Text Mining and Social Media Mining

THS. LÊ NHẬT TÙNG

Topic modeling

THS. LÊ NHẬT TÙNG

Topic Modeling là gì?

- Mô hình thống kê
- Được sử dụng để xác định các chủ đề trừu tượng xuất hiện trong bộ sưu tập tài liệu
- Topic modeling (không giám sát) vs topic classification (có giám sát)
- Có giá trị trong việc tìm kiếm cấu trúc ngữ nghĩa ẩn trong nội dung văn bản, phân loại tài liệu theo các chủ đề được phát hiện và sau đó phân tích văn bản

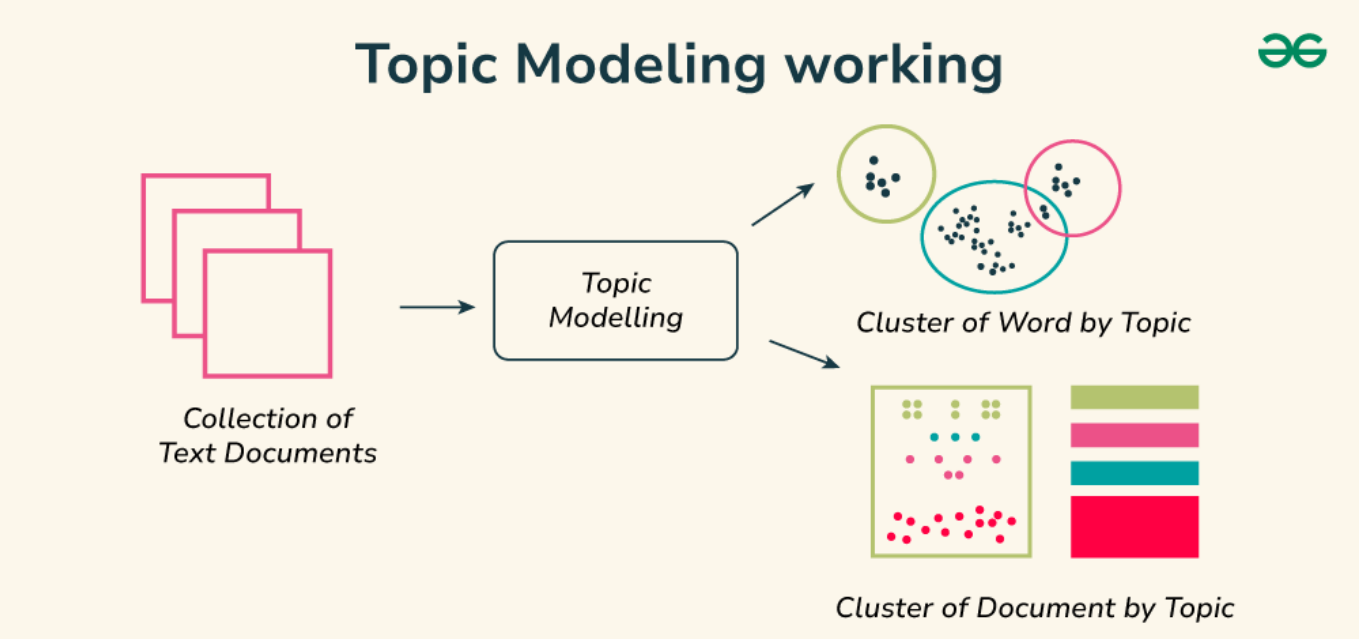
Ứng dụng của Topic Modeling

- Email
- Chats
- Survey responses
- Social media posts
- Official claims

Topic Modeling

Tìm một nhóm từ từ bộ sưu tập tài liệu có vẻ đại diện tốt nhất cho thông tin trong bộ sưu tập

Thu được các mẫu từ lặp lại trong tài liệu văn bản



Topics

- Thực tế, chúng là dạng cụm các từ tương tự nhau
- Thống kê về các từ trong mỗi cụm
- Document's balance of topics
- Cũng được áp dụng trong việc phát hiện các cấu trúc hữu ích trong dữ liệu

Topic Model

Topic model: một thuật toán quét qua các tài liệu (corpus), kiểm tra cách các từ và cụm từ xuất hiện cùng nhau trong tài liệu, và sau đó học các nhóm từ đặc trưng nhất cho các tài liệu này - những tập hợp này thường đại diện cho một chủ đề mạch lạc

Giả định mạnh của mọi thuật toán Topic Modeling

- Tài liệu của chúng ta có số lượng chủ đề cố định
- Đánh giá cấu trúc cơ bản của các từ đằng sau dữ liệu và tìm các nhóm phù hợp với corpus
- Một từ có thể được gán cho nhiều topics với tỷ lệ khác nhau hoặc được gán cho một topic duy nhất

