100 XP

소개

3분

텍스트 분석은 텍스트 내용에 대한 인사이트를 얻기 위해 문서 또는 구의 다양한 측면을 평가하는 프로세스입니다. 대개 사람들은 일부 텍스트를 읽고 이면의 의미를 이해할 수 있습니다. 텍스트를 작성한 언어의 문법 규칙을 고려하지 않아도 텍스트에서 구체적인 인사이트를 파악할 수 있습니다.

예를 들어 일부 텍스트를 읽고 텍스트의 주요 논점을 나타내는 몇 가지 핵심 문구를 파악할 수 있습니다. 또한 인물 이름이나 에펠탑처럼 잘 알려진 랜드마크를 인지할 수도 있습니다. 어려울 때도 있지만, 경우에 따라서는 글쓴이가 어떤 느낌(감정)으로 텍스트를 작성했는지 느낄 수도 있습니다.

텍스트 분석 기법

텍스트 분석은 컴퓨터에서 실행되는 AI(인공 지능) 알고리즘이 텍스트에서 동일한 특성을 평가하여 구체적인 인사이트를 알아내는 프로세스입니다. 일반적으로 사람들은 인사이트를 얻기 위해 자신의 경험과 지식을 활용합니다. 컴퓨터에도 작업을 수행하려면 비슷한 정보가 제공되어야 합니다. 다음을 포함하여 텍스트를 분석하는 소프트웨어를 작성하는 데 사용할 수 있는 자주 사용되는 몇 가지 기술이 있습니다.

- 텍스트에 사용된 용어의 통계 분석입니다. 예를 들어 일반적인 "중지 단어"(텍스트에 대한 의미론적 정보를 거의 나타내지 않는 "the" 또는 "a"와 같은 단어)를 제거하고 나머지 단어의 '빈도 분석'(각 단어가 나오는 횟수 세기)을 수행하면 텍스트의 주요 주제에 대한 단서를 제공할 수 있습니다.
- 빈도 분석을 일반적으로 'N-그램'(두 단어 구는 '바이-그램', 세 단어 구는 '트라이-그램' 등)이라고 부르는 다중 용어구 분석으로 확장합니다.
- '형태소 분석' 또는 '기본형 분석' 알고리즘을 적용하여 단어를 계산하기 전에 단어를 표준화합니다. 예를 들어, "power", "powered" 및 "powerful" 등의 단어는 동일한 단어로 해석됩니다.
- 언어적 구조 규칙을 적용하여 문장을 분석합니다. 예를 들어, '명사', '동사', '형용사' 등을 포함하는 '명사구'와 같은 트리 형태의 구조로 문장을 나눕니다.
- 기계 학습 모델을 학습하는 데 사용할 수 있는 수치로 단어 또는 용어를 인코딩합니다. 예를 들어 포함된 용어를 기준으로 텍스트 문서를 분류합니다. 이 기술은 문서를 긍정적 또는 부정적으로 분류하는 '감정 분석'을 수행하는 데 자주 사용됩니다.
- 단어를 n차원 공간의 위치에 할당하여 단어 간의 의미론적 관계를 포착하는 '벡터화' 모델을 만듭니다. 예를 들어 이 모델링 기술은 "꽃"과 "식물"이라는 단어에는 서로 가까이 위치

하도록 하는 값을 할당하고, "스케이트보드"에는 훨씬 더 멀리 위치하도록 하는 값을 줄 수 있습니다.

이러한 기술은 뛰어난 효과를 발휘할 수 있지만 프로그래밍은 복잡할 수 있습니다. Microsoft Azure에서 **Text Analytics** 인지 서비스는 다음을 수행할 수 있는 미리 학습된 모델을 사용하여 애플리케이션 개발을 간소화하는 데 도움이 될 수 있습니다.

- 문서 또는 텍스트의 언어를 판단합니다(예: 프랑스어 또는 영어).
- 텍스트에 대한 감정 분석을 수행하여 긍정적 또는 부정적 감정을 판단합니다.
- 텍스트에서 주요 논점을 나타낼 수 있는 핵심 구를 추출합니다.
- 텍스트에서 엔터티를 식별하고 분류합니다. 엔터티는 사람, 장소, 조직 또는 일상적인 항목(예: 날짜, 시간, 수량 등)일 수 있습니다.

이 모듈에서는 이러한 기능 중 일부를 살펴보고 다음과 같은 애플리케이션에 해당 기능을 적용하는 방법을 이해할 수 있습니다.

- 정치적 캠페인이나 시장에서 제품을 둘러싼 감정을 검색하는 소셜 미디어 피드 분석기
- 카탈로그에 있는 문서의 주요 주제 요약을 지원하기 위해 핵심 문구를 추출하는 문서 검색 애플리케이션
- 식별을 위해 문서 또는 다른 텍스트에서 브랜드 정보나 회사 이름을 추출하는 도구

위의 예시는 Text Analytics가 도움이 될 수 있는 여러 영역 중 일부입니다.

다음 단원: Azure에서 Text Analytics 시작하기

계속 >