HUTECH
Đại học Công nghệ Tp.HCM

HUTECH University

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

TÓM TẮT VĂN BẢN (SUMMARIZATION)

TS. Lê Thị Ngọc Thơ

ltn.tho@hutech.edu.vn

NỘI DUNG

- Giới thiệu
- Các phương pháp
  - Tóm tắt văn bản bằng trên xếp hạng trên đồ thị
  - Tóm tắt văn bản bằng cách phân lớp dữ liệu

GIỚI THIỆU

- Định nghĩa: Tìm những điểm quan trọng trong văn bản và biểu diễn lại dưới dạng cô đọng súc tích.
- Quy trình:
  - Biểu diễn văn bản;
  - Rút trích các thông tin có liên quan (chủ đề của nguồn thông tin)
  - Cô đọng các thông tin rút trích và tạo ra biểu diễn tóm tắt;
  - Trình bày biểu diễn tóm tắt cho người đọc bằng ngôn ngữ tự nhiên.

GIỚI THIỆU

- Các loại tóm tắt:
  - Rút trích (Extract) là tóm tắt bao gồm toàn bộ vật liệu được sao chép từ tài liệu gốc (ví dụ: trích xuất 25% tài liệu gốc).
  - Tóm tắt trừu tượng (Abstract) chứa các câu từ không có trong tài liệu đầu vào.
  - Tóm tắt chỉ thị (indicative) giúp chúng ta quyết định có nên đọc tài liệu hay không.
  - Thông tin hữu ích (Informative) bao gồm tất cả các thông tin nổi bật trong nguồn (thay thế toàn bộ tài liệu).

GIỚI THIỆU

- Rút trích và trừu tượng hóa: The Gettysburg Address

Rút trích

Four score seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract.

Trừu tượng

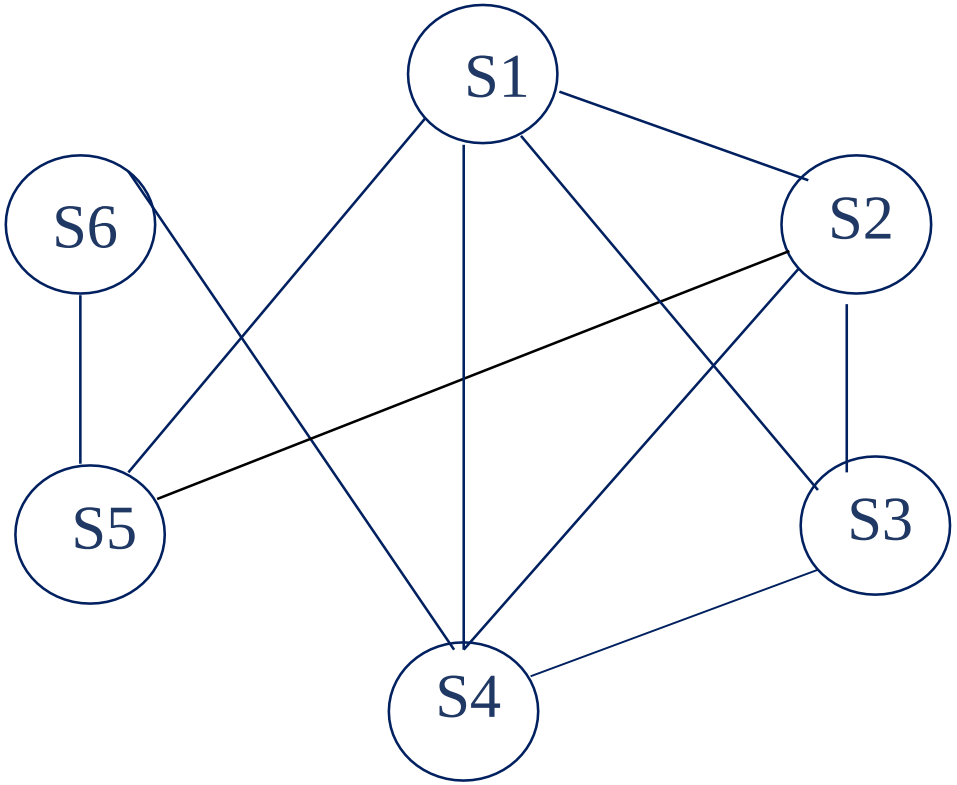
The speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.

