BERT Keyword Extraction of TED Talk Videos

Melissa Viator

CS688 Term Project

GOAL OF THE PROJECT

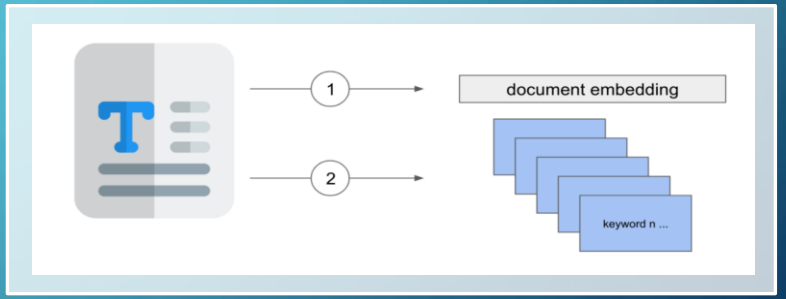
- The goal of this project is to extract meaningful keywords from TED Talk videos
- To do so, the project utilizes the package, **KeyBERT**, because it had the ability to consider the semantic aspects of the full document
- NLP keyword extraction is a useful real-word application because it automatically provides an overview of the content of a text

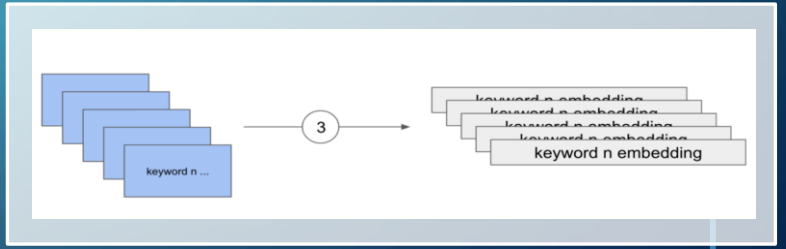
BERT KEYWORD EXTRACTION

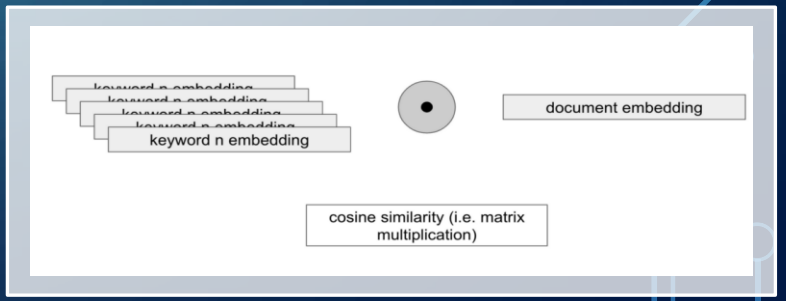
KeyBERT is a package by Maarten Grootendorst that leverages the BERT language model and relies on the transformers library

STEPS:

1. The input document is embedded using a pre-trained BERT model (i.e. turns the text into a fixed-sized vector representing the semantics of the document)
2. Keywords and expressions (n-grams) are extracted from the same document using techniques such as count vectorizer or TF-IDF vectorizer
3. Each keyword is embedded into a fixed-size vector with the same model used to embed the document, providing a list of keyword embeddings
4. Computes a cosine similarity between the keyword embeddings and the document embedding. Extracts the most similar keywords (with the highest cosine similarity score)







BERT KEYWORD DIVERSITY

- KeyBERT includes two methods to introduce diversity in the resulting keywords: Max Sum Similarity (MSS) and Maximal Marginal Relevance (MMR)
- For the project, I utilized MMR as it tries to minimize redundancy and maximize the diversity of results in text summarization tasks
- MMR starts by selecting the keywords that are the most similar to the document. Then, it iteratively selects new candidates that are both similar to the document and not similar to the already selected keywords

```
def keywords(txt_file):
    kw_model = KeyBERT()
    keywords = kw_model.extract_keywords(txt_file, keyphrase_ngram_range=(1,2),
                                                    stop_words="english",
                                                    use_mmr = True, diversity= 0.6)
    return keywords
```

COSINE SIMILARITY

- To analyze the results of the NLP, we will use **cosine similarity** to understand the similarity between:
  1. BERT keywords and the TED Talk video
  2. Keywords groups
- Precautions:
  - Averaging of the token vectors
  - Insensitive to the order of the words

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

PROCESS FOR KEYWORD EXTRACTION OF TED TALK VIDEOS

1. CONVERT
   Video → Text
2. NLP
   BERT Keywords from Transcript
3. NLP
   BERT Keywords from Summary
4. ANALYZE
   BERT Keywords vs Human Keywords

TED Talk 1

THE INGREDIENT IN ALMOST EVERYTHING YOU EAT

TED Talk 1
THE INGREDIENT IN ALMOST EVERYTHING YOU EAT

HUMAN KEYWORDS:

TED Talk 1

THE INGREDIENT IN ALMOST EVERYTHING YOU EAT

TRANSCRIPT BERT KEYWORDS	
soy allergy	36%
heart disease	8%
food overwhelming	22%
unhealthy eating	4%
asian cuisine	4%

