Unsupervised and Unstructured Machine Learning

BA820 – Mohannad Elhamod

BOSTON UNIVERSITY

Boston University Questrom School of Business

Midterm Review

BOSTON UNIVERSITY

Boston University Questrom School of Business

Start Stop Continue

BOSTON UNIVERSITY

Boston University Questrom School of Business

Please fill this out

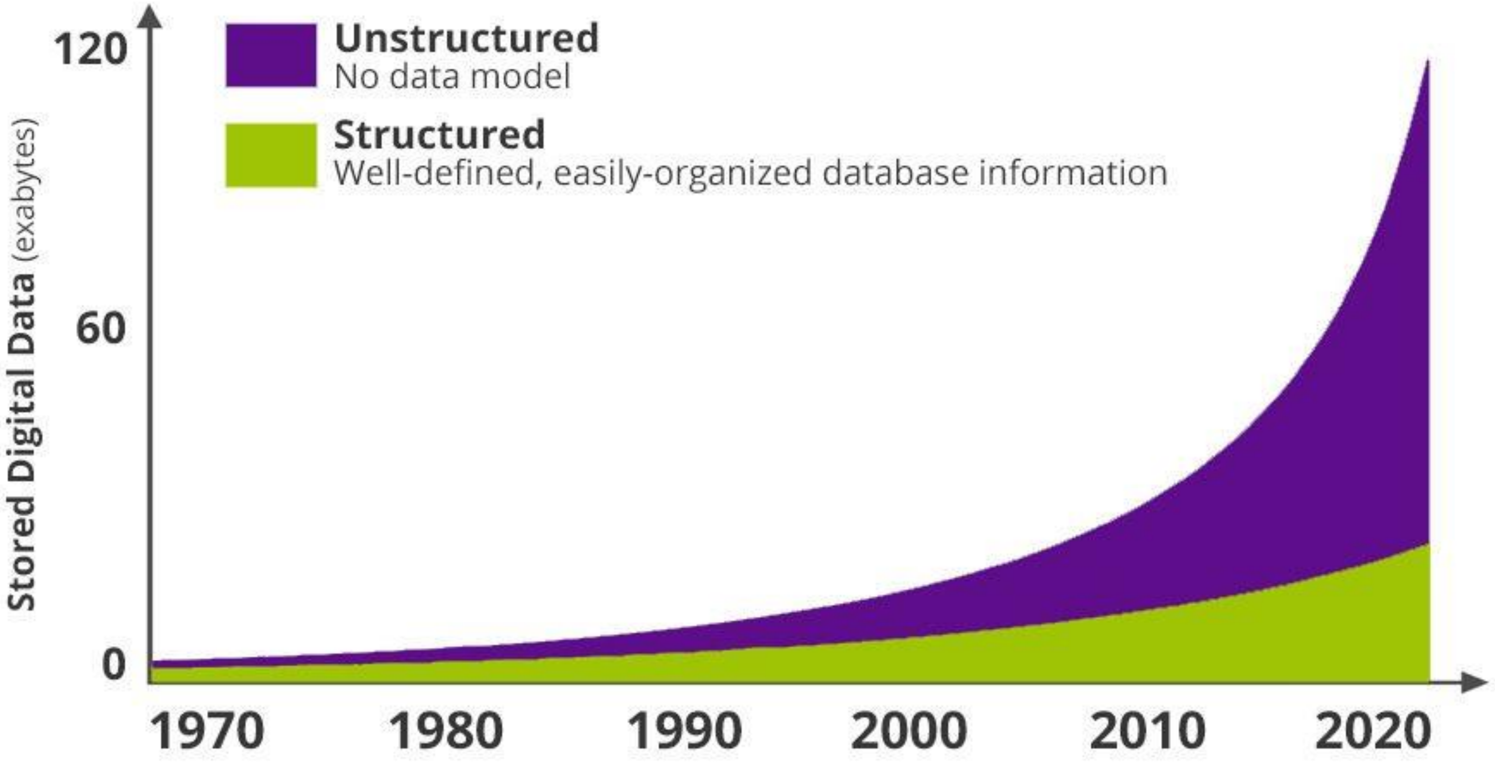
- https://forms.gle/VuHJUjtzLrrr1TNY8

Text Mining

BOSTON UNIVERSITY

Boston University Questrom School of Business

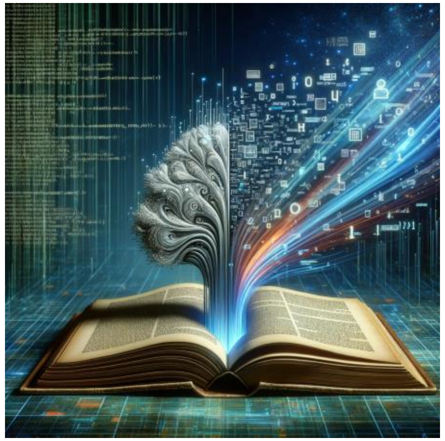
Most Data is No Longer Structured…



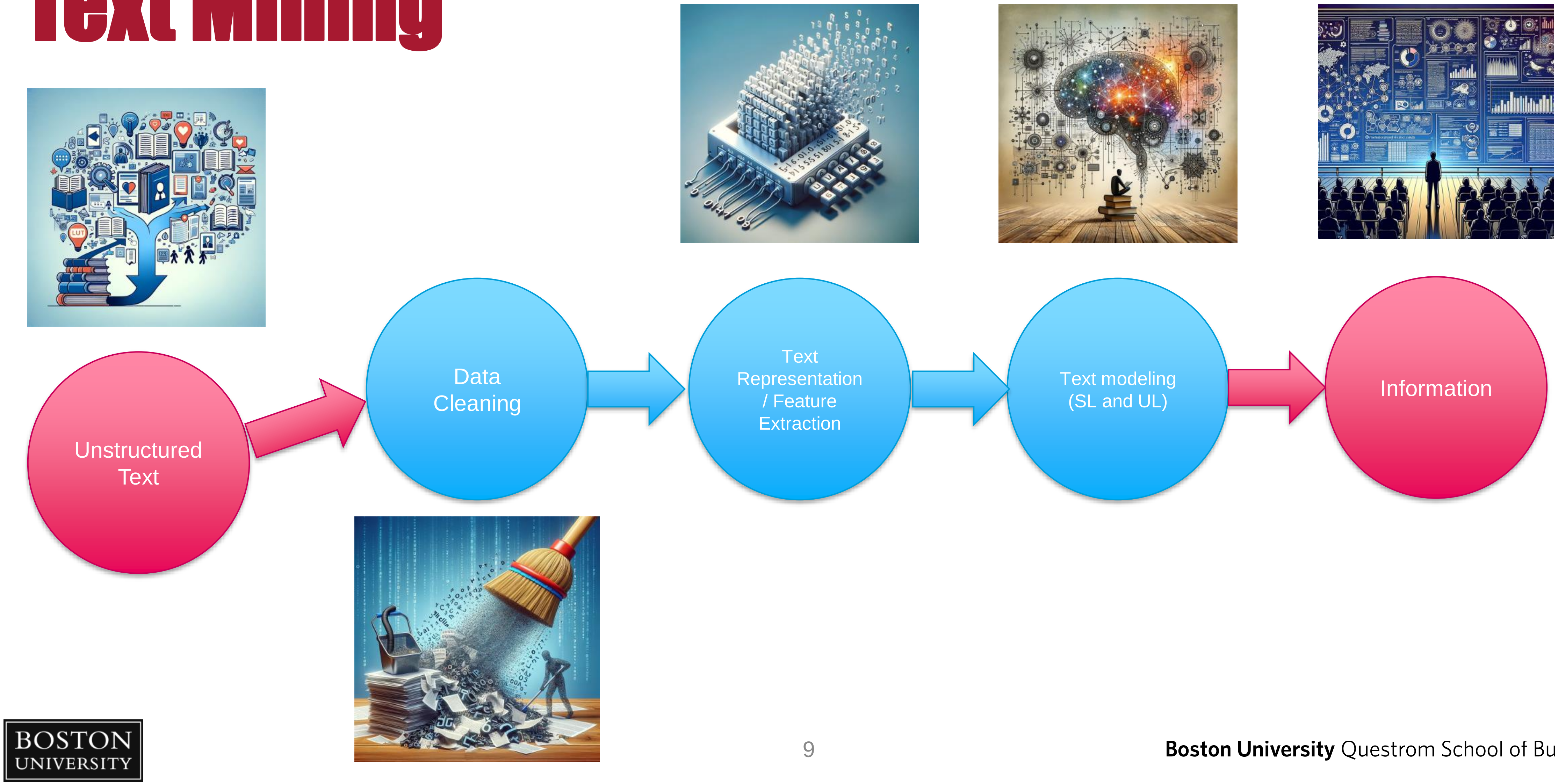
https://www.komprise.com/

What can we do with this data?

- Text classification
- Text generation
- Text summarization
- Music recommendation
- Image categorization
- ….



Text Mining



Unstructured Text

Data Cleaning

Text Representation / Feature Extraction

Text modeling (SL and UL)

Information

Collecting Unstructured Data

- How can we extract this data?
  - Datasets found publicly (UCI, Kaggle, Common Crawl, Wikimedia Downloads, etc.).
  - Using APIs (e.g., twitter API, News API).
  - Web scrapping (e.g., BeautifulSoup, Selenium)
  - Own private data.

