

# ThaiQCor 2.0

Thai Query Correction via Soundex and  
Word Approximation

# Outline

- **Background**
- **Problems**
- **Overview of ThaiQCor 2.0 process**
- **Experimental results**
- **Conclusion and future works**
- **Q & A**

# Background

The number of people using **internet search engines** is increasing year on year.

**6 billion** searches are made every day worldwide. (as of April 2017)

amazon

ebay™

twitter

Baidu 百度

Google

wongnai

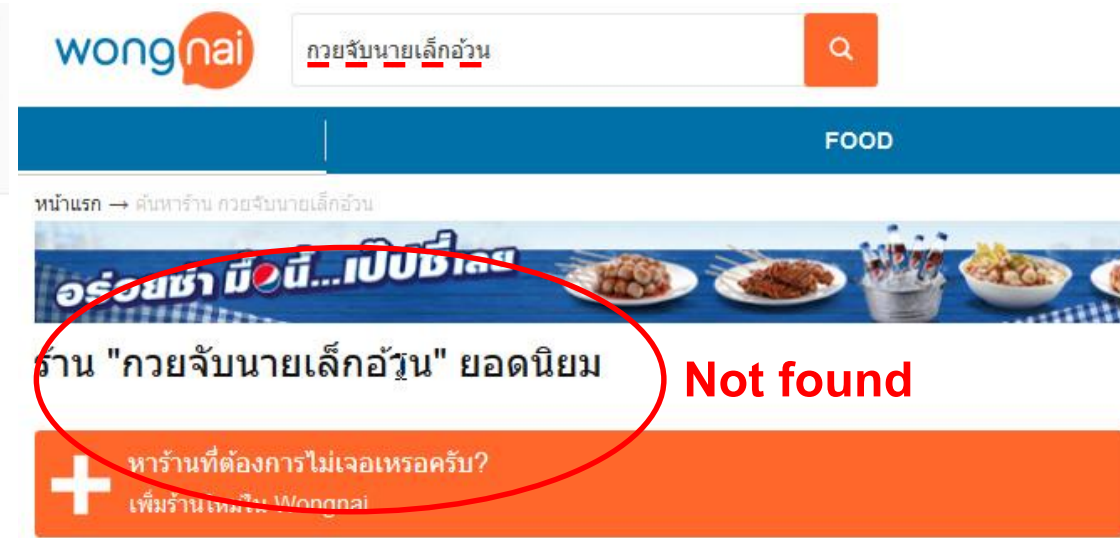
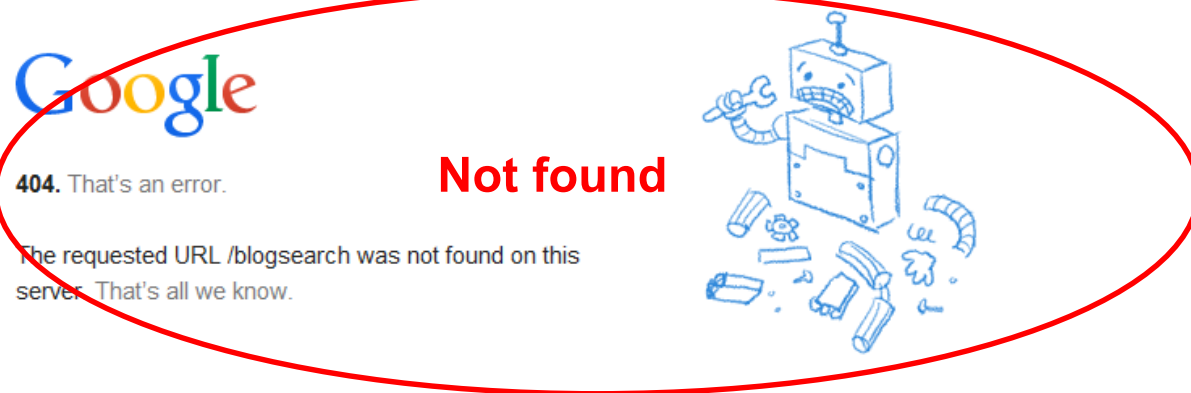
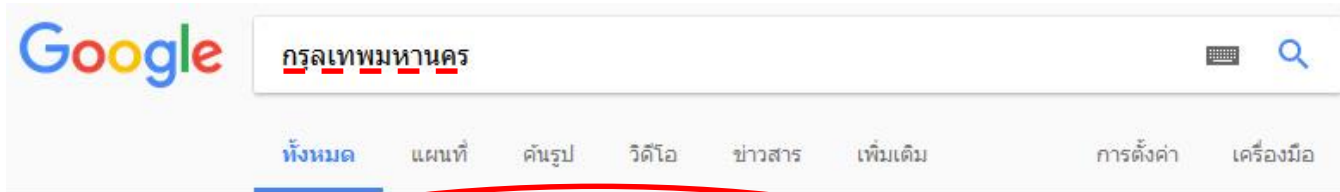
YAHOO!®

