DEEP LEARNING for Search

Tommaso Teofili

Foreword by Chris Mattmann

박 경 규

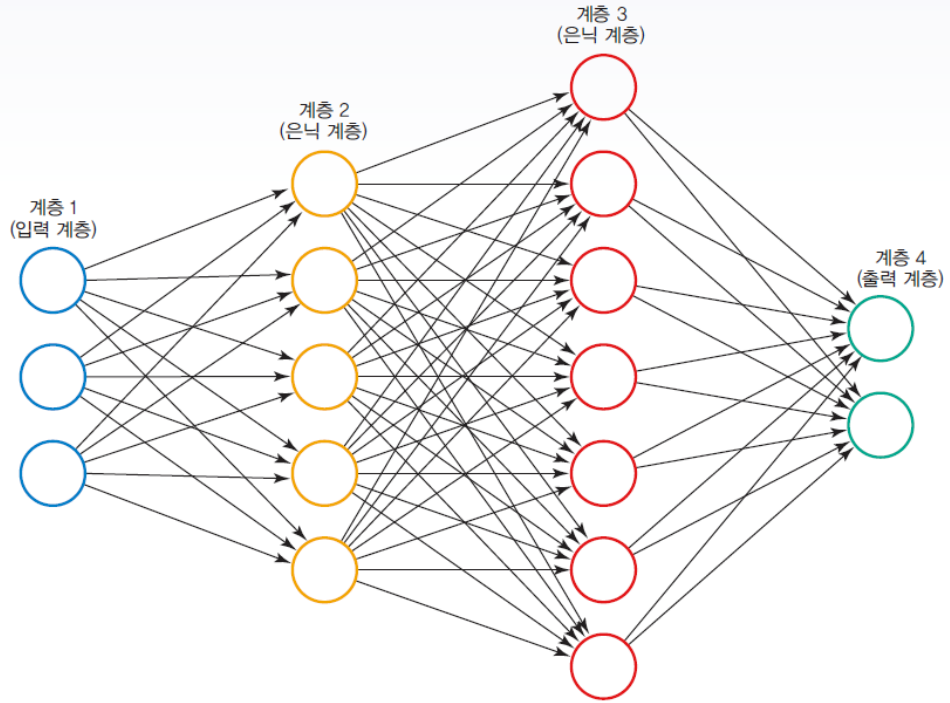
Contents

1. Neural search
(신경망을 이용한 검색)

- 검색에 관한 기초 내용
- 검색시에 당면할 수 있는 중요 문제
- 신경망이 검색엔진에 더 효과적일 수 있는 이유

심층신경망(Deep Neural Networks)

- Providing a representation of textual data that captures word and document semantics, allowing a machine to say which words and documents are semantically similar.
- Generating text that's meaningful in a certain context: for example, useful for creating chatbots.
- Providing representations of images that pertain not to the pixels but rather to their composing objects. This allows us to build efficient face/object-recognition systems.
- Performing machine translation efficiently.
- Under certain assumptions, approximating any function.
- There's theoretically no limit to the kinds of tasks deep neural networks can achieve.



