석 자세히 알아보기

담론분석 하기

분석결과 - 1-mode

0	무료수강	발전	온라인	
1	진학	청년	한양대학교	협약
2	개발	공부	교류	기회
3	교육철학	대학원	성장	
4	이념	이상	인터넷	자립기회	전공
5	취득	코로나	협력	협력관계	
6	교육	온라인대학	전문가	
7	미래	사이버	언택트	

❶

분석결과 - 밀도값 (1-mode)

	Block_1	Block_2	Block_3	Block_4	Block_5	Block_6	Block_7	Block_8
Block_1	1	1	1	1	1	1	1	1
Block_2	1	1	1	1	1	1	1	1
Block_3	1	1	1	1	1	1	1	1
Block_4	1	1	1	1	1	1	1	1
Block_5	1	1	1	1	1	1	1	1
Block_6	1	1	1	1	1	1	1	1
Block_7	1	1	1	1	1	1	1	1
Block_8	1	1	1	1	1	1	1	1

❷

❶ 결과

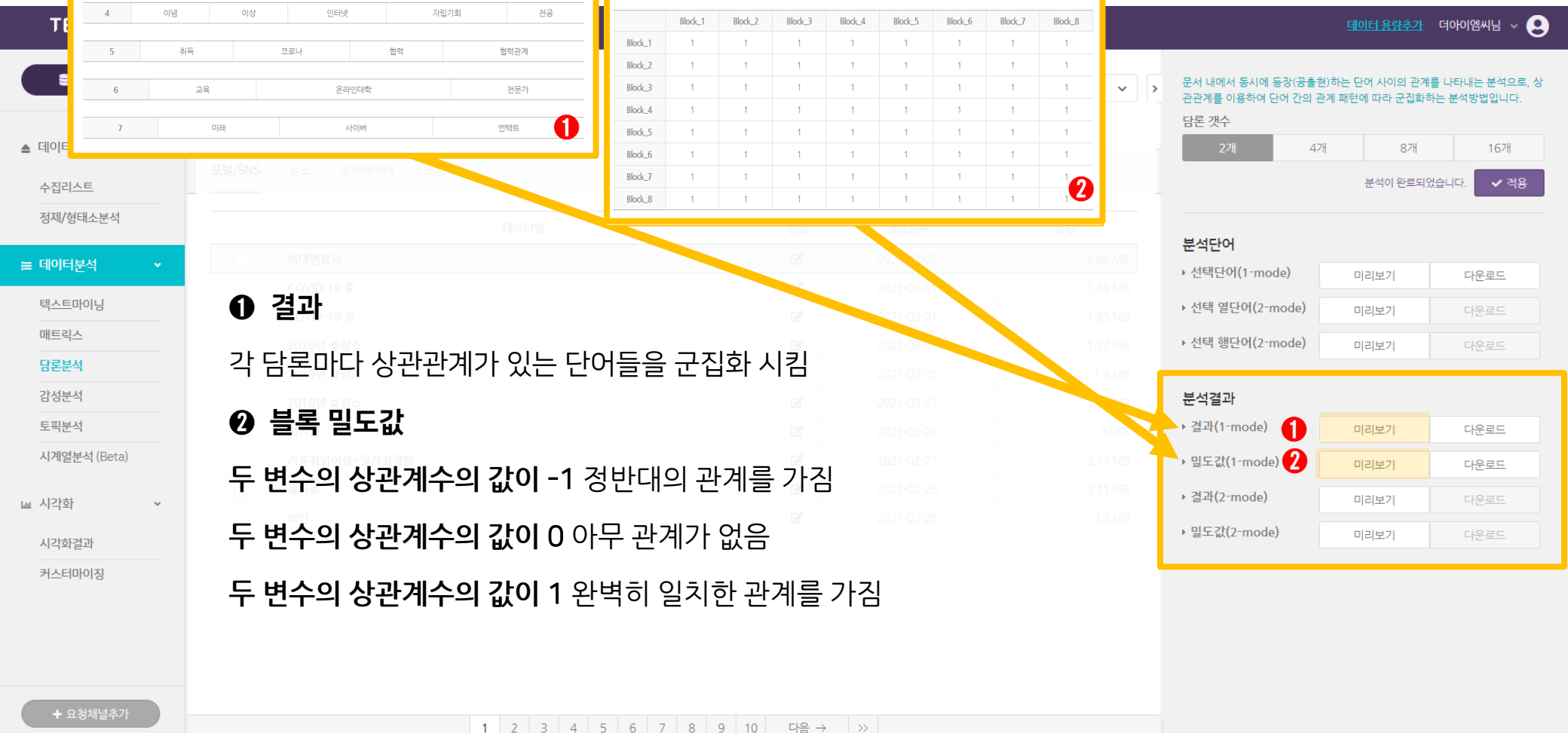
각 담론마다 상관관계가 있는 단어들을 군집화 시킴

❷ 블록 밀도값

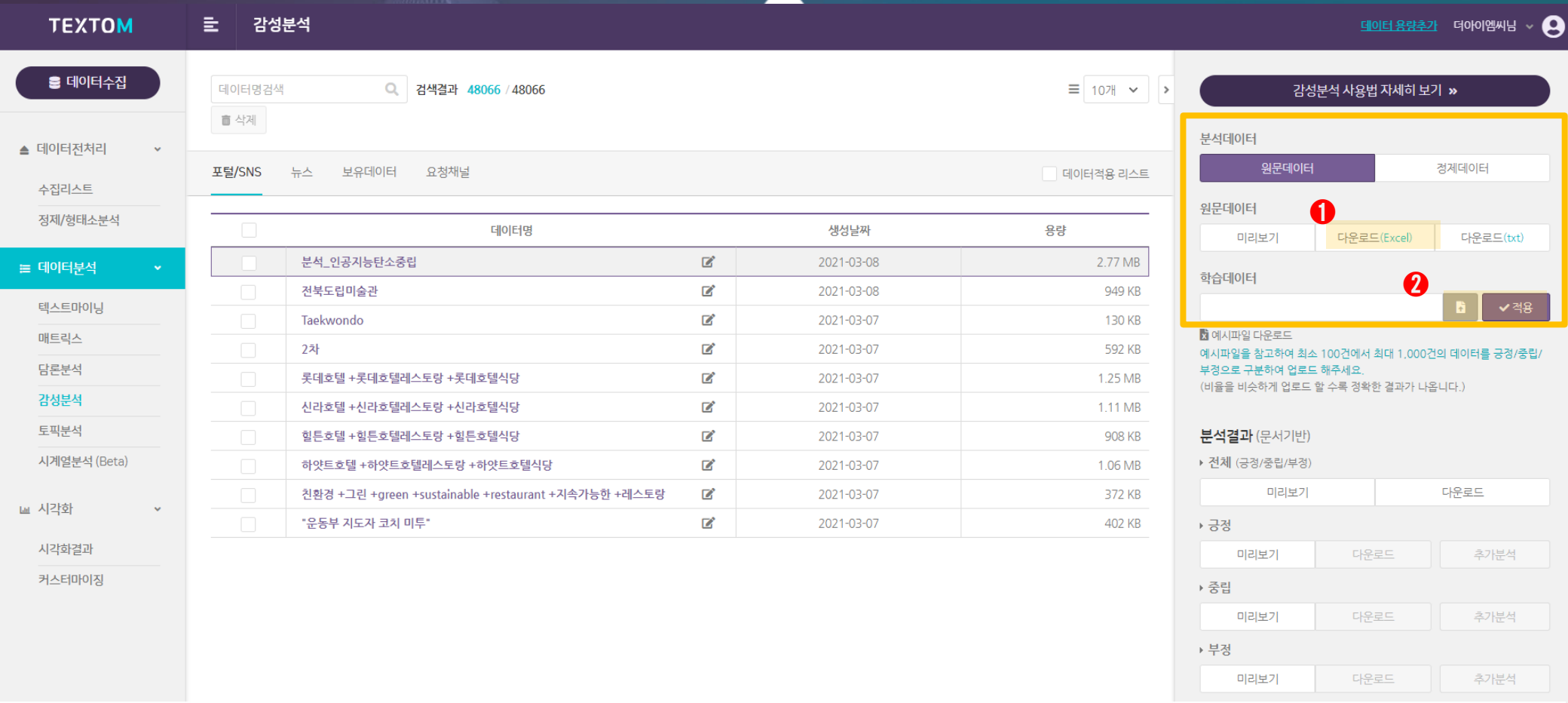
두 변수의 상관계수의 값이 -1 정반대의 관계를 가짐

두 변수의 상관계수의 값이 0 아무 관계가 없음

두 변수의 상관계수의 값이 1 완벽히 일치한 관계를 가짐



감성분석 하기



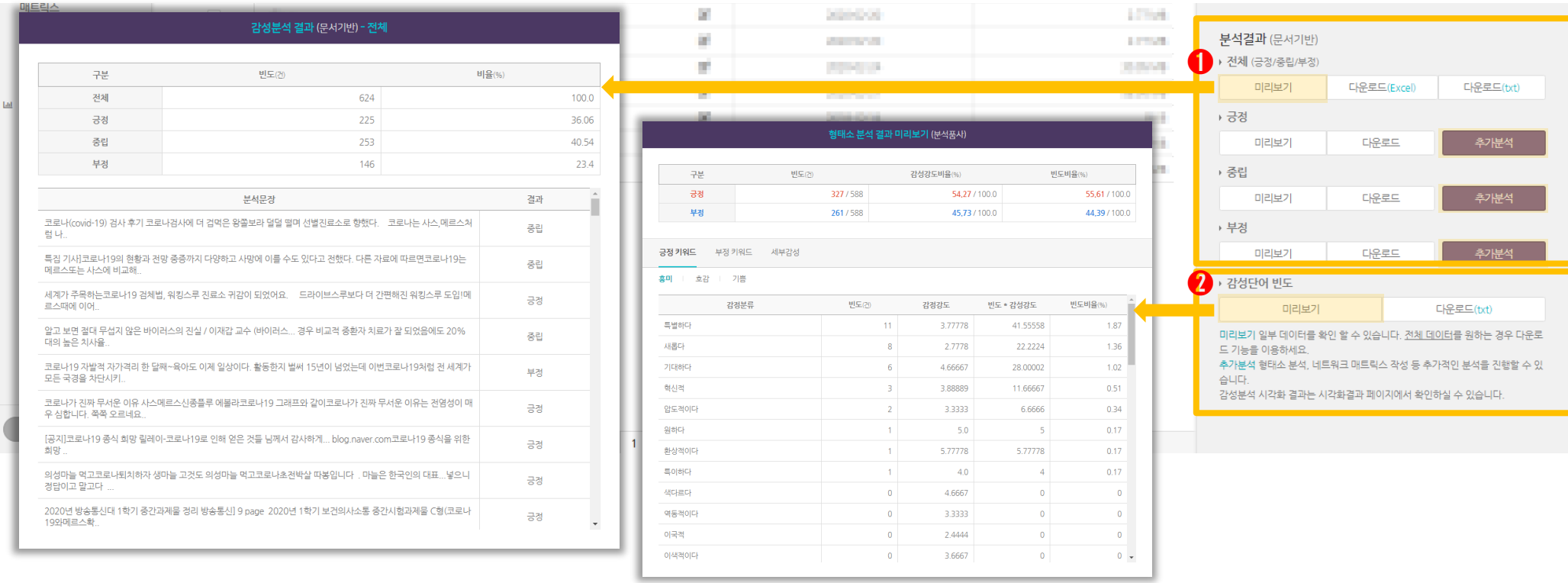
원문데이터에 직접 극성을 부여한 학습데이터를 업로드하여, **문서의 극성 분석결과**와 **감성단어 빈도 분석결과**를 확인할 수 있습니다