You Tube

bing

# Problems

- **26% of all queries** entered into web search engines are **inaccurate typing**. (Wang et al. 2003)
- Search results are **incorrect** or **not found**.



# Inaccurate Typing

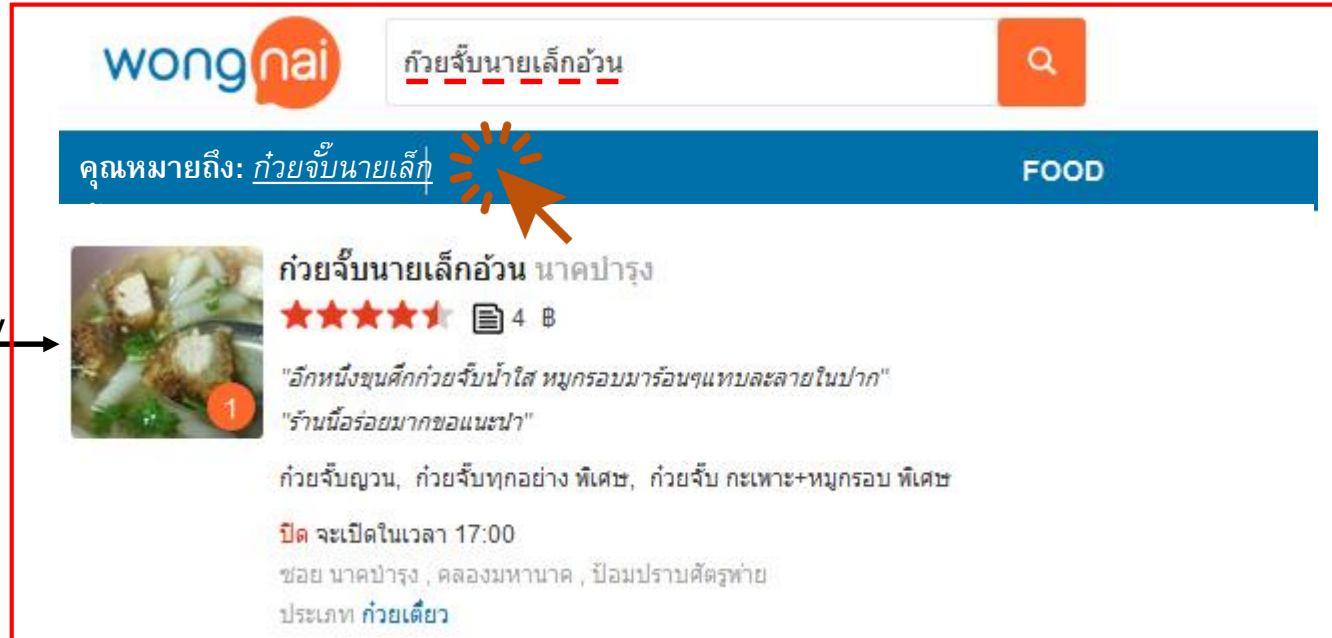
- **Typographical errors**
  - giraffe -> guraffe
  - อิทธิพัทธ์ -> อิทธิพัทธ์ท์
- **Cognitive or Phonetic errors**
  - there -> theire
  - พันธุ์ศักดิ์ -> พันธุ์ศักดิ์

# ThaiQCor 2.0

- To develop a new version of **Thai query correction program**
- To solve **typographical and phonetic errors**
- To help users obtain **satisfactory search results**



Query



wongnai

ก๋วยจั๊บน้ำใส เล็ก อ้วน

คุณหมายถึง: ก๋วยจั๊บน้ำใส เล็ก อ้วน

FOOD

ก๋วยจั๊บน้ำใส เล็ก อ้วน นาคปารุง

★★★★★ 4.8

"อีกหนึ่งซนตึกก๋วยจั๊บน้ำใส หมูกรอบมาร้อนๆแทบละลายในปาก"

