Recognizing Similar Texts

G.DIVYA SRI - 192211883
A.NEHA - 192225026

OBJECTIVE

Recognizing similar texts is a common Natural Language Processing (NLP) project that involves determining the degree of similarity between two or more text documents. Students working on this project can gain a variety of skills and insights

WHAT IS TEXT

- A book or other written or printed work regarded in terms of its content rather than its physical form.
- In the 1980s, the emergence of SMS (Short Message Service) revolutionized communication by introducing text messaging capabilities to cellular networks.
- The first-ever text message, however, was sent on December 3, 1992, by Neil Papworth, a 22-year-old engineer, to Richard Jarvis of Vodafone.
- Texting, also known as SMS (Short Message Service), refers to the exchange of short written messages between individuals using mobile devices or computers.

WHY TEXT SIMILARITY?

- Text similarity is used to discover the most similar texts. It used to discover similar documents such as finding documents on any search engine such as Google.
- It used to discover similar documents such as finding documents on any search engine such as Google. We can also use text similarity in document recommendations.
- Text similarity with FastText FastText is another excellent library for efficiently learning word representations and sentence classification. It can be used to find out how similar two pieces of text are by representing each piece of text as a vector and comparing the vectors using a similarity metric like cosine similarity.

EXAMPLES OF TEXT SIMILARITIES

- Text similarity with NLTK
- Text similarity with Scikit-Learn
- Text similarity with BERT
- Text similarity with RoBERTa
- Text similarity with FastText
- Text similarity with PyTorch

CLASSICAL LITERAT

CODE

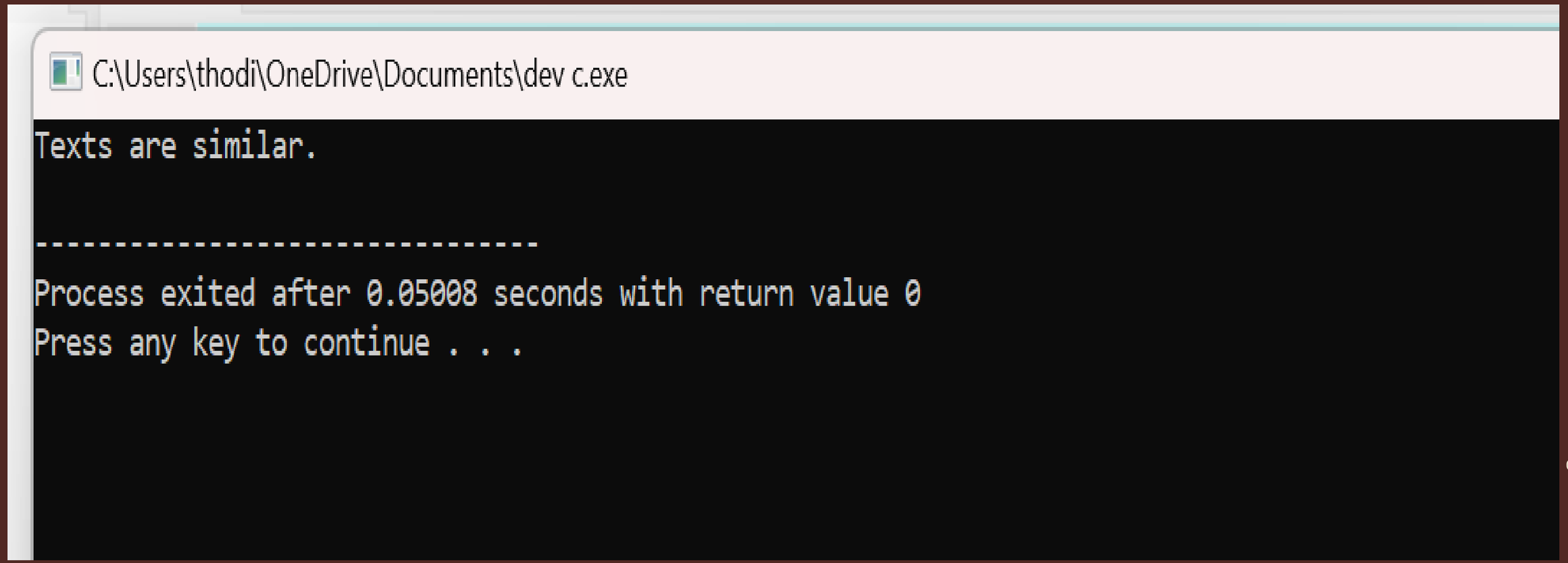
```
#include <stdio.h>
#include <string.h>
int min(int a, int b, int c) {
    return a < b ? (a < c ? a : c) : (b < c ? b : c);
}
int levenshtein_distance(const char *s1, const char *s2) {
    int len1 = strlen(s1);
    int len2 = strlen(s2);

    int dp[len1 + 1][len2 + 1];

    for (int i = 0; i <= len1; ++i)
        dp[i][0] = i;
    for (int j = 0; j <= len2; ++j)
        dp[0][j] = j;
```

```
 for (int i = 1; i <= len1; ++i) {
        for (int j = 1; j <= len2; ++j) {
            int cost = (s1[i - 1] == s2[j - 1]) ? 0 : 1;
            dp[i][j] = min(dp[i - 1][j] + 1,
                           dp[i][j - 1] + 1,
                           dp[i - 1][j - 1] + cost);
        }
    }

    return dp[len1][len2];
}

int are_similar(const char *s1, const char *s2, int threshold) {
    int distance = levenshtein_distance(s1, s2);
    int max_len = strlen(s1) > strlen(s2) ? strlen(s1) : strlen(s2);
```


*

```
    float similarity_ratio = 1.0 - (float)distance / max_len;
    return similarity_ratio >= (float)threshold / 100.0;
}

int main() {
    const char *text1 = "hello world!";
    const char *text2 = "Hello world!";

    int threshold = 80;

    if (are_similar(text1, text2, threshold)) {
        printf("Texts are similar.\n");
    } else {
        printf("Texts are not similar.\n");
```

```
    }

    return 0;
}
```


C:\Users\thodi\OneDrive\Documents\dev c.exe
Texts are similar.