❶ **분석데이터** 원문데이터 혹은 정제데이터를 선택한 후, **다운로드(Excel)**하여 학습데이터를 만들어야 합니다

❷ **학습데이터** 전체 데이터를 분류하기 위한 기준이 되는 데이터로, 각 극성의 비율을 비슷하게 태깅하여 업로드해줍니다

학습데이터 만드는 법 자세히 알아보기

감성분석 하기



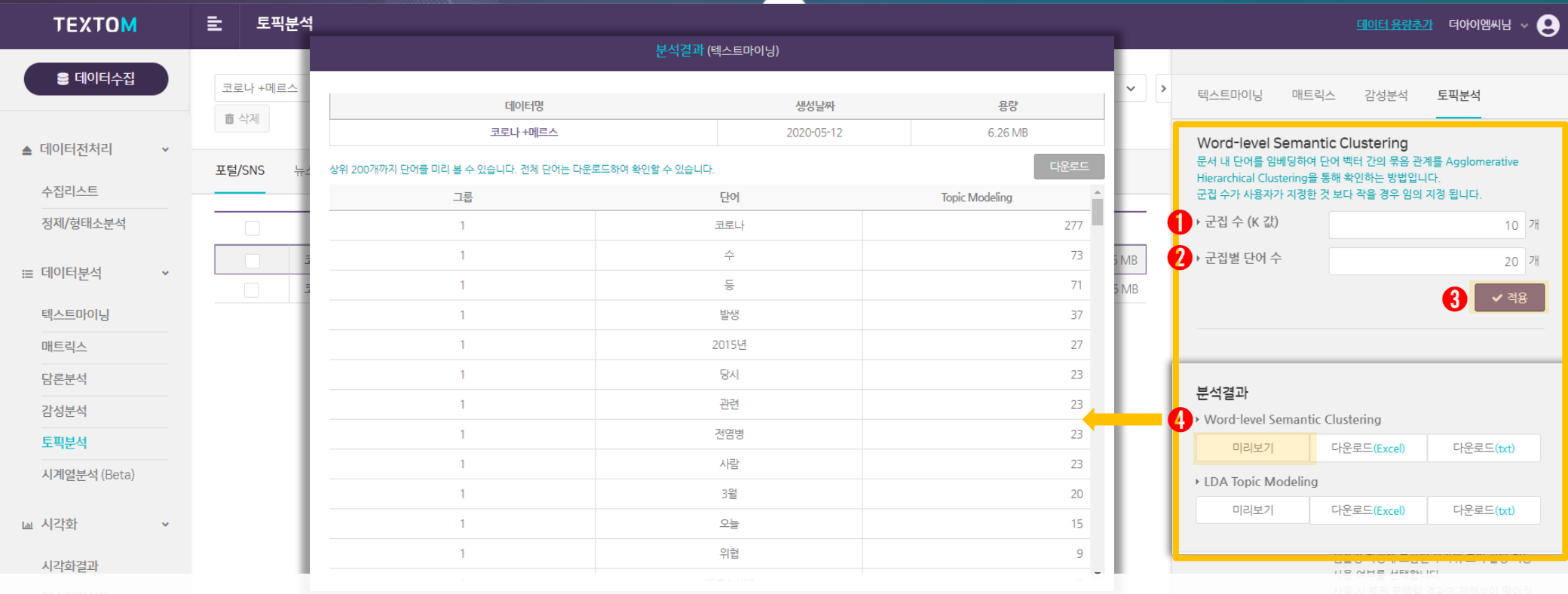
분석결과 학습데이터가 적용되면 분석결과를 확인할 수 있습니다

❶ **문서의 극성 분석** 학습데이터 기준으로 분류된 전체 문서의 극성을 확인할 수 있으며, 긍정·부정·중립으로 분류된 단어들에 대하여 개별적인 분석을 하고 싶은 경우, **추가분석** 기능으로 가능합니다

※ **감성단어**들은 감성분석에 탑재된 형태소 분석기로 출력된 키워드이므로, 텍스트마이닝 페이지에서 직접 편집한 데이터와 차이가 있습니다

※ **감성강도**는 세부 감성(흥미, 호감, 기쁨, 통증, 슬픔, 분노, 두려움, 놀람, 거부감) 안에서 표현의 세기를 의미하며, 7점 만점입니다

토픽분석 하기



Word-level Semantic Clustering 문서 내 단어들의 공출현 관계를 기준으로 벡터화하여 인접 단어를 같은 군집으로 묶어줍니다

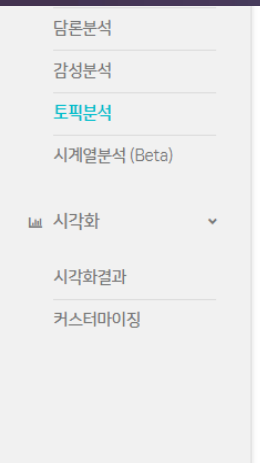
❶ **군집 수(K값)** 만들고 싶은 군집의 개수를 적어줍니다

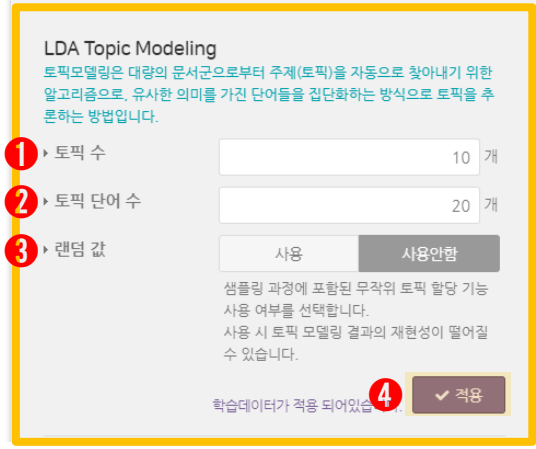
❷ **군집별 단어 수** 군집 내에 들어갈 단어의 개수를 적어줍니다

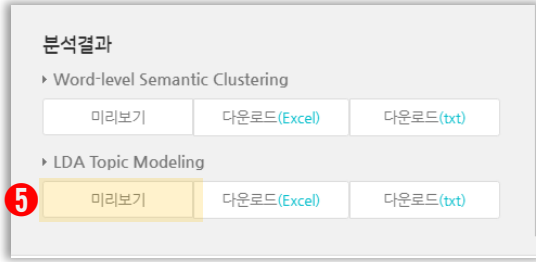
❸ **적용** 군집 수(K값)과 군집별 단어 수를 설정 후 적용 버튼을 클릭합니다

❹ **분석결과** 적용이 완료되면 연관성이 높은 단어들끼리 군집으로 분류가 된 결과를 확인할 수 있습니다

토픽분석 하기







LDA Topic Modeling 대량의 문서군으로부터 주제(토픽)을 자동으로 찾아내기 위한 알고리즘으로, 유사한 의미를 가진 단어들을 집단화 합니다

❶ **토픽 수** 만들고 싶은 그룹의 개수를 적어줍니다

❷ **토픽 단어 수** 그룹 내에 들어갈 단어의 개수를 적어줍니다

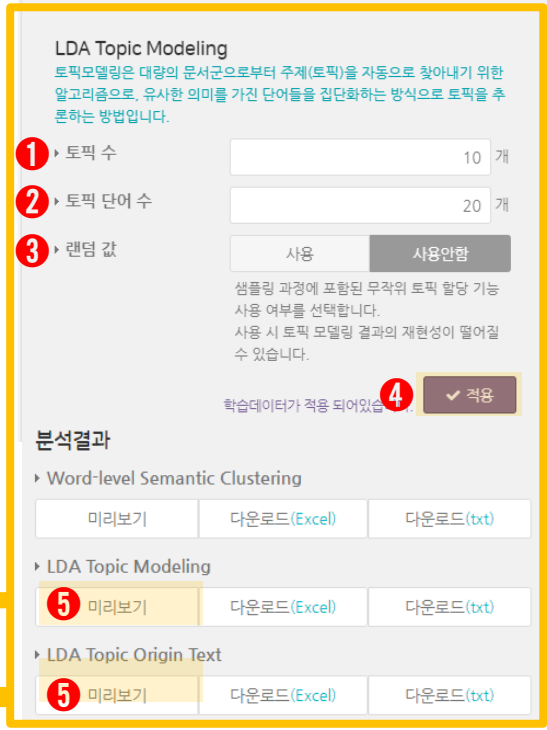
❸ **랜덤 값** 무작위 할당을 진행할 경우 분석 결과의 재현성이 떨어지기 때문에, 같은 데이터로 분석을 진행하더라도 결과값이 달라질 수 있습니다. 분석 결과의 재현성을 확보하고 싶은 경우 **사용안함**을 선택해주세요