"ร้านนี้อร่อยมากขอแนะนำ"

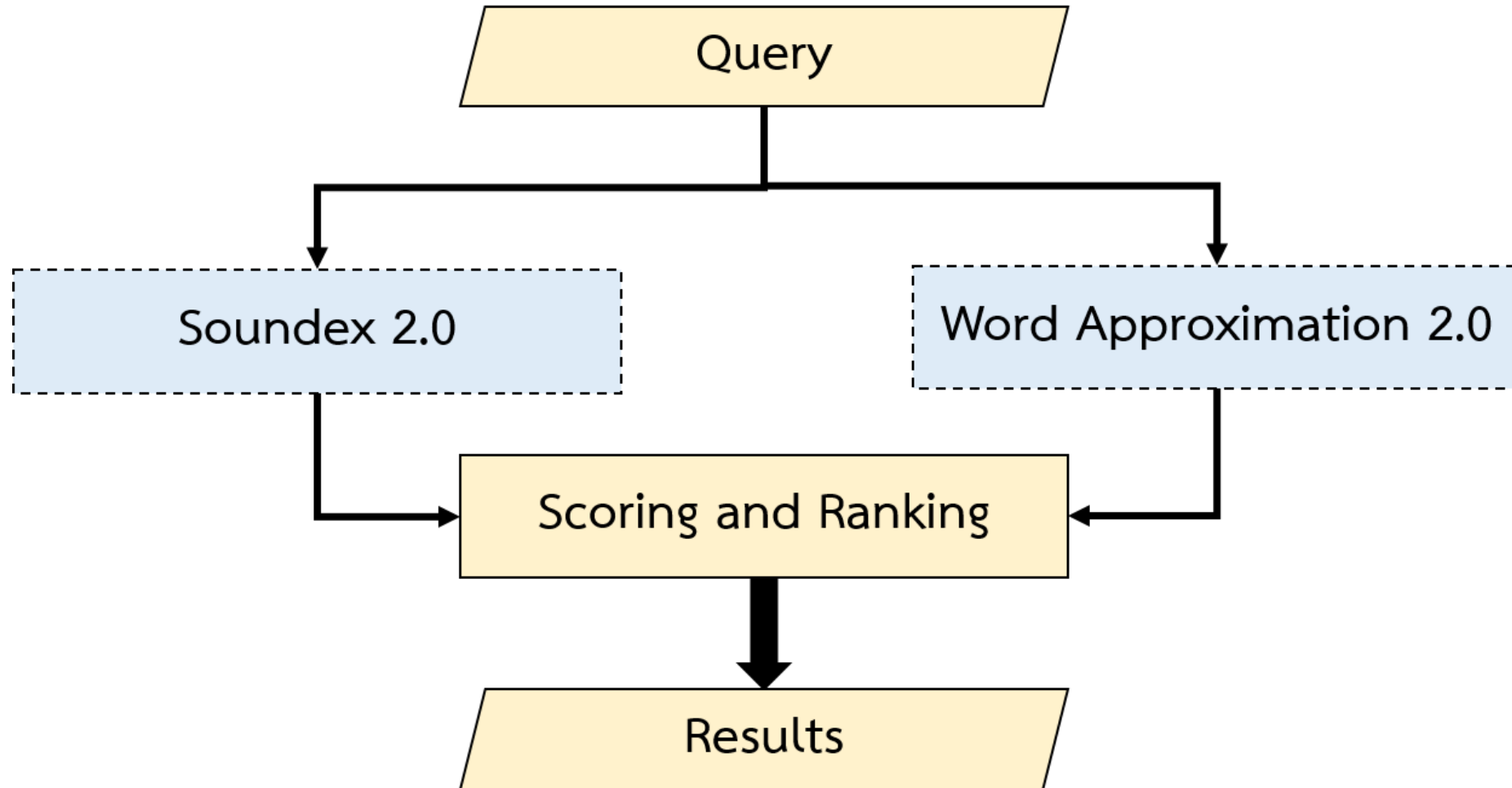
ก๋วยจั๊บน้ำใส, ก๋วยจั๊บน้ำใสทุกอย่าง พิเศษ, ก๋วยจั๊บน้ำใส กะเพรา+หมูกรอบ พิเศษ

