BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM**

**BÁO CÁO MÔN HỌC**

**XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**ĐỀ TÀI:**

**TÓM TẮT VĂN BẢN**

Ngành: **CÔNG NGHỆ THÔNG TIN**

Giảng viên hướng dẫn: **TS. Lê Thị Ngọc Thơ**

Lớp: 23SCT21

Học viên thực hiện:

**Vũ Công Tiến** *MSHV: 2341862010*

# LỜI CAM ĐOAN

Tôi xin cam đoan đề tài: tóm tắt văn bản được nghiên cứu độc lập, dưới sự tìm hiểu theo cá nhân. Ngoài ra không có bất cứ sự sao chép của người khác. Đề tài, nội dung báo cáo thực tập là sản phẩm tôi đã nỗ lực nghiên cứu trong quá trình học tập. Các số liệu, kết quả trình bày trong báo cáo là hoàn toàn trung thực, được tham khảo từ nhiều nguồn uy tín, tôi xin chịu hoàn toàn trách nhiệm, kỷ luật của bộ môn và nhà trường đề ra nếu như có vấn đề xảy ra.

***Học viên thực hiện***

# NHẬN XÉT CỦA GIẢNG VIÊN

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

..................................................................................................................................................

.......................................................................

Điểm

TP.HCM, ngày…tháng…năm 2024

Giảng viên hướng dẫn

**TS. Lê Thị Ngọc Thơ**

MỤC LỤC

[LỜI CAM ĐOAN 2](#_Toc157374482)

[NHẬN XÉT CỦA GIẢNG VIÊN 3](#_Toc157374483)

[DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT 2](#_Toc157374484)

[CHƯƠNG 1: TỔNG QUAN 3](#_Toc157374485)

[1.1 Mục tiêu của đề tài: 3](#_Toc157374486)

[1.2 Dữ liệu đầu vào: 3](#_Toc157374487)

[1.3 Dữ liệu đầu ra: 3](#_Toc157374488)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 3](#_Toc157374489)

[2.1 Một số khái niệm: 3](#_Toc157374490)

[2.2 Phương pháp tiếp cận bài toán: 8](#_Toc157374491)

[CHƯƠNG 3 THỰC NGHIỆM 9](#_Toc157374492)

[3.1 Dữ liệu đầu vào: 9](#_Toc157374493)

[3.2 Kết quả: 10](#_Toc157374494)

[CHƯƠNG 3: KẾT LUẬN 11](#_Toc157374495)

# DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| **TF** | Term Frequency |
| **IDF** | Inverse Document Frequency |
| **TF-IDF** | Term Frequency – inverse Document Frequency |
| **NLP** | Natural Language Processing |
|  |  |
|  |  |

# CHƯƠNG 1: TỔNG QUAN

## Mục tiêu của đề tài:

Trong quá trình học tập, nghiên cứu các chủ đề của môn xử lý ngôn ngữ tự nhiên, được sự hướng dẫn tận tình của cô Lê Thị Ngọc Thơ. Tôi lựa chọn đề tài: “Tóm tắt văn bản” Bài toán đặt ra yêu cầu phải hiểu được cách tính độ tương đồng ngữ nghĩa của các câu trong văn bản, xác định được độ quan trọng của câu và sau cùng là rút trích các câu để tạo ra một bản văn bản tóm tắt.

## Dữ liệu đầu vào:

Là tập tin nguồn: một tập hợp các câu sẽ được tóm tắt nội dung. Định dạng của một câu như sau:

*<s docid="AP880915-0003" num="45" wdcount="17"> Earlier Wednesday Gilbert was classified as a Category 5 storm, the strongest ad deadliest type of hurricane.</s>*

Trong đó:

* docid: là tên tài liệu được rút trích.
* Num: vị trí của câu trong docid.
* Wdcount: số từ trong câu
* *Earlier Wednesday Gilbert was classified as a Category 5 storm, the strongest ad deadliest type of hurricane:* Nội dung của câu

Dữ liệu đầu vào để test: <https://drive.google.com/drive/folders/1ZCbIjio89hQ5sWb8MKQ07Ne5zogd68hb>

Là Tập tin chứa các stopword.