❹ **적용** 토픽 수와 토픽 단어 수, 랜덤 값 설정 후 적용 버튼을 클릭합니다

❺ **분석결과** 적용이 완료되면 키워드가 어떤 그룹으로 분류가 되었는지 확인할 수 있습니다

※ Topic Modeling 숫자값은 해당하는 토픽에 단어가 들어갈 확률을 백분율 형태로 표시한 값입니다

토픽분석 하기



LDA Topic Modeling 대량의 문서군으로부터 주제(토픽)을 자동으로 찾아내기 위한 알고리즘으로, 유사한 의미를 가진 단어들을 집단화 합니다

❶ **토픽 수** 만들고 싶은 그룹의 개수를 적어줍니다

❷ **토픽 단어 수** 그룹 내에 들어갈 단어의 개수를 적어줍니다

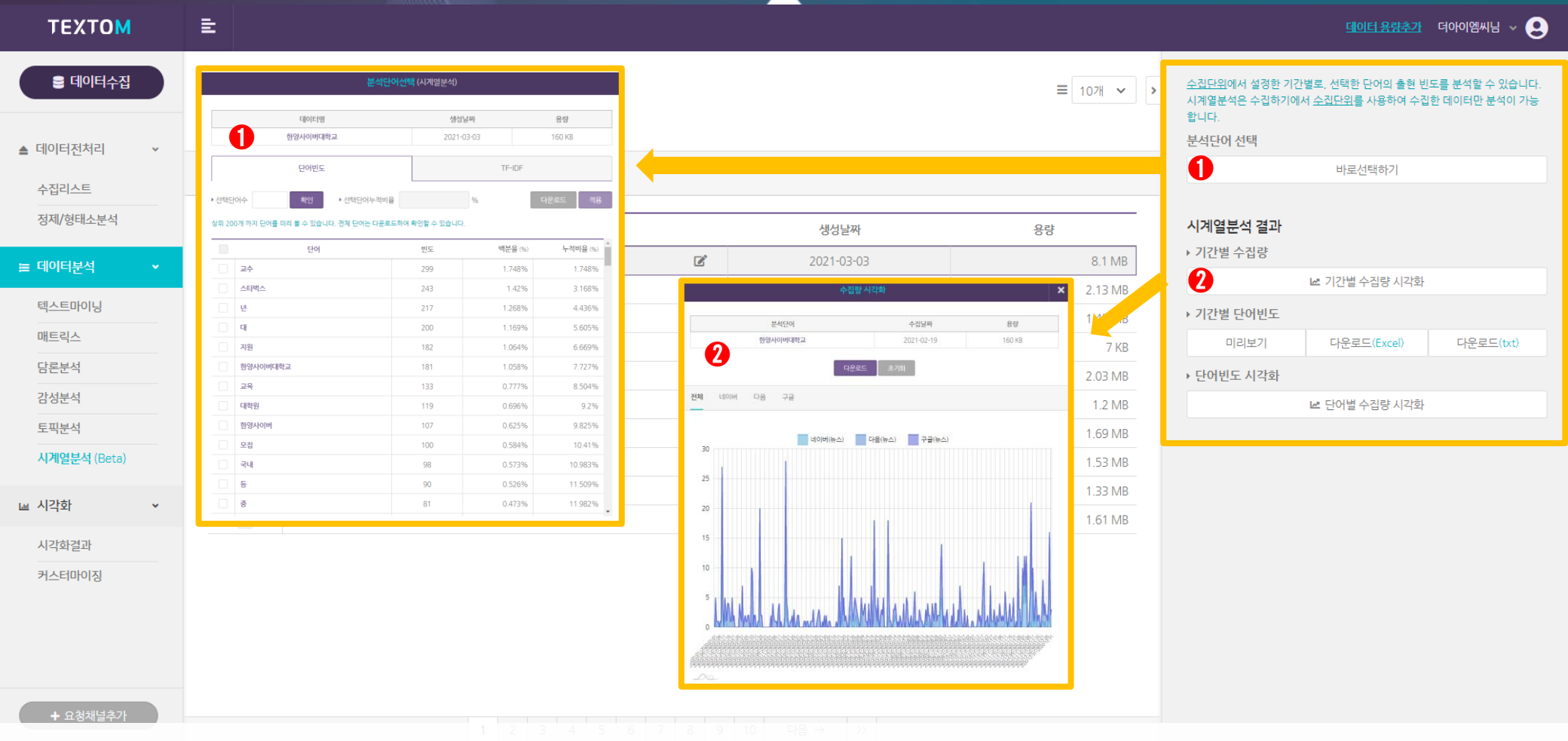
❸ **랜덤 값** 무작위 할당을 진행할 경우 분석 결과의 재현성이 떨어지기 때문에, 같은 데이터로 분석을 진행하더라도 결과값이 달라질 수 있습니다. 분석 결과의 재현성을 확보하고 싶은 경우 **사용안함**을 선택해주세요

❹ **적용** 토픽 수와 토픽 단어 수, 랜덤 값 설정 후 적용 버튼을 클릭합니다

❺ **분석결과** 적용이 완료되면 키워드가 어떤 그룹으로 분류가 되었는지 확인할 수 있습니다

※ Topic Modeling 숫자값은 해당하는 토픽에 단어가 들어갈 확률을 백분율 형태로 표시한 값입니다

시계열분석 하기 (Beta)

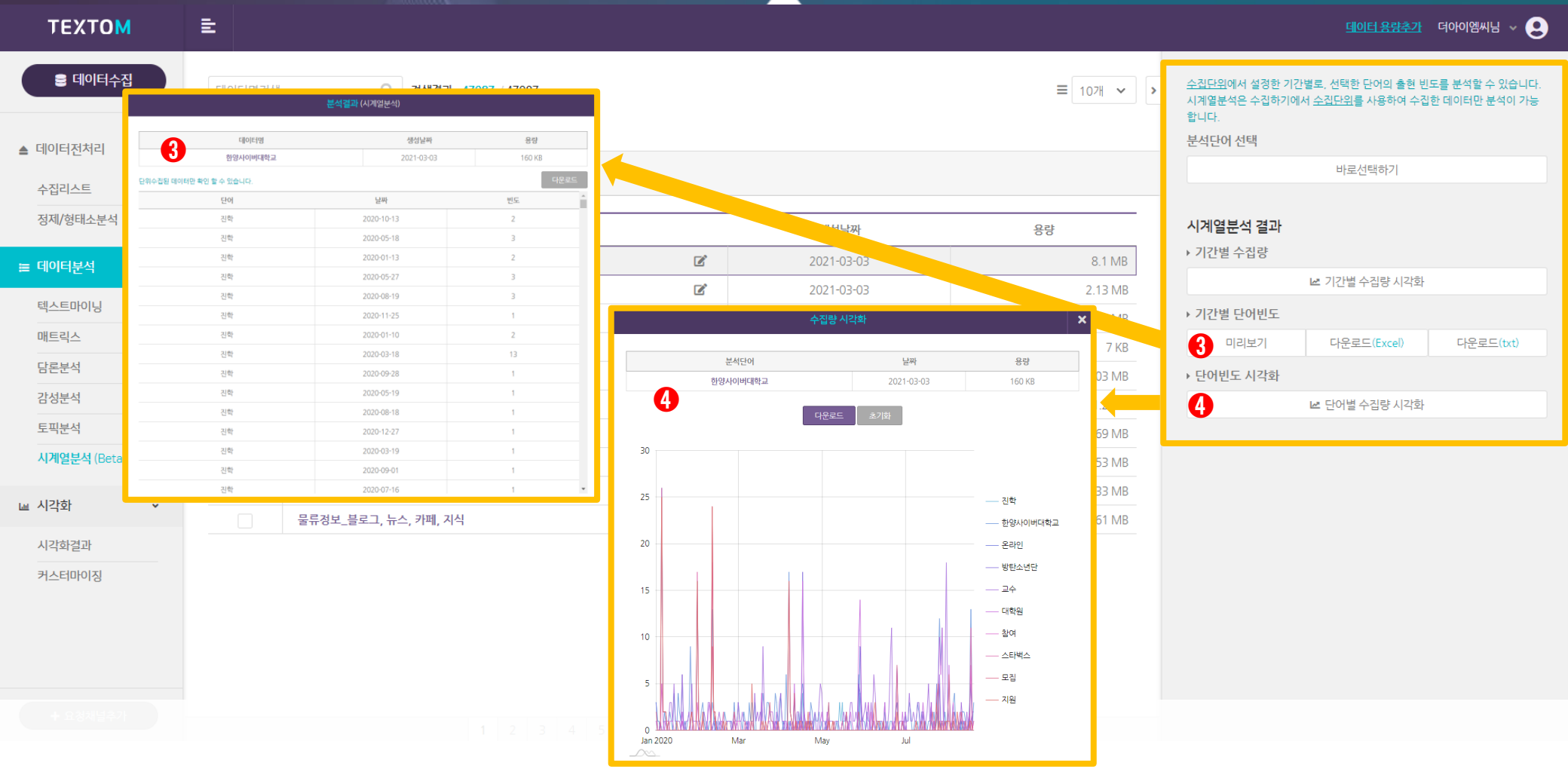


❶ **분석단어 선택** 분석하고 싶은 상위 단어를 적용시켜줍니다.(최대 10개)

❷ **기간별 수집량** 기간별로 수집한 각 채널의 수집량을 시각화 차트로 보여줍니다.

※ **수집단위**를 이용하여 수집을 진행해야 시계열 분석 사용이 가능합니다.