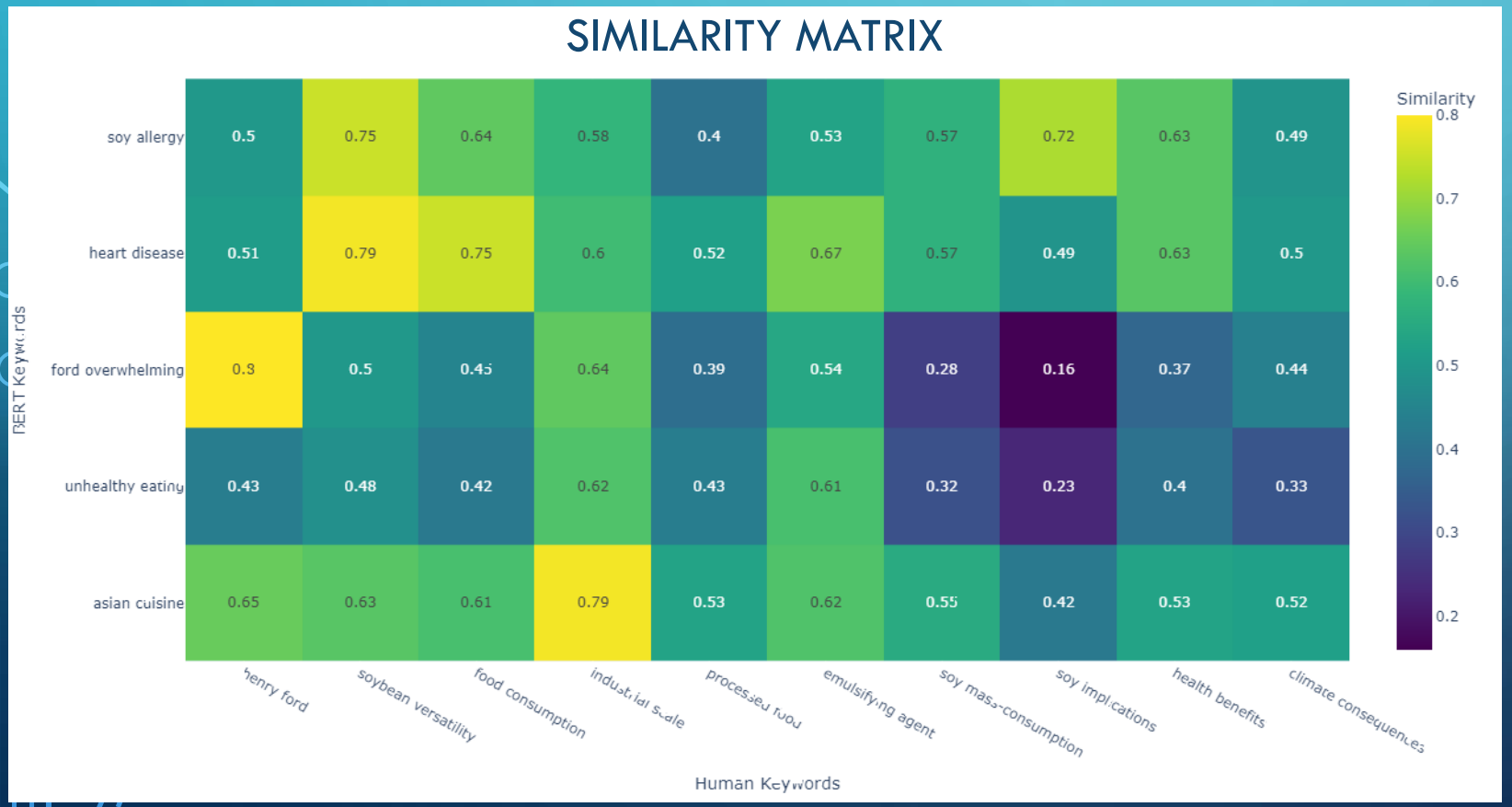
70%

Similarity Between Transcript & Summary BERT Keywords

SUMMARY BERT KEYWORDS	
makes soybeans	42%
global obsession	17%
biodegradable plastic	7%
cultivated asia	14%
soybeans versatile	41%