ปิด จะเปิดในเวลา 17:00

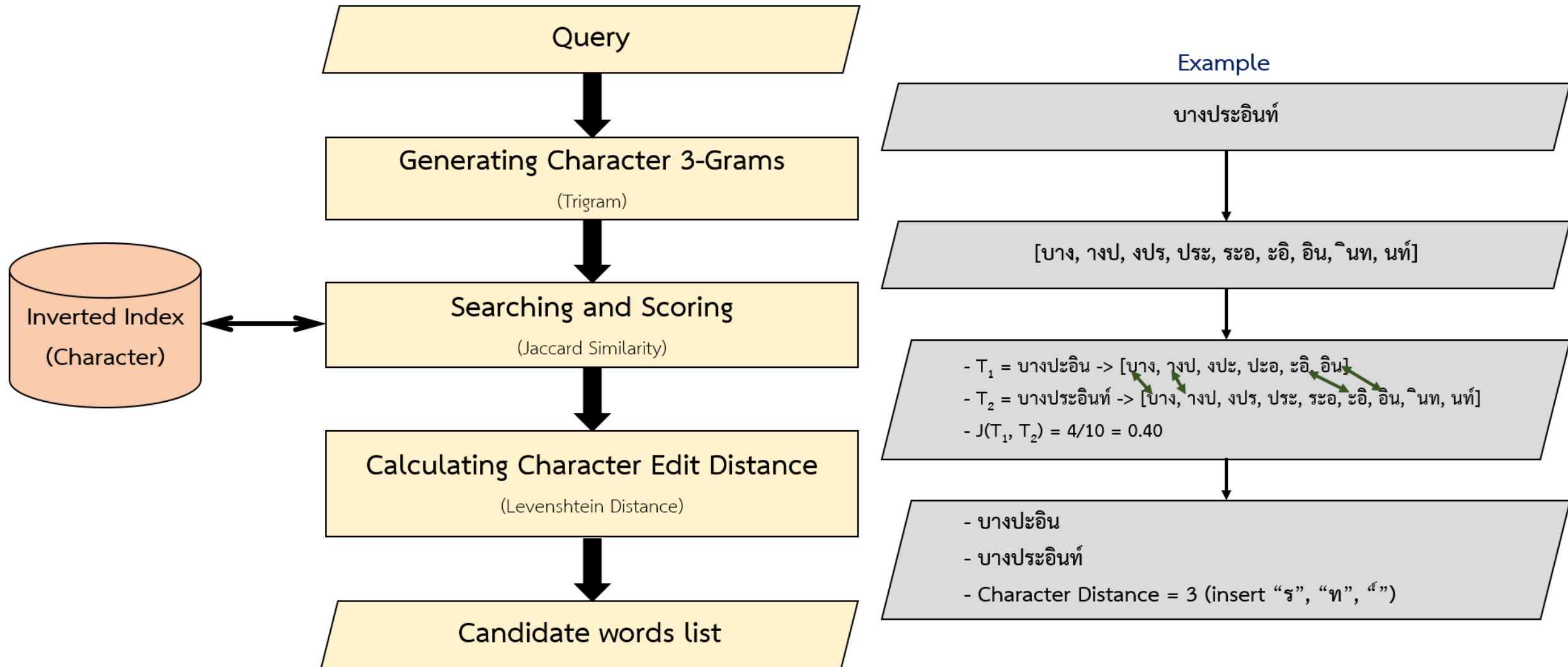
ซอย นาคปารุง, คลองมหานาค, ป้อมปราบศัตรูพ่าย