시계열분석 하기 (Beta)

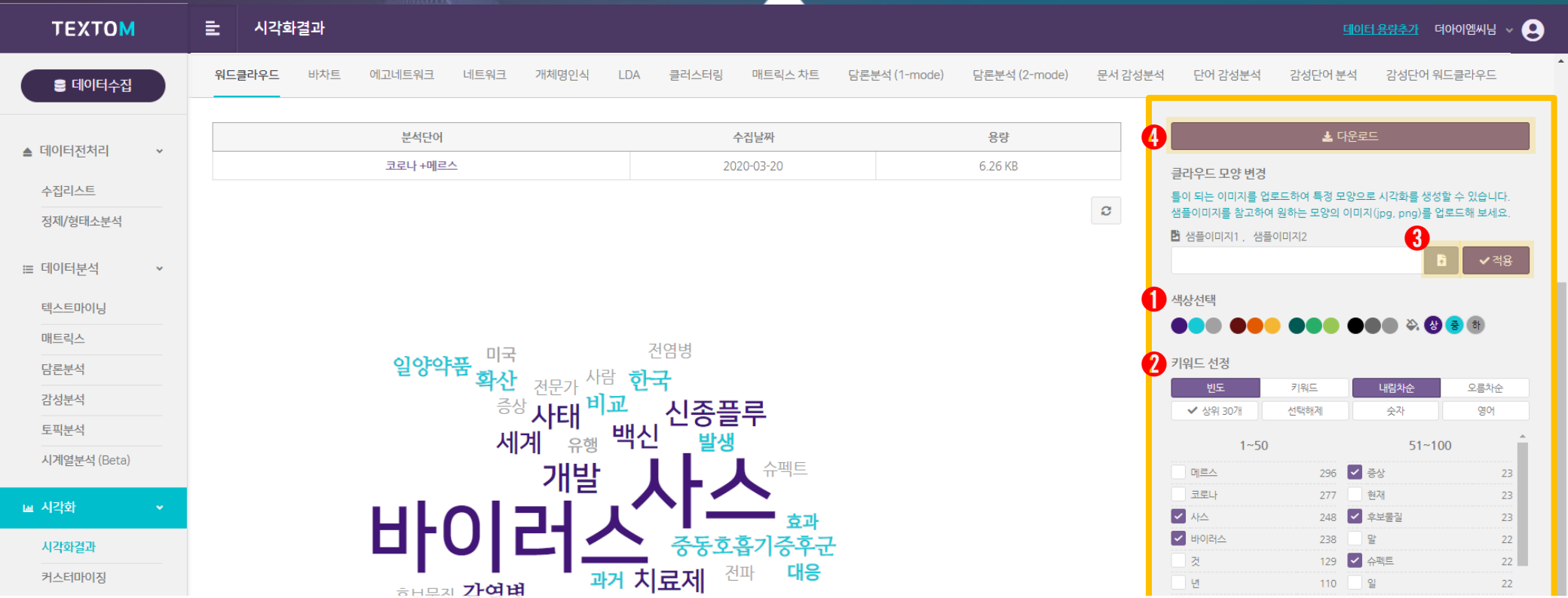


❸ **기간별 단어빈도** 기간별 단어의 빈도를 보여줍니다.

❹ **단어빈도 시각화** 기간별 단어빈도를 시각화 차트로 보여줍니다.

※ **수집단위**를 이용하여 수집을 진행해야 시계열 분석 사용이 가능합니다.

시각화 결과 확인하기



워드클라우드는 문서의 키워드를 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 시각화로, 단어빈도 결과값을 사용합니다

❶ **색상선택** 4가지의 기본 색상 조합 또는 우측의 **상, 중, 하**에서 직접 원하는 색상을 선택할 수 있습니다

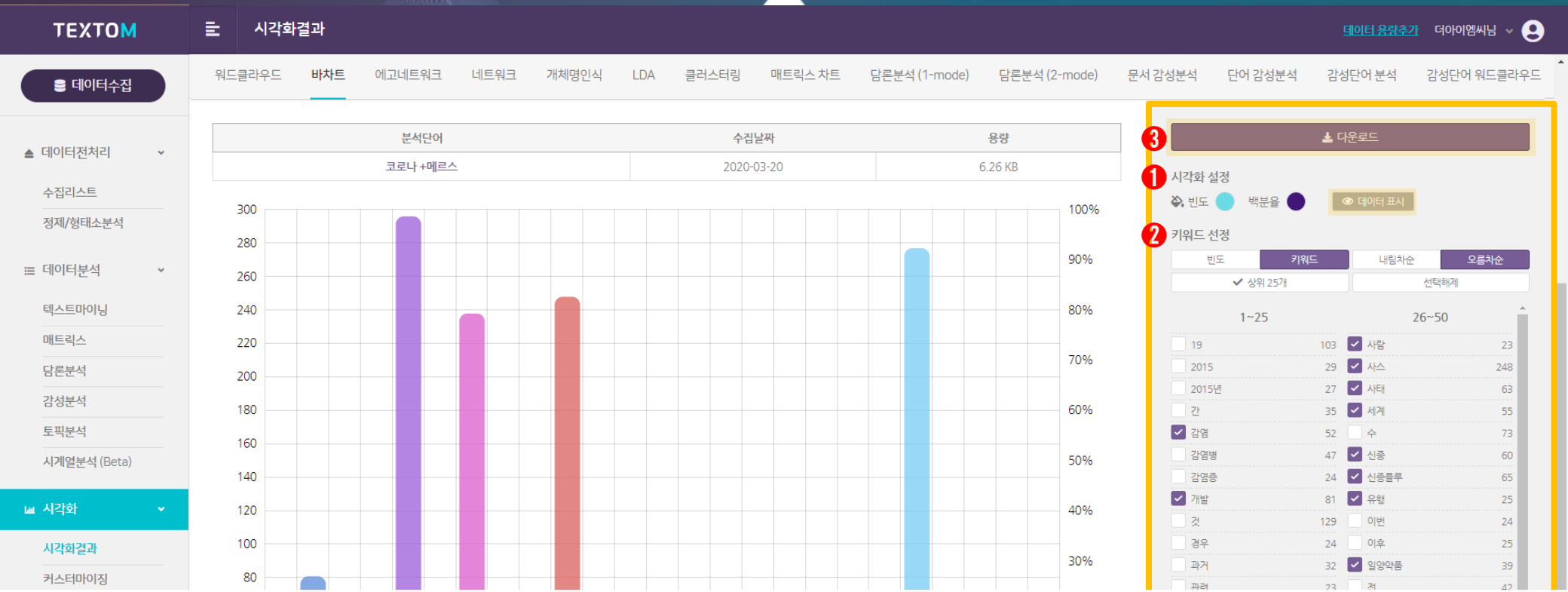
❷ **키워드 선정** **빈도** 또는 **키워드**(가나다) 기준으로 정렬할 수 있으며 중요하다고 생각되는 키워드들을 선택해줍니다

※ 상위 빈도 순으로 최대한 유의미한 키워드들을 선택하는 것이 좋습니다

※ 선택한 키워드 중 숫자나 영어가 있을 시, **숫자** 또는 **영어** 버튼을 클릭하면 해당 키워드가 강조되며 다른 키워드들은 회색조 처리가 됩니다

❸ **클라우드 모양 변경** 배경이 없는 jpg, png 이미지를 업로드하면 업로드한 모양의 워드클라우드가 만들어집니다

시각화 결과 확인하기



바차트는 문서의 단어빈도를 빈도에 비례하는 길이의 막대로 나타낸 차트입니다

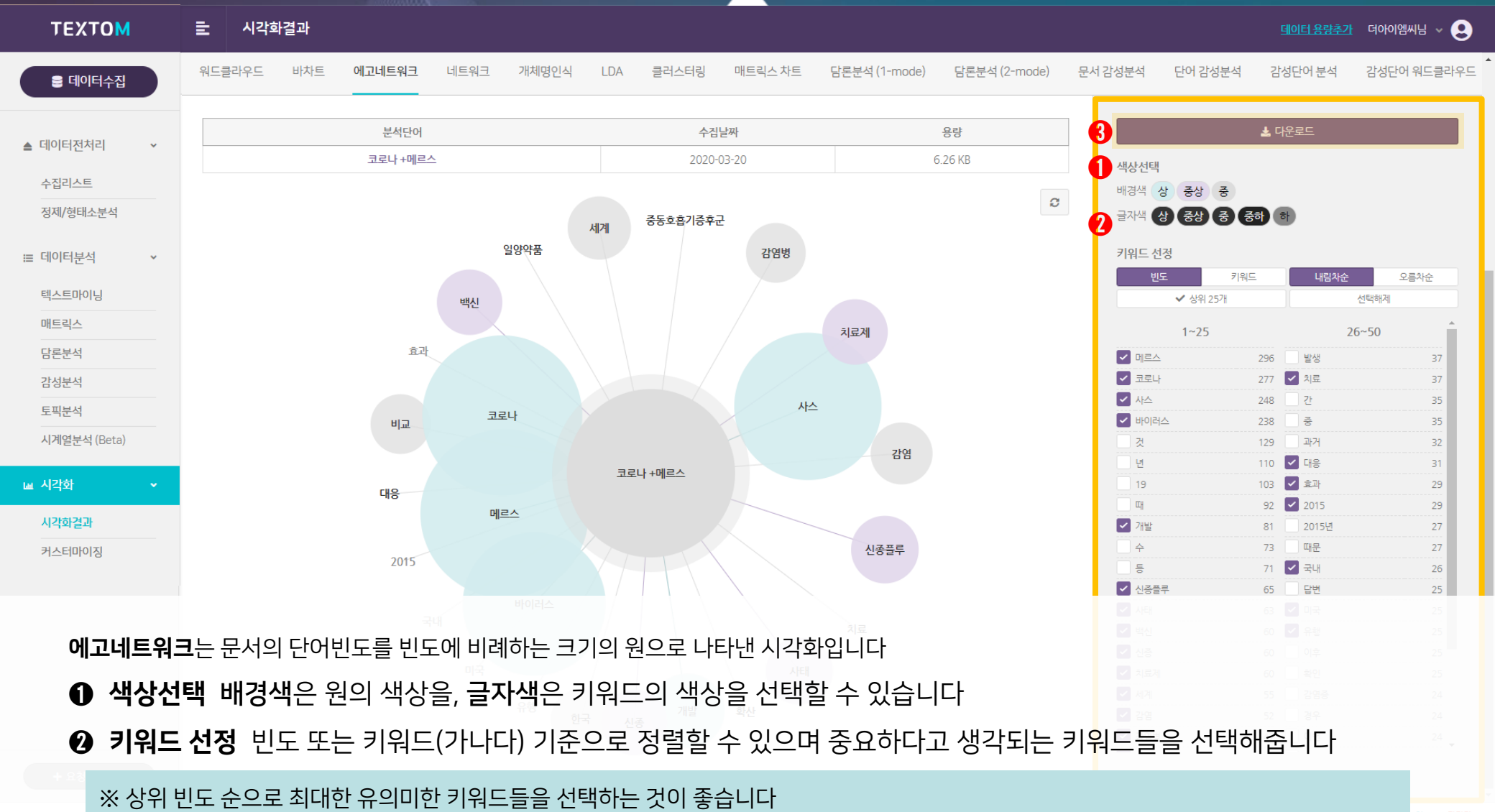
❶ **시각화 설정** **빈도**는 막대의 색상을, **백분율**은 점의 색상을 선택할 수 있으며, **데이터 표시**는 차트 위 빈도, 백분율 값 표시 여부를 선택할 수 있습니다