TED Talk 1
THE INGREDIENT IN ALMOST EVERYTHING YOU EAT

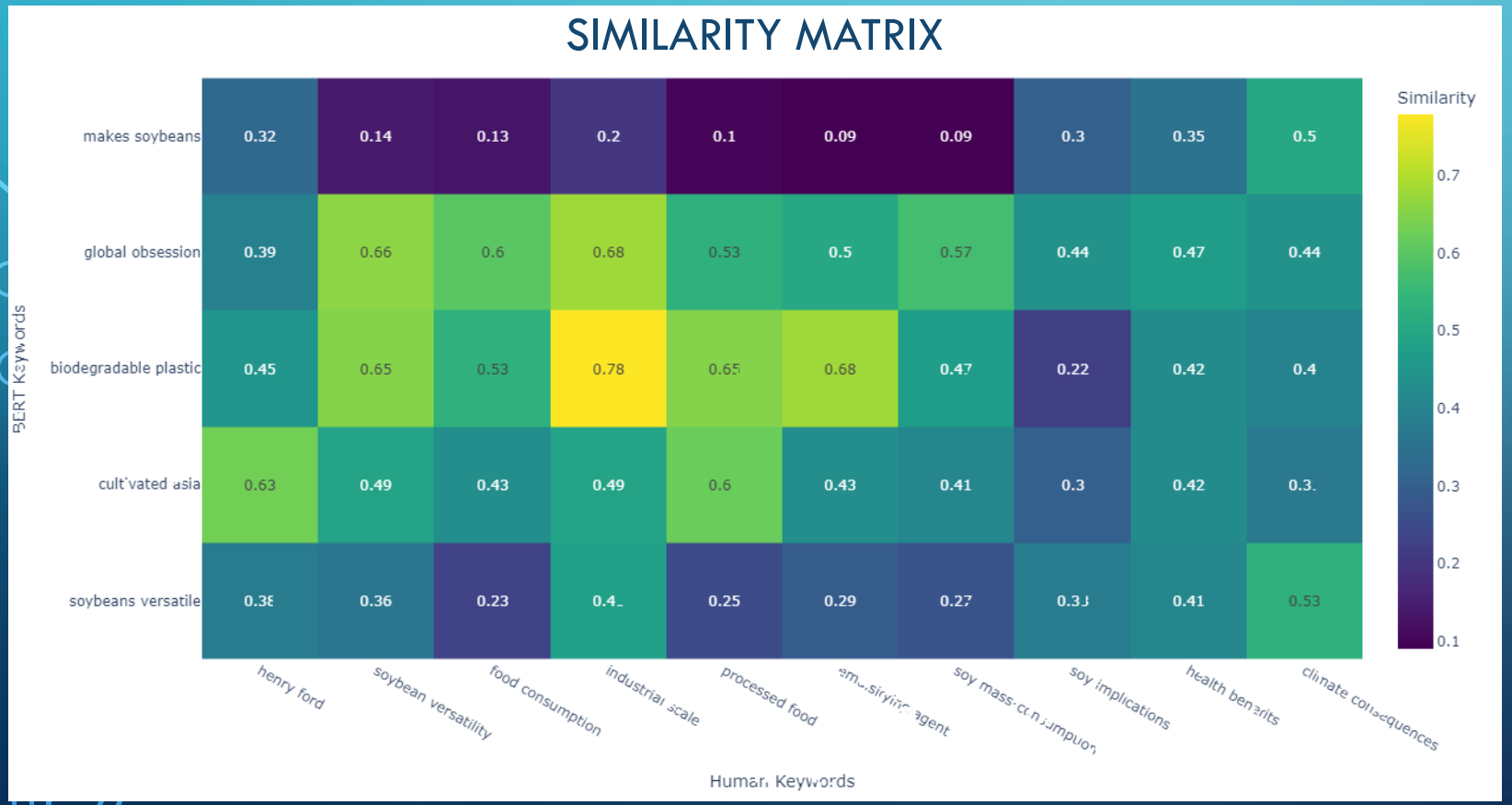
HUMAN KEYWORDS VS TRANSCRIPT BERT KEYWORDS