ประเภท ก๋วยเตี๋ยว

# ThaiQCor 2.0

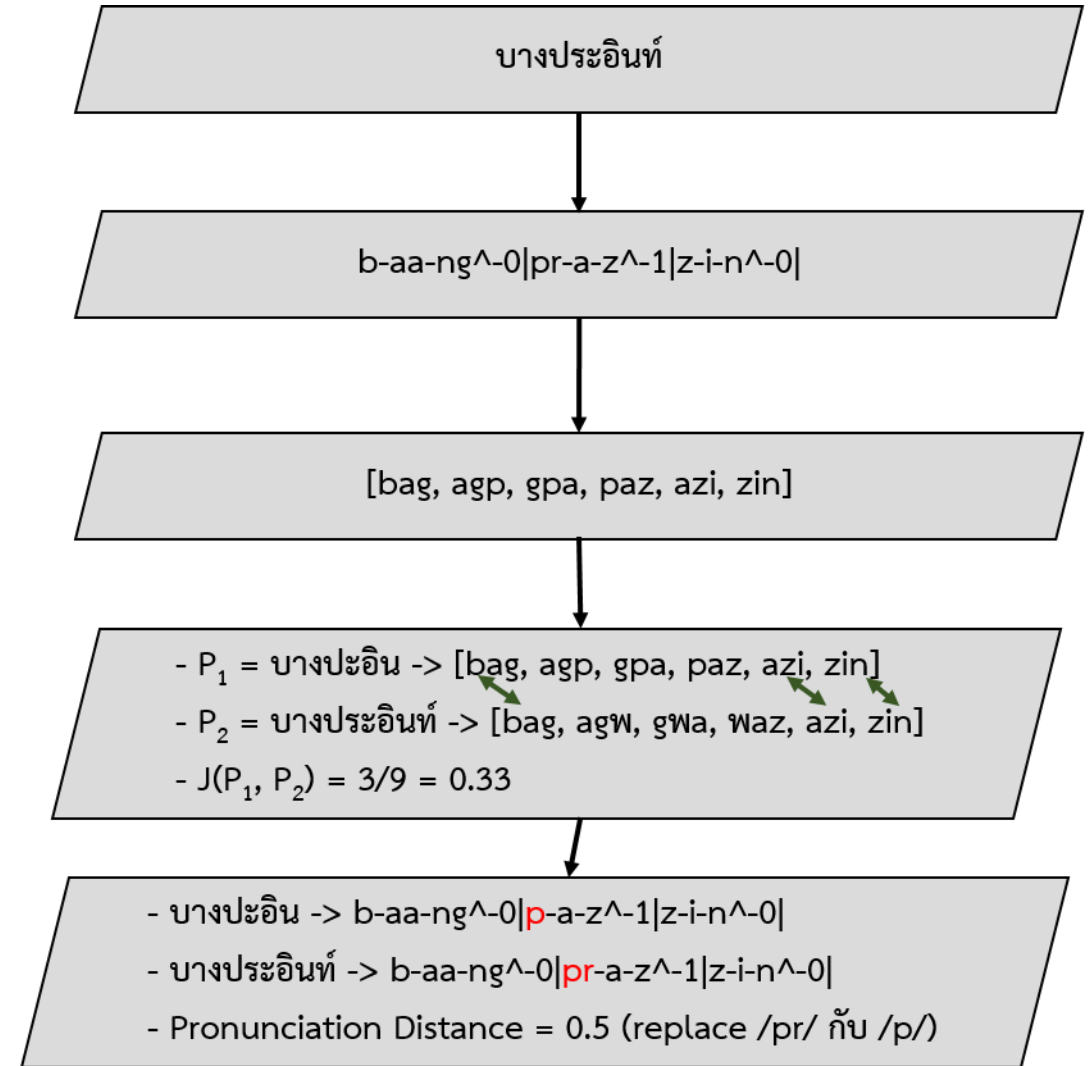
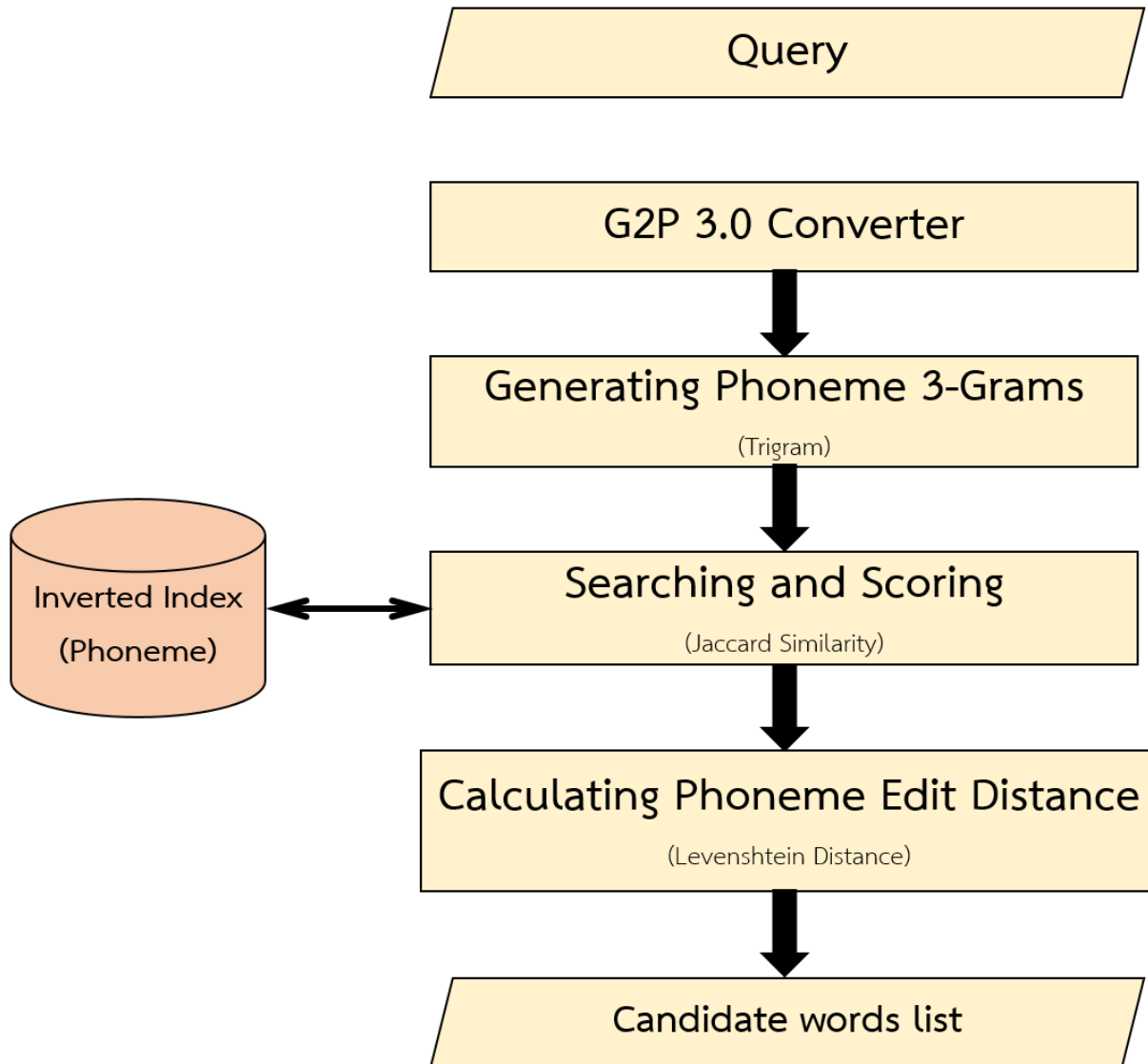