검색엔진에 딥러닝 적용 장점

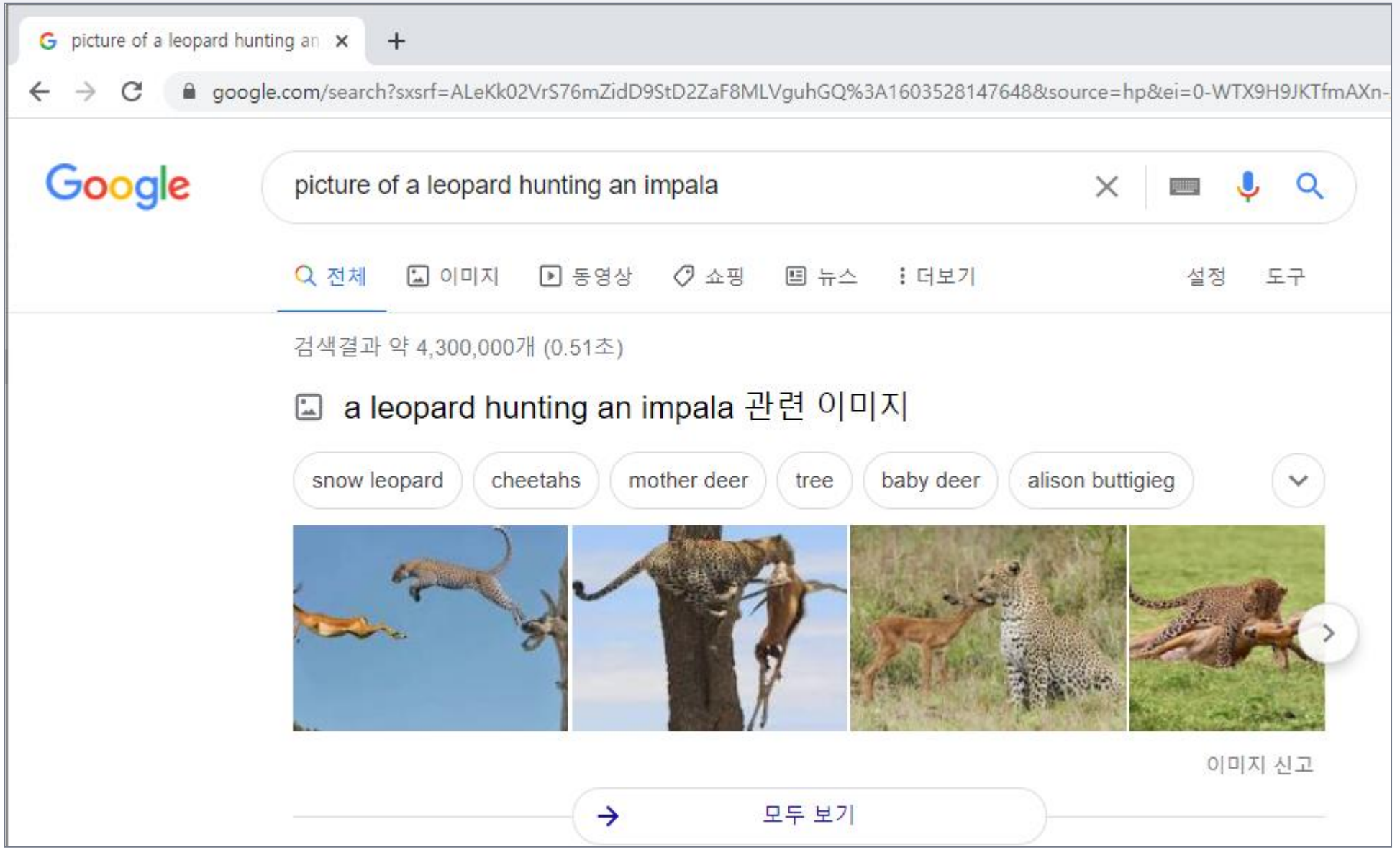
- Provide more-relevant results to its end users, increasing user satisfaction.

- Search through binary content like images the same way we search text.
  Think of this as being able to search for an image with the phrase
  “picture of a leopard hunting an impala” (and you’re not Google).

- Serve content to users speaking different languages, allowing more users
  to access the data in the search system.

- Generally become more sensitive to the data it serves,
  which means less chance for queries that give no results.

검색엔진에 딥러닝 적용 장점



검색엔진(Search engine)

- 색인화(Indexing)

  데이터를 효율적으로 수립하고 저장해둠으로써 빠르게 검색해 볼 수 있도록 한다.

