Data Retrieval via Natural Language Processing

Marie Cho, Changxu Liu
Institute for Computing in Research
August 4, 2022

Introduction & Background

- Data Processing
- Data Retrieval
- Results and Observations

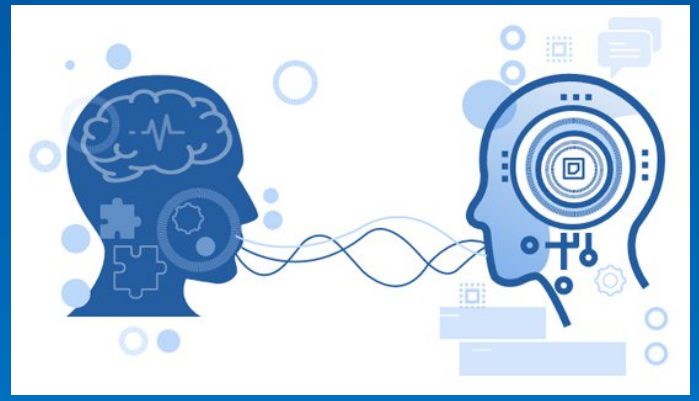
INTRODUCTION

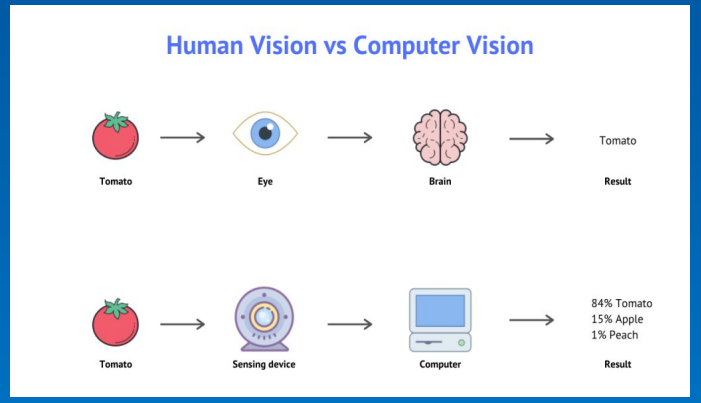
Many modern devices today implement a software to search for file names on the computer. However, most do not allow searches for the content within files themselves.

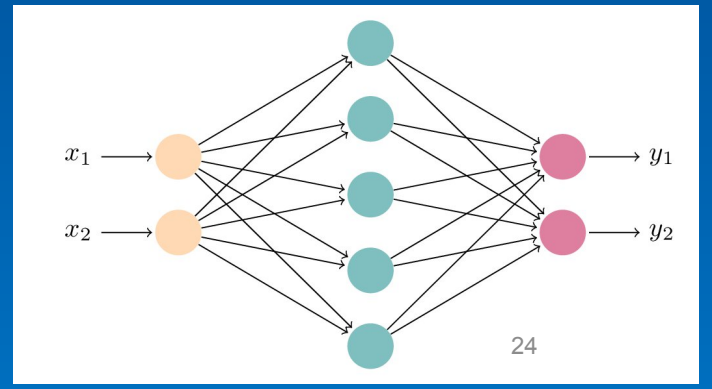
In this project, we create a data retrieval system that will search through the contents within .TXT files and images and return the most similar files.

Artificial Intelligence

- Artificial Intelligence (AI): simulation of human intelligence onto machines
- Natural Language Processing (NLP): the field of AI concerned with computer and human language interaction
  - Transformer Models: adopts self-attention, the ability to weigh the significance of a part of a data
- Neural Networks (NN): a series of computer algorithms that mimics the way the human brain — an artificial system of neurons
  - Recurrent Neural Networks (RNN): text recognition
  - Convolution Neural Networks (CNN): image analysis
- Computer Vision (CV): image and video analysis







| Introduction & Background

| Data Processing

| Data Retrieval
| Results and Observations

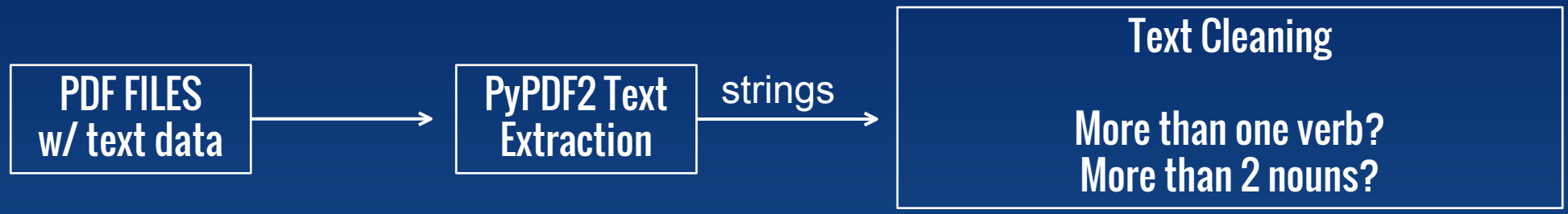
TEXT EXTRACTION & CLEANING

PDF FILES
w/ text data

TEXT EXTRACTION & CLEANING

PyPDF2: extract text from .PDF files

TEXT EXTRACTION & CLEANING



PyPDF2: extract text from .PDF files

To clean up the text:

1. Any sentences without at least 2 nouns and a verb are removed.
2. Any symbols, including dashes and '\n' are removed.

```
for sent in doc.sents:
    if sent[0].is_title and sent[-1].is_punct:
        has_noun = 2
        has_verb = 1

        for token in sent:
            if token.pos_ in ["NOUN", "PROPN", "PRON"]:
                has_noun -= 1

            elif token.pos_ == "VERB":
                has_verb -= 1

        if has_noun < 1 and has_verb < 1:
            sentence_list.append(sent.text)
```

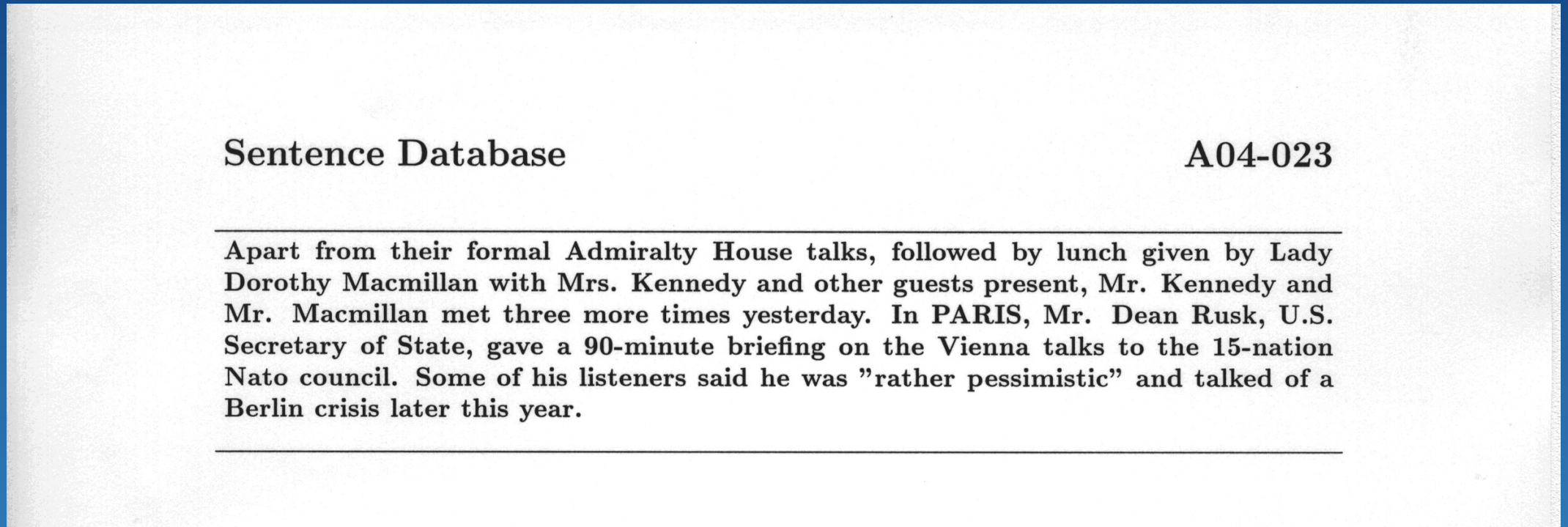
```
for i in range(0, len(sentence_list)):
    sentence_list[i] = sentence_list[i].strip()
    sentence_list[i] = re.sub("@\S+", "", sentence_list[i])
    sentence_list[i] = re.sub("#", "", sentence_list[i])
    sentence_list[i] = re.sub("\n", "", sentence_list[i])
    sentence_list[i] = re.sub("-", "", sentence_list[i])
    sentence_list[i] = re.sub("[\(\[].*?[\)\]]", "", sentence_list[i])
    if sentence_list[i].find(' ') == -1:
        sentence_list[i] = None
```

OPTICAL CHARACTER RECOGNITION

For PDFs with no text data, the Pytesseract module is used, which performs optical character recognition (OCR).

```
72
73              if len(sentence_list) == 0:
74                  page_img = convert_from_path(file)[page]
75                  text = pytesseract.image_to_string(page_img)
76                  sentence_list = clean_text(text)
77
```

Apart from their formal Admiralty House talks, followed by lunch given by Lady Dorothy Macmillan with Mrs. Kennedy and other guests present, Mr. Kennedy and Mr. Macmillan met three more times yesterday. In PARIS, Mr. Dean Rusk, U.S. Secretary of State, gave a 90-minute briefing on the Vienna talks to the 15-nation Nato council. Some of his listeners said he was ”rather pessimistic” and talked of a Berlin crisis later this year.