❷ **키워드 선정** 빈도 또는 키워드(가나다) 기준으로 정렬할 수 있으며 중요하다고 생각되는 키워드들을 선택해줍니다

※ 상위 빈도 순으로 최대한 유의미한 키워드들을 선택하는 것이 좋습니다

※ 막대 위에 마우스를 올리시면 데이터 값을 확인할 수 있습니다

시각화 결과 확인하기



에고네트워크는 문서의 단어빈도를 빈도에 비례하는 크기의 원으로 나타낸 시각화입니다

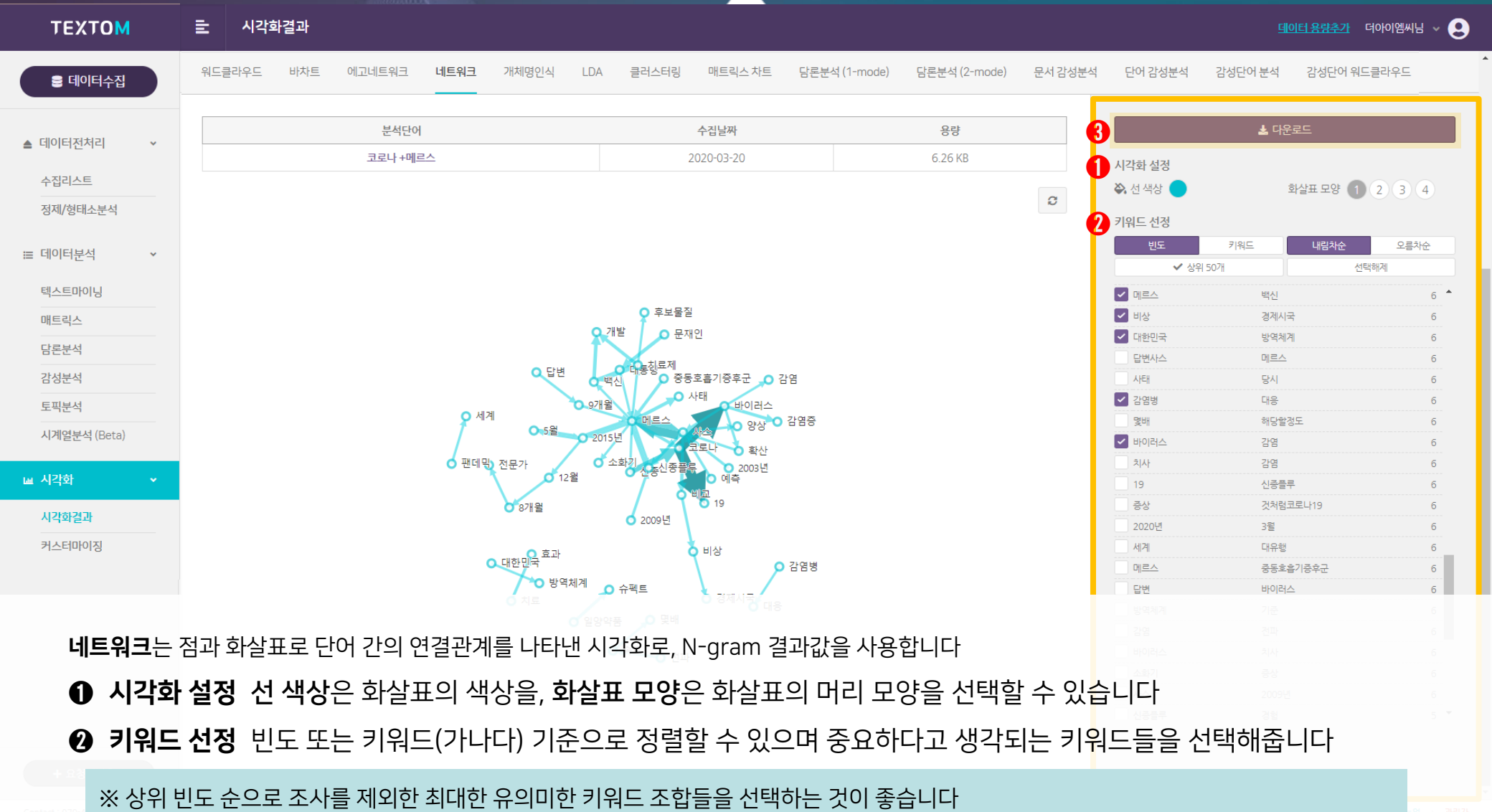
❶ **색상선택** **배경색**은 원의 색상을, **글자색**은 키워드의 색상을 선택할 수 있습니다

❷ **키워드 선정** 빈도 또는 키워드(가나다) 기준으로 정렬할 수 있으며 중요하다고 생각되는 키워드들을 선택해줍니다

※ 상위 빈도 순으로 최대한 유의미한 키워드들을 선택하는 것이 좋습니다

※ 원들이 한곳에 모아져 겹쳐서 보이는 경우, 가운데 원을 잡고 흔들어 주시면 고르게 분산됩니다

시각화 결과 확인하기



네트워크는 점과 화살표로 단어 간의 연결관계를 나타낸 시각화로, N-gram 결과값을 사용합니다

❶ **시각화 설정** **선 색상**은 화살표의 색상을, **화살표 모양**은 화살표의 머리 모양을 선택할 수 있습니다

❷ **키워드 선정** 빈도 또는 키워드(가나다) 기준으로 정렬할 수 있으며 중요하다고 생각되는 키워드들을 선택해줍니다

※ 상위 빈도 순으로 조사를 제외한 최대한 유의미한 키워드 조합들을 선택하는 것이 좋습니다

※ 원문에서 한 단어였었던 키워드들이 각각의 키워드로 많이 나타난다면, 데이터 편집을 더 해주셔야 합니다

시각화 결과 확인하기

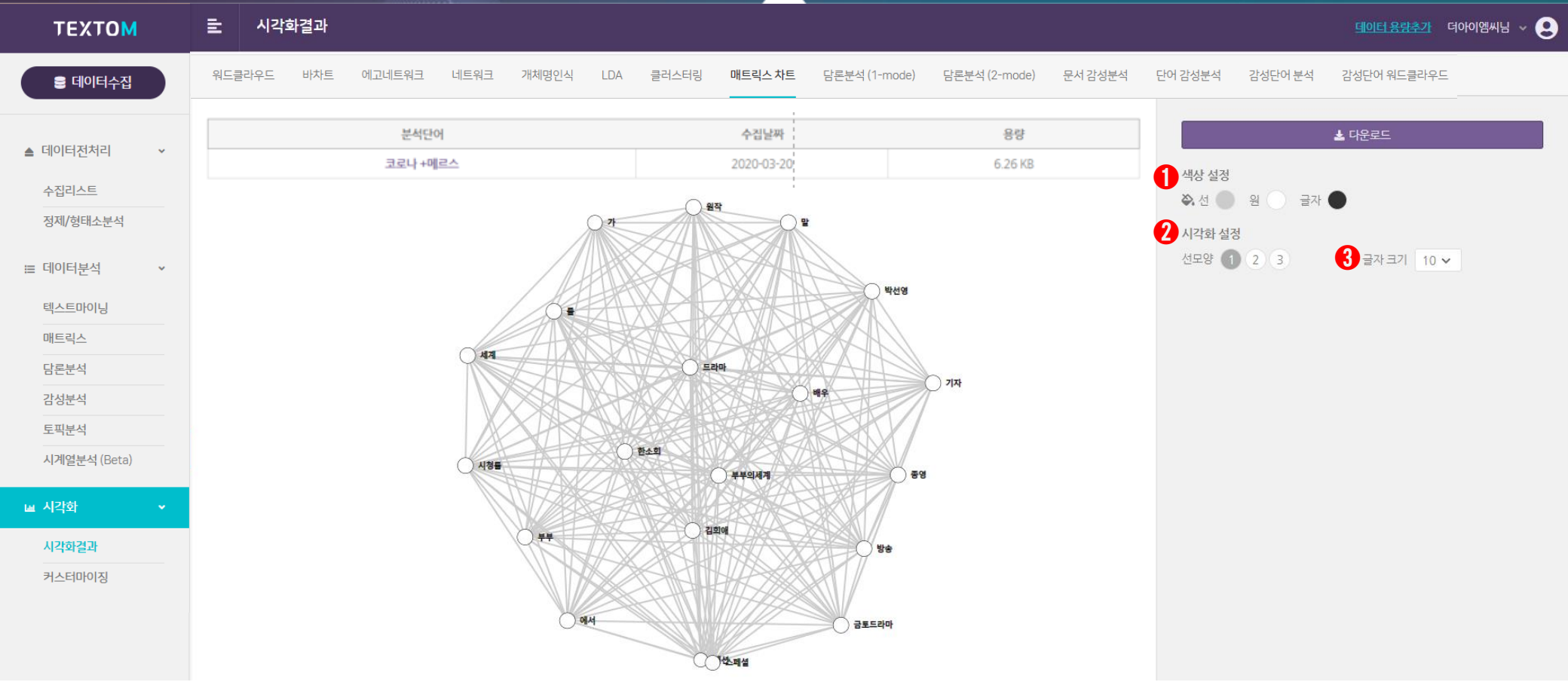
개체명인식(트리맵)은 형태소 분석한 데이터를 미리 정의된 14개의 개체명 범주로 분류하여 트리맵으로 나타낸 시각화입니다

❶ **개체명** 사람, 학문, 대상물, 기관, 지역, 문명, 날짜, 시간 숫자, 사건/사고, 동물, 식물, 금속, 용어 개체명 인식 자세히 알아보기

