Từ dừng(Stopwords) và Rút gọn từ về gốc(stemming)

Stopwords và **stemming** – hai bước rất quan trọng trong quá trình **tiền xử lý văn bản** trong lĩnh vực **Xử lý ngôn ngữ tự nhiên (NLP)**:

Từ dừng(Stopwords**)**

Là gì?

Stopwords là những từ **rất phổ biến trong ngôn ngữ** nhưng **ít mang ý nghĩa nội dung** khi phân tích, ví dụ như:

Tiếng Việt	**Tiếng Anh**
"là", "của", "và", "những", "đã", "đang", "sẽ", "tôi", "anh", "chị", "một", "này", "kia", "đó", "với"...	is, the, a, an, in, at, of, with, was...

Tại sao cần loại bỏ?

- Giúp **giảm nhiễu** khi phân tích nội dung.
- **Tiết kiệm tài nguyên xử lý** (bộ nhớ, thời gian).
- Tập trung vào **từ khóa mang ý nghĩa chính**.

Ví dụ 1:

Câu gốc:

> "Tôi đang học lập trình với Python tại trường đại học."

Các stopwords:

> "Tôi", "đang", "với", "tại"

Sau khi loại bỏ stopwords:

>

> "học lập trình Python trường đại học"

➡ Kết quả giúp mô hình NLP tập trung vào **từ khóa chính**.

Ví dụ 2:

Câu gốc:

> "Anh ấy đã mua một chiếc xe mới vào hôm qua."

Stopwords loại bỏ:

> "Anh ấy", "đã", "một", "vào"

Kết quả:

> "mua chiếc xe mới hôm qua"

Rút gọn từ về gốc(Stemming)

Là gì?

Stemming giúp đưa các từ về **gốc từ** bằng cách **cắt bỏ tiền tố, hậu tố**.
Với tiếng Việt, điều này phức tạp hơn tiếng Anh do từ có thể mang nhiều thành phần.

Ví dụ tiếng Anh:

Từ gốc (Stem)	Từ biến thể
run	running, runs, ran
connect	connected, connecting
develop	developing, developed

Tất cả sẽ được đưa về từ gốc: `run`, `connect`, `develop`.

Ví dụ 1:

Các từ biến thể:

> "học", "học sinh", "học tập", "học hành", "học hỏi"

Stem (từ gốc):

> "học"

Giải thích:

- "học sinh" → người đi học
- "học tập", "học hành", "học hỏi" → các biến thể của hành động "học"

➡ Khi phân tích nội dung, ta có thể gom các từ này về cùng một chủ đề: **“học”**

Ví dụ 2:

Câu gốc:

> "Người lao động đang làm việc chăm chỉ để hoàn thành dự án."

Các từ cần stem:

- "lao động" → "lao động" (giữ nguyên)
- "làm việc" → "làm"
- "hoàn thành" → "thành"

Kết quả sau khi stemming:

> "người lao động làm chăm chỉ thành dự án"

➡ Có thể không tự nhiên trong văn nói, nhưng rất hữu ích cho phân loại văn bản, tìm kiếm hoặc phân tích ngữ nghĩa.

Công cụ phổ biến:

- `Porter Stemmer` (tiếng Anh)
- `Snowball Stemmer`

- Với tiếng Việt: công cụ **VnCoreNLP**, `pyvi`, `underthesea`,...

Ví dụ tiếng Việt:

Câu: "*Học sinh đang học bài học mới.*"
→ Sau stemming: "**học sinh học bài học mới**"
(Từ `"học"` được giữ nguyên; `"học sinh"` và `"bài học"` không cần tách vì vẫn giữ nghĩa)

Tổng kết mối quan hệ:

Kỹ thuật	Mục đích	Kết quả
Tokenization	Tách văn bản thành từ/token	["Tôi", "đã", "đi", "đến"...]
Stopword Removal	Loại bỏ từ phổ biến không cần	["Tôi", "đi", "trường", "sáng"]
Stemming	Đưa từ về gốc	["Tôi", "đi", "trường", "sáng"] (hoặc gốc thêm nếu có)

Dùng thư viện xử lý ngôn ngữ tự nhiên (NLP)

Ví dụ bằng Python + thư viện **Underthesea**

```
from underthesea import word_tokenize


# Câu ví dụ
sentence = "Tôi đang học lập trình Python tại trường đại học."

# Tách từ
words = word_tokenize(sentence, format="text")
print("Các từ trong câu:", words)
```

Loại bỏ stopwords thủ công:

```
stopwords = ["tôi", "đang", "tại", "là", "và", "của", "một", "những", "này", "đó"]
filtered_words = [word for word in words.split() if word.lower() not in stopwords]
print("Sau khi loại bỏ stopwords:", filtered_words)
```

Ví dụ nhận diện stemming (rút gọn từ)

```
words = ["học", "học sinh", "học tập", "học hành", "học hỏi"]
stem = "học"
for word in words:
    if word.startswith(stem):
```

```
    print(f"Từ '{word}' có thể được quy về gốc '{stem}'")
```

Nhận biết thủ công (khi không dùng code)

- Với **stopwords**: Tìm các từ **quá phổ biến, không mang nhiều nội dung riêng biệt**.
  Ví dụ: "và", "là", "của", "một", "đã", "đang", "với", "cho", "như", "này", "kia"...
- Với **stemming**: Tìm các từ **liên quan đến cùng một gốc từ**, dù khác nhau về hậu tố hoặc dạng sử dụng.
  Ví dụ: "chơi", "chơi đùa", "chơi game", "chơi thể thao" → chung một gốc là "chơi"

Một số công cụ gợi ý

Công cụ	Tính năng chính	Ngôn ngữ
Underthesea	Tách từ, POS tagging, nhận diện từ gốc	Tiếng Việt
VnCoreNLP	Tách từ, phân tích ngữ pháp, NER	Tiếng Việt
spaCy + pyvi	Kết hợp để xử lý văn bản tiếng Việt	Tiếng Việt
NLTK	Hỗ trợ stopwords tiếng Anh mạnh mẽ	Tiếng Anh

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #1
Được tạo 4 tháng 5 2025 04:50:36 bởi Đỗ Ngọc Tú
Được cập nhật 4 tháng 5 2025 05:10:06 bởi Đỗ Ngọc Tú