GIỚI THIỆU

- Các loại cô đọng thông tin:
  - Headlines (tiêu đề)
  - Outlines (phác thảo)
  - Minutes (biên bản)
  - Biographies (tiểu sử)
  - Abridgments (trích yếu)

TÓM TẮT VĂN BẢN DỰA TRÊN ĐỒ THỊ

- Biểu diễn văn bản dưới dạng đồ thị



TÓM TẮT VĂN BẢN DỰA TRÊN ĐỒ THỊ

- Tóm tắt dựa trên trung tâm (Radev)
- Giả sử: Trung tâm của đỉnh chỉ ra sự quan trọng của nó.
- Biểu diễn: Ma trận liên kết dựa trên độ đo cosine trong câu.
- Cơ chế rút trích
  - Tính điểm PageRank cho mỗi câu u

$$PageRank(u) = \frac{(1-d)}{N} + d \sum_{v \in adj[u]} \frac{PageRank(v)}{\deg(v)}$$

N là số đỉnh trong đồ thị

  - Rút trích k câu có số điểm PageRank cao nhất

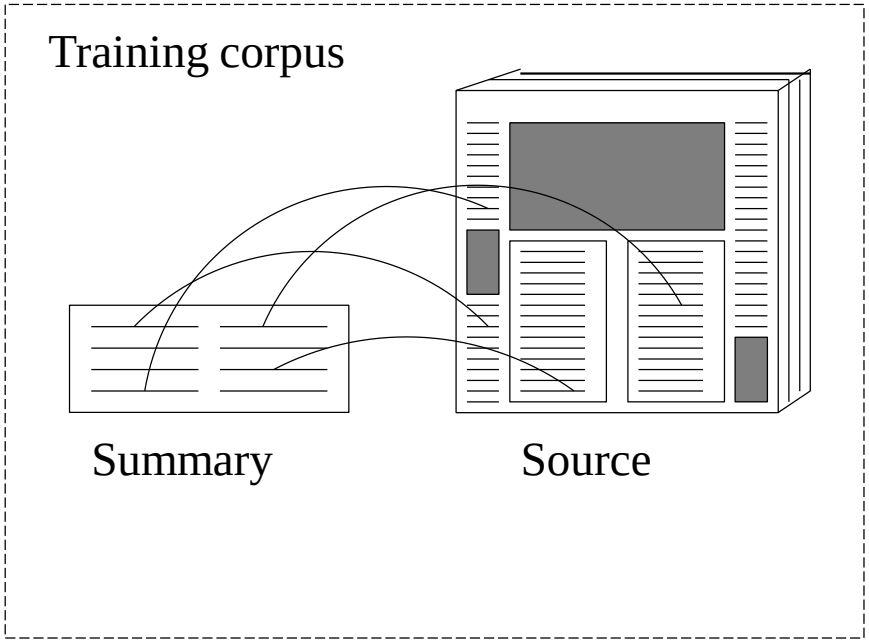
TÓM TẮT VĂN BẢN DỰA TRÊN ĐỒ THỊ

- Phương pháp này hoạt động có tốt?
- Đánh giá: So sánh với tóm tắt do con người tạo ra
- Độ đo ROUGH: Sự trùng khớp của N-gram có trọng số (tương tự BLEU)

Phương pháp	**ROUGE score**
Random	0.3261
Lead	0.3575
Degree	0.3595
PageRank	0.3666

TÓM TẮT VĂN BẢN BẰNG PHÂN LỚP

- Tóm tắt văn bản bằng cách phân lớp câu **rút trích**:
  - Cho một tập tài liệu và các tóm tắt tương ứng;
  - Gán nhãn cho mỗi câu trong tài liệu là có phải câu tóm tắt không.