Sentence Database A04-023

Apart from their formal Admiralty House talks, followed by lunch given by Lady Dorothy Macmillan with Mrs. Kennedy and other guests present, Mr. Kennedy and Mr. Macmillan met three more times yesterday. In PARIS, Mr. Dean Rusk, U.S. Secretary of State, gave a 90-minute briefing on the Vienna talks to the 15-nation Nato council. Some of his listeners said he was "rather pessimistic" and talked of a Berlin crisis later this year.

Developmental prycfia(agi:{ Jean Pinget studced childrens' cognihive olevelopment
- ntevest came Fom woréeng with childrens' Countal ackivikes velated inkelligence tests in Pavis (wwk Binet) hf'fif@,fi;fiT£A4Aféwmw'
< childpens' minds gre not just "mini ™ adulé byains
- chitdrens' mind flwlap through rerits of Stages

Praget's core idea (s that fhe driving force behinol jnteflecival progression (5_an_unteaghg stuggle fo
Mok Songe of our Cperitnces,
= our bran builds SCHEMAS : concepts /mental molds iné whicd we pour our experiences

Developmental psychologist Jean Piaget studied childrens' cognitive development
- interest came from working with childrens' intelligence tests in Paris (with Binet)
- childrens' minds are not just "mini" adult brains
- childrens' mind develop through series of stages

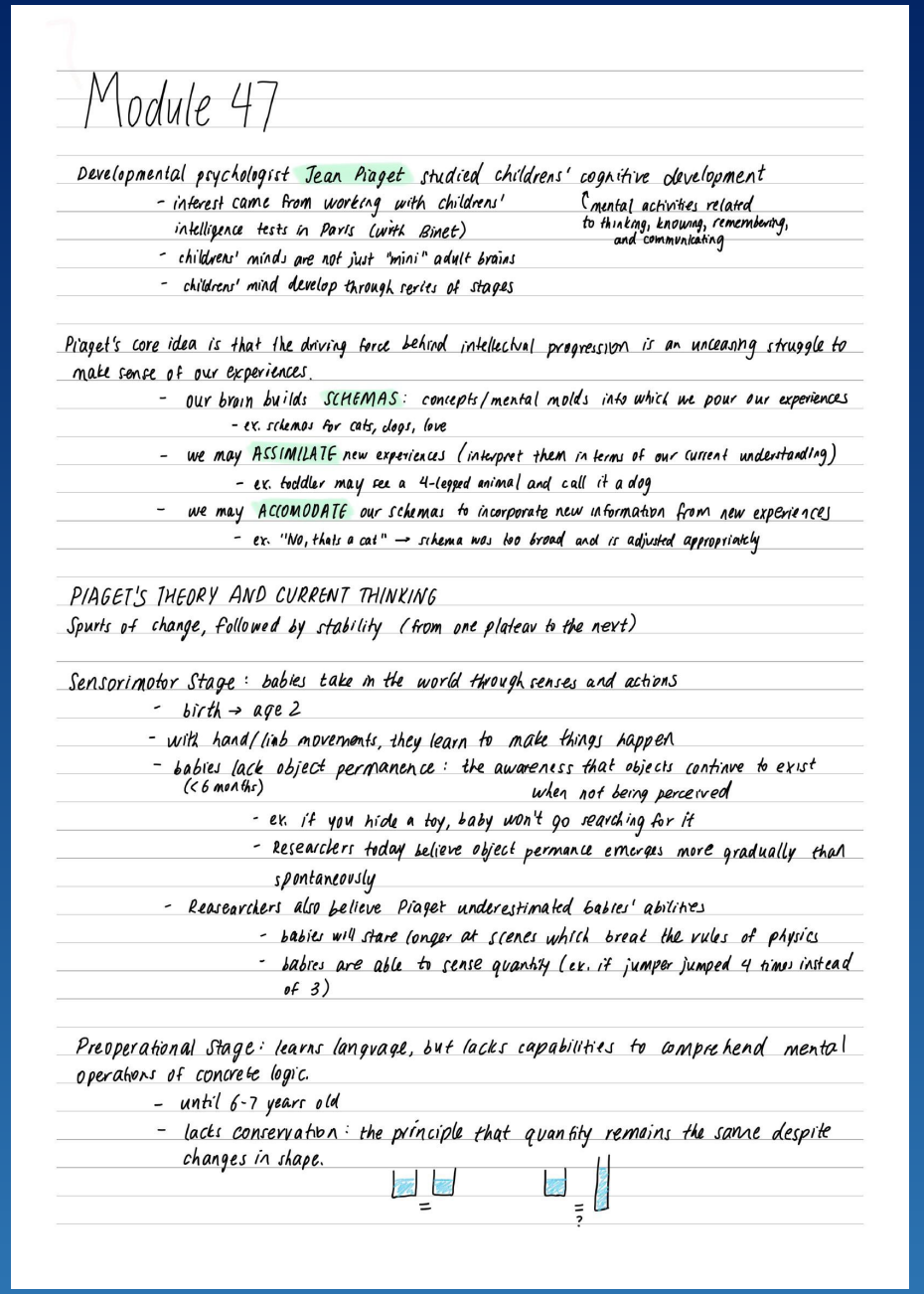
↑ mental activities related to thinking, knowing, remembering, and communicating

Piaget's core idea is that the driving force behind intellectual progression is an unceasing struggle to make sense of our experiences.
- our brain builds SCHEMAS: concepts/mental molds into which we pour our experiences
  - ex. schemas for cats, dogs, love

HANDWRITTEN TEXT RECOGNITION

But first, where is the text?

EAST: An Efficient and Accurate Scene Text Detector within OpenCV



Module 47

Developmental psychologist Jean Piaget studied childrens' cognitive development (mental activities related to thinking, knowing, remembering, and communicating)
- interest came from working with childrens' intelligence tests in Paris (with Binet)
- childrens' minds are not just "mini" adult brains
- childrens' mind develop through series of stages

Piaget's core idea is that the driving force behind intellectual progression is an unceasing struggle to make sense of our experiences.
- our brain builds SCHEMAS: concepts/mental molds into which we pour our experiences
  - ex. schemas for cats, dogs, love
- we may ASSIMILATE new experiences (interpret them in terms of our current understanding)
  - ex. toddler may see a 4-legged animal and call it a dog
- we may ACCOMODATE our schemas to incorporate new information from new experiences
  - ex. "No, thats a cat" → schema was too broad and is adjusted appropriately

PIAGET'S THEORY AND CURRENT THINKING
Spurts of change, followed by stability (from one plateau to the next)

Sensorimotor Stage: babies take in the world through senses and actions
- birth → age 2
- with hand/limb movements, they learn to make things happen
- babies lack object permanence: the awareness that objects continue to exist when not being perceived (<6 months)
  - ex. if you hide a toy, baby won't go searching for it
  - Researchers today believe object permance emerges more gradually than spontaneously
- Reasearchers also believe Piaget underestimated babies' abilities
  - babies will stare longer at scenes which break the rules of physics
  - babies are able to sense quantity (ex. if jumper jumped 4 times instead of 3)

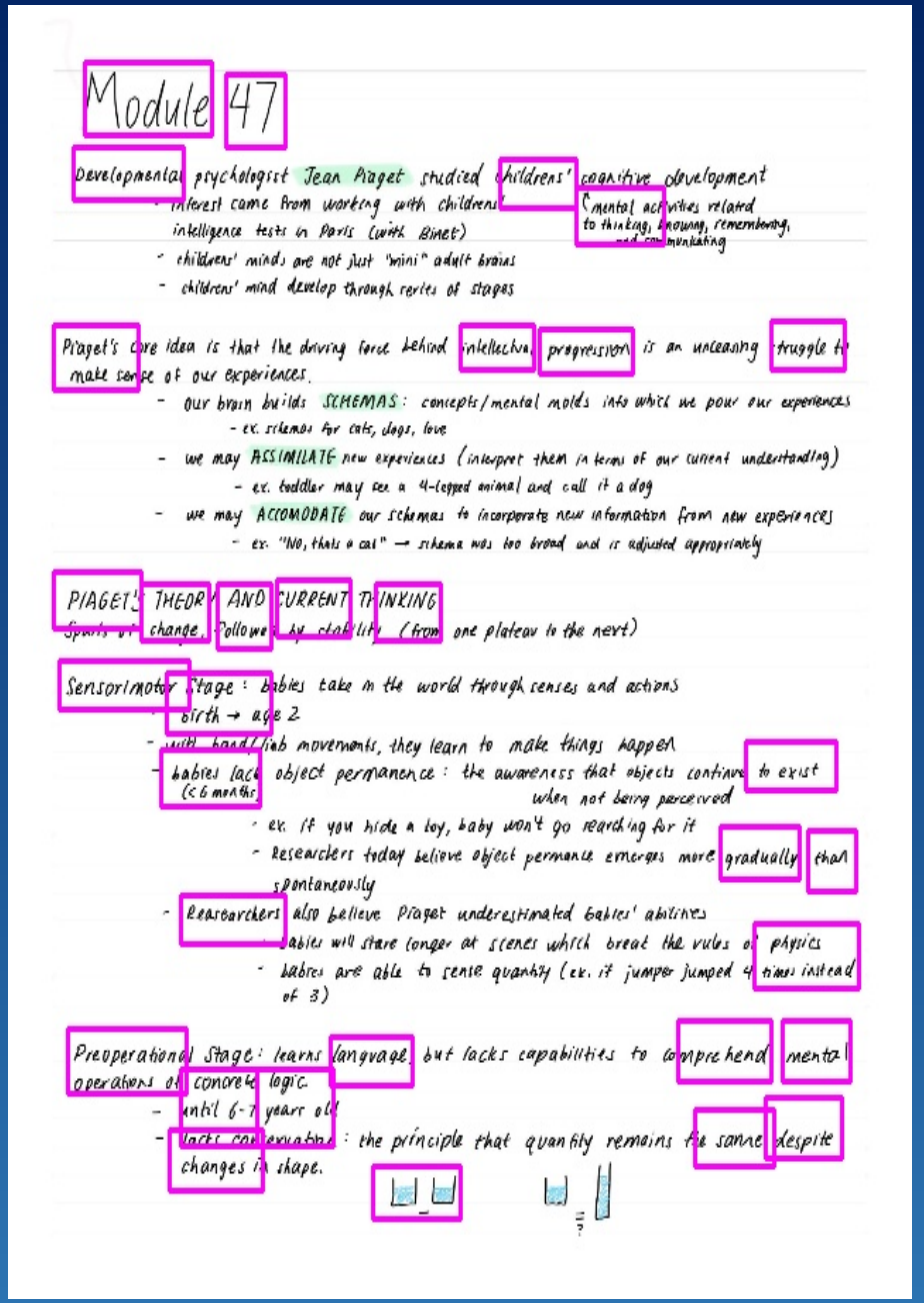
Preoperational Stage: learns language, but lacks capabilities to comprehend mental operations of concrete logic.
- until 6-7 years old
- lacks conservation: the principle that quantity remains the same despite changes in shape.