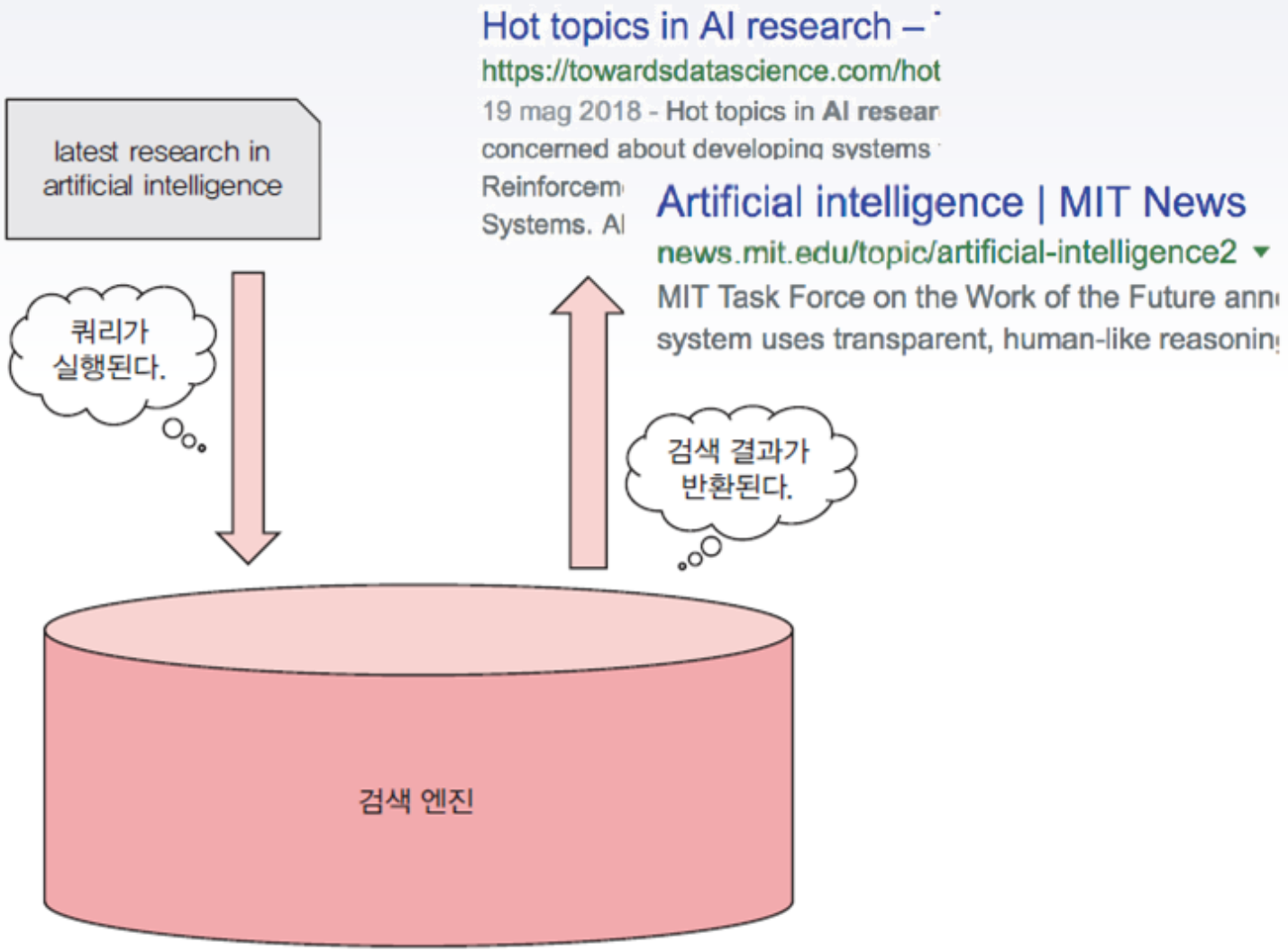
- 쿼리처리(Querying)

  사용자가 정보를 찾아볼 수 있게 검색 기능을 제공한다.

- 순위지정(Ranking)

  사용자의 정보 요구를 가장 잘 충족하기 위해 특정 지표에 맞춰 순위를 지정해 결과를 표시한다.

검색엔진(Search engine)



텍스트 분석 파이프라인(Text analysis pipeline)

- Tokenizers

  Components that break a stream of text into words, phrases, symbols, or other units called tokens

  The tokenizer will split tokens every time it encounters a whitespace character.

- Token filters

  Components that accept a stream of tokens (from a tokenizer or another filter) and can modify, delete, or add new tokens

  The token filter will remove tokens that match a certain blacklist (also known as a stopword list).

"I like search engines"

"the brown fox jumped over the lazy dog"