Process exited after 0.05008 seconds with return value 0
Press any key to continue . . .

MEASURES OF TEXT SIMILARITY

- **JACCARD SIMILARITY**
- **COSINE SIMILARITY**
  - **COSINE SIMILARITY USING SPACY**
  - **COSINE SIMILARITY USING SCIPY**

“

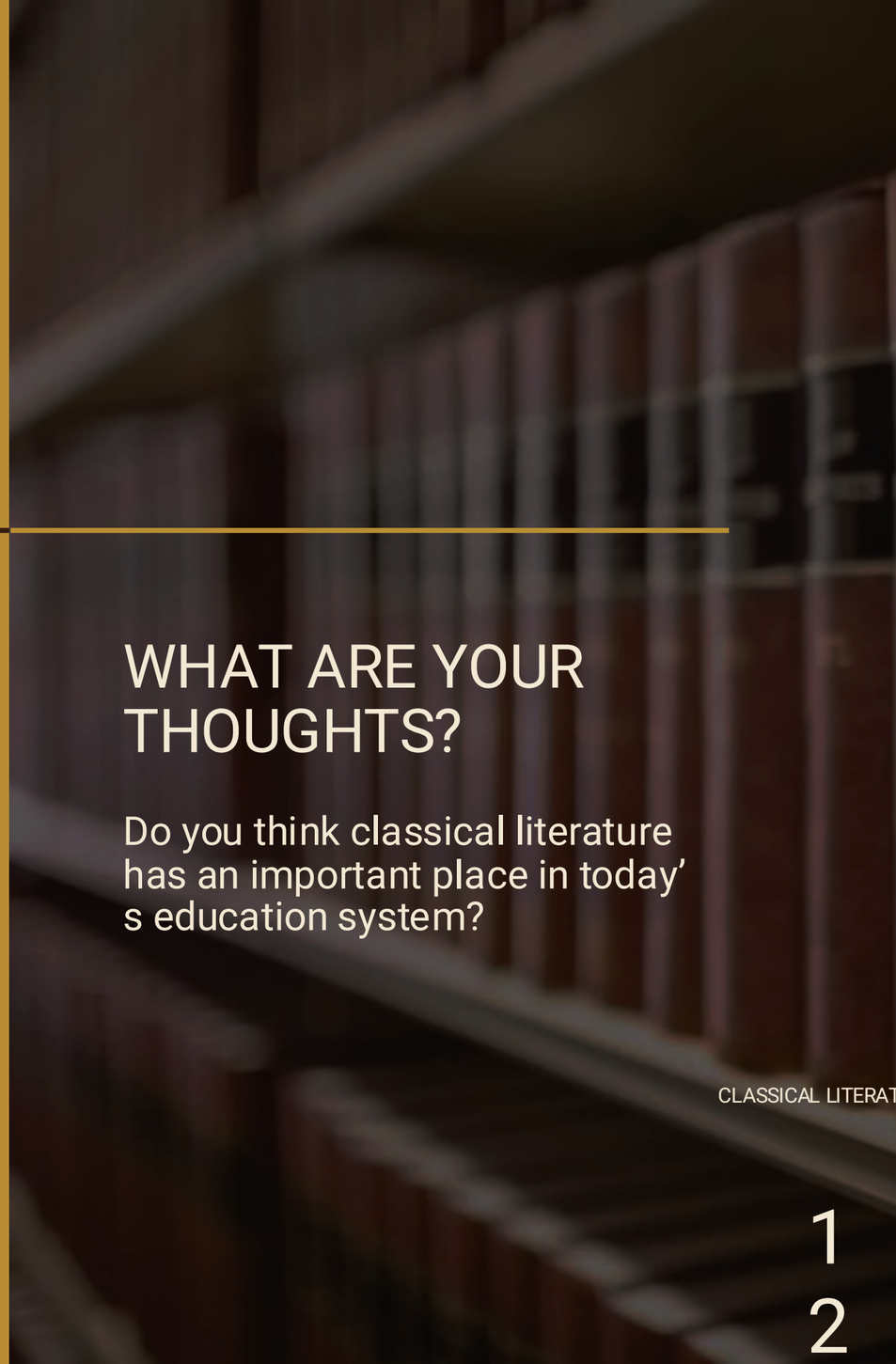
Texting is a brilliant way to miscommunicate how you feel, and misinterpret what other people mean

”

- HOMER, THE ODYSSEY

ANCIENT TEXT IN THE 21ST CENTURY

- Opinions vary on whether classical literature should be included in contemporary studies
- Some believe it is integral to helping students understand complex ideas
- Others believe more modern texts should be focused on due to their relevance



WHAT ARE YOUR THOUGHTS?

Do you think classical literature has an important place in today's education system?

OUTCOMES

- Text Similarity Metrics
- Feature Extraction:
- Vectorization Methods
- Similarity Metrics

SOURCES

WEBSITES

- www.contoso.com
- www.relecloud.com

TEXTS

- The Odyssey
- The Iliad
- Tarikh-i Beyhaqi
- Georgics
- Metamorphoses


Thank
you!!