Data Retrieval using Natural Language Processing

Marie H. Cho, Changxu Liu

Institute for Computing in Research

August XX, 2022

Abstract

Modern devices allow users to search for certain files, but solely utilize the names of the files, instead of the contents stored within. As humanity progresses technologically, the increasing amounts of data will serve as a barrier to a practical application of the latter task in terms of time and cost. This will necessitate accurate methods for finding information in a timely and effective manner. In this work, we create a system to search for relevant content within texts and images in respect to a user's queries utilizing NLP that is both productive and accurate when tested on over 300,000 sources of data.

Background

i. AI and its uses in the world

Artificial Intelligence, otherwise known as AI, refers to machines and systems that imitates human intelligence by performing and improving certain tasks. Although AI has rapidly developed in recent years, the idea of mechanizing human intelligence has existed since long before. In the modern world, AI is found in all aspects of life and is continuing to develop today: autonomous cars, voice and facial recognition, and surveillance cameras are only a few prime examples. Under the broad term of AI are six main branches: robotics, expert systems, fuzzy logic, machine learning, neural networks, and natural language processing.

ii. Branches of AI: NLP - Language models, Neural Networks - CNN/RNN/FNN

There are many branches in the field of AI. Natural language processing and computer vision are present in our work.

Natural Language Processing (NLP) is concerned with the ability for computers to understand human speech and text. The field of NLP began after realizing the importance of interlingual communication and hoped for an automatic translator. As technology advanced, NLP diverged into several divisions, each regarding a different aspect of language, such as the lingual interaction between humans and computers to mimic a person to person conversation. Today, NLP is used in translators, grammar and spell check, and virtual assistants.

Recurrent neural networks are commonly used in tasks with sequential data, such as that which is found in natural language processing. Introduced in the 1980s, these are neural networks that contain a short term memory by being able to factor in the result of a previous element of a sequence into the processing the next.

<RNN pic>

A model developed more recently in 2017 are transformers. In contrast to recurrent neural networks, which processes sequential data in element by element, transformers analyze data as a whole.

Computer vision (CV) is concerned with the ability for computers to understand and process visual data. The field of computer vision started with the detection of basic edges and geometric shapes, but has expanded to fill many roles in society. Applications of this technology include self driving vehicles, facial recognition, optical character recognition, and other visual tasks.

When dealing with visual data, convolutional neural networks are frequently used to extract their features. Convolutional neural networks function by passing filters, called kernels, through an input. These kernels represent various features that may be present in the data. By performing matrix multiplication, activation values are created to represent how well different sections of an image match up with the kernel, and the feature it represents. The results may be passed through multiple layers to find increasingly sophisticated features.

<CNN pic?>

Data Processing

“To get the data into a format that we can apply natural language processing to”

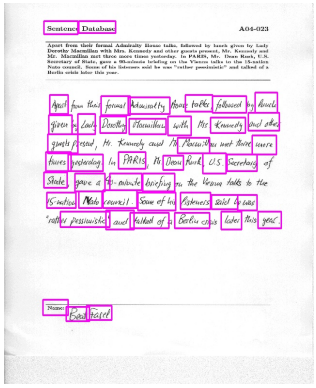
i. Initial PyPDF2 Extraction

ii. OCR (Tesseract Library, HTR)

In many cases, PDF files may be scanned documents, in which case there will only be image data present in the document, which is not searchable. In order to extract the text data within, optical character recognition (OCR) is used.

Initially, we used the Pytesseract OCR module to quickly pull printed text data out of an image. However, it was unable to properly read handwritten text data accurately with the vast amount of variation present. .

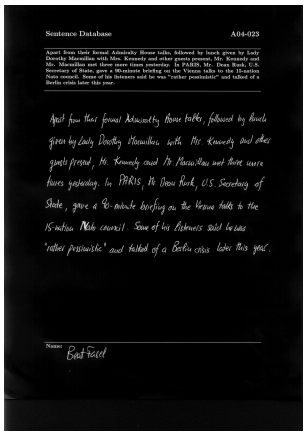
A separate system was implemented to convert handwritten text into a retrievable format. To be able to properly extract handwritten text data, the handwritten text must be located and isolated within each page before being fed into the model. This was initially done using the Efficient and Accurate Scene Text Detector (EAST) model within the OpenCV library, made by [list research paper peoples here from who wrote that paper].