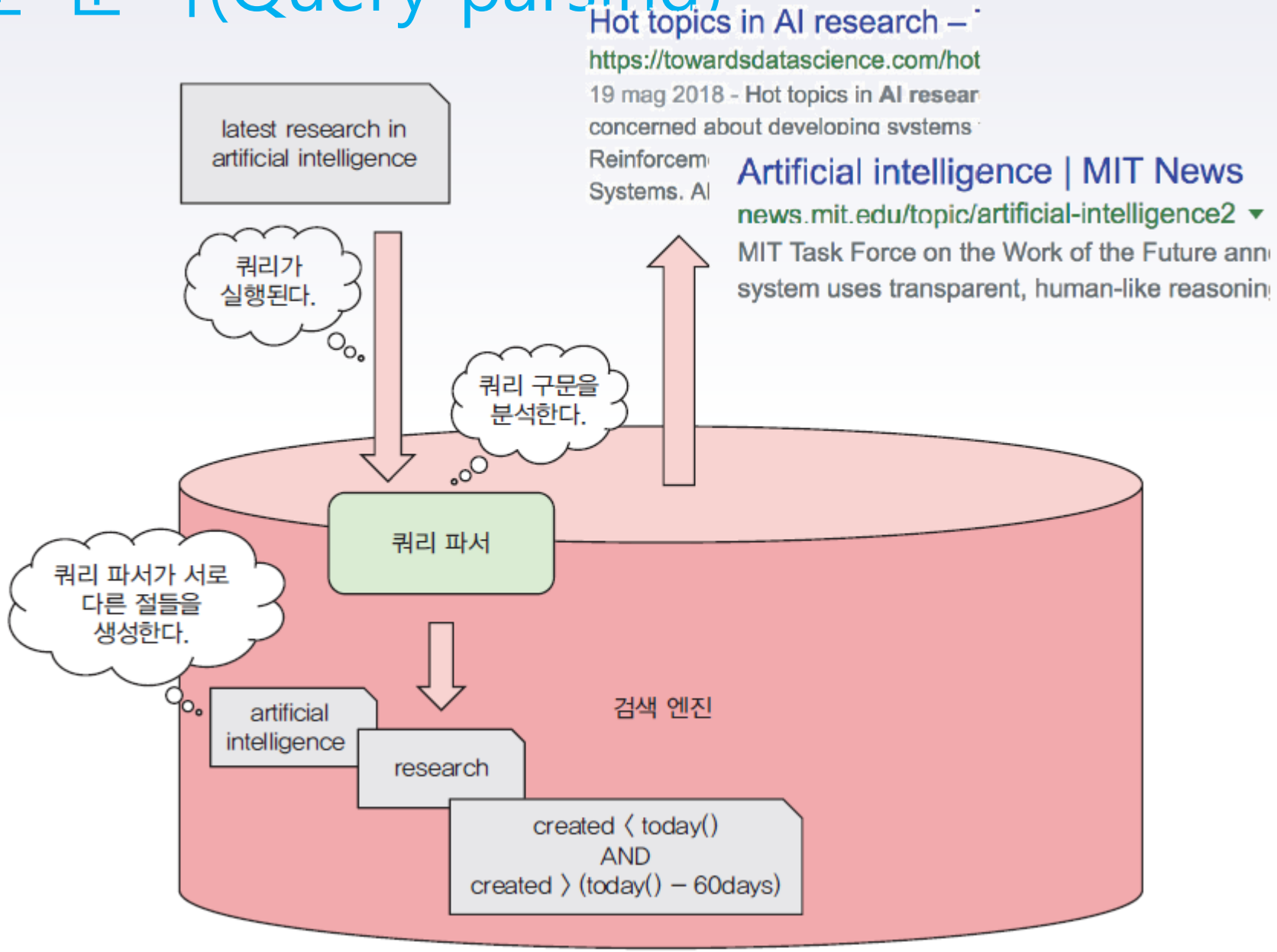
색인화(INDEXING)

- “the brown fox jumped over the lazy dog” (document 1)
- “a quick brown fox jumps over the lazy dog” (document 2)