## Dữ liệu đầu ra:

Là tập tin chứa nội dung tóm tắt từ dữ liệu đầu vào. Tổng số câu trong output file là 10% valid input.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Một số khái niệm:

* TF-IDF: **là một thống kê số dùng trong truy xuất thông tin và khai thác văn bản để đánh giá sự quan trọng của một từ trong một tài liệu so với một bộ sưu tập các tài liệu (corpus). TF-IDF thường được sử dụng để phân tích văn bản, phân loại tài liệu và truy xuất thông tin**. **Các thành phần:**
  + Tần suất từ (TF): Đo lường mức độ một từ (từ) xuất hiện trong một tài liệu. Nó được tính là tỷ lệ giữa số lần một từ xuất hiện trong một tài liệu và tổng số từ trong tài liệu đó. Ý tưởng là một từ xuất hiện càng nhiều lần trong một tài liệu, nó càng quan trọng đối với tài liệu đó.



* + Nghịch đảo tần suất tài liệu (IDF): Đo lường sự quan trọng của một từ trên toàn bộ corpus. Các từ phổ biến trong nhiều tài liệu có giá trị IDF thấp, trong khi các từ hiếm có giá trị IDF cao. Công thức IDF là logarithm cơ số 2 của tỷ lệ giữa tổng số tài liệu và số tài liệu chứa từ đó.



* + TF-IDF: Điểm số TF-IDF cho một từ trong một tài liệu là tích của Tần suất từ và Nghịch đảo tần suất tài liệu. Điểm số này phản ánh cả hai yếu tố là tần suất xuất hiện của từ trong tài liệu và tính độc đáo của nó trong toàn bộ corpus.

**TF-IDF(*t*,*d*,*D*)=TF(*t*,*d*)×IDF(*t*,*D*)**

* + Điểm số TF-IDF càng cao cho một từ trong một tài liệu, thì từ đó càng quan trọng đối với tài liệu đó. TF-IDF thường được sử dụng để xếp hạng các từ trong một tài liệu dựa trên sự liên quan của chúng đối với tài liệu và thường được áp dụng trong nhiều ứng dụng xử lý ngôn ngữ tự nhiên (NLP), bao gồm tóm tắt văn bản, phân nhóm tài liệu và truy xuất thông tin.
* Vector TF-IDF:
  + là một biểu diễn số học của một tài liệu trong không gian vector dựa trên các giá trị TF-IDF của các từ trong tài liệu đó. Mỗi thành phần của vector đại diện cho một từ cụ thể và chứa thông tin về tần suất xuất hiện của từ đó trong tài liệu cũng như độ quan trọng của từ đó trong toàn bộ corpus
  + Cách xây dựng vector TF-IDF:
    - Xây dựng từ điển: Tạo ra một danh sách các từ duy nhất xuất hiện trong toàn bộ bộ sưu tập tài liệu (corpus).
    - Tính toán TF-IDF cho mỗi từ trong mỗi tài liệu: Đối với mỗi từ trong từ điển, tính giá trị TF-IDF tương ứng với tài liệu đang xét.
    - Tạo vector TF-IDF: Cho mỗi tài liệu, xây dựng một vector trong không gian n chiều, với n là số lượng từ trong từ điển. Mỗi thành phần của vector là giá trị TF-IDF của từ tương ứng.
* Ma trận trọng số:
  + Ma trận trọng số thường ánh xạ từng tài liệu trong một bộ sưu tập thành một vector TF-IDF tương ứng và tổng hợp tất cả các vector này thành một ma trận. Mỗi hàng của ma trận thường đại diện cho một tài liệu và mỗi cột đại diện cho một từ trong từ điển.
    - Ví dụ, giả sử chúng ta có bộ sưu tập văn bản với 3 tài liệu và một từ điển gồm các từ ["cat", "dog", "fish"]. Ta tính giá trị TF-IDF cho mỗi từ trong mỗi tài liệu và xây dựng ma trận trọng số. Dưới đây là một ví dụ giả định về giá trị TF-IDF:
    - Tài liệu 1:
      * cat: 0.2
      * dog: 0.4
      * fish: 0.1
    - Tài liệu 2:
      * cat: 0.1
      * dog: 0.3
      * fish: 0.2
    - Tài liệu 3:
      * cat: 0.3
      * dog: 0.2
      * fish: 0.1
    - Khi biểu diễn dưới dạng ma trận trọng số, mỗi hàng của ma trận sẽ là vector TF-IDF tương ứng với mỗi tài liệu, và mỗi cột sẽ là giá trị TF-IDF của mỗi từ trong từ điển:

Tài liệu | cat | dog | fish

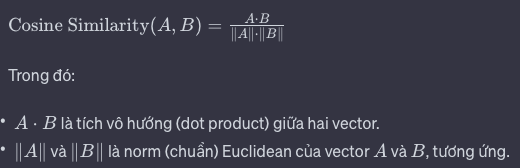
--------------------------------

1 | 0.2 | 0.4 | 0.1

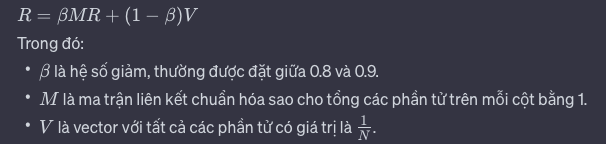
2 | 0.1 | 0.3 | 0.2

3 | 0.3 | 0.2 | 0.1

* + - Trong ma trận trọng số này, mỗi hàng biểu diễn một tài liệu, mỗi cột biểu diễn một từ trong từ điển, và mỗi phần tử là giá trị TF-IDF của từ tương ứng trong tài liệu đó. Ma trận trọng số thường được sử dụng trong các nhiệm vụ như phân loại văn bản, phân nhóm, và tìm kiếm thông tin trong các ứng dụng xử lý ngôn ngữ tự nhiên
* Tương đồng Cosine
  + là một phép đo độ tương đồng giữa hai vector trong không gian vector bằng cách sử dụng cosin của góc giữa chúng. Trong ngữ cảnh của xử lý ngôn ngữ tự nhiên (NLP) và vector TF-IDF, tương đồng Cosine thường được sử dụng để đánh giá sự tương đồng giữa các tài liệu hoặc từ vựng.
    - Cho 2 vector A và B, Tương đồng Cosine được tính bằng công thức:



* + - Giá trị của tương đồng Cosine nằm trong khoảng [−1,1], với 1 thể hiện hai vector hoàn toàn giống nhau, −1 thể hiện hai vector hoàn toàn ngược nhau, và 0 thể hiện sự không tương đồng
* Ma trận kề (Adjacency Matrix):
  + được sử dụng trong đồ thị vô hướng, trong đó mối quan hệ giữa các đỉnh là đối xứng. Nếu đỉnh i kề với đỉnh j, thì ma trận tại vị trí (i, j) và (j, i) sẽ có giá trị là 1, ngược lại là 0. Nếu đồ thị có hướng, ma trận kề không còn đối xứng và có thể chứa các giá trị 1 hoặc 0 tùy thuộc vào có hoặc không có cạnh nối giữa các đỉnh.
* Thuật toán Pagerank
  + là một thuật toán được phát triển bởi Larry Page và Sergey Brin tại Đại học Stanford và được sử dụng chủ yếu bởi Google để xếp hạng các trang web trong kết quả tìm kiếm. Nó là một phương pháp đánh giá tầm quan trọng của các trang web dựa trên mối liên kết giữa chúng.
  + Cách hoạt động:
    - Mô hình đồ thị: Các trang web được mô hình hóa thành đồ thị, trong đó mỗi trang web là một đỉnh, và có các liên kết giữa các đỉnh đại diện cho các liên kết giữa các trang web
      * Trong bài toán tóm tắt tài liệu, mỗi câu trong tài liệu là một đỉnh, và độ tương đồng Cosine giữa (i,j) > 0,5 là một kết nối vô hướng (cạnh).
    - Ma trận liên kết: PageRank sử dụng ma trận liên kết, trong đó mỗi phần tử Aij của ma trận biểu diễn liên kết từ câu i đến câu j. Nếu câu i có liên kết đến câu j thì có gía trị 1, ngược lại là 0.
    - Vector trạng thái ban đầu: một vector trạng thái ban đầu R0 được tạo, trong đó, mỗi phần tử đại diện cho 1 câu và có giá trị bắt đầu là 1/N, với N là số lượng câu trong tập tài liệu.
    - Cập nhật trạng thái: Vector trạng thái R được cập nhật theo công thức:



* + - Lặp lại quá trình: Quá trình cập nhật trạng thái được lặp lại cho đến khi vector trạng thái hội tụ hoặc một số vòng lặp tối đa được đạt đến
    - Kết quả PageRank: Các trang web được xếp hạng theo giá trị của phần tử trong vector trạng thái cuối cùng.
  + PageRank giúp đánh giá tầm quan trọng của một câu dựa trên số lượng và chất lượng của các liên kết đến câu đó. Các câu với nhiều liên kết từ các câu quan trọng sẽ có xếp hạng cao hơn.
* Tóm tắt dựa trên trung tâm trung tâm Radev
  + Radev (Rada Mihalcea và Paul Tarau, 2004) là một phương pháp tóm tắt dựa trên trung tâm (center-based summarization) được thiết kế để tạo ra các tóm tắt ngắn gọn từ các văn bản dài. Dưới đây là một tóm tắt về phương pháp này
    - Xác định từ khoá:
      * Sử dụng mô hình ngôn ngữ và đồng tham chiếu để xác định các từ khóa quan trọng trong văn bản
      * Ưu tiên các từ khóa xuất hiện nhiều và có mối liên kết với các từ khác
    - Xác định trung tâm:
      * Sử dụng các đặc trưng như từ khóa và vị trí câu để xác định các trung tâm trong văn bản.
      * Các trung tâm thường là các câu chứa nhiều từ khóa và có mối liên kết mạnh mẽ với các câu khác
    - Phân loại câu:
      * Phân loại các câu thành các nhóm dựa trên đặc trưng như độ dài và số lượng từ khóa
      * Ưu tiên việc chọn các câu thuộc nhóm quan trọng.
    - Tạo tóm tắt:
      * Kết hợp các câu thuộc các nhóm quan trọng để tạo thành tóm tắt cuối cùng.
      * Đồng thời giữ cho tóm tắt ngắn gọn và bao quát.
    - Đánh giá và tối ưu
      * Đánh giá hiệu suất của tóm tắt bằng cách sử dụng các độ đo như ROUGE.
      * Tối ưu hóa các tham số của mô hình để cải thiện chất lượng tóm tắt.

## Phương pháp tiếp cận bài toán:

* Ý tưởng chính của phương pháp tiếp cận: Tóm tắt văn bản dựa trên đồ thị, với thuật toán pagarank để rút trích 10% các câu quan trọng nhất từ văn bản đầu vào.
* Chi tiết thực hiện:
  + Tiền xử lý dữ liệu:
    - Đọc nội dung các file văn bản đầu, xử lý theo từng dòng dữ liệu:
      * Sử dụng định dạng như bên dưới để tách dòng dữ liệu thành những thuộc tính:
        + DocumentID
        + Số từ trong câu
        + Nội dung câu