  - Học câu nào có thể sẽ được đưa vào tóm tắt
  - Cho một tài liệu chưa biết (tài liệu thử nghiệm), phân loại các câu trong văn bản đó có phải là câu tóm tắt hay không.

TÓM TẮT VĂN BẢN BẰNG PHÂN LỚP

Blue: tóm tắt
Red: Thường

- Ví dụ:

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate — we can not consecrate — we can not hallow — this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract.

The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced.

TÓM TẮT VĂN BẢN BẰNG PHÂN LỚP

- Trong quá trình huấn luyện, mỗi câu có một điểm quan trọng / "sự khúc chiết".
- Trong quá trình thử nghiệm trích xuất các câu có điểm số cao nhất theo nguyên văn như trích xuất.
- Nhưng làm thế nào chúng ta tính điểm này?
  - Đầu tiên, gán câu tóm tắt trong tài liệu.
  - Sau đó, tinh giản các câu thành các đặc trưng quan trọng.
  - Mỗi câu được biểu diễn dưới dạng một vector của các đặc trưng.

TÓM TẮT VĂN BẢN BẰNG PHÂN LỚP

- Các đặc trưng:

Sentence Length Cut-off Feature	true if sentence > 5 words
Fixed-Phrase Feature indicator	true if sentence contains phrases: *this letter, in conclusion*
Paragraph Feature	initial, final, medial
Thematic Word Feature	true if sentence contains frequent words
UppercaseWord Feature	true if sentence contains proper names: *the American Society for Testing and Materials*

TÓM TẮT VĂN BẢN BẰNG PHÂN LỚP

Training Data	Test Data
⇓	⇓
[1,0,INITIAL,1,0]	[0,0,MEDIAL,0,0]
[0,0,INITIAL,1,1]	[0,0,INITIAL,1,1]
[1,1,MEDIAL,0,0]	??
[1,1,MEDIAL,1,1]	
[0,0,MEDIAL,0,0]	
[1,1,INITIAL,1,1]	
[0,0,INITIAL,1,1]	

red: not in summary, blue: in summary

TÓM TẮT VĂN BẢN BẰNG PHÂN LỚP

- Kết hợp những đặc trưng câu: Kupiec, Pedersen, Chen. *"A trainable document summariser,"* SIGIR 1995

$$P(s \in S|F_1, \dots, F_k) = \frac{P(F_1, \dots, F_k|s \in S)P(s \in S)}{P(F_1, \dots, F_k)}$$

$$\approx \frac{P(s \in S)\prod_{j=1}^{k} P(F_j|s \in S)}{\prod_{i=1}^{k} P(F_j)}$$

  - $P(s \in S|F_1, \dots, F_k)$: xác suất mà s từ văn bản gốc được tóm tắt S, cho các giá trị đặc trưng
  - $P(s \in S)$: xác suất s từ văn bản gốc được tóm tắt là S không có điều kiện
  - $\prod_{j=1}^{k} P(F_j|s \in S)$: xác suất cặp giá trị-tính có trong câu nằm trong bản tóm tắt
  - $P(F_j)$: xác suất mà cặp tính năng-giá trị F_j xảy ra vô điều kiện

TÓM TẮT VĂN BẢN BẰNG PHÂN LỚP

- Đánh giá:
  - Ngữ liệu của 85 bài viết trong 21 tạp chí
  - Baseline: chọn các câu đầu của tài liệu
  - Độ nén rất cao làm cho công việc này khó khăn hơn

Feature	Individual Sents Correct	Cumulative Sents Correct
Paragraph	163 (33%)	163 (33%)
Fixed Phrases	145 (29%)	209 (42%)
Length Cut-off	121 (24%)	217 (44%)
ThematicWord	101 (20%)	209 (42%)
Baseline	24%	