HANDWRITTEN TEXT RECOGNITION

But first, where is the text?

EAST: An Efficient and Accurate Scene Text Detector within OpenCV

Results were inconsistent.



HANDWRITTEN TEXT RECOGNITION

To find the text...

Module 47

Developmental psychologist Jean Piaget studied childrens' cognitive development (mental activities related to thinking, knowing, remembering, and communicating)
- interest came from working with childrens' intelligence tests in Paris (with Binet)
- childrens' minds are not just "mini" adult brains
- childrens' mind develop through series of stages

Piaget's core idea is that the driving force behind intellectual progression is an unceasing struggle to make sense of our experiences.
- our brain builds SCHEMAS: concepts/mental molds into which we pour our experiences
  - ex. schemas for cats, dogs, love
- we may ASSIMILATE new experiences (interpret them in terms of our current understanding)
  - ex. toddler may see a 4-legged animal and call it a dog
- we may ACCOMODATE our schemas to incorporate new information from new experiences
  - ex. "No, thats a cat" → schema was too broad and is adjusted appropriately

PIAGET'S THEORY AND CURRENT THINKING

Spurts of change, followed by stability (from one plateau to the next)

Sensorimotor Stage: babies take in the world through senses and actions
- birth → age 2
- with hand/limb movements, they learn to make things happen
- babies lack object permanence: the awareness that objects continue to exist when not being perceived (<6 months)
  - ex. if you hide a toy, baby won't go searching for it
  - Researchers today believe object permance emerges more gradually than spontaneously
- Reasearchers also believe Piaget underestimated babies' abilities
  - babies will stare longer at scenes which break the rules of physics
  - babies are able to sense quantity (ex. if jumper jumped 4 times instead of 3)

Preoperational Stage: learns language, but lacks capabilities to comprehend mental operations of concrete logic.
- until 6-7 years old
- lacks conservation: the principle that quantity remains the same despite changes in shape.

HANDWRITTEN TEXT RECOGNITION

To find the text...

1. Convert the image to grayscale and invert colors.

HANDWRITTEN TEXT RECOGNITION

To find the text...

1. Convert the image to grayscale and invert colors.
2. Apply a threshold to the image.

Module 47

Developmental psychologist Jean Piaget studied childrens' cognitive development (mental activities related to thinking, knowing, remembering, and communicating)
- interest came from working with childrens' intelligence tests in Paris (with Binet)
- childrens' minds are not just "mini" adult brains
- childrens' mind develop through series of stages

Piaget's core idea is that the driving force behind intellectual progression is an unceasing struggle to make sense of our experiences.
- our brain builds SCHEMAS: concepts/mental molds into which we pour our experiences
  - ex. schemas for cats, dogs, love
- we may ASSIMILATE new experiences (interpret them in terms of our current understanding)
  - ex. toddler may see a 4-legged animal and call it a dog
- we may ACCOMODATE our schemas to incorporate new information from new experiences
  - ex. "NO, thats a cat" → schema was too broad and is adjusted appropriately

PIAGET'S THEORY AND CURRENT THINKING
Spurts of change, followed by stability (from one plateau to the next)

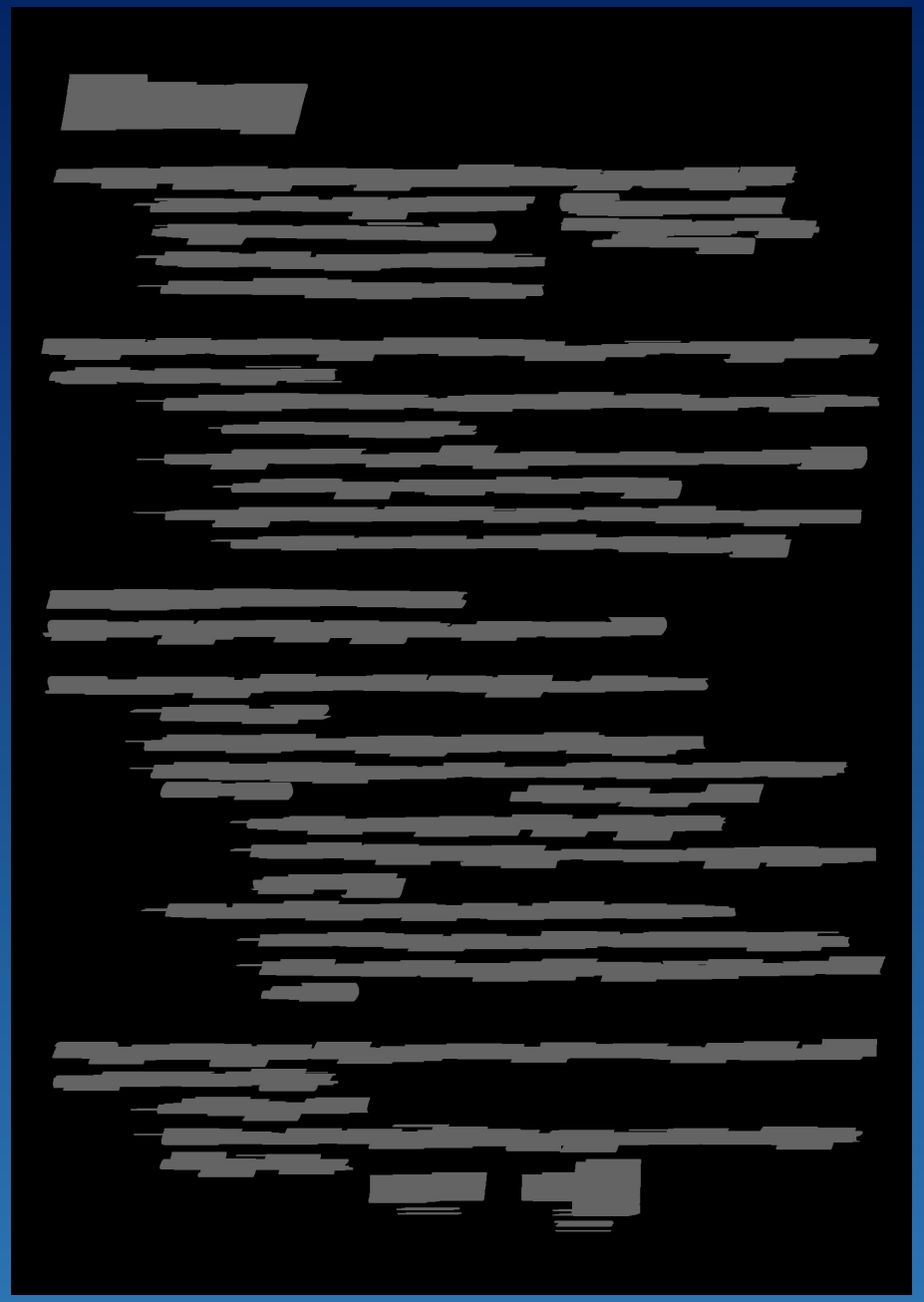
Sensorimotor Stage: babies take in the world through senses and actions
- birth → age 2
- with hand/limb movements, they learn to make things happen
- babies lack object permanence: the awareness that objects continue to exist when not being perceived (< 6 months)
  - ex. if you hide a toy, baby won't go searching for it
  - Researchers today believe object permanence emerges more gradually than spontaneously
- Reasearchers also believe Piaget underestimated babies' abilities
  - babies will stare longer at scenes which break the rules of physics
  - babies are able to sense quantity (ex. if jumper jumped 4 times instead of 3)

Preoperational Stage: learns language, but lacks capabilities to comprehend mental operations of concrete logic.
- until 6-7 years old
- lacks conservation: the principle that quantity remains the same despite changes in shape.

HANDWRITTEN TEXT RECOGNITION

To find the text...

1. Convert the image to grayscale and invert colors.
2. Apply a threshold to the image.
3. Dilate the remaining contours in the x direction so that the words come together.



