CƠ SỞ LÝ THUYẾT

2.1. KMeans Clustering

KMeans Clustering là một thuật toán phân cụm (clustering) không giám sát phổ biến, được sử dụng để nhóm các điểm dữ liệu (trong không gian vector) thành *k* cụm sao cho các điểm trong cùng một cụm có độ tương đồng cao hơn so với các cụm khác.

Nguyên lý hoạt động của KMeans:

- Khởi tạo *k* điểm làm tâm cụm ban đầu (centroids).
- Gán mỗi điểm dữ liệu vào cụm có tâm gần nhất (thường dựa trên khoảng cách Euclidean).
- Cập nhật lại tâm cụm bằng cách tính trung bình các điểm thuộc cụm đó.
- Lặp lại bước 2–3 cho đến khi các tâm cụm hội tụ hoặc không thay đổi nhiều.

Áp dụng vào tóm tắt văn bản:

- Mỗi câu được biểu diễn thành một vector bằng cách trung bình các vector từ của các từ trong câu (sử dụng Word2Vec).
- Áp dụng thuật toán KMeans để nhóm các câu thành k cụm.
- Từ mỗi cụm, chọn câu gần tâm cụm nhất làm đại diện, tạo thành bản tóm tắt.

Ưu điểm:

- Đơn giản, hiệu quả.
- Có thể điều chỉnh độ dài tóm tắt thông qua tham số k.

Nhược điểm:

- Cần xác định trước số cụm k.
- Phân cụm tuyến tính, không thể hiện rõ quan hệ ngữ nghĩa sâu sắc giữa các câu.

2.2. TextRank

TextRank là một thuật toán xếp hạng dựa trên PageRank, ban đầu được đề xuất bởi Rada Mihalcea và Paul Tarau (2004), thích hợp cho các bài toán trích xuất thông tin, đặc biệt là **tóm tắt văn bản**.

Nguyên lý hoạt động:

- Xây dựng đồ thị: mỗi nút là một câu trong văn bản.
- Tính độ tương đồng giữa các câu (thường là cosine similarity), đặt làm trọng số cạnh giữa các nút.
- Áp dụng thuật toán PageRank để tính điểm quan trọng của từng câu.
- Sắp xếp và chọn n câu có điểm cao nhất làm bản tóm tắt.

Công thức PageRank cơ bản:

$$PR(A) = (1 - d) + d \cdot \sum_{i \in In(A)} \frac{PR(i)}{Out(i)}$$

Trong đó:

- ***PR(A):*** điểm PageRank của nút ***A***.
- ***d***: hệ số giảm (thường là 0.85).
- ***In(A)***: tập các nút trỏ đến ***A***.
- ***Out(i)***: số liên kết từ nút ***i***.

Ưu điểm:

- Không yêu cầu dữ liệu huấn luyện.
- Xét đến mối liên kết giữa các câu nên cho ra tóm tắt có tính tổng quát cao.

Nhược điểm:

- Phụ thuộc nhiều vào cách xây dựng đồ thị và đo độ tương đồng.
- Có thể xảy ra lỗi hội tụ (như PowerIterationFailedConvergence) khi áp dụng trên ma trận quá nhỏ hoặc suy biến.

2.3. LSA (Latent Semantic Analysis)

Latent Semantic Analysis (LSA) là một kỹ thuật phân tích ngữ nghĩa dựa trên thống kê, thường dùng trong khai phá văn bản và tóm tắt tự động.

Nguyên lý hoạt động:

- Biểu diễn văn bản dưới dạng ma trận thuật ngữ – câu (Term-Sentence Matrix), thường sử dụng TF-IDF.
- Áp dụng phân tích giá trị suy biến (SVD – Singular Value Decomposition):

$$A = U \sum V^T$$

+ A: ma trận gốc (TF-IDF).

+ U, Σ, V^T: các ma trận thành phần thu được từ SVD.

- Giữ lại k thành phần chính (thường chọn các giá trị kỳ dị lớn nhất), giúp làm nổi bật các chủ đề chính (latent topics).
- Dựa vào các vector biểu diễn câu trong không gian giảm chiều, chọn những câu có mức đóng góp lớn vào các chủ đề chính làm tóm tắt.

Ưu điểm:

- Phân tích được cấu trúc ngữ nghĩa sâu của văn bản.
- Giảm nhiễu và mối quan hệ mờ nhạt giữa các từ.

Nhược điểm:

- Yêu cầu xử lý ma trận, phức tạp về tính toán.

- Không phản ánh tốt ngữ cảnh động (thời gian, vị trí trong văn bản).

2.4. Word2Vec

Word2Vec là một mô hình học nhúng từ (word embedding) do nhóm của Tomas Mikolov tại Google phát triển. Mục tiêu của mô hình là học các vector từ sao cho các từ có ngữ nghĩa tương tự nằm gần nhau trong không gian vector.

Hai kiến trúc chính:

- CBOW (Continuous Bag of Words): Dự đoán từ hiện tại từ ngữ cảnh xung quanh.
- Skip-gram: Dự đoán ngữ cảnh xung quanh từ hiện tại.

Đặc điểm:

- Word2Vec học các vector từ kích thước cố định (thường 100–300 chiều).
- Mô hình dùng trong đề tài là pretrained cc.vi.300.bin (FastText tiếng Việt), sau đó giảm chiều xuống còn 100 để tối ưu hiệu suất xử lý (cc.vi.100.vec).

Biểu diễn câu:

- Vector câu thường được tính bằng cách trung bình các vector từ trong câu.
- Dữ liệu sau khi vector hóa sẽ được sử dụng cho các thuật toán phân cụm hoặc tóm tắt (KMeans, LSA...).

Ưu điểm:

- Biểu diễn ngữ nghĩa tốt, nhẹ, có thể dùng cho nhiều tác vụ NLP.
- Có thể biểu diễn ngữ pháp và quan hệ từ vựng (ví dụ: “vua - nam + nữ ≈ nữ hoàng”).

Nhược điểm:

- Không xử lý tốt từ mới (OOV).
- Không nắm bắt được ngữ cảnh động như các mô hình BERT.

2.5 Tài liệu tham khảo

[1] https://www.ibm.com/think/topics/k-means-clustering

[2] https://www.ibm.com/think/topics/latent-semantic-analysis

[3] https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

[4] https://machinelearningcoban.com/tabml_book/ch_embedding/word2vec.html

[5] https://fasttext.cc/docs/en/crawl-vectors.html