※ 개체명 인식기에 기본 탑재된 형태소 분석기로 출력된 키워드이므로, 텍스트마이닝 페이지에서 직접 편집한 데이터와는 차이가 있습니다

※ 해당 문서에 14개 개체명에 속하는 키워드가 없다면 개체명이 나타나지 않습니다

시각화 결과 확인하기



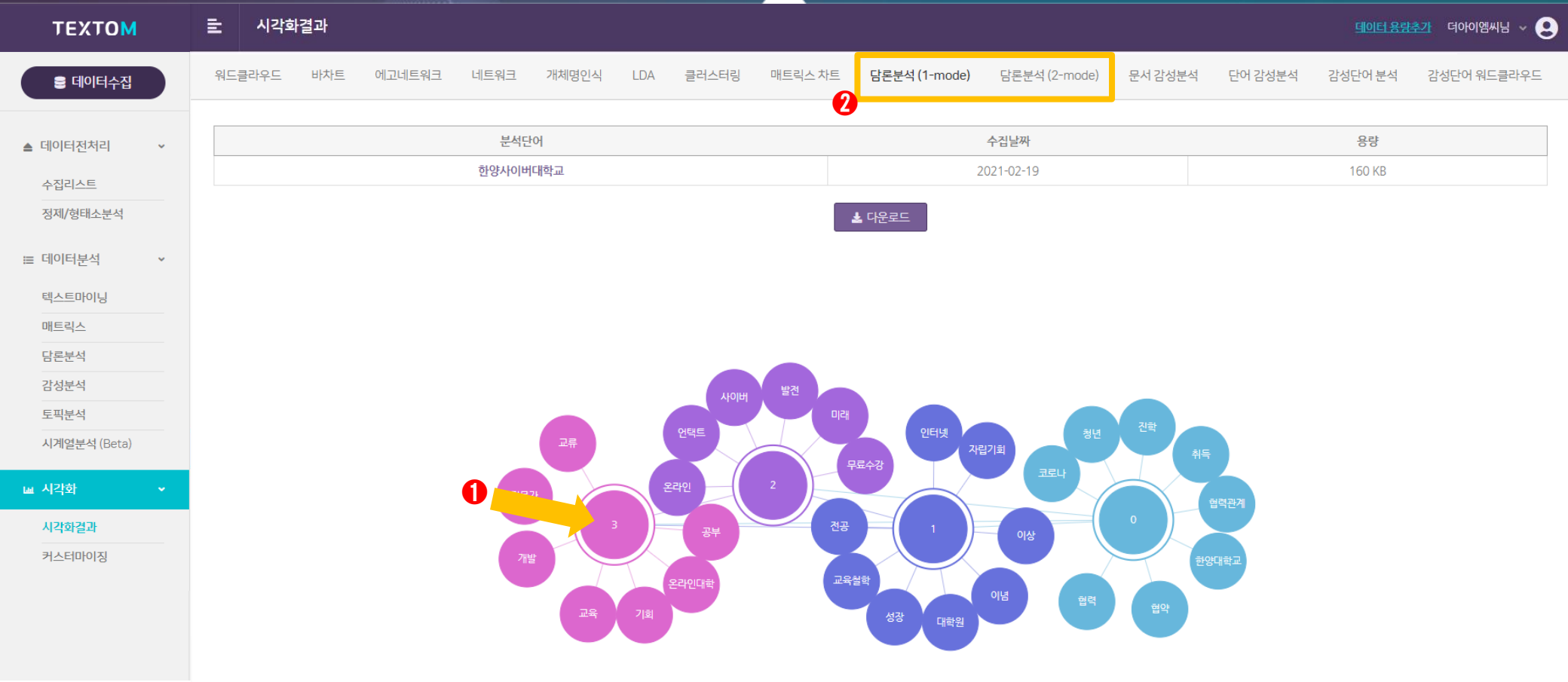
매트릭스 차트는 매트릭스 데이터(공출현 빈도)의 결과값을 사용합니다. (매트릭스 분석을 진행해야 시각화 결과를 확인할 수 있습니다)

❶ **색상 설정** 선 색상, 원의 색상, 글자의 색상을 설정할 수 있습니다

❷ **시각화 설정** 선의 모양을 설정할 수 있습니다

❸ **글자크기 설정** 글자의 크기를 설정할 수 있습니다

시각화 결과 확인하기

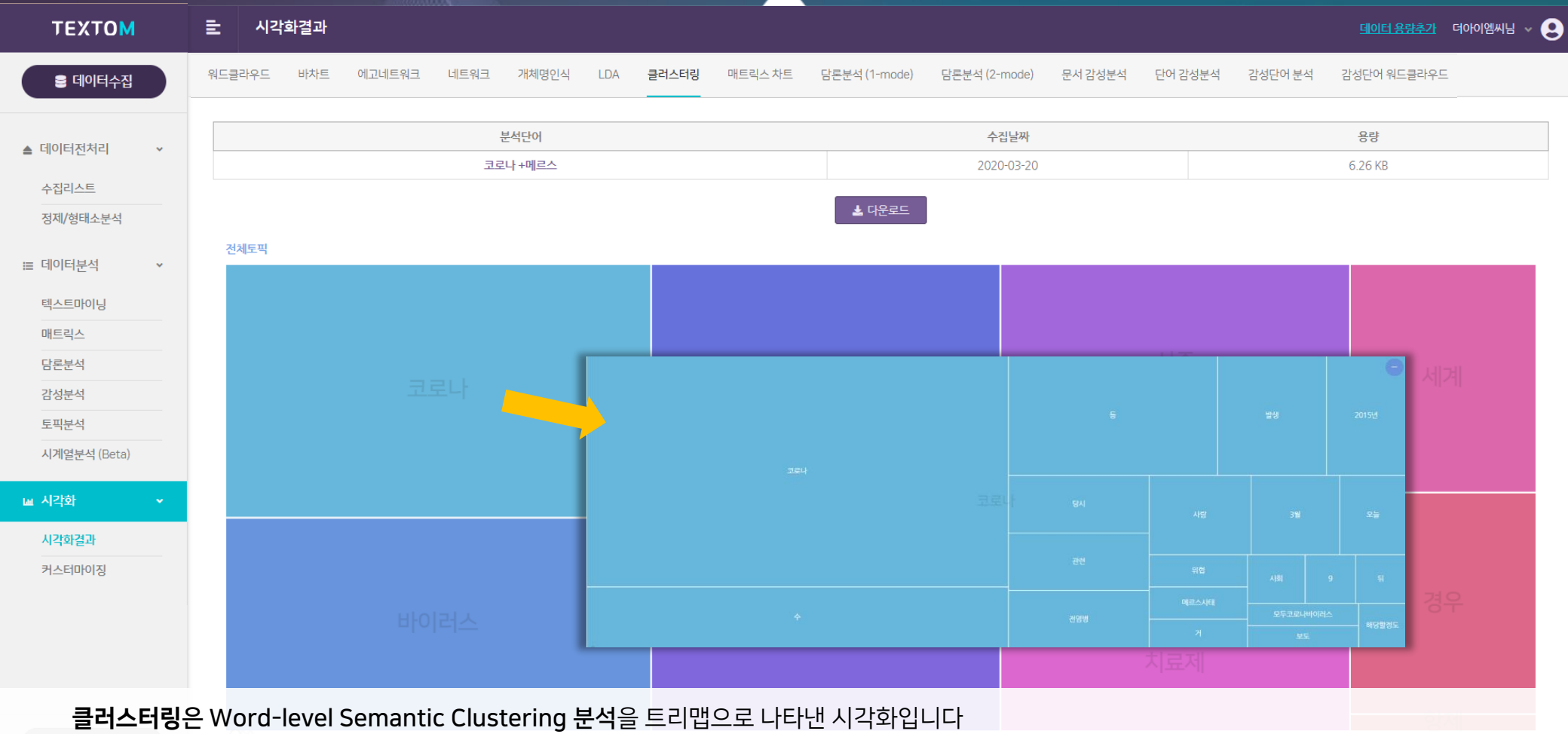


담론분석의 시각화는 매트릭스 데이터(공출현 빈도)의 결과값을 사용합니다. (담론 분석을 진행해야 시각화 결과를 확인할 수 있습니다)

❶ **시각화 설정** 군집을 이동하여 위치를 변경 할 수 있습니다

❷ **1-mode, 2-mode**를 분석하여 각각 시각화 결과를 볼 수 있습니다

시각화 결과 확인하기



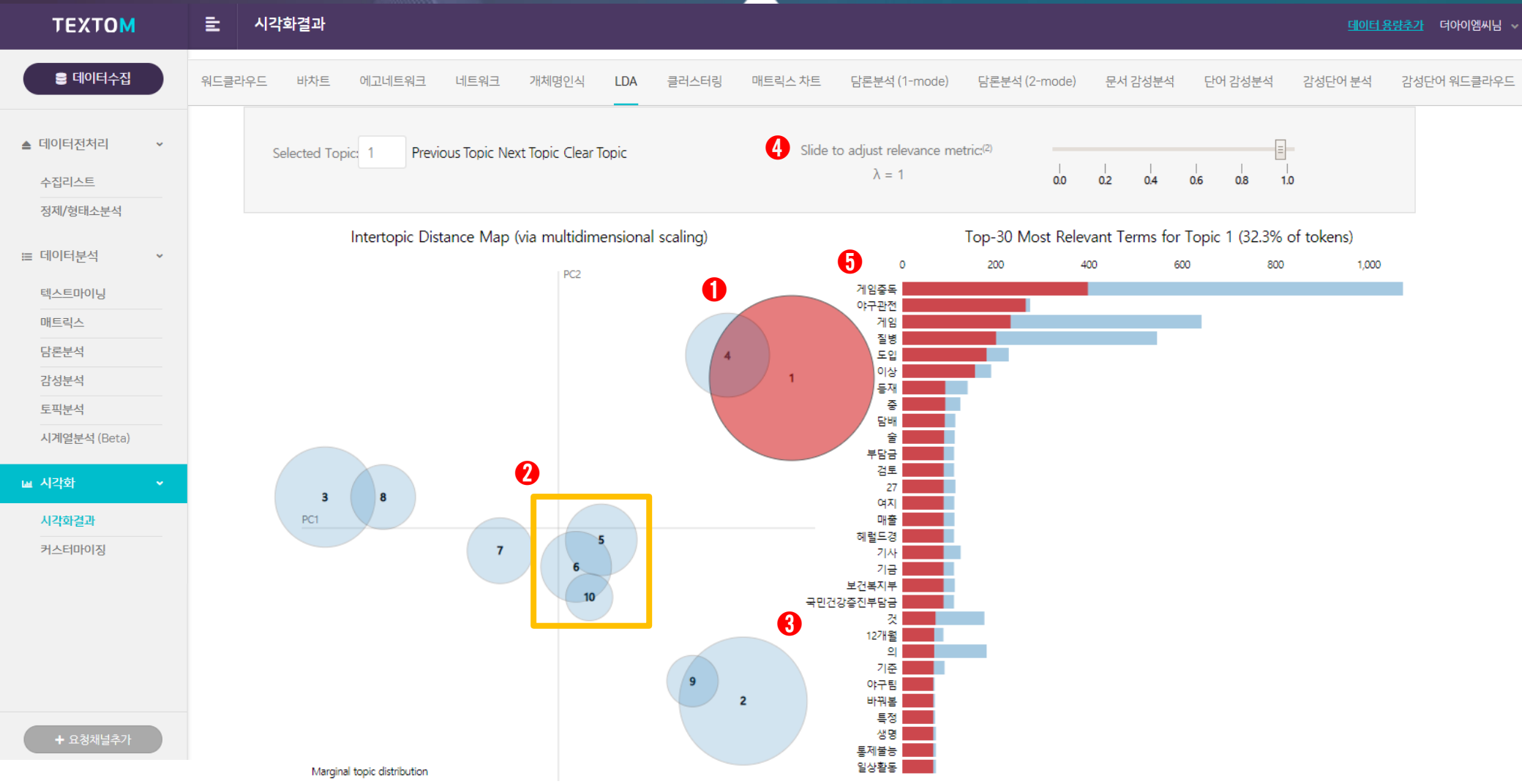
클러스터링은 **Word-level Semantic Clustering 분석**을 트리맵으로 나타낸 시각화입니다
(Word-level Semantic Clustering 분석을 진행해야 시각화 결과를 확인할 수 있습니다)