- childrens' minds are not just "mini" adult brain

intelligence tests in Paris (with Binet)

velopmental psychologist Jean Piaget studied chil

(mental activities related

- interest came from working with childrens'

childrens' mind develop through series of stages

Module 47

Developmental psychologist Jean Piaget studied childrens' cognitive development
- interest came from working with childrens'
  - intelligence tests in Paris (with Binet)
- childrens' minds are not just "mini" adult brains
- childrens' mind develop through series of stages

(mental activities related to thinking, knowing, remembering, and communicating

Piaget's core idea is that the driving force behind intellectual progression is an unceasing struggle to make sense of our experiences.
- our brain builds SCHEMAS: concepts/mental molds into which we pour our experiences
  - ex. schemas for cats, dogs, love
- we may ASSIMILATE new experiences (interpret them in terms of our current understanding)
  - ex. toddler may see a 4-legged animal and call it a dog
- we may ACCOMODATE our schemas to incorporate new information from new experiences
  - ex. "No, that's a cat" → schema was too broad and is adjusted appropriately

PIAGET'S THEORY AND CURRENT THINKING
Spurts of change, followed by stability (from one plateau to the next)

Sensorimotor Stage: babies take in the world through senses and actions
- birth → age 2
- with hand/lib movements, they learn to make things happen
- babies lack object permanence: the awareness that objects continue to exist when not being perceived (<6 months)
  - ex. if you hide a toy, baby won't go searching for it
  - researchers today believe object permance emerges more gradually than spontaneously
- Reasearchers also believe Piaget underestimated babies' abilities
  - babies will stare longer at scenes which break the rules of physics
  - babies are able to sense quantity (ex. if jumper jumped 4 times instead of 3)

Preoperational Stage: learns language, but lacks capabilities to comprehend mental operations of concrete logic.
- until 6-7 years old
- lacks conservation: the principle that quantity remains the same despite changes in shape.

HANDWRITTEN TEXT RECOGNITION

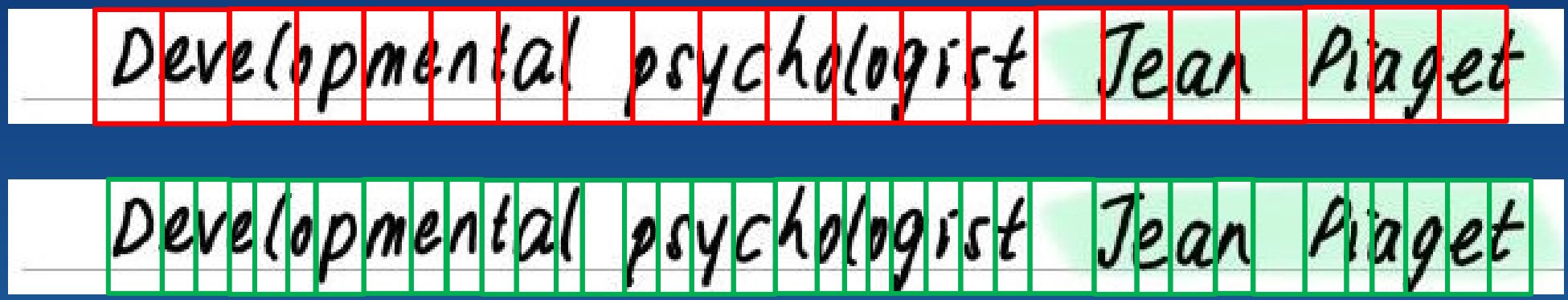
For each line...

HANDWRITTEN TEXT RECOGNITION

CTC Decoder
and CTC Loss

Connectionist Temporal Classification (CTC) is used when a sequence isn't perfectly spaced out.

HANDWRITTEN TEXT RECOGNITION



Connectionist Temporal Classification (CTC) is used when a sequence isn't perfectly spaced out.

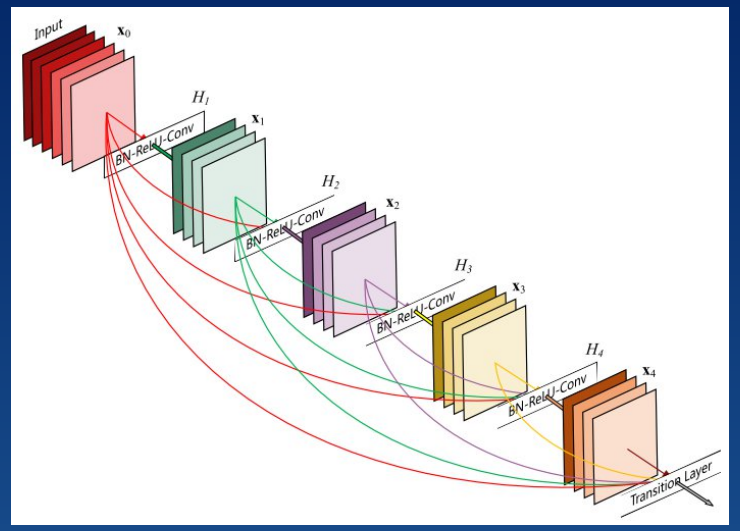
RESULTS?

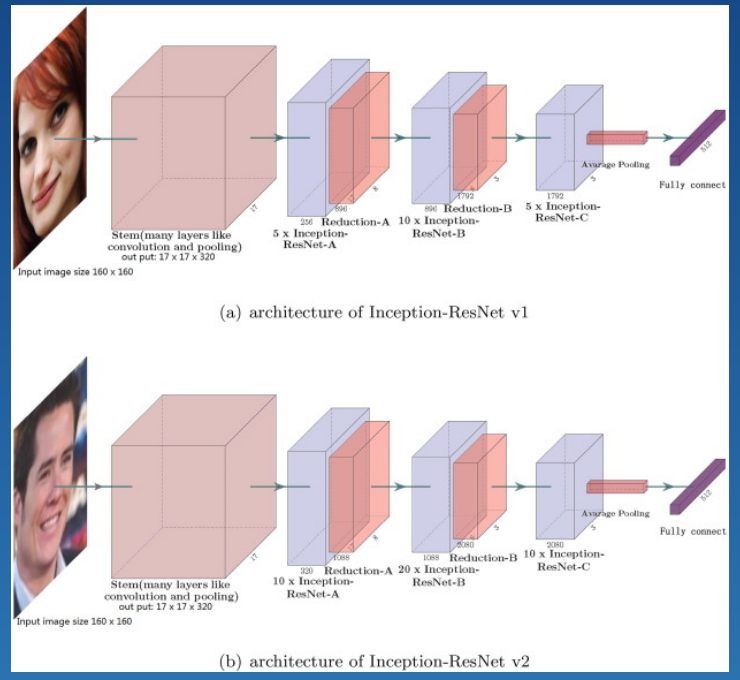
IMAGE CAPTIONING

Generate a caption for pure images and photographs.

Resnet Model: a 50 layer Convolutional Neural Network to extract image features

Passed to Recurrent Neural Network that decodes features into a caption for description





(a) architecture of Inception-ResNet v1

(b) architecture of Inception-ResNet v2

IMAGE CAPTIONING

Preparing the images...

IMAGE CAPTIONING

Preparing the images...

1. Resizing to (299, 299)

IMAGE CAPTIONING

Preparing the images...

1. Resizing to (299, 299)
2. Random Cropping to (224, 224)

IMAGE CAPTIONING

Preparing the images...

1. Resizing to (299, 299)
2. Random Cropping to (224, 224)
3. Random Horizontal Flip

IMAGE CAPTIONING

Preparing the images...

1. Resizing to (299, 299)
2. Random Cropping to (224, 224)
3. Random Horizontal Flip
4. Convert to Tensor

```
[[[0.7333, 0.7294, 0.7294,  ..., 0.8039, 0.8039, 0.8039],
  [0.7333, 0.7294, 0.7294,  ..., 0.8039, 0.8039, 0.8039],
  [0.7333, 0.7294, 0.7294,  ..., 0.8039, 0.8039, 0.8039],
   ...,
  [0.5294, 0.5333, 0.5373,  ..., 0.5804, 0.5765, 0.5765],
  [0.5333, 0.5373, 0.5412,  ..., 0.5961, 0.5961, 0.5922],
  [0.5333, 0.5333, 0.5373,  ..., 0.5804, 0.5804, 0.5804]],

 [[0.7333, 0.7294, 0.7294,  ..., 0.8039, 0.8039, 0.8039],
  [0.7333, 0.7294, 0.7294,  ..., 0.8039, 0.8039, 0.8039],
  [0.7333, 0.7294, 0.7294,  ..., 0.8039, 0.8039, 0.8039],
   ...,
  [0.5451, 0.5490, 0.5529,  ..., 0.6157, 0.6118, 0.6118],
  [0.5490, 0.5529, 0.5569,  ..., 0.6314, 0.6314, 0.6275],
  [0.5490, 0.5490, 0.5529,  ..., 0.6157, 0.6157, 0.6157]],

 [[0.7333, 0.7294, 0.7294,  ..., 0.7961, 0.7961, 0.7961],
  [0.7333, 0.7294, 0.7294,  ..., 0.7961, 0.7961, 0.7961],
  [0.7333, 0.7294, 0.7294,  ..., 0.7961, 0.7961, 0.7961],
   ...,
  [0.5412, 0.5451, 0.5490,  ..., 0.6039, 0.6000, 0.6000],
  [0.5451, 0.5490, 0.5529,  ..., 0.6196, 0.6196, 0.6157],
  [0.5451, 0.5451, 0.5490,  ..., 0.6039, 0.6039, 0.6039]]]
```