Data Cleaning with *Regex*

- Just like structured/tabular data, we generally need to clean up the text to make it more useful.
  - Examples: case-sensitivity, punctuation, etc.

- *Regex* is used for finding string matches and formatting text:
  - Playground: https://www.w3schools.com/python/python_regex.asp
  - Cheat sheet: https://www.debuggex.com/cheatsheet/regex/python

How to Represent Text?

- Computers do not understand text…
- We need to represent text in a language they understand… Numbers!

- Simple proposals (we are just brainstorming):
  - Each sentence is represented in terms of the words it contains…
    - This is called ***Tokenization.***
  - Each word is represented by a number…
    - This is called ***Vectorization***.

Tokenization: Some Terminology

- ***Document:*** A body of text (e.g., a tweet, a pdf, an article, etc.).
- ***A token:*** The building block of a document.
  - Examples: character-level, word-level, …
- ***A separator:*** Special tokens that split a document into tokens.
  - Examples: punctuation, spaces,
- Demo

Need for Advanced Text Pre-Processing

- Simply extracting tokens does not preserve meaning/semantics.
- Some words occur too frequently in any text. These are called *stop words* and are generally removed.
- Issues:
  - ***Stemming:*** big, bigger, biggest
  - ***Lemmatization:*** drive drove driven
  - ***Homonyms:*** bank (river or money?)
  - ***Synonym:*** Yes, sure.
- We will come back to this later…

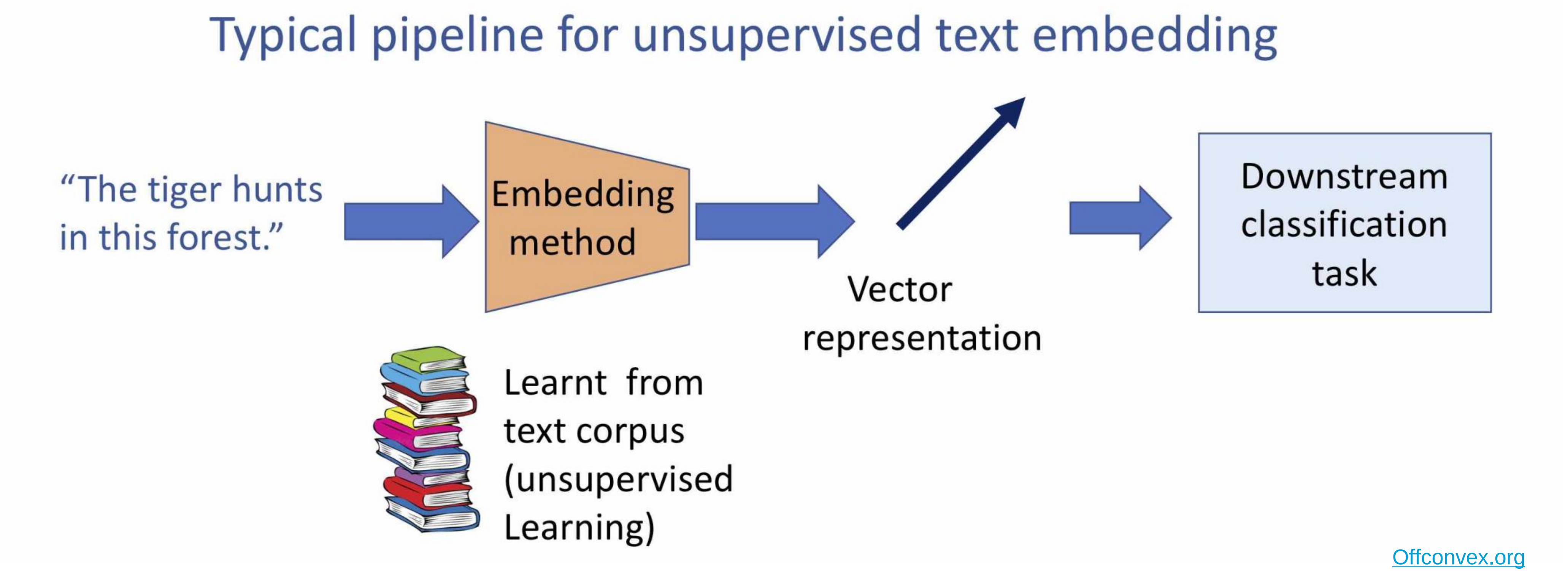
Text Modeling

- Once text data is in the proper representation (i.e., tokenized and vectorized), we can apply the methods we have learned so far:
    - Unsupervised ML (e.g., dimensionality reduction, clustering, etc.).
    - Supervised ML (e.g., classification, translation, etc.).
- **Demo**

Text Vectorization

BOSTON UNIVERSITY

Vectorization: Text as Numbers



Typical pipeline for unsupervised text embedding

“The tiger hunts in this forest.”