※ 전체토픽 위에 마우스를 올리면 해당 토픽에 속해있는 키워드들의 빈도 총합을 확인할 수 있습니다

※ 토픽 키워드를 클릭하면 해당 토픽에 속한 키워드를 확인할 수 있습니다

※ 토픽분석 페이지에서 개수를 적용하지 않으면 시각화가 나타나지 않습니다

시각화 결과 확인하기



LDA는 LDA Topic Modeling 분석의 결과값을 시각화로 제공해줍니다.
(LDA Topic Modeling 분석을 진행해야 시각화 결과를 확인할 수 있습니다)

시각화 결과 확인하기

LDA Topic Modeling 시각화 해석

❶ **토픽선택** 토픽분포도에서 토픽을 클릭하거나 토픽번호를 'Selected Topic'에 직접 입력하여 토픽을 선택하면
토픽을 구성하는 30개(사용자 설정 개수) 단어를 확인할 수 있습니다

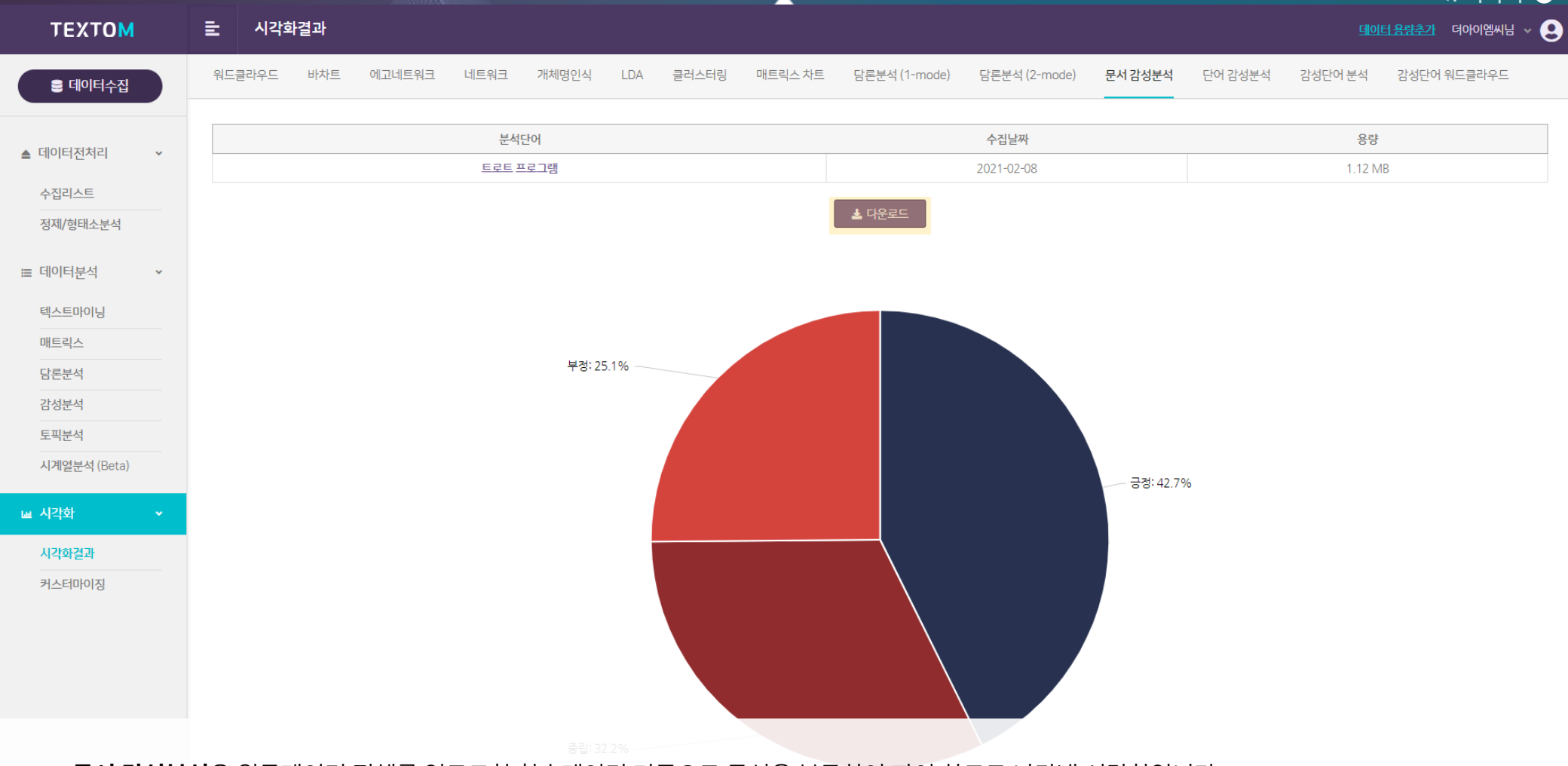
❷ **토픽 간의 거리** 토픽 간의 거리가 멀 수록 판별 타당도가 높고 주제가 뚜렷하게 구분됩니다
토픽 간의 거리가 가깝거나 겹쳐져 있다면 판별 타당도가 낮음으로 비슷한 주제를 나타냅니다

❸ **토픽의 크기** 토픽 원의 크기가 클 수록 높은 빈도수의 단어들로 구성되어 있습니다
가장 큰 원이 메인 토픽이라고 해석할 수 있습니다

❹ **λ(람다) 값 설정** λ(람다) 값을 조절하여 토픽을 구성하는 단어의 출현 조건을 설정할 수 있습니다

❺ **토픽 구성 단어** 토픽을 구성하는 단어들을 확인할 수 있으며, 파란 막대그래프는 전체 단어의 빈도를,
빨간 막대그래프는 해당 토픽에서의 빈도를 보여줍니다