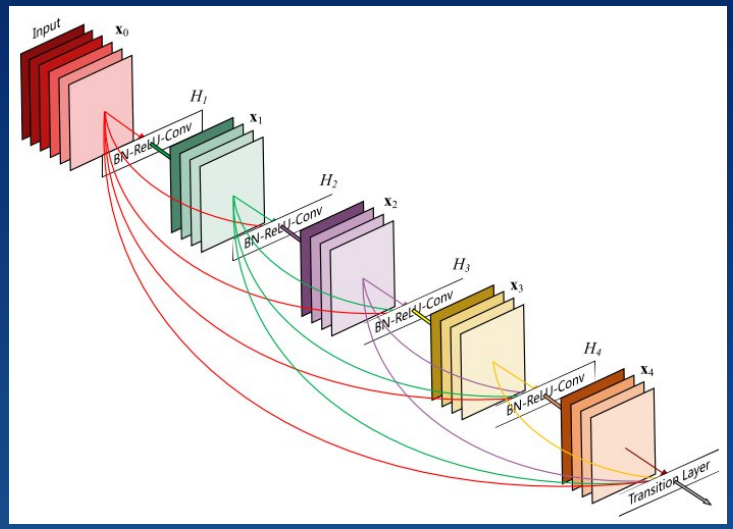
IMAGE CAPTIONING

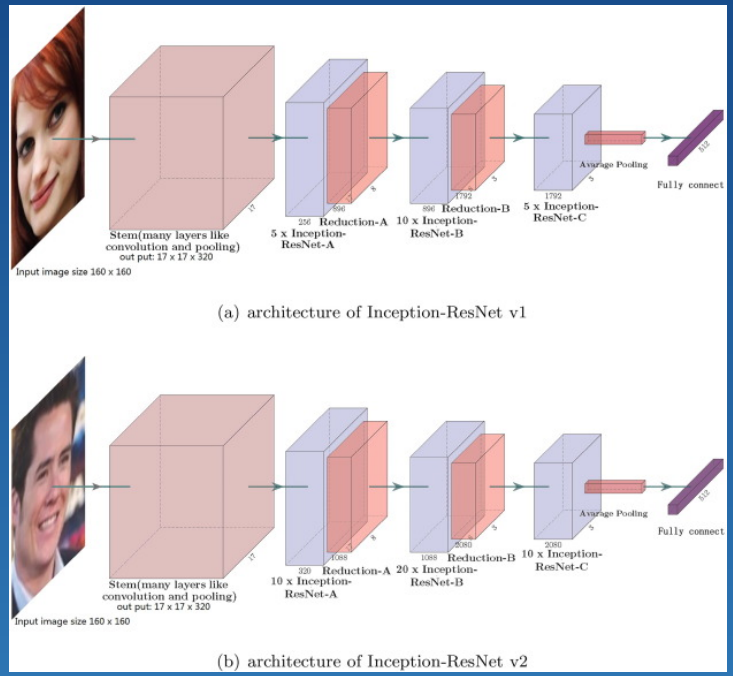
Preparing the images...

1. Resizing to (299, 299)
2. Random Cropping to (224, 224)
3. Random Horizontal Flip
4. Convert to Tensor
5. Normalization

Image Captioning

RESULTS





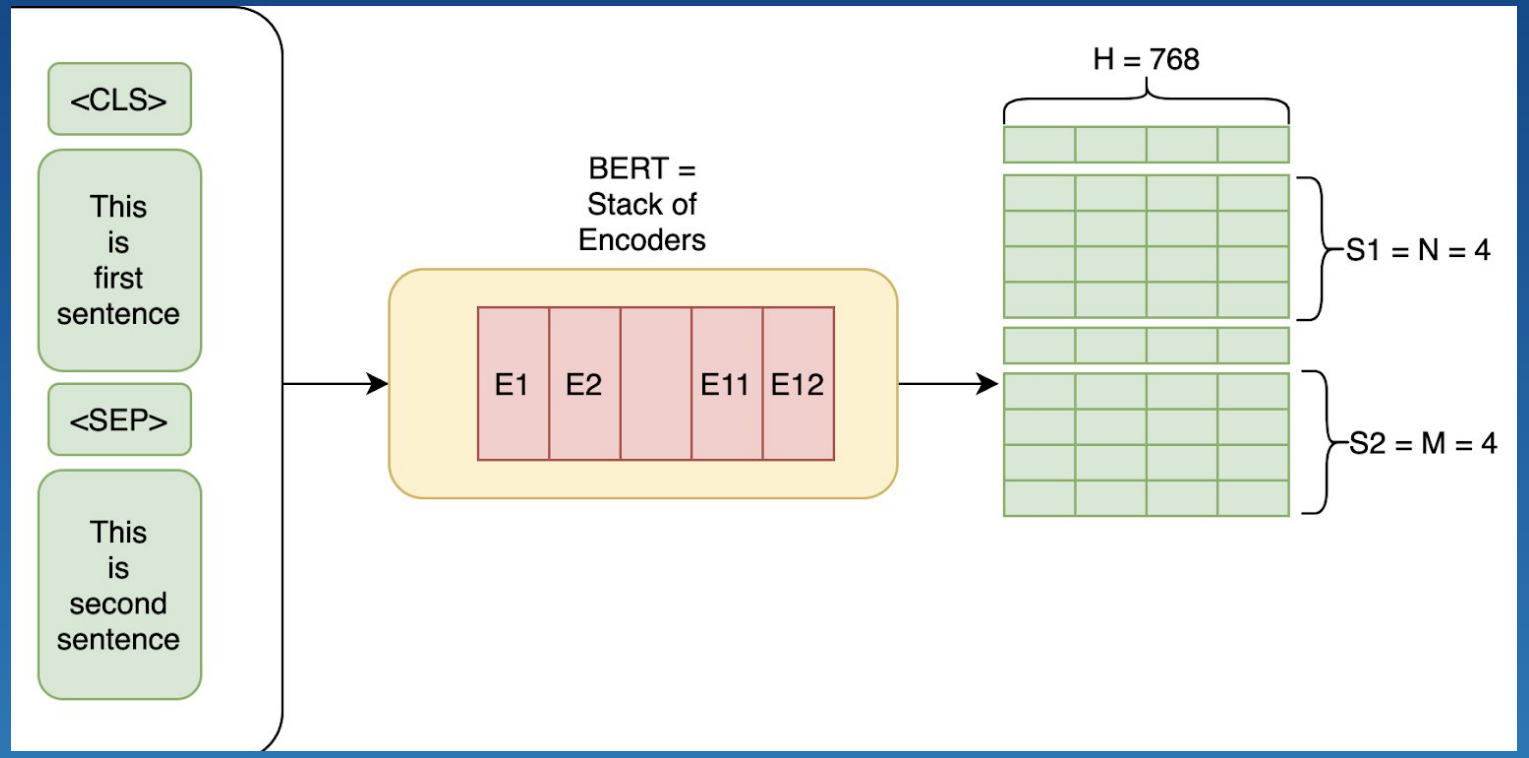
| Introduction & Background
| Data Processing

| Data Retrieval

| Results and Observations

The BERT Model

- Bidirectional Encoder Representations from Transformers (BERT): an unsupervised Transformer-based model for NLP
- Bridge between human language and semantic embeddings for numerical comparisons
- Text => model => tensor vector

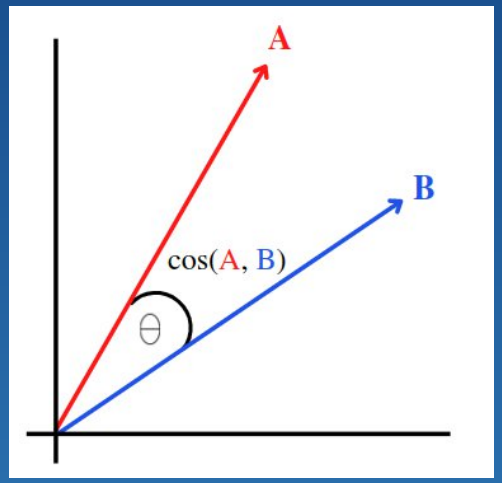


Means of Determining Proximity

Cosine Similarity

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

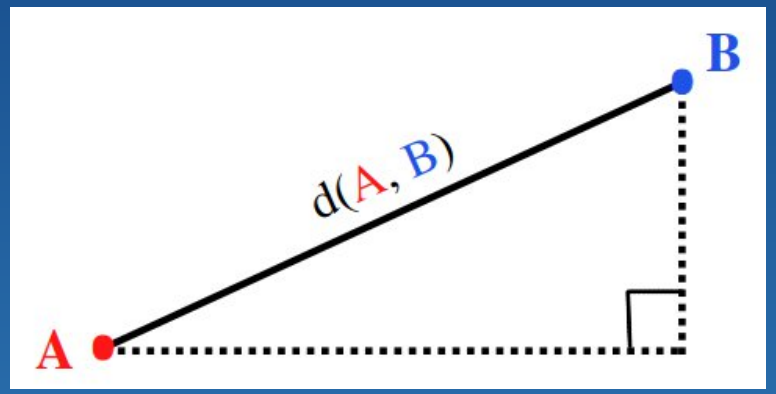
- Measures angle between vectors
- Orientation and rotation
- Independent of magnitude and weight



Euclidean Distance

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

- Measures distance between two points on vectors
- Dependent on magnitude and weight