Overall Similarity → 82%

TED Talk 1
THE INGREDIENT IN ALMOST EVERYTHING YOU EAT

HUMAN KEYWORDS VS SUMMARY BERT KEYWORDS



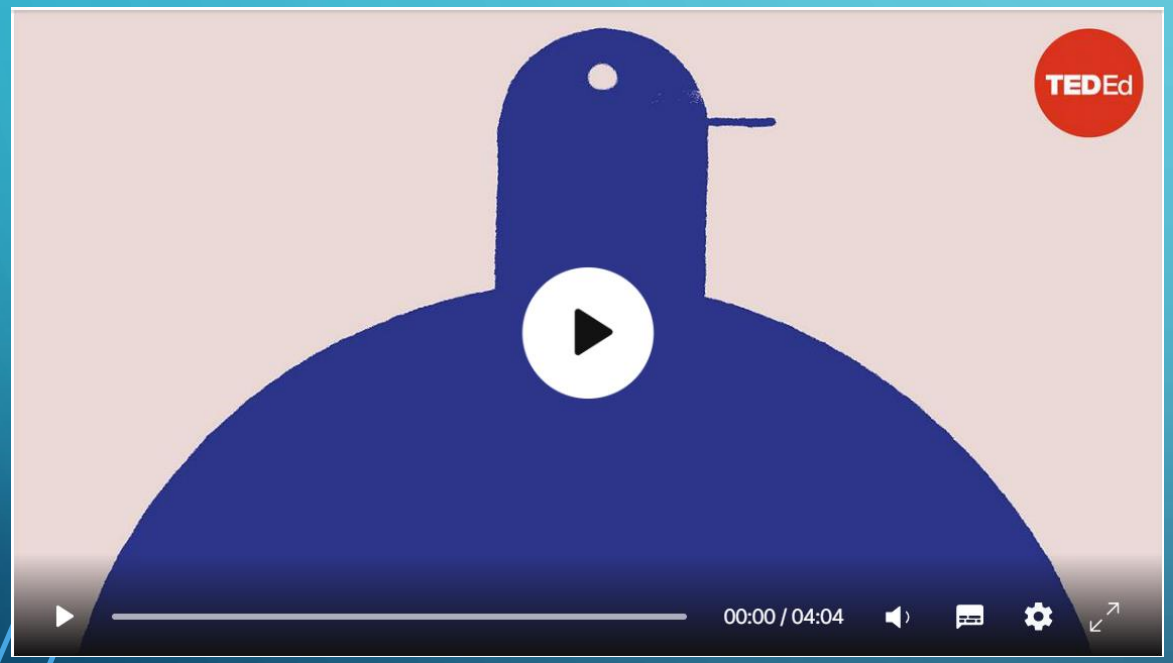
Overall Similarity

68%

TED Talk 2

HOW PIGEONS TOOK OVER THE WORLD

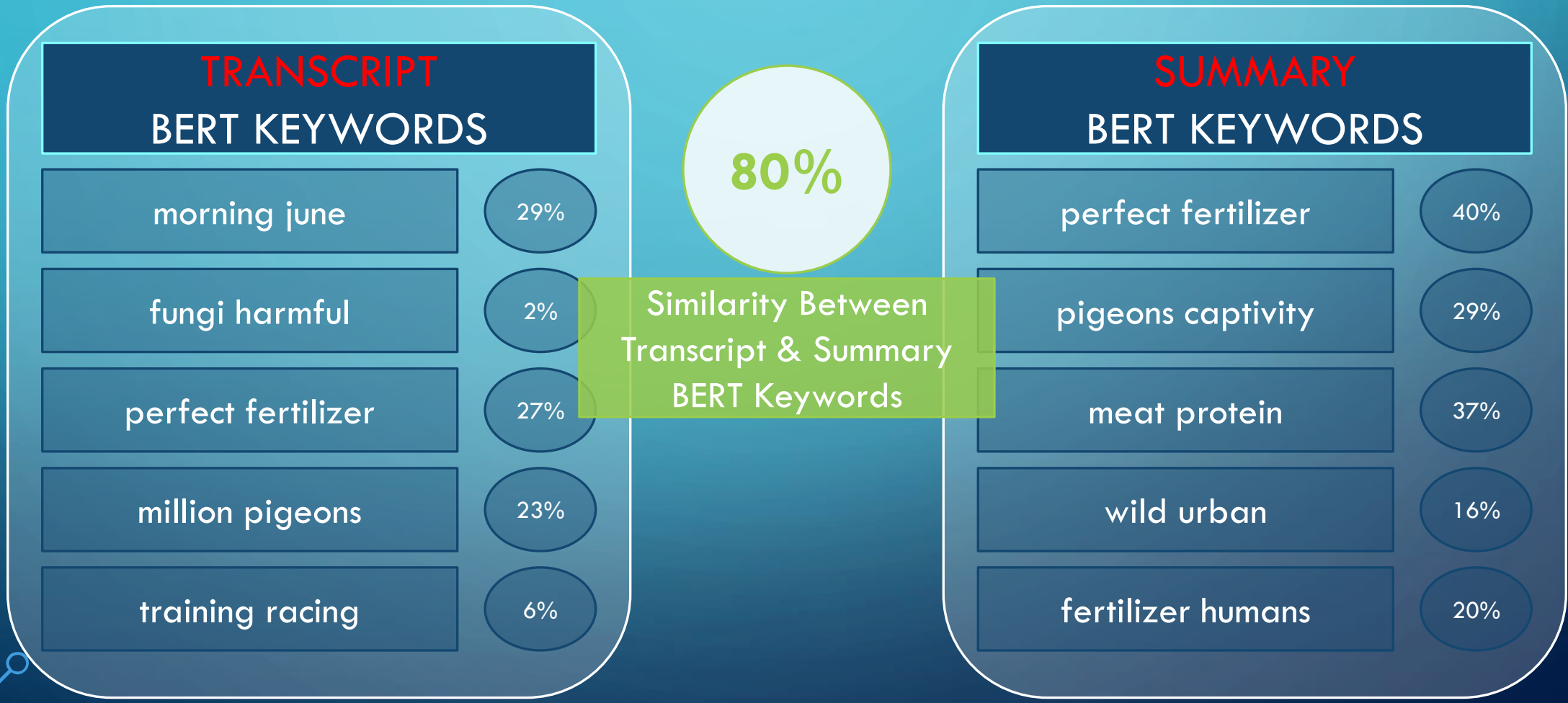
TED Talk 2
HOW PIGEONS TOOK OVER THE WORLD



HUMAN KEYWORDS:

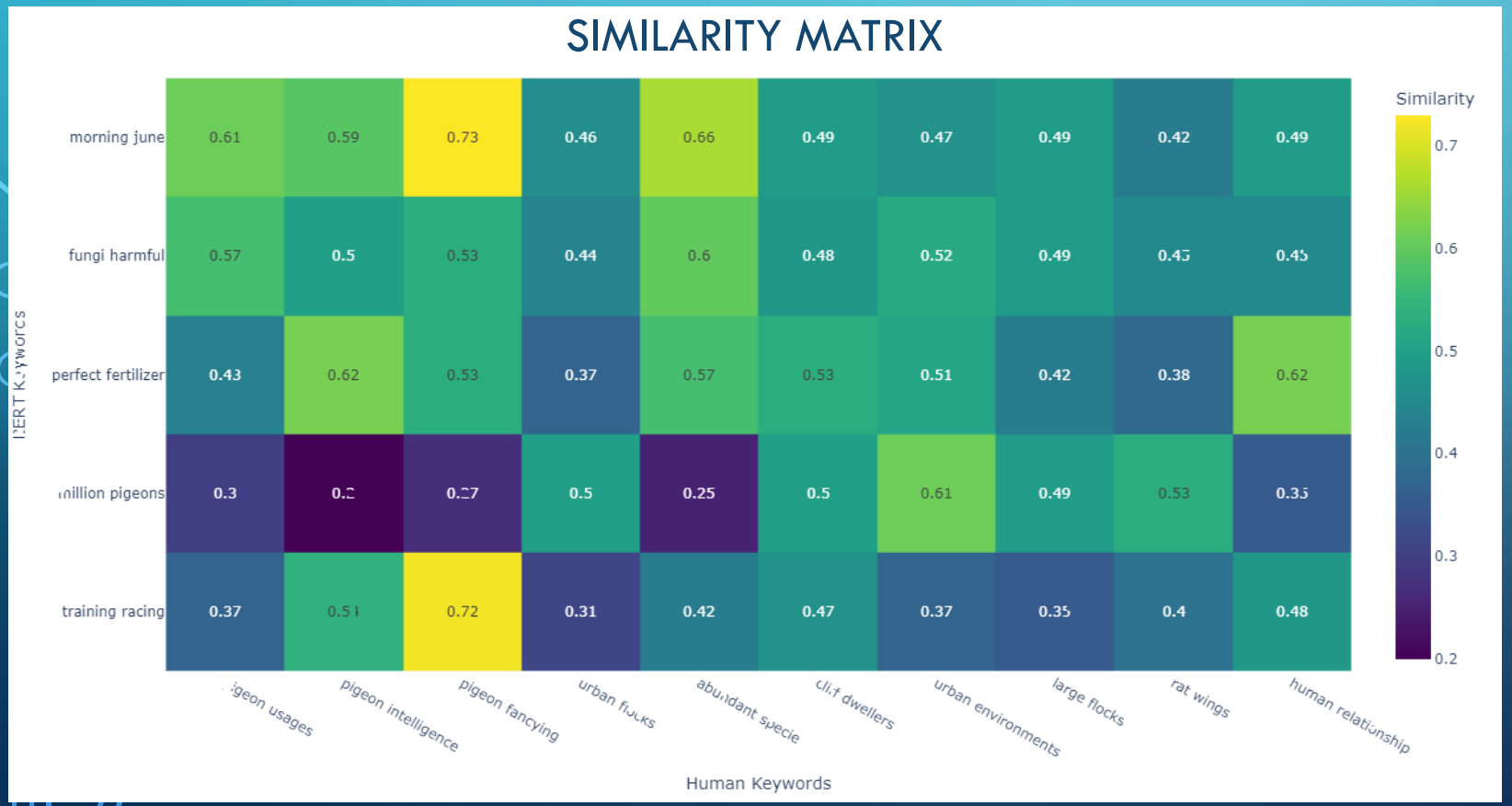
TED Talk 2

HOW PIGEONS TOOK OVER THE WORLD



TED Talk 2
HOW PIGEONS TOOK OVER THE WORLD

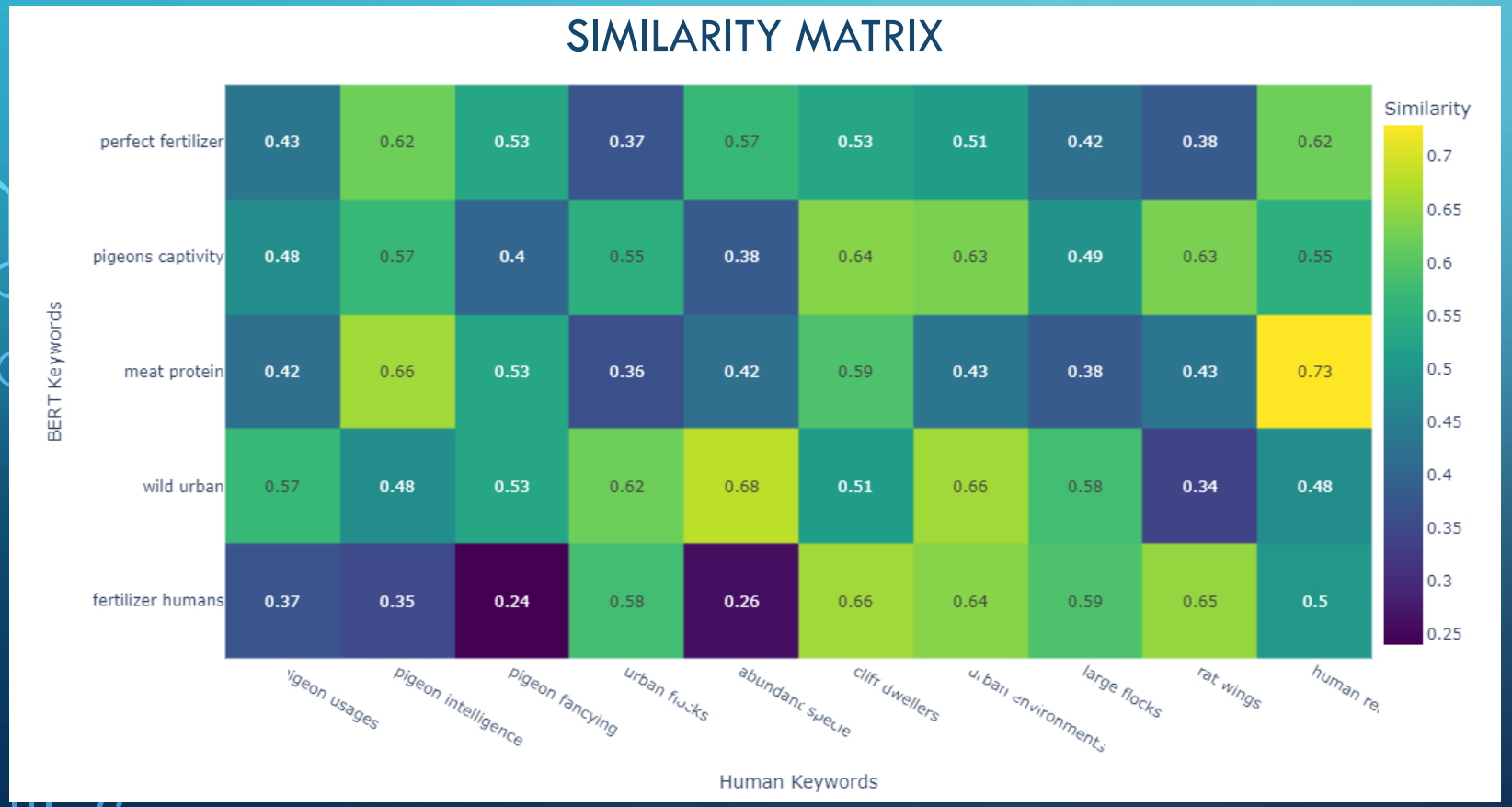
HUMAN KEYWORDS VS TRANSCRIPT BERT KEYWORDS