- Inverted index table

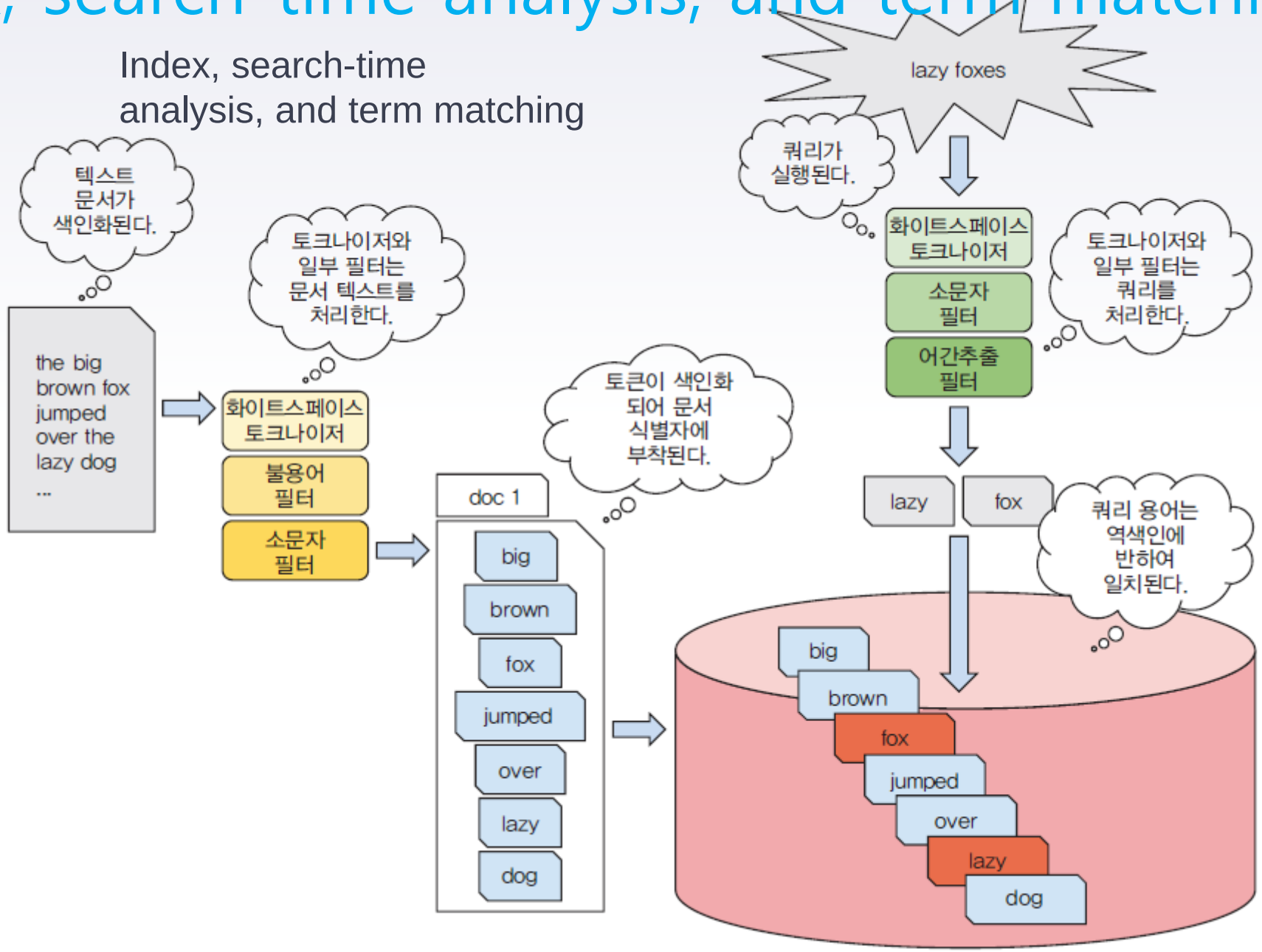
Term	Document IDs
brown	1, 2
fox	1, 2
jumped	1
over	1, 2
lazy	1, 2
dog	1, 2
quick	2
jumps	2

쿼리 구문 분석(Query parsing)