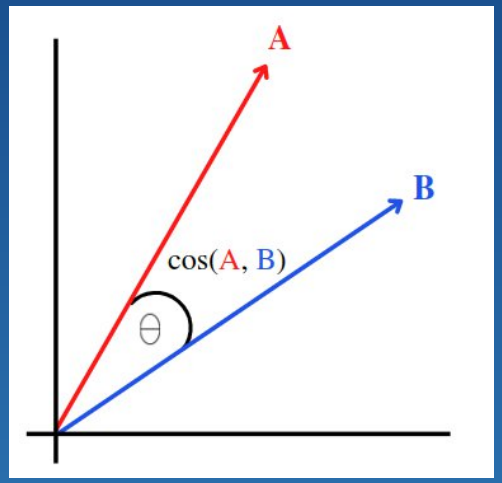


Means of Determining Proximity

Cosine Similarity

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

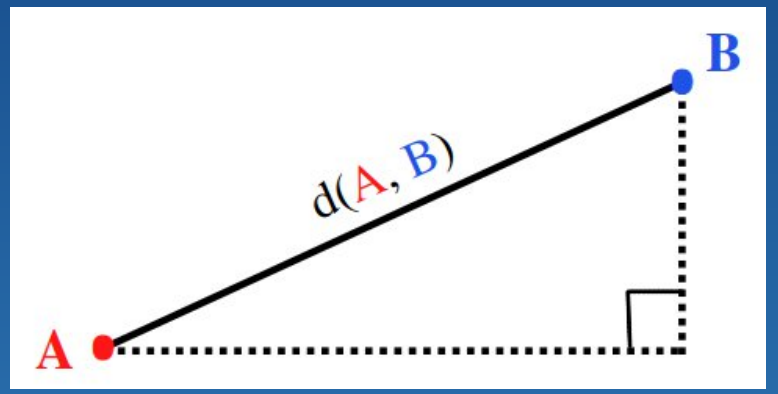
- Measures angle between vectors
- Orientation and rotation
- Independent of magnitude and weight



Euclidean Distance

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

- Measures distance between two points on vectors
- Dependent on magnitude and weight

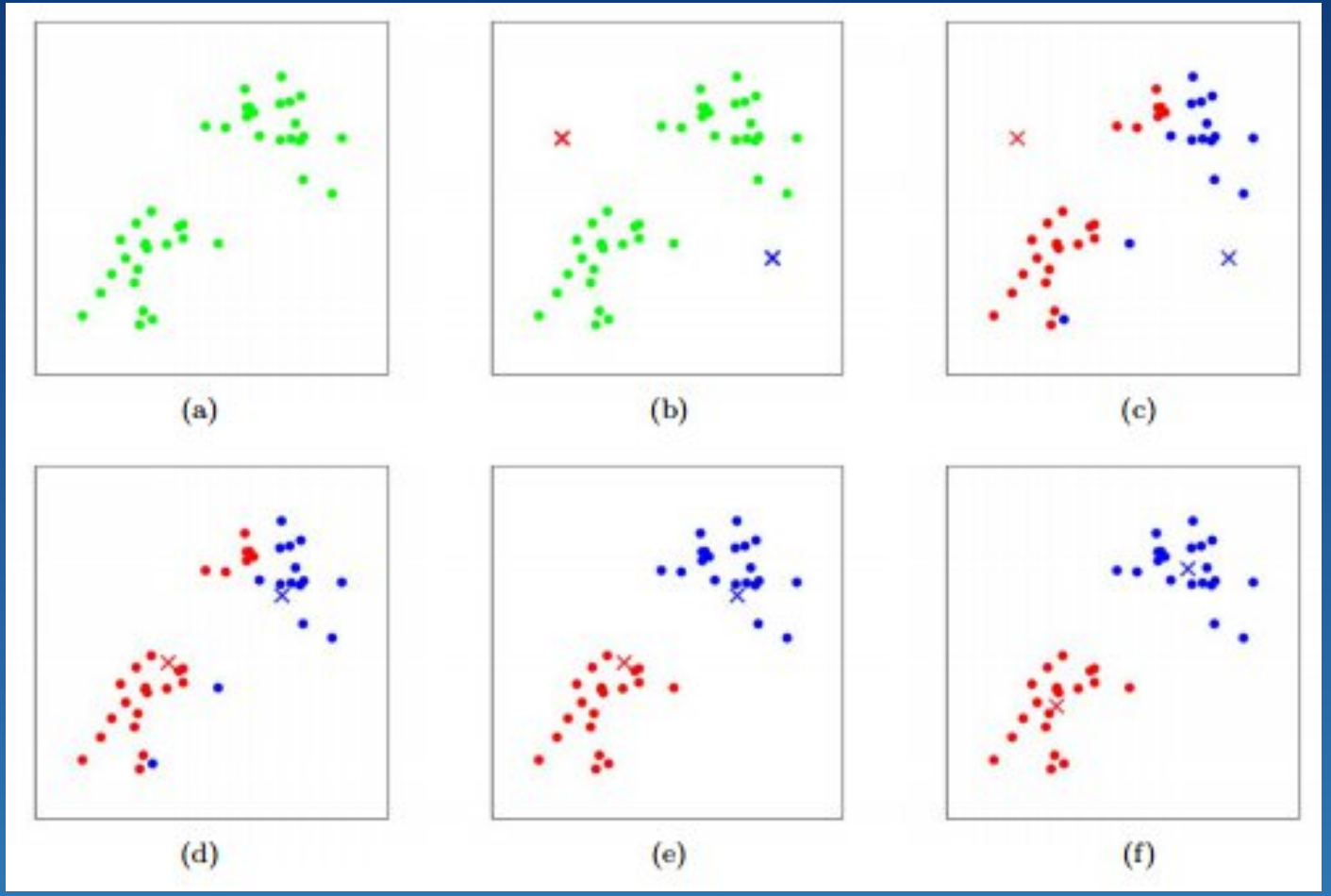


K-Means Clustering

- K-Means Clustering: an unsupervised machine learning algorithm for vector organization

- Vectors in a space separated into k groups — form k centroids and assign each vector to the closest centroid labeled 1, 2, 3 ... k

- Centroids adjust by taking average of the vectors in its cluster until centroids do not move or the adjustment is minimal



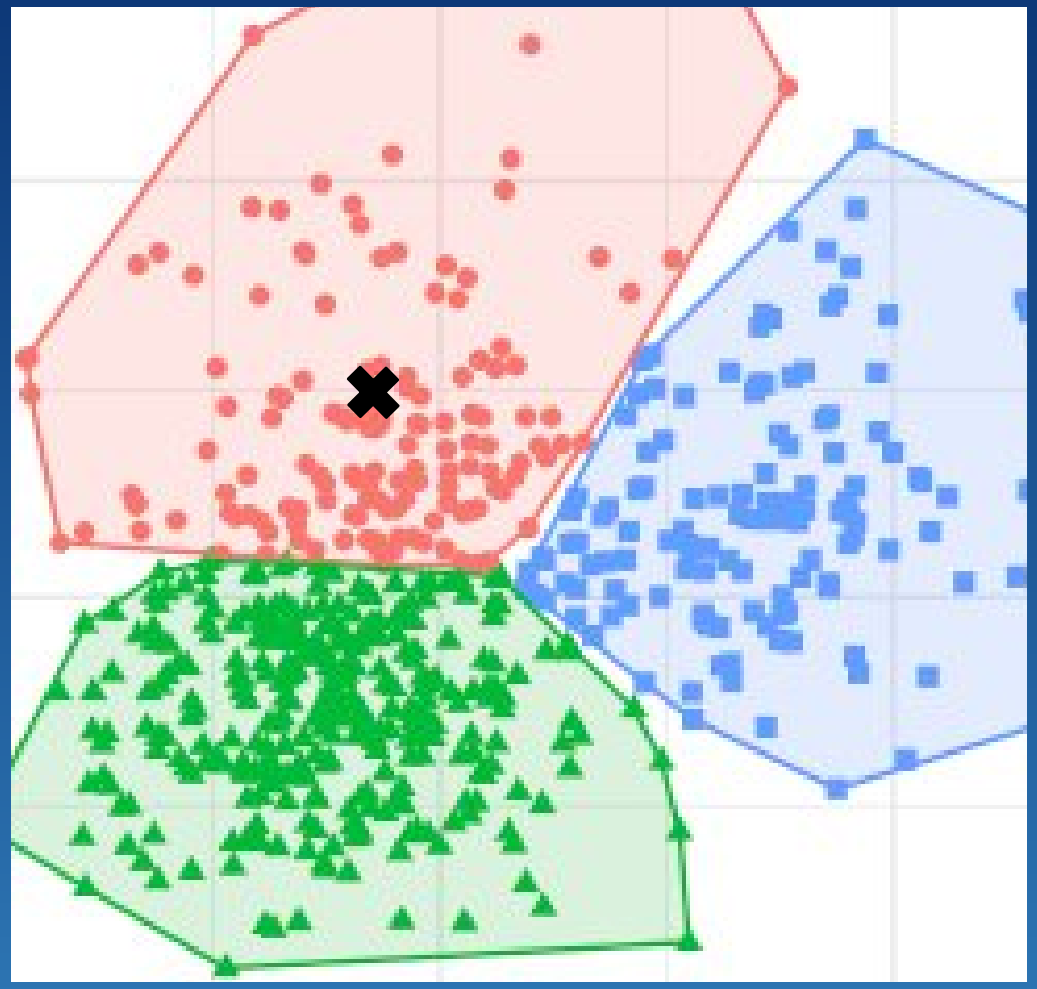
Acceleration Through K-Means Clustering

- Higher cosine similarity = lower Euclidean distance
- Take cosine similarity of all k vectors and query
- Search in only the closest cluster

[0.8, 0.3, 0.6, 0.2, 0.5] => [0.8, 0.6, 0.5, 0.3, 0.2]