<s docid=\"%100[^\"]\" num=\"%d\" wdcount=\"%d\">%[^\n]s</s>

* + - Từ nội dung của câu, xử lý theo từng bước sau
      * Loại bỏ các câu có số từ nhỏ thua 5.
      * Loại bỏ các stop word trong câu.
      * Chuyển đổi câu thành vector các từ dựa trên ký tự tách chuỗi: (.,:;!?`)
    - Kết quả của quá trình tiền xử lý dữ liệu là một ma trận W chứa các word đã được xử lý từ tập tài liệu.
  + Tạo ma trận trọng số TF-IDF từ ma trận W:
    - Bước 1: tạo túi từ/từ điển (Bag of word) từ ma trận W.
    - Bước 2: tính TF-IDF của mỗi từ trong câu dựa trên túi từ.
    - Bước 3: lưu tất cả các giá trị ở bước 2 vào ma trận trọng số TF-IDF.
  + Tạo ma trận tương đồng Cosine từ ma trận trọng số TF-IDF:
    - Xét câu thứ i từ ma trận TF-IDF
      * Với mỗi cặp vector [i,j], ta tính được giá trị tương đồng Cosine cho vector [i].
    - Lặp cho đển khi kết thúc, ta có ma trận vuông Cosine [i,i].
  + Tạo ma trận liên kết từ ma trận Cosine:
    - Xét tất cả các giá trị [i,j] trong ma trận Cosine.
      * Với mỗi giá trị [i,j] > 0.5 => đỉnh i và j có liên kết, và giá trị [i,j] của ma trận liên kết là 1.
      * Ngược lại, giá trị [i,j] của ma trận liên kết là 0.
    - Lặp cho đến khi kết thúc, ta có ma trận liên kết L (đồ thị vô hướng).
  + Tính xếp hạng PageRank:
    - Chạy thuât toán PageRank với các thông số:
      * Số lần lặp tốt đa: 100.
      * Giá trị epsilon = 0.00000000056
        + Là giá trị để so sánh kết quả 2 lần chạy PageRank liên tiếp, nếu hiệu 2 lần chạy bé hơn espilon, ta có thể kết luận pageRank hội tụ và dừng thuật toán.
      * Damping factor: 0.85
        + Ngưỡng quyết định liên kết khi sử dụng PageRank để xác định mức độ quan trọng của các câu trong một đồ thị có thể thay đổi tùy thuộc vào yêu cầu cụ thể và đặc tính của đồ thị đó. PageRank sử dụng một hệ số damping (thường được đặt giá trị là khoảng 0.85) để ổn định thuật toán và tránh hiện tượng "dead ends" (các câu không có liên kết ra) hoặc "spider traps" (các chuỗi liên kết mà không có liên kết ra).
        + Nếu muốn giữ lại một số câu quan trọng nhất, bạn có thể chọn một ngưỡng thấp hơn. Ngược lại, nếu bạn muốn giữ lại ít câu hơn nhưng có ảnh hưởng lớn, bạn có thể chọn ngưỡng cao hơn.
    - Sau khi thuật toán dừng, ta thu được một vector chứa các điểm PagaRank cho mỗi câu.
  + Xuất kết quả ra file:
    - Dựa trên vector PageRankScore và ma trận W, ta tính được vector map (Vector result) giữa nội dung câu và giá trị PagRankScore tương ứng.
    - Sắp xếp vector result theo thứ tự giảm dần.
    - Tính số lượng câu sẽ lưu ra file (10%), và lưu ra file.

# CHƯƠNG 3 THỰC NGHIỆM

## Dữ liệu đầu vào:

* Dữ liệu file test:

<s docid="AP900103-0077" num="8" wdcount="26"> Former East German leader Erich Honecker may be moved to a monastery </s>

<s docid="AP900103-0077" num="8" wdcount="26"> test Former East German leader Erich Honecker may beastery </s>

<s docid="AP900103-0077" num="8" wdcount="26"> test Former East German leader Erich Honecker may be moved to a monastery </s>

<s docid="AP900103-0077" num="8" wdcount="26"> test Former Erich Honecker may be moved to a monastery </s>

<s docid="AP900103-0077" num="8" wdcount="26"> test Former East German leader Erichmoved to a monastery </s>

<s docid="AP900103-0077" num="8" wdcount="26"> test Former East German leader Erich Honecker </s>

<s docid="AP900129-0071" num="12" wdcount="8"> All four have been expelled from the party.</s>

<s docid="AP900129-0071" num="13" wdcount="16"> Mielke was Honecker's security chief, Mittag oversaw the economy and Herrmann headed the party's propaganda department.</s>

<s docid="AP900129-0071" num="14" wdcount="12"> Eleven members of Honecker's ousted Politburo already are in prison awaiting trial.</s>

<s docid="AP900129-0071" num="15" wdcount="32"> Honecker's lawyer, Wolfgang Vogel, asked the chief prosecutor to free the former leader from Rummelburg prison based on a doctor's statement that he is too ill to stay in jail, ADN said.</s

## Kết quả:

* Đồ thị (ma trận liên kết):