Overall Similarity → 67%

TED Talk 2
HOW PIGEONS TOOK OVER THE WORLD

HUMAN KEYWORDS VS SUMMARY BERT KEYWORDS



Overall Similarity

74%

TED Talk 3

THE LIFE CYCLE OF A CUP OF COFFEE

TED Talk 3
THE LIFE CYCLE OF A CUP OF COFFEE

HUMAN KEYWORDS:

TED Talk 3

THE LIFE CYCLE OF A CUP OF COFFEE

TRANSCRIPT BERT KEYWORDS	
tons coffee	45%
local forest	24%
fruit pizza	29%
flavorful seas	11%
dockworkers unload	8%

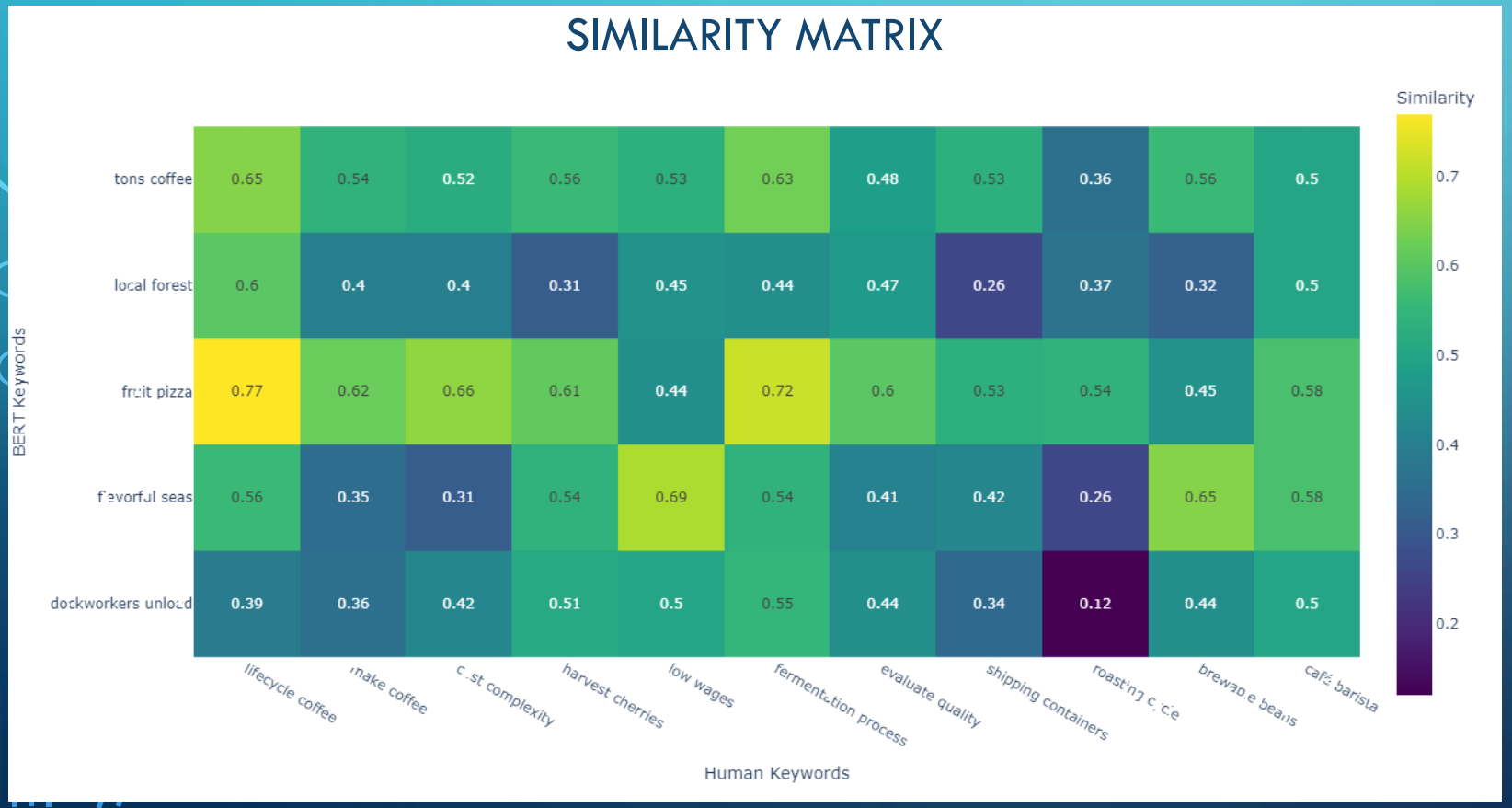
72%

Similarity Between Transcript & Summary BERT Keywords

SUMMARY BERT KEYWORDS	
cup coffee	51%
walk quick	25%
aj jacobs	6%
cost complexity	16%
elixir seed	31%

TED Talk 3
THE LIFE CYCLE OF A CUP OF COFFEE

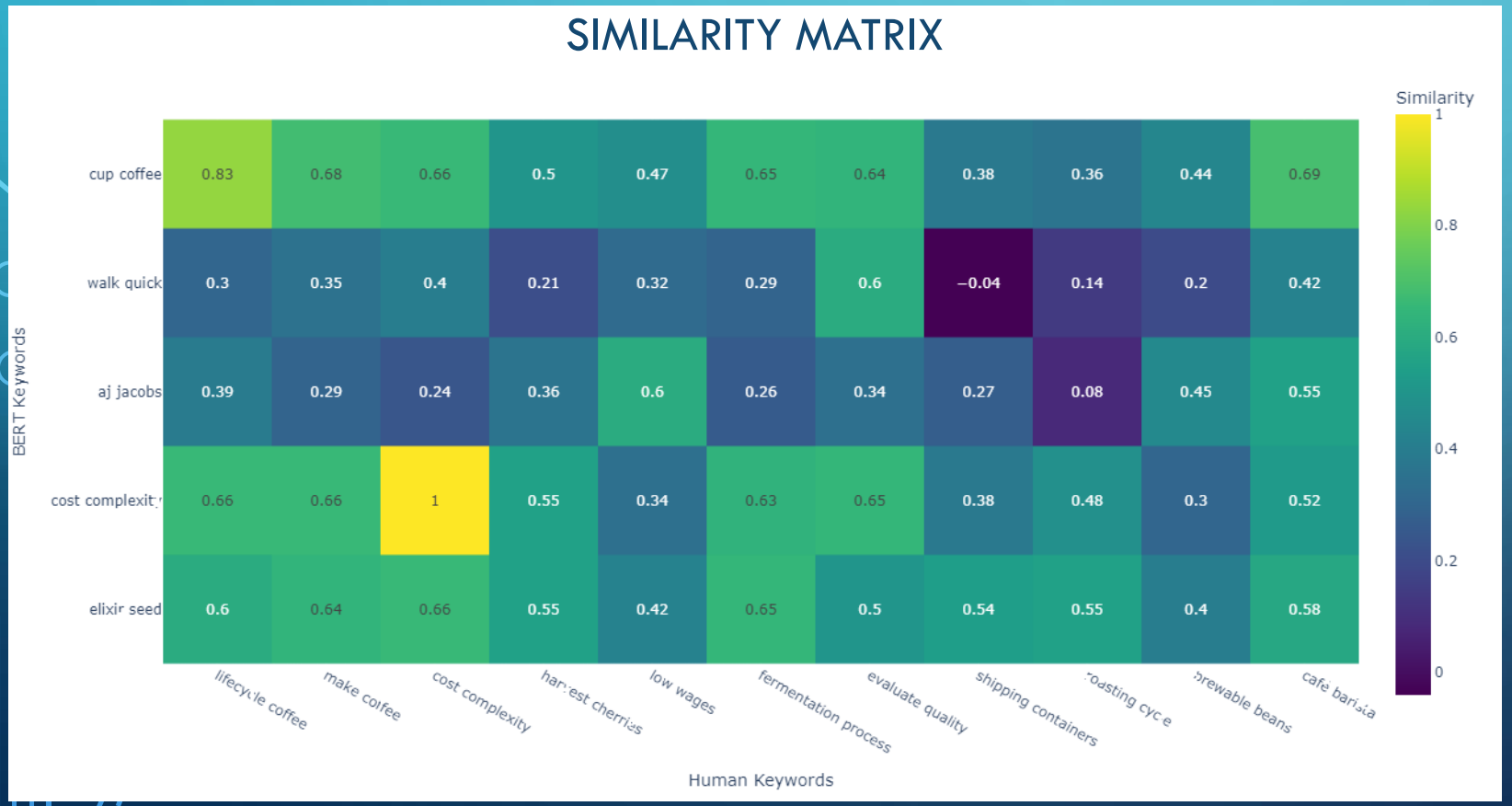
HUMAN KEYWORDS VS TRANSCRIPT BERT KEYWORDS