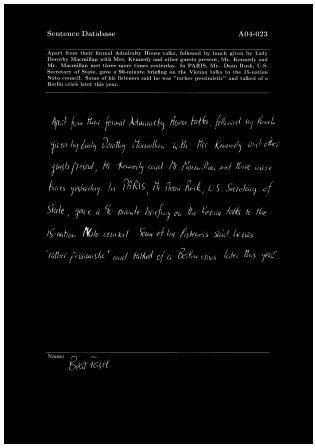
The results were found to be inconsistent. Furthermore, it would be better to have text be segmented on a line-by-line, rather than word-by-word, basis to better maintain the formatting of the document.

A simpler and more effective method involved transforms in the OpenCV library to highlight text contours in scanned data.

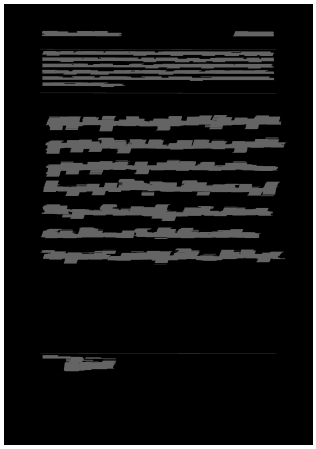
1. The scanned image is converted to grayscale and its colors are inverted.



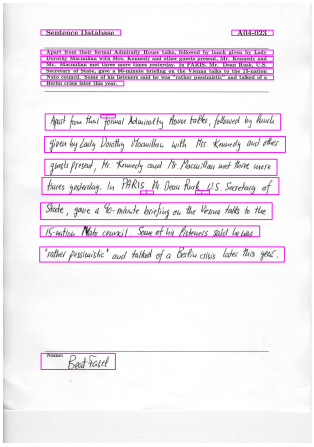
2. A threshold is applied to convert each pixel to either 0 or its max value. This helps improve the contrast, makes the lighting consistent, and reduces any noise.



3. The remaining contours on the image are dilated in the x direction to consolidate each line of text into one shape.



4. A bounding box is created around the remaining contours, which mostly consists of each line. Visualization of each bounding box on original image:



This method presents limitations to the format of image data in which text may be detectable, however since our main focus is on scanned documents for data retrieval, this should be adequate.

Then, with each line of text cropped, it can be fed into a <something this weekend - note: still spitting out gibberish> 5 layer convolutional neural network to perform feature extraction. After retrieving the

iii. Image Captioning

For images and photographs that don't contain text, generating a caption to describe the image will need to be done to make it searchable.

To do this, a convolutional neural network is utilized to extract features from the input image. We used the pre-trained 50 layer deep ResNet model. This model was originally developed and introduced in the paper "[]". Unique for its residuals which pass data past certain layers, it helped reduce the degradation problem, where the accuracy of neural networks decreased with a greater number of layers due to the model already solving everything before the end of the network.

Because classification isn't needed, the final fully-connected layer is removed from the network, and the features is then passed into a recurrent neural network, which decodes the feature vector into a basic caption that can be used to describe the image.

This model was trained on the Flickr8K dataset for X epochs.

<example images and captions>

Data Retrieval

i. BERT

Within the vast field of NLP are various language models that have been developed to make such innovations possible. One such model is the Bidirectional Encoder Representations from Transformers (BERT) Model, an unsupervised machine learning algorithm designed by Google to allow computers to understand the meanings of specific text. This serves as the bridge between human language and semantic embeddings, allowing the text files and queries to be comparable numerically.

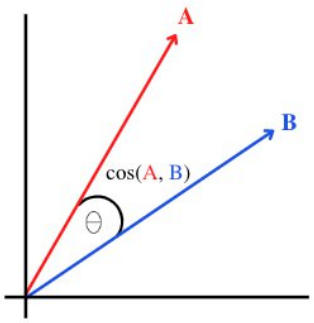
When a text, in our case a single sentence, is inputted into BERT, the model will create a tensor vector with numerical values equivalent to the meaning of the text. The same process goes for input queries.

ii. Determining Proximity

Two important methods of measuring the proximity between vector points in vector space are the cosine similarity and the Euclidean distance. Although each has their own advantage and uses for specific scenarios and the two methods are both widely used, they take different approaches of identifying the closest set of points.

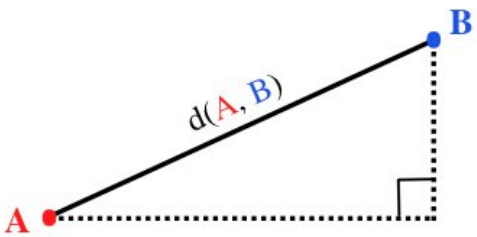
Cosine Similarity is a measure of proximity between two vectors by determining the cosine of the angle between them. Thus, the smaller the angle difference, the higher the similarity score. Two vectors of equal orientation (0°) have a maximum similarity of 1, two vectors that are orthogonal (90°) have a similarity of 0, and two vectors that lie directly opposed to each other (180°) have a minimum similarity of -1. This method is therefore a judgement of orientation and direction, independent of the vectors' magnitude and weight.

$$cos(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$



In contrast, Euclidean Distance is the distance between two points in space, or the length of a line segment connecting those two points. The greater the difference between the points, the greater the distance. Thus, this method does consider a vector's magnitude and weight when calculating the proximity.