Kết quả đầu ra

- Term-topic matrix
  - Topics dưới dạng các thành phần từ của chúng
- Document-topic matrix
  - Documents dưới dạng các topics của chúng

Các thuật toán Topic Modeling

- Early topic model (1998)
- Probabilistic latent semantic analysis (PLSA, 1999)
- Latent Dirichlet allocation (LDA, 2002)
- Extensions of the LDA
  - Pachinko allocation

Name of The Methods	Characteristics
Latent Semantic Analysis (LSA)	* LSA can get from the topic if there are any synonym words. * Not robust statistical background.
Probabilistic Latent Semantic Analysis (PLSA)	* It can generate each word from a single topic; even though various words in one document may be generated from different topics. * PLSA handles polysemy.
Latent Dirichelet Allocation (LDA)	* Need to manually remove stop-words. * It is found that the LDA cannot make the representation of relationships among topics.
Correlated Topic Model (CTM)	* Using of logistic normal distribution to create relations among topics. * Allows the occurrences of words in other topics and topic graphs.

Name of The Methods	Limitations
Latent Semantic Analysis (LSA)	- It is hard to obtain and to determine the number of topics. - To interpret loading values with probability meaning, it is hard to operate it.
Probabilistic Latent Semantic Analysis (PLSA)	- At the level of documents, PLSA cannot do probabilistic model.
Latent Dirichelet Allocation (LDA)	- It becomes unable to model relations among topics that can be solved in CTM method.
Correlated Topic Model (CTM)	- Requires lots of calculation - Having lots of general words inside the topics.

Alghamdi R., Alfalqi K. (2015)

Latent Dirichlet Allocation (LDA)

- Thuật toán lặp xác định một tập hợp các topics liên quan đến một tập hợp documents
- Dựa trên sparse Dirichlet prior distributions - các véc-tơ xác suất của các từ, chỉ ra mức độ liên quan của chúng với text corpus
- Prior distributions được thực hiện trên document-topic và topic-word distributions
- Trực quan: documents thường đề cập đến một số lượng nhỏ các topics; topics thường dựa trên một số lượng nhỏ các từ

Tham số LDA

- Alpha - document-topic density; alpha càng cao, documents được tạo thành từ nhiều topics hơn
- Beta - topic-word density; beta càng cao, topics được tạo thành từ nhiều từ hơn trong corpus
- Number of topics - được trích xuất từ corpus
- Để có số lượng tối ưu, xem Kullback Leibler Divergence Score
- Number of topic terms - được tạo thành trong một topic duy nhất
  - Decided according to the requirement

Cơ chế LDA đơn giản

•Mỗi document được coi là hỗn hợp các topics xuất hiện trong corpus
•Mô hình LDA đề xuất rằng mỗi từ trong document có thể gán cho một trong các topics đã xác định
•Chúng ta có thể chỉ định rằng chúng ta đang tìm kiếm k different topics
•LDA sẽ đi qua từng từ xuất hiện trong văn bản, ngẫu nhiên gán nó cho một trong k topics, và tính toán special score cho từ này dựa trên xác suất từ này sẽ được tìm thấy trong topic cụ thể này trong tập tài liệu
•LDA sau đó gán từ này cho một topic khác và tính toán cùng score

Cơ chế LDA đơn giản

•Sau nhiều iterations, chúng ta có được danh sách các từ trong mỗi topics với probabilities
•Đối với mỗi topic, chúng ta có thể chọn top n words với highest probability of belonging to that particular topic và chúng ta sẽ có mô tả khá tốt về nội dung topic thông qua sự kết hợp của các từ có highest probability
•Những từ này có xu hướng co-occur together in the same context. Các từ có high frequency sẽ chiếm vị trí nổi bật hơn trong mỗi topic
•Tất cả các documents trong một collection chia sẻ cùng một tập hợp các topics, nhưng mỗi document thể hiện các topics đó với different proportion

Ví dụ (Nair, 2016)

Document 1: Tôi đã ăn bánh sandwich bơ đậu phộng cho bữa sáng.

Document 2: Tôi thích ăn hạnh nhân, đậu phộng và hạt óc chó.

Document 3: Hàng xóm của tôi đã có một con chó nhỏ vào ngày hôm qua.

Document 4: Mèo và chó là kẻ thù không đội trời chung.

Document 5: Bạn không được cho chó ăn đậu phộng.

Topic 1: 30% peanuts, 15% almonds, 10% breakfast... (food)

Topic 2: 20% dogs, 10% cats, 5% peanuts... (pets or animals)

Documents 1 và 2: 100% Topic 1

Documents 3 và 4: 100% Topic 2

Document 5: 70% Topic 1, 30% Topic 2