Embedding method

Vector representation

Downstream classification task

Learnt from text corpus (unsupervised Learning)

Offconvex.org

How Would You Represent a Document?

- Let’s start simple. Represent a document simply as *a collection of tokens*.
  - Vectorization by document (not token).
- This approach is called ***Bag of Words (BoW).***
- **Demo**

Sentence	hockey	fun	i	like	golf
I like golf!			1	1	1
I like hockey.	1		1	1	
Hockey and golf are fun!	1	1			1

Bag of Words (BoW)

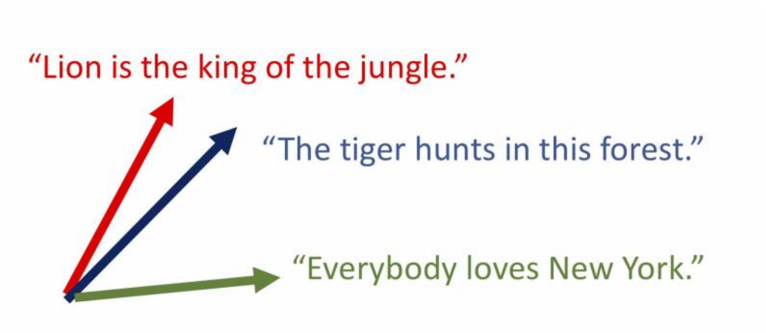
- Cons:
  - Disregards word order.
  - Number of features can be exhaustive.
  - Frequency bias.
    - (e.g., If the word “space” appears in a children’s book, it carries more significance than when it appears in an article about galaxies.
- Pros:
  - Simple to implement

	I	hate	love	golf	soccer
I hate golf and love soccer	1	1	1	1	1
I hate soccer and love golf	1	1	1	1	1

Document Similarity

- How do we measure if two documents are similar?
  - We need a metric like Euclidean distance.
  - But… What if documents have different lengths?

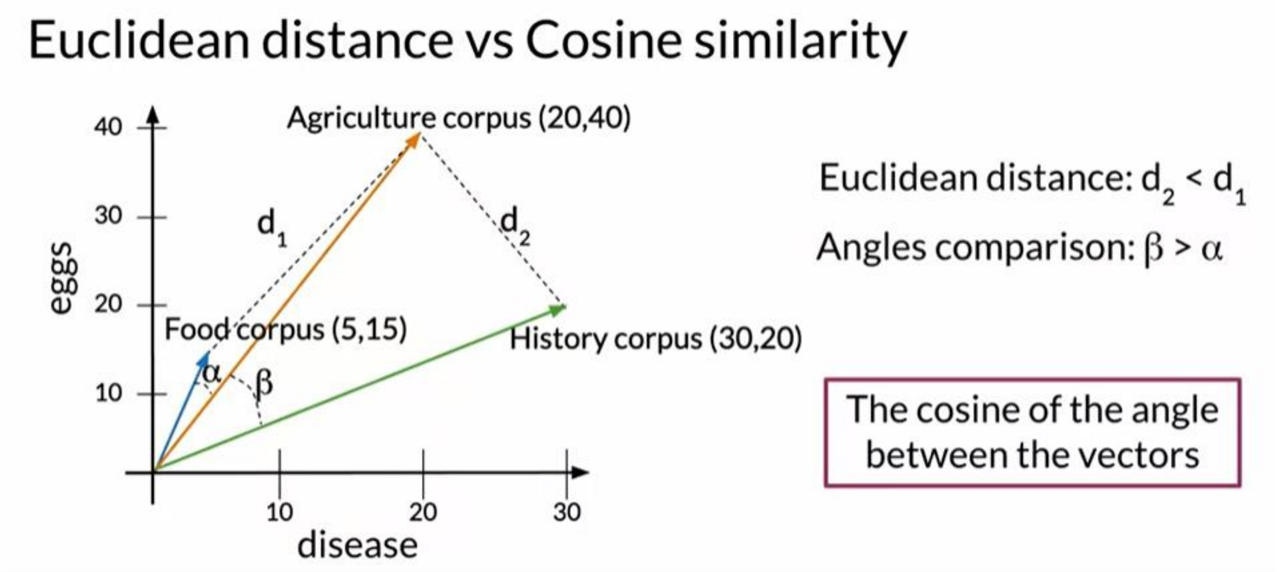
- We need a metric that is robust to differences in document size…
  - Enter *Cosine Similarity.*


"Lion is the king of the jungle."
"The tiger hunts in this forest."
"Everybody loves New York."

Offconvex.org

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

Document Similarity



Euclidean distance vs Cosine similarity

Euclidean distance: $d_2 < d_1$

Angles comparison: $\beta > \alpha$

The cosine of the angle between the vectors

medium.com

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