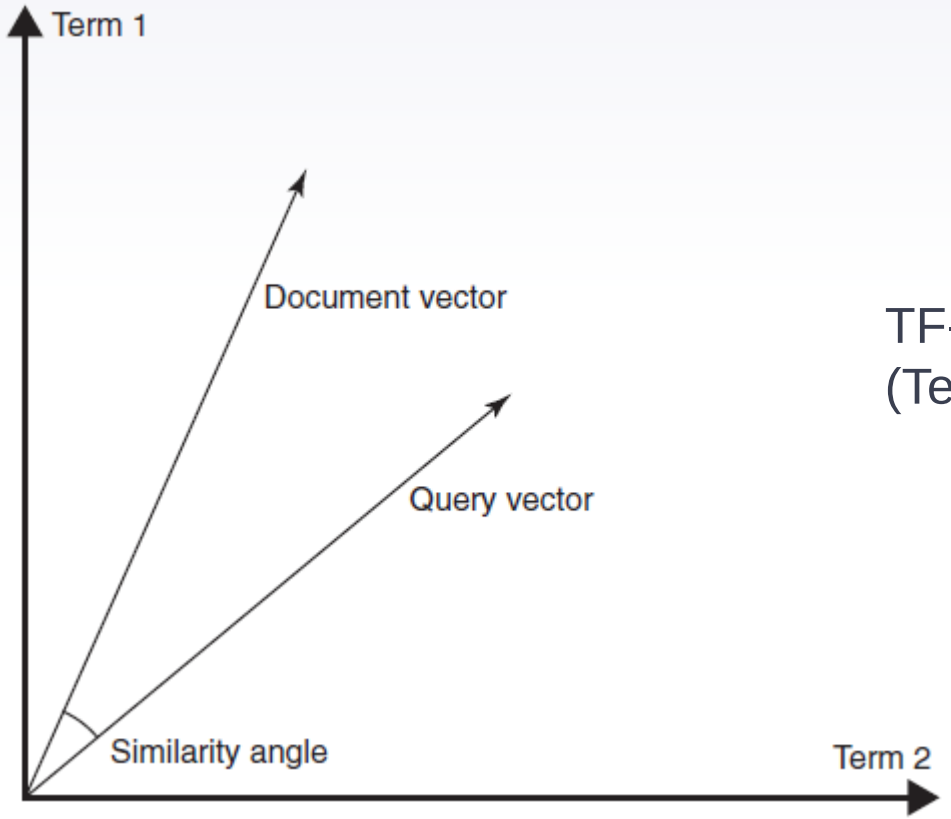
Index, search-time analysis, and term matching

Index, search-time
analysis, and term matching



고전적인 검색모델(Classic retrieval models)

- VSM(Vector Space Model)에 따른 문서 및 쿼리 벡터간의 유사도



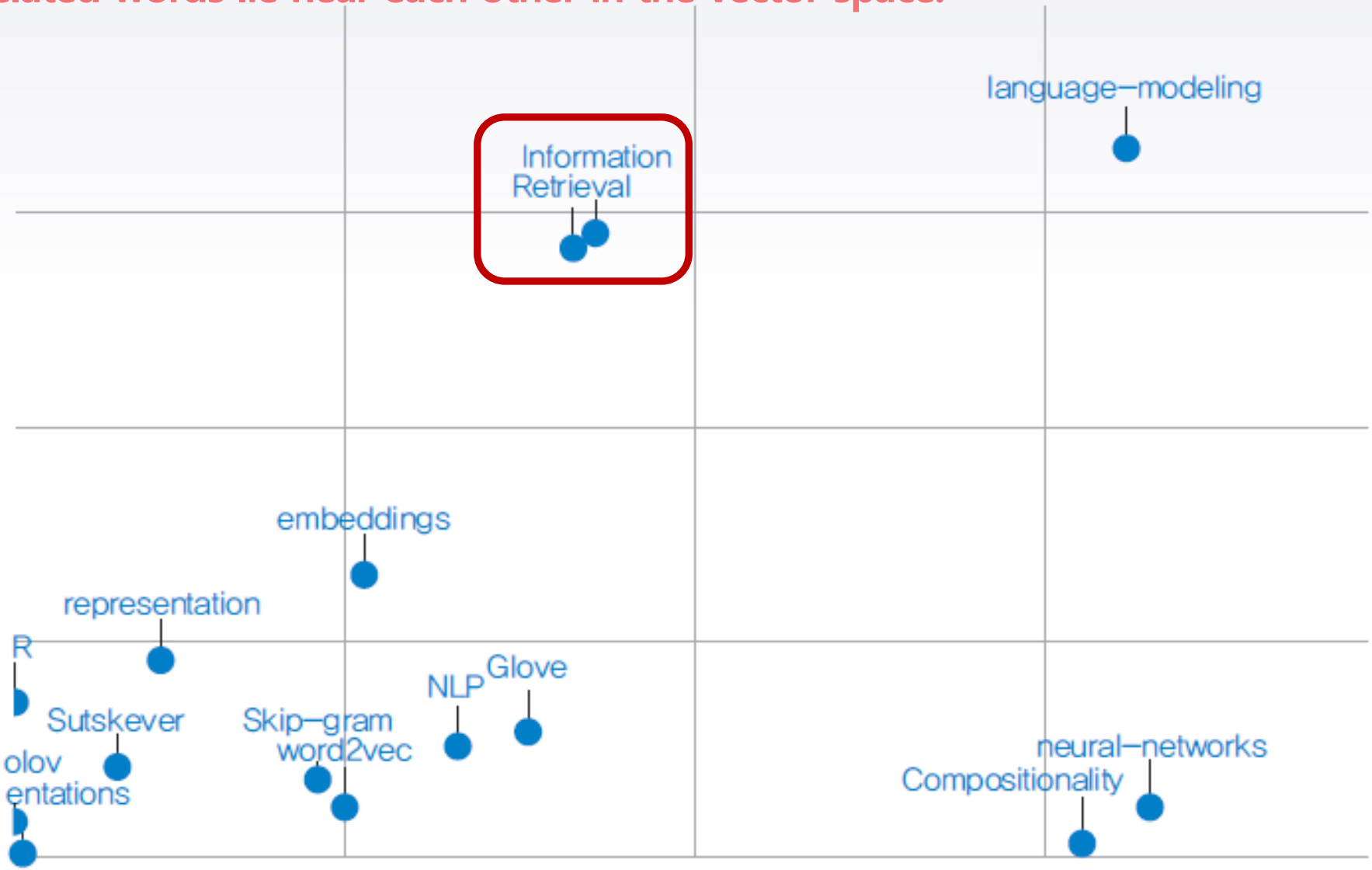
TF-IDF
(Term Frequency–Inverse Document Frequency)