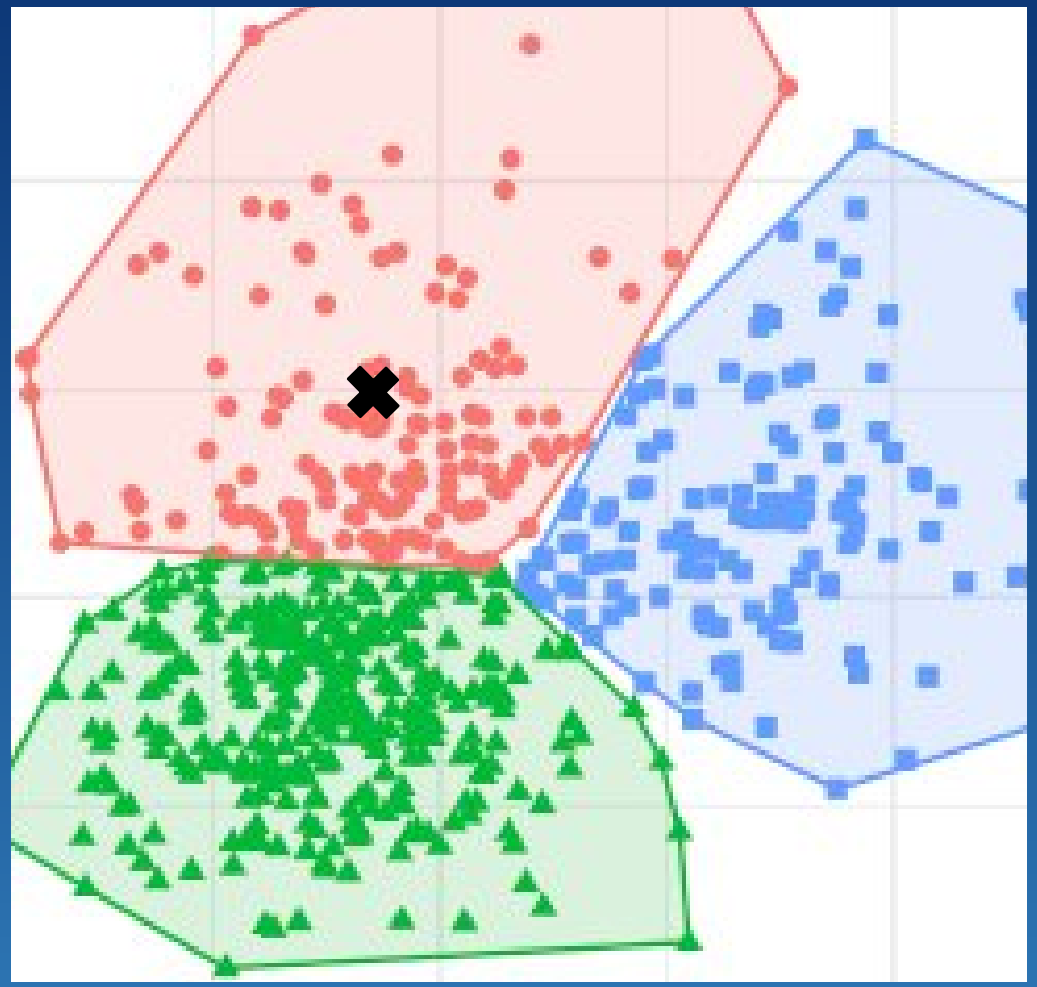
similarity(query, sentence) = 0.4 > 0.2	similarity(query, sentence) = 0.1 < 0.2
[0.8, 0.6, 0.5, 0.4, 0.3]	No change