Overall Similarity → 80%

TED Talk 3
THE LIFE CYCLE OF A CUP OF COFFEE

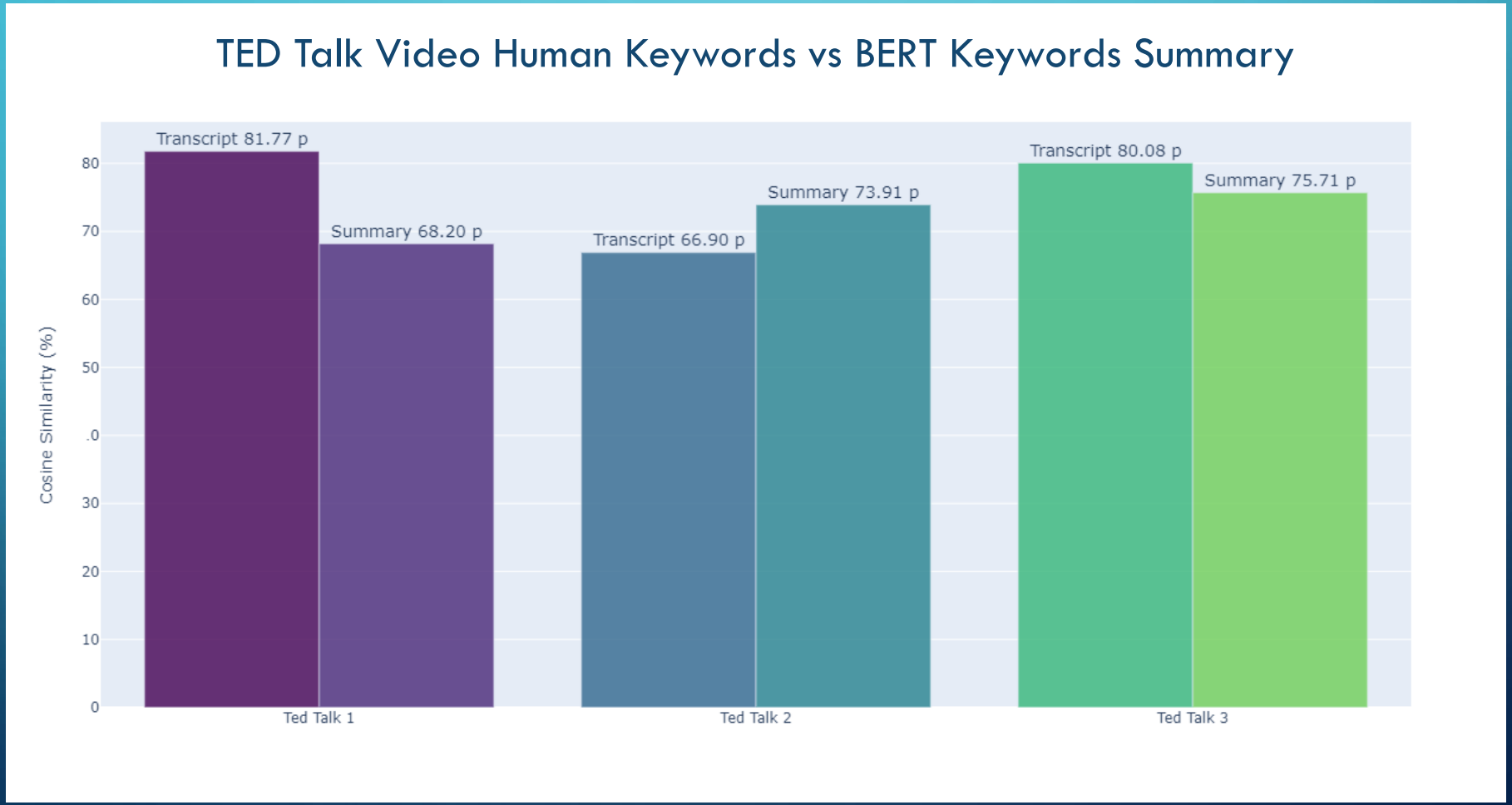
HUMAN KEYWORDS VS SUMMARY BERT KEYWORDS



Overall Similarity

76%

RESULT SUMMARY



THANK YOU!

TED TALKS

- TED Talk 1: The ingredient in almost everything you eat
- TED Talk 2: How pigeons took over the world
- TED Talk 3: The life cycle of a cup of coffee

RESOURCES

- https://towardsdatascience.com/transcribing-interview-data-from-video-to-text-with-python-5cdb6689eea1
- https://towardsdatascience.com/keyword-extraction-with-bert-724efca412ea
- https://towardsdatascience.com/how-to-extract-relevant-keywords-with-keybert-6e7b3cf889ae
- https://stackoverflow.com/questions/66919407/calculating-words-similarity-score-in-python
- https://betterprogramming.pub/the-beginners-guide-to-similarity-matching-using-spacy-782fc2922f7c