비슷한 쿼리 비교하기(Comparing similar queries)

Query	First result title
Latest breakthroughs in artificial intelligence	Academic papers for “latest breakthroughs in artificial intelligence” (Google Scholar)
Latest advancements in artificial intelligence	Google advancements artificial intelligence push with 2 top hires
Latest advancements on artificial intelligence	Images related to “latest advancements on artificial intelligence” (Google Images)
Latest breakthroughs in AI	Artificial Intelligence News—ScienceDaily
Più recenti sviluppi di ricerca sull’intelligenza artificiale	Intelligenza Artificiale (Wikipedia)

word2vec

When using a neural network algorithm to learn word representations within a set of text documents, closely related words lie near each other in the vector space.



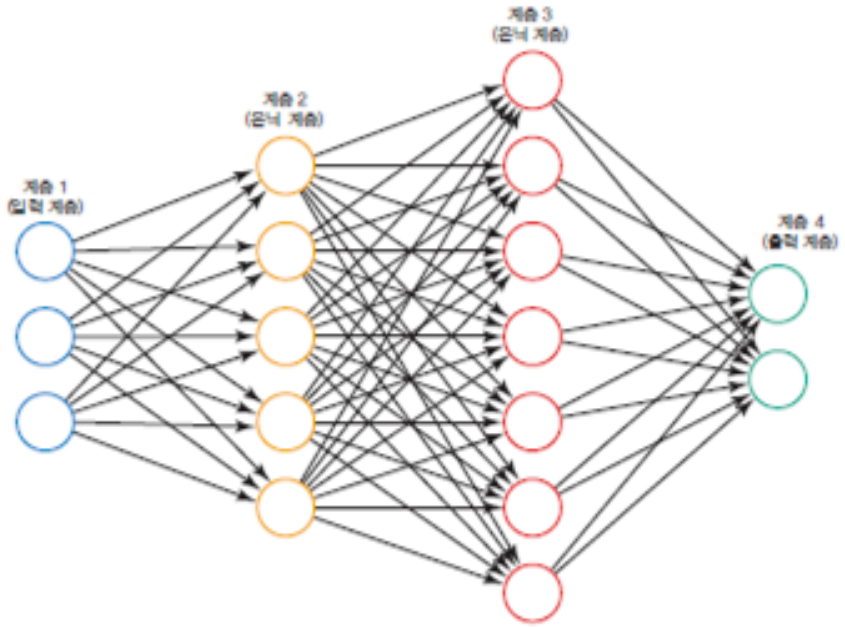
Neural search application

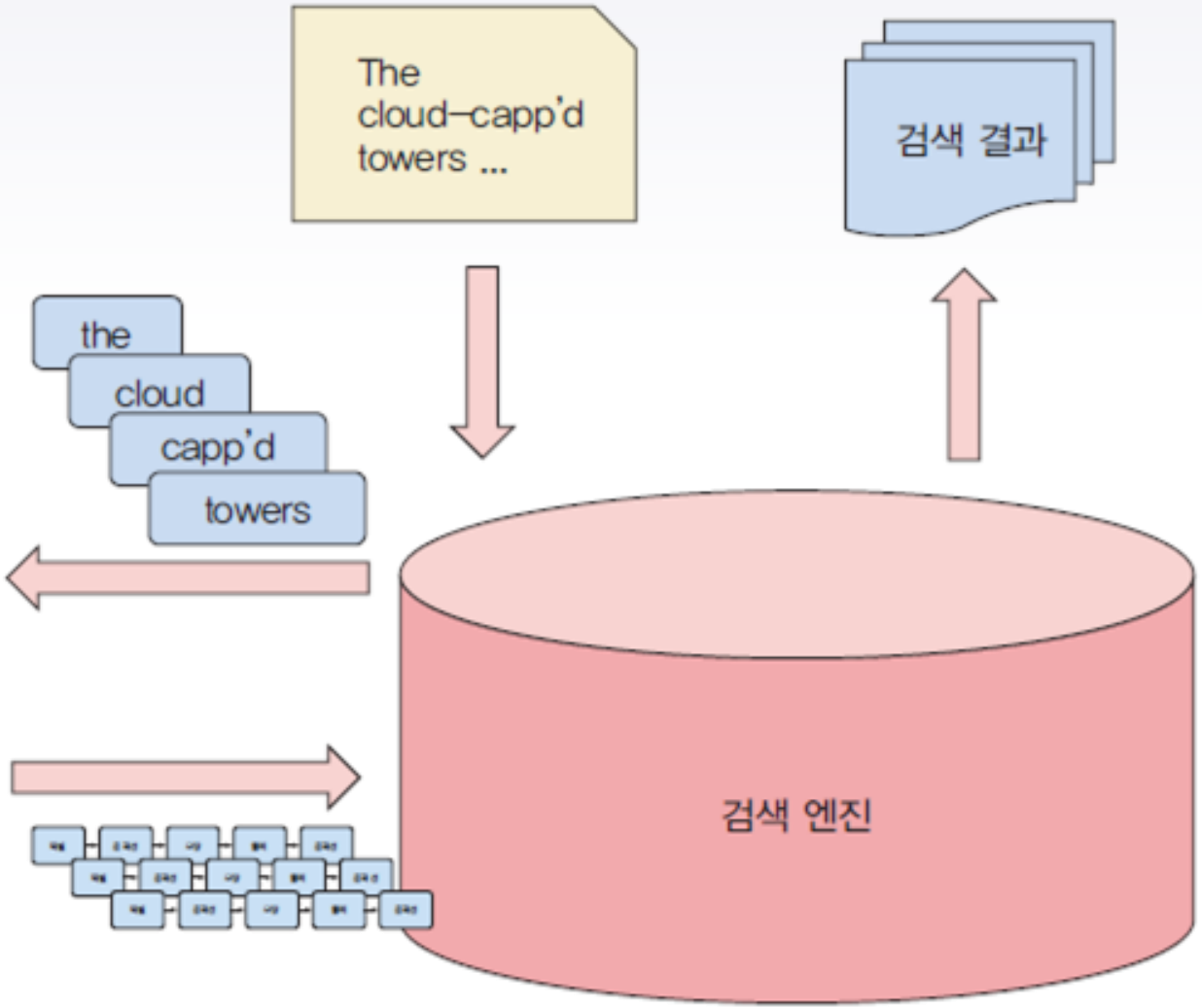
One of the key ideas of neural search is to use such representations to improve the effectiveness of search engines.

Word embeddings

Document embeddings

Nearest neighbors





Summary

- Search is a hard problem.
- Text analysis is an important task in search(Indexing and Search phases)
- Relevance(연관도) is the fundamental measure of search engine
- Deep neural networks to learn representations of content (words, sentences, paragraphs, images) that can capture semantically relevant similarity measures.
- Neural search stands as a bridge between search and deep neural networks, with the goal of using deep learning to help improve different tasks related to search.

2. Generating synonyms
(동의어 생성)

- 검색시 동의어가 사용되는 이유 및 방법
- 아파씨 루씬 소개
- Word2vec을 사용해 동의어 생성

THANKS!

Any questions?

You can find me at:

- kgpark88@gmail.com
- https://github.com/kgpark88