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Rather than finding the physical distance between two points, finding the angle measure between vectors is more accurate when measuring document similarity in text analysis. Because texts can vary in terms of size and length yet can contain the same meaning as another, thus the same orientation, cosine similarity is more effective in catching the semantics and therefore more practical in this scenario.

Acceleration through Cosine Similarity and K-Means Clustering

In preparation for the experiment, we used L2 normalization to make the lengths of all sentence, reference, and query embedding vectors equal to 1. Two different modes are used: one without optimization and the other with reference vectors for optimization using k-means clustering to identify the most effective reference points at a low cost. All loading and query times are recorded for numerical representations of each reference point's efficiency.

<model>

Running document retrieval int he base mode means taking the cosine similarity of every sentence with the query and sorting a list with over 300,000 elements greatest to least, and finally retrieving only the first 5 elements. Using this method will result in 100\% accuracy, but will take a longer amount of time depending on the scale of data.

For the optimization process, the list of normalized vectors are inputted into {k}-means clustering with various numbers of centroids (1, 2, 4, 8, 16, 32). From each centroid, 5 random vectors within its respective cluster are stored into a list for later use.

The query vector, Q, is compared with each of the centroids and takes the centroid with the greatest similarity score. The program will then use the correlating list with the 5 random vectors from its cluster as its starting base (instead of [0, 0, 0, 0, 0] for greater optimization and skips) to build the final list with the top 5 nearest neighbors. Finally, for all the sentences in the nearest cluster, it will take the cosine similarity between each said sentence and the query, which then will append to a list and sort, taking only the top 5 - similar to what is seen in the base mode.

When {k}=1, it will essentially go through the same process as the base mode, as every sentence is included in the one same cluster, and will result in the same outcomes. In contrast, when {k}>1, the search times will be reduced by {k}x because we are only examining one out of {k} clusters, but in exchange for a chance that results may be not 100\% accurate as not all possible sentences are included in the similarity process. Despite this, accuracy rates are still very high while exponentially reducing query times.

Results

Query: Cognitive science is the study of the human mind and brain

Result	Similarity	File	Page	Text
1	0.873885	KrillPlatekGoetzShackelford2007.txt	1	Abstract: Cognitive neuroscience, the study

				of brain-behavior relationships, has long attempted to map the brain.
2	0.86339	MITECS.txt	40	As the science of the representation and processing of information by organisms, psychology (particularly cognitive psychology) forms part of the core of cognitive science.
3	0.851145	MITECS.txt	616	Cognitive science, on the other hand, is concerned with mechanism, with how humans reason.
4	0.849385	VerplaetseDeshrijverVannesteBraeckman2009.txt	260	Cognitive neuroscientists' central focus of attention is the question how the brain enables the mind.
5	0.845173	Fodor1981.txt	2	The psychologist frequently applies the experimental methods of the physical sciences to the study of the mind.

Conclusion/Future Work/Discussion

Throughout this project, we have been able to produce a data retrieval system through the use of natural language processing in addition to machine learning and neural networks. While our program is comparable to existing ones, we believe that in the future further steps could be taken to make this project more advanced. These are multiple next steps that can be taken to build on these findings:

1) An even larger mass of data would be helpful in proving the ability to search through huge amounts of documents and images
2) Attention Mechanism in image cap, More detailed image cap (How?)
3) Format recognition in OCR
4) Finding a more advanced method of identifying practical and effective R vectors would allow us to experiment with more vectors
5) A practical application of this program can help improve the usability in the real world

Acknowledgments

We would like to sincerely thank Optiver for providing us their space during this internship, Mark Galassi and Rhonda Crespo for running the ICR and giving us this research opportunity, our fellow interns for being supportive along the way, and finally our mentor Jim Davies for guiding us through this project.

References

Resnet paper

Pytorch

Numpy/Pandas?

LSTM Paper

BERT Paper

Tesseract OCR

OpenCV

IAM Dataset

Flickr8k Dataset

https://www.g2.com/articles/history-of-artificial-intelligence

https://cs.stanford.edu/people/eroberts/courses/soco/projects/2004-05/nlp/overview_history.html#:~:text=NLP%20%2D%20overview&text=The%20field%20of%20natural%20language,this%20sort%20of%20translation%20automatically