시각화 결과 확인하기

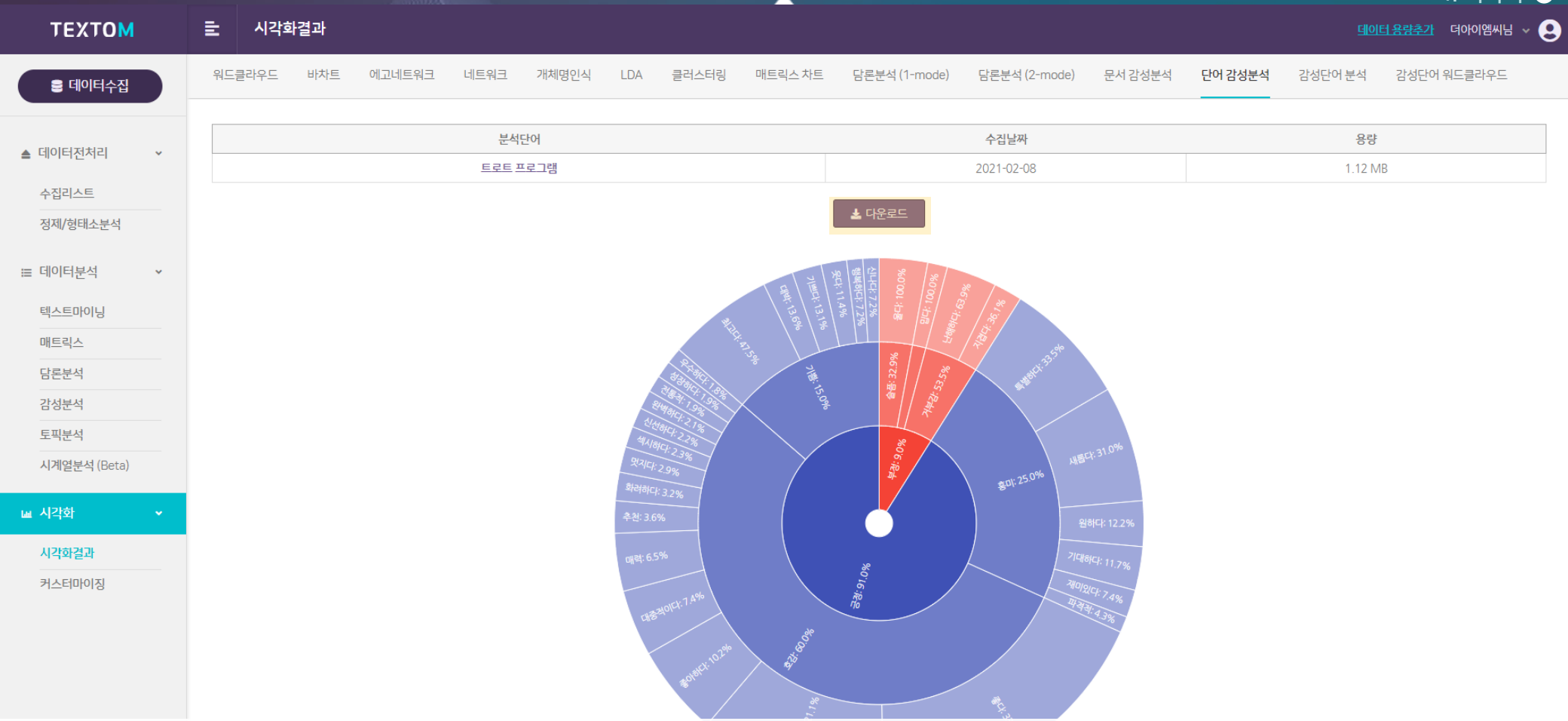


문서 감성분석은 원문데이터 전체를 업로드한 학습데이터 기준으로 극성을 분류하여 파이 차트로 나타낸 시각화입니다

※ 차트 위에 마우스를 올리시면 문서 건수와 백분율 값을 확인할 수 있습니다

※ 학습데이터를 업로드하지 않으면 시각화가 나타나지 않습니다

시각화 결과 확인하기

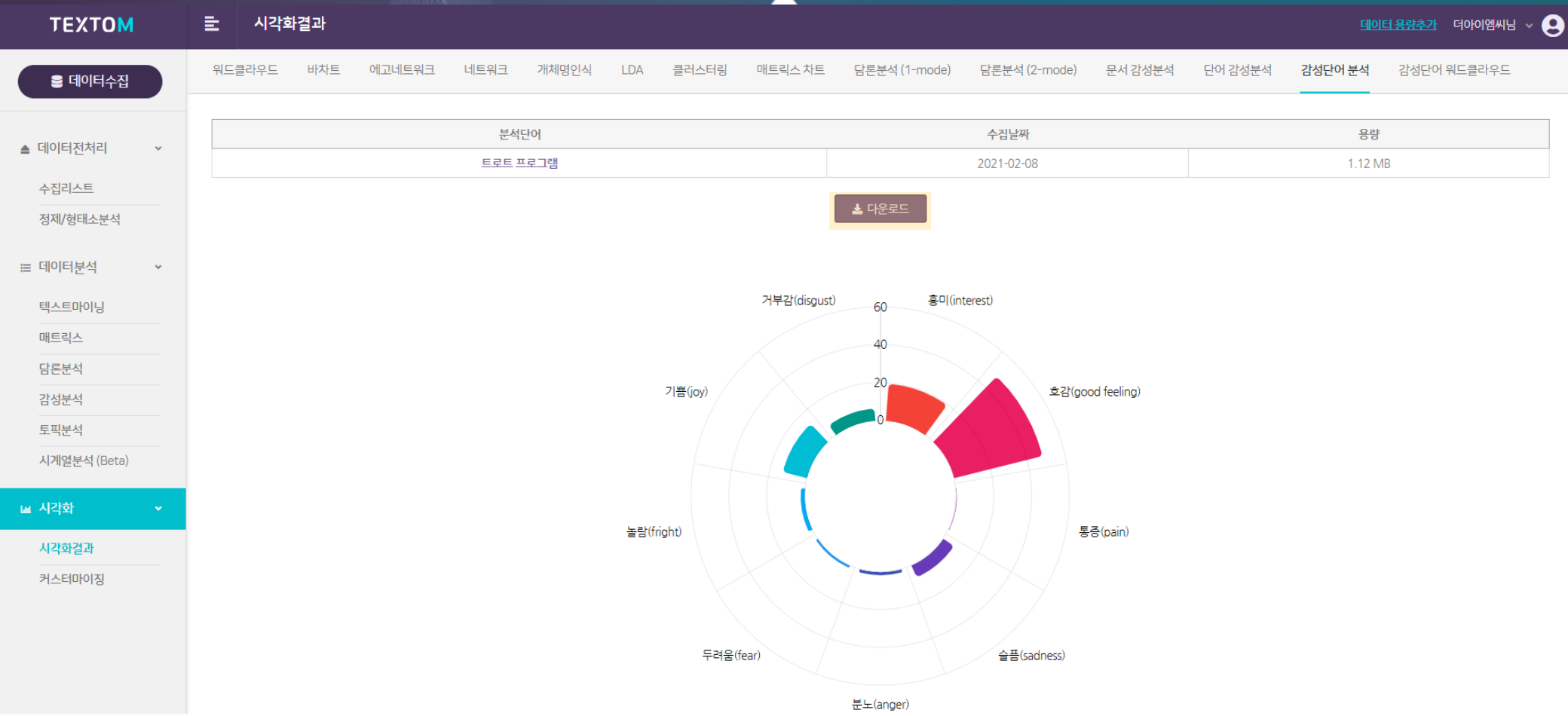


키워드 감성분석은 원문데이터에 있는 감성과 관련된 키워드의 감성강도를 파이 차트로 나타낸 시각화입니다

※ 차트 위에 마우스를 올리시면 감성강도의 백분율 값을 확인할 수 있습니다

※ 학습데이터를 업로드하지 않으면 시각화가 나타나지 않습니다

시각화 결과 확인하기



어휘 감성분석은 원문데이터에 있는 감성과 관련된 키워드의 세부감성비율을 레이다 차트로 나타낸 시각화입니다

※ 차트 계열 위에 마우스를 올리시면 세부감성비율 값을 확인할 수 있습니다

※ 학습데이터를 업로드하지 않으면 시각화가 나타나지 않습니다

시각화 결과 확인하기

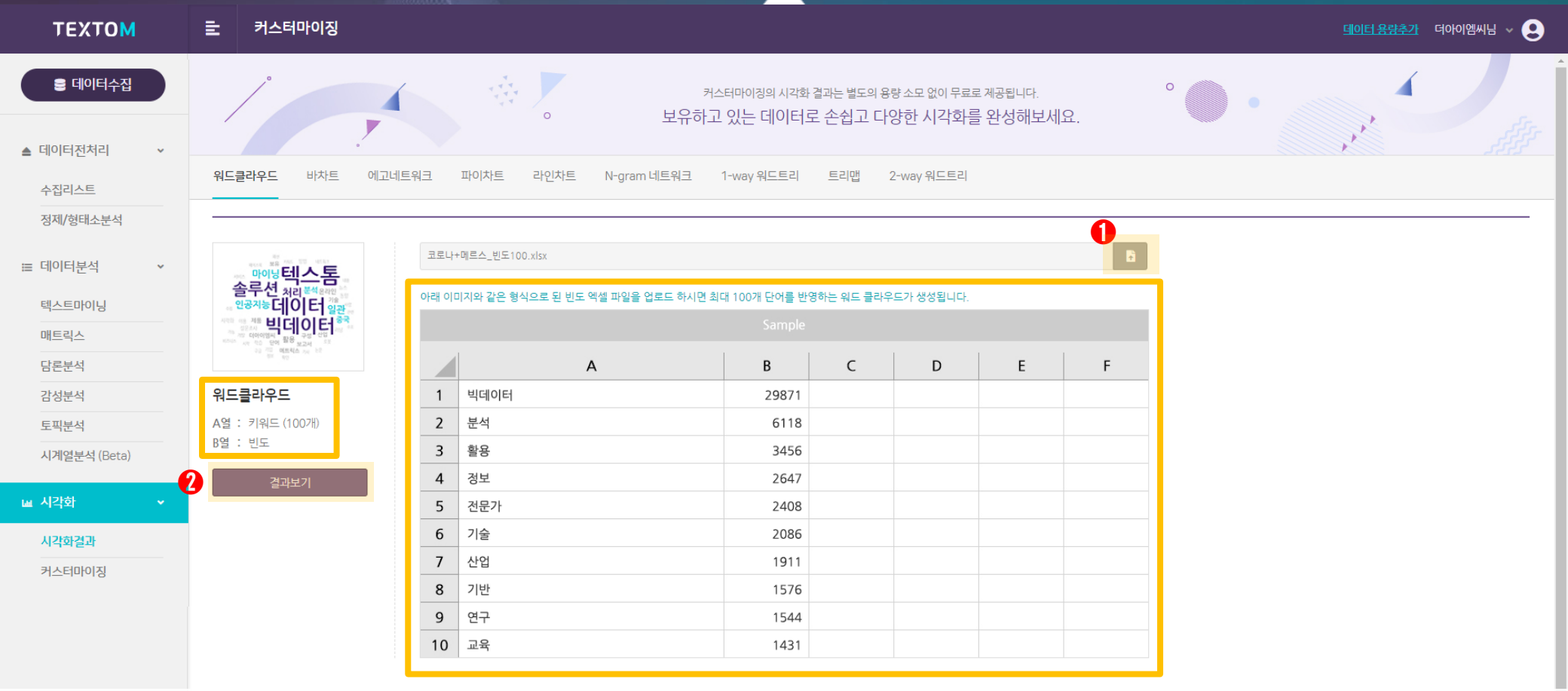
감성단어 워드클라우드는 감성단어 빈도분석의 감성 키워드의 결과값을 사용합니다.

❶ **색상선택** 4가지의 기본 색상 조합 또는 우측의 **상, 중, 하**에서 직접 원하는 색상을 선택할 수 있습니다

❷ **키워드 선정** **빈도** 또는 **키워드**(가나다) 기준으로 정렬할 수 있으며 중요하다고 생각되는 키워드들을 선택해줍니다

❸ **클라우드 모양 변경** 배경이 없는 jpg, png 이미지를 업로드하면 업로드한 모양의 워드클라우드가 만들어집니다

시각화 커스터마이징하기



	A	B	C	D	E	F
1	빅데이터	29871				
2	분석	6118				
3	활용	3456				
4	정보	2647				
5	전문가	2408				
6	기술	2086				
7	산업	1911				
8	기반	1576				
9	연구	1544				
10	교육	1431				

시각화 커스터마이징은 무료이며, 9개의 각 시각화별 양식에 맞는 파일을 업로드만 하면 원하는 시각화 결과물을 바로 확인할 수 있는 기능입니다

❶ **파일업로드** 좌측 네모 박스 안의 정보와 우측 네모 박스 안의 Sample 데이터를 참고하여 각 시각화에 맞게 적합한 데이터를 엑셀 파일로 업로드해줍니다

❷ **결과보기** 업로드한 파일의 시각화 결과물을 확인할 수 있습니다