Print matrix[10][10]

1 1 1 1 0 1 0 0 0 0

1 1 1 1 0 1 0 0 0 0

1 1 1 1 0 1 0 0 0 0

1 1 1 1 0 1 0 0 0 0

0 0 0 0 1 1 0 0 0 0

1 1 1 1 1 1 0 0 0 0

0 0 0 0 0 0 1 0 0 0

0 0 0 0 0 0 0 1 0 0

0 0 0 0 0 0 0 0 1 0

0 0 0 0 0 0 0 0 0 1

PageRank stop at: 22

* Xếp hạng từ trong đồ thị theo mức độ quan trọng (factor = 0.3):

[PageRanks Score=0.0824321][DocID=AP900103-0077\_8]

[PageRanks Score=0.0824321][DocID=AP900103-0077\_8]

[PageRanks Score=0.0824321][DocID=AP900103-0077\_8]

[PageRanks Score=0.0824321][DocID=AP900103-0077\_8]

[PageRanks Score=0.0747838][DocID=AP900103-0077\_8]

[PageRanks Score=0.0846756][DocID=AP900103-0077\_8]

[PageRanks Score=0.0721649][DocID=AP900129-0071\_12]

[PageRanks Score=0.0721649][DocID=AP900129-0071\_13]

[PageRanks Score=0.0721649][DocID=AP900129-0071\_14]

[PageRanks Score=0.0721649][DocID=AP900129-0071\_15]

* Tóm tắt được văn bản:
  + Văn bản tóm tắt không quy định % số câu output

[PageRankScore = 0.0824321][docID=AP900103-0077\_8]former east german leader erich honecker may moved monastery

[PageRankScore = 0.0721649][docID=AP900129-0071\_15]honecker's lawyer wolfgang vogel asked chief prosecutor free former leader rummelburg prison based doctor's statement ill stay jail adn said

[PageRankScore = 0.0721649][docID=AP900129-0071\_14]eleven members honecker's ousted politburo already prison awaiting trial

* + Văn bản tóm tắt theo chương 2:

[PageRankScore = 0.0824321][docID=AP900103-0077\_8]former east german leader erich honecker may moved monastery

# CHƯƠNG 3: KẾT LUẬN

Như vậy để đánh giá độ chính xác của thuật toán PagaRank trong tóm tắt văn bản, ta cần xem xét hiều yếu tố khi thực hiện chạy thuật toán:

* Giá trị của damping factor: việc lựa chọn giá trị này ảnh hưởng đến độ chính xác của thuật toán.
* Cấu trúc đồ thị: nếu đồ thị có nhiều đỉnh không kết nối, và giá trị damping factor đủ thấp, số lần lặp để tìm ra điểm hội tụ sẽ nhiều hơn, hoặc có thể không tìm được điểm hội tụ.

Cải tiến thuật toán:

* Đồ thị: {0,1} là 2 giá trị được sử dụng trong đồ thị và chỉ nói lên rằng đỉnh i,j có kết nối với nhau, không phân biệt độ quan trọng của nó, nên nếu chúng ta sử dụng đồ thị có trọng số thì thiết nghĩ kết quả sẽ khả quan hơn rất nhiều.
* Kết hợp thuật toán pageRank và độ tương đồng Cosine đến trung tâm của đỉnh:
  + Tìm vector trung tâm của đỉnh.
  + Tính toán độ tương đồng Cosine từ các vector trọng số TF-IDF đến trung tâm của đỉnh, sau đó kết hợp với giá trị PageRank Score theo công thức:
    - Value = alpha \* CosineVal + (1 – alpha) \* PageRankScore[i]
  + Sắp xếp lại giá trị và chọn ra những giá trị cao nhất là kết quả của bài toán tóm tắt văn bản.

Quá trình tiền xử lý:

* Khi tách từ, mỗi từ sẽ tương úng với một chiều của vector, có n từ thì sẽ có vector n chiều, dẫn đến tốn tài nguyên cho việc lưu trữ và tính toán.
* Nếu có thể thì nên tối ưu hoá số chiều của vector trước khi thực hiện xử lý tính toán.