Acceleration Through K-Means Clustering

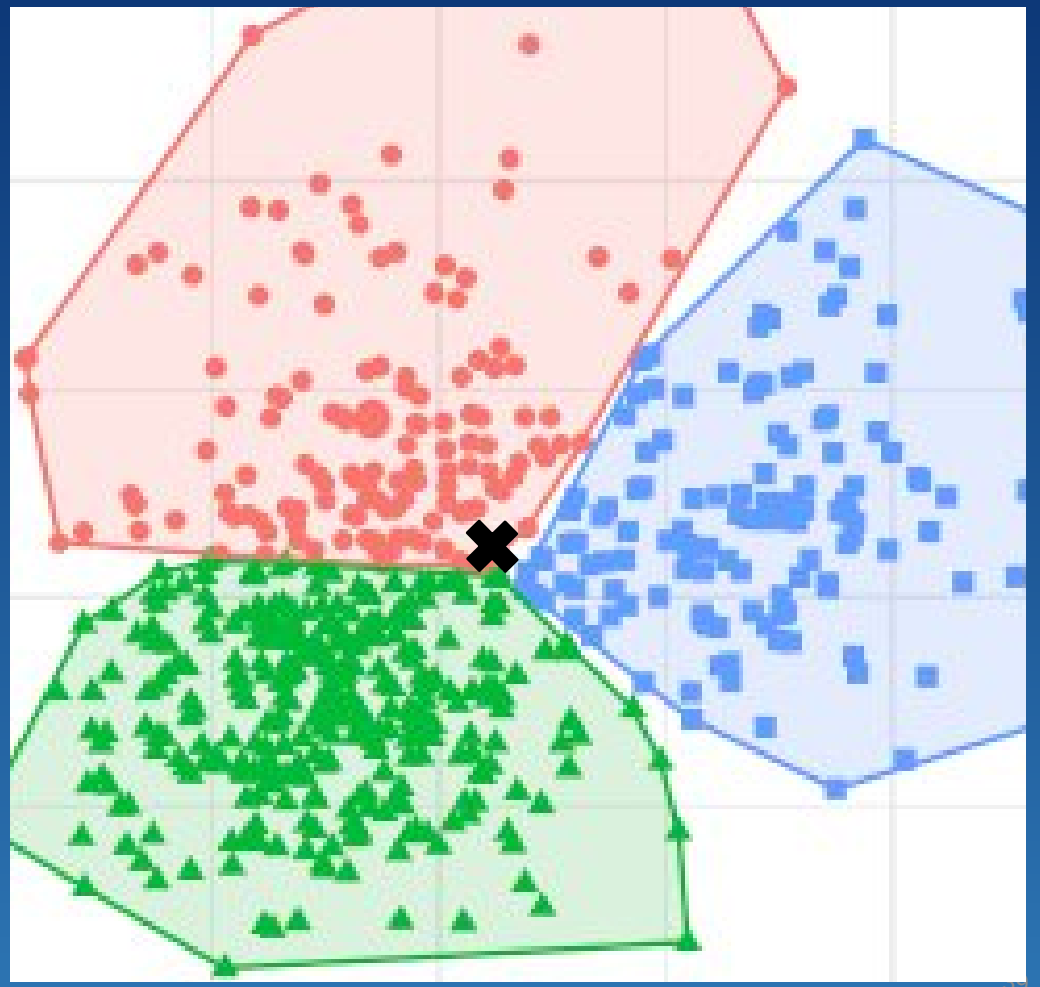


Acceleration Through K-Means Clustering

No risk of error

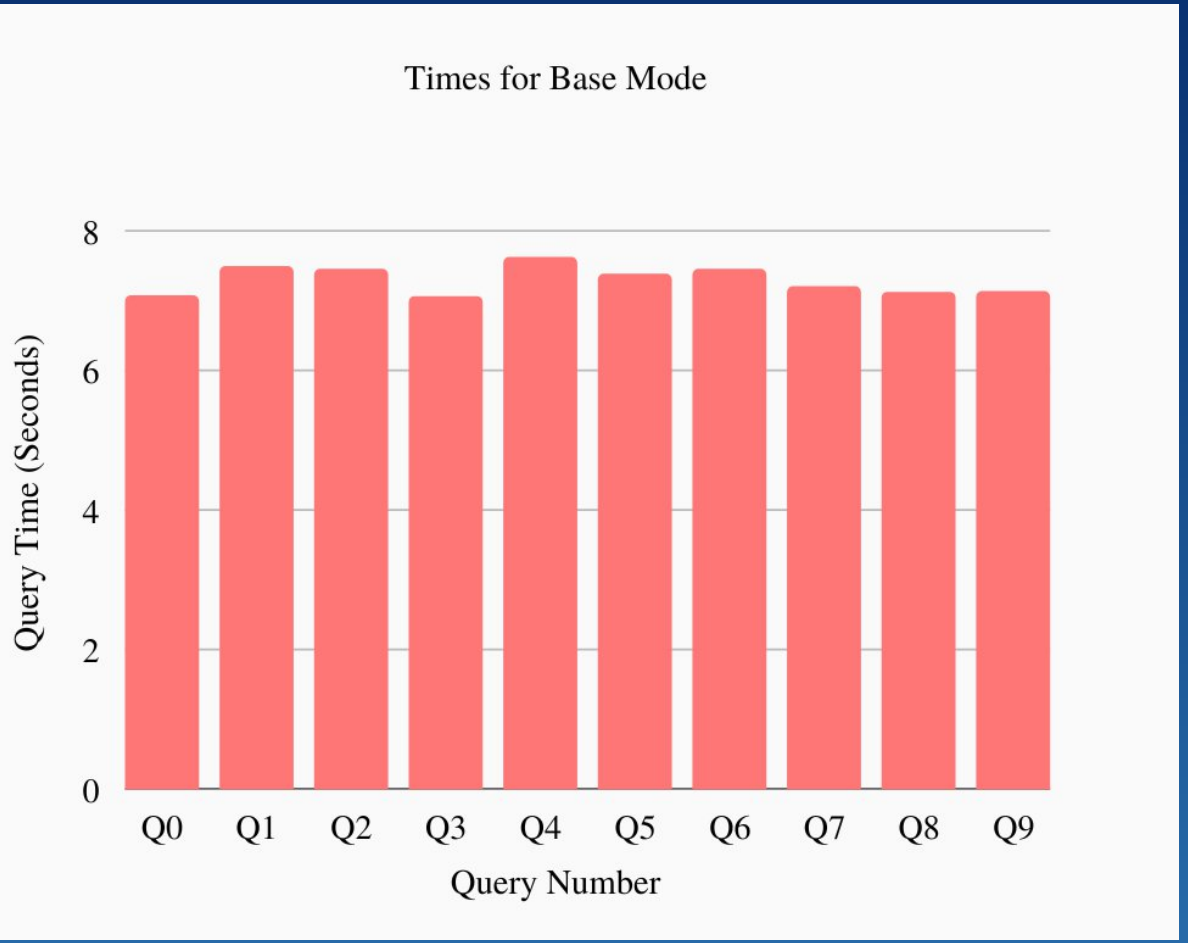


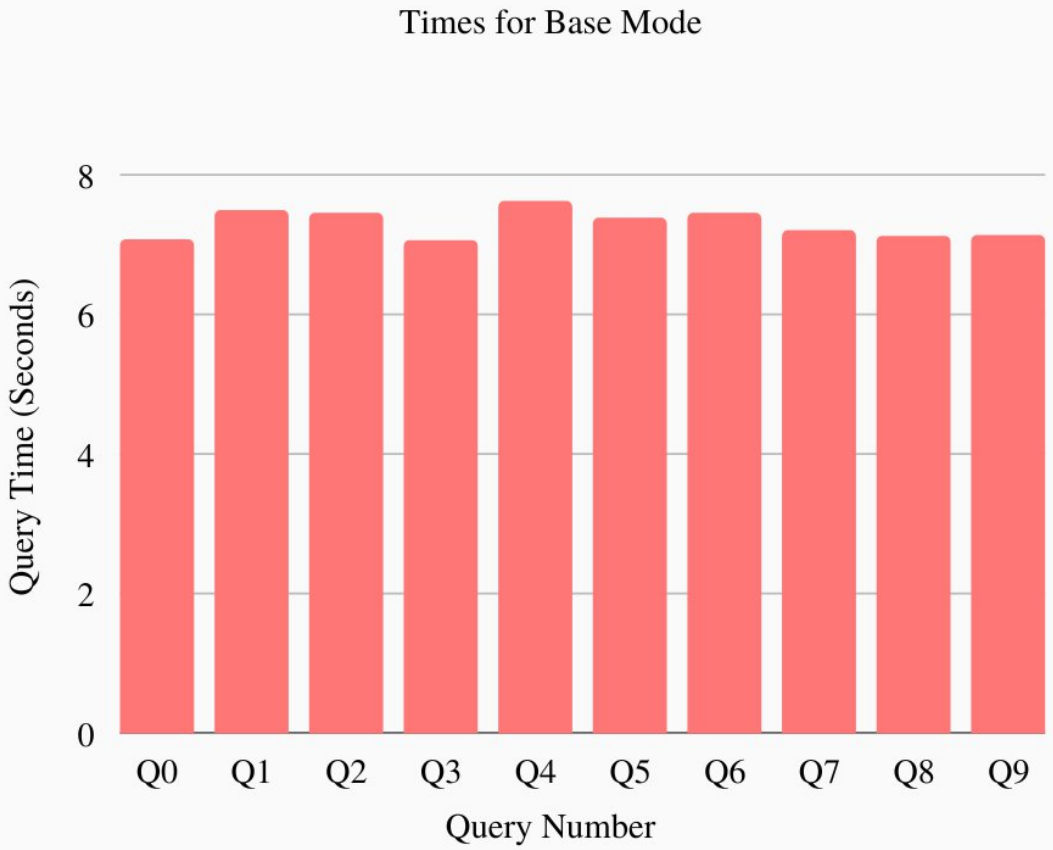
Risk of error



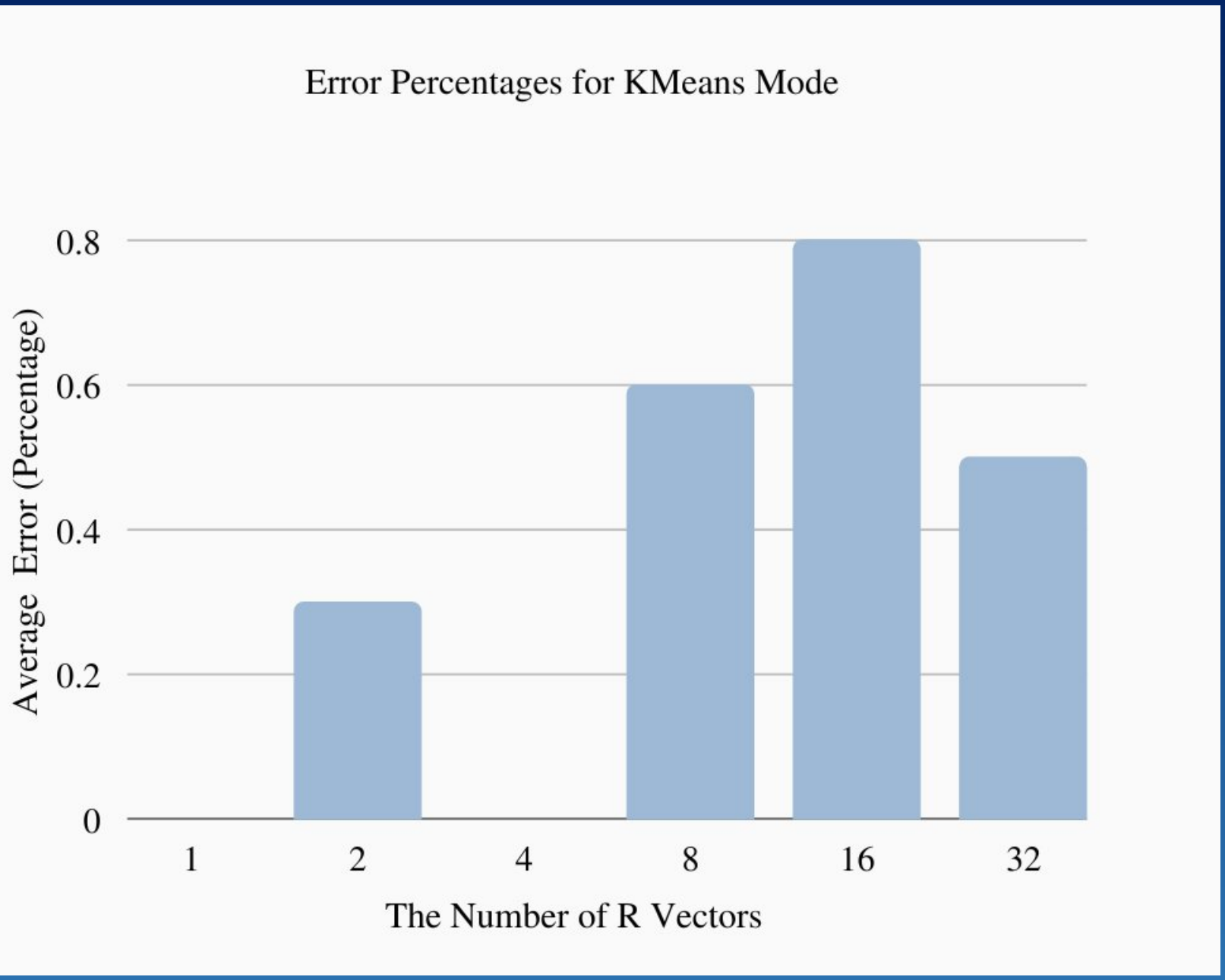
| Introduction & Background
| Data Processing
| Data Retrieval

| Results & Observations


Times for Base Mode
Query Time (Seconds)
8
6
4
2
0
Q0
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Query Number


Times for Base Mode
Query Time (Seconds)
8
6
4
2
0
Q0
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Query Number


The Number of R Vectors
1
2
4
8
16
32
Query Time (Seconds)
10
7.5
5
2.5
0
Q0
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Query Number


Error Percentages for KMeans Mode
Average Error (Percentage)
0.8
0.6
0.4
0.2
0
1
2
4
8
16
32
The Number of R Vectors

Results

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

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	0.842273	MansvelderVerhoogGoriounova2019.txt	5	One of the major goals in Neuroscience is to understand the neuronal basis of human learning and memory.
3	0.838349	Fodor1981.txt	1	Psychological explanations of behavior refer liberally to the mind and to states, operations and processes of the mind.
4	0.830323	NorenzayanHeine2005.txt	2	Cognitive science has relied heavily on the idea that the human mind is analogous to the computer (Block,1995).
5	0.83019	Chalmers2011.txt	2	The mathematical theory o f computation in the abstract is well -understood, but cognitive science and artificial intelligence

Query: Cognitive science is the study of the human mind and brain (K=8)

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	0.842273	MansvelderVerhoogGoriounova2019.txt	5	One of the major goals in Neuroscience is to understand the neuronal basis of human learning and memory.
3	0.838349	Fodor1981.txt	1	Psychological explanations of behavior refer liberally to the mind and to states, operations and processes of the mind.
4	0.826757	Ritter2004.txt	18	Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.
5	0.826301	SwodenSchyns2006.txt	8	Decoding the visual and subjective contents of the human brain.

Future Work

These are multiple next steps that can be taken to build on these findings:

- Implementing the ability to search and sort through a even larger mass of files
- Improving the accuracy of search results
- Attention Mechanism in image cap, More detailed image cap (How?)
- Format recognition in OCR

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
- Jim Davies, our mentor, for guiding us through this project.