# Word Approximation





# Soundex



# Words Scoring and Ranking

$$Score_{(t,T_{all})} = \left\{ \begin{array}{l} N_{approx} - rank_t, \{t|t \in W_{approx} \wedge t \in W_{sound}\} \\ N_{sound} - rank_t, \{t|t \notin W_{approx} \wedge t \in W_{sound}\} \\ \frac{(N_{approx} - rank_t) + (N_{sound} - rank_t)}{2} + 1, \{t|t \in W_{approx} \wedge t \in W_{sound}\} \end{array} \right\} \quad (1)$$

Where

$Score_{(t,T_{all})}$  is the score of each term in  $T_{all}$ .

$N_{approx}$  is the number of all words in word approximation 2.0 approach.

$N_{sound}$  is the number of all words in soundex 2.0 approach.

$Rank_t$  is the position of the term.

# Corpus

- **Words corpus**
  - Place name: **61,738** words
  - Person name: **2,890** words
- **Test set (wrong words)**
  - Place name: **868** words
  - Person name: **254** words

# Evaluation

- We use accuracy as metric to identify which approach yields the best accuracy.

$$N - \text{Best Accuracy} = \frac{\text{Number of words where approach can search correctly}}{\text{Number of words in the test set}}$$

# ThaiQCor 1.0 VS. ThaiQCor 2.0

Approaches	Place Name (#query 868)		Person Name (#query 254)	
	nBest=1	nBest=5	nBest=1	nBest=5
ThaiQCor 1.0	52.07	89.97	12.59	51.57
ThaiQCor 2.0	88.82	97.11	75.19	89.76

# Discussions

- Word approximation approach is also suitable in the domain of **place names**.
- Soundex approach yielded the good accuracy in the domain of **person names**.

# Conclusion and Future Works

- We designed and developed a new Thai query correction program called ThaiQCor 2.0.
- Our program consists of two main approaches: word approximation and soundex approaches.
- ThaiQCor 2.0 yielded better query correction than ThaiQCor 1.0 in the all test sets.
- We plan to calculate the proximity between Thai keyboard keys by adjusting the cost of edit operations.
- In soundex approach, we will improve the performance of generating phones on G2P conversion based on LSTM.

**Thank you**