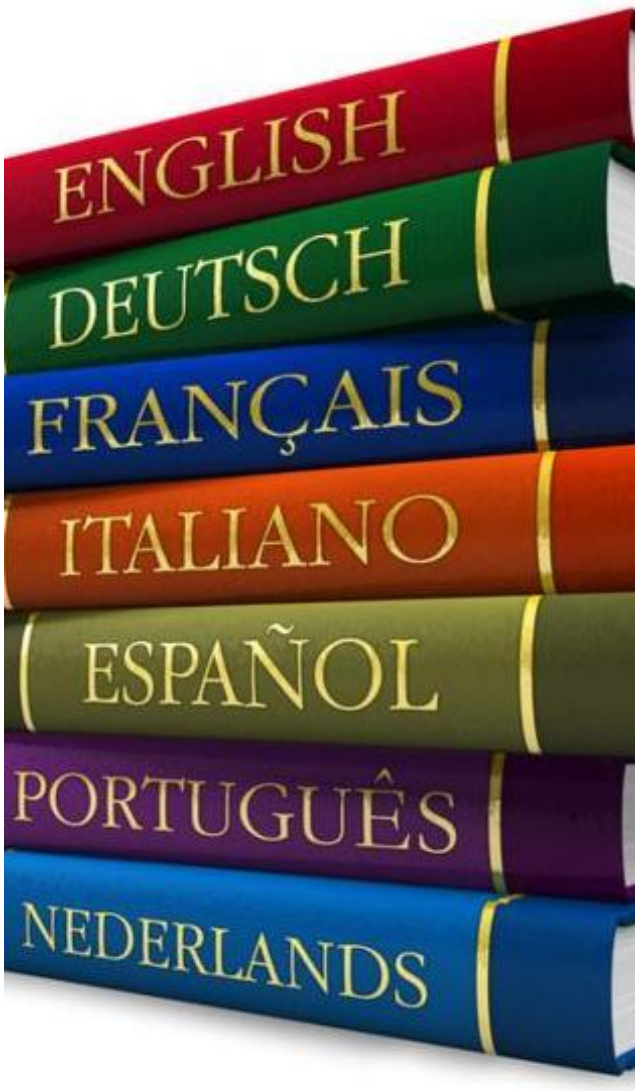
Text Mining and Social Media Mining

THS. LÊ NHẬT TÙNG

Natural language processing

THS. LÊ NHẬT TÙNG



Natural language

sản phẩm của quá trình phát triển lịch sử, trái ngược với các ngôn ngữ nhân tạo

có cú pháp phức tạp, nhiều điểm mơ hồ; vẫn đang thay đổi và phát triển

sử dụng nó đòi hỏi kiến thức về thế giới xung quanh

phương tiện chứa đựng kiến thức, thông tin và giao tiếp của con người

Natural language processing (NLP)

Xử lý thông tin viết bằng ngôn ngữ tự nhiên

Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)

Sự phát triển của NLP (Xử lý ngôn ngữ tự nhiên):

- Phân tích ngữ pháp: Các nhà Khắc kỷ (thế kỷ 3 TCN), Grimm, Rask (thế kỷ 19), Chomsky (thế kỷ 20)
- Phân tích thống kê:
  - - Các phương pháp ngẫu nhiên
  - - Mô hình xác suất
  - - Kho ngữ liệu ngôn ngữ
  - - Học máy

Natural language processing (NLP)

Tiền xử lý văn bản: "mã hóa" thông tin văn bản - từ văn bản sang số

Phân tích văn bản: phát hiện mối quan hệ và mẫu

Trực quan hóa: biểu diễn kết quả bằng đồ họa

Tiền xử lý văn bản (Text preprocessing)

1. Phân tích cú pháp

- Phân tích cú pháp, quá trình phân tích một chuỗi ký tự trong ngôn ngữ tự nhiên, tuân theo quy tắc của ngữ pháp hình thức

- Xác định các đơn vị văn bản: đoạn văn, câu, cụm từ

2. Phân đoạn

- Xác định các token, đơn vị văn bản nhỏ nhất (thường là: từ)

3. Loại bỏ stopword

- Bỏ qua các từ và cụm từ không quan trọng xuất hiện thường xuyên nhưng không mang ý nghĩa trong phân tích (thực hiện bằng stoplist)

Tiền xử lý văn bản (Text preprocessing)

4. Stemming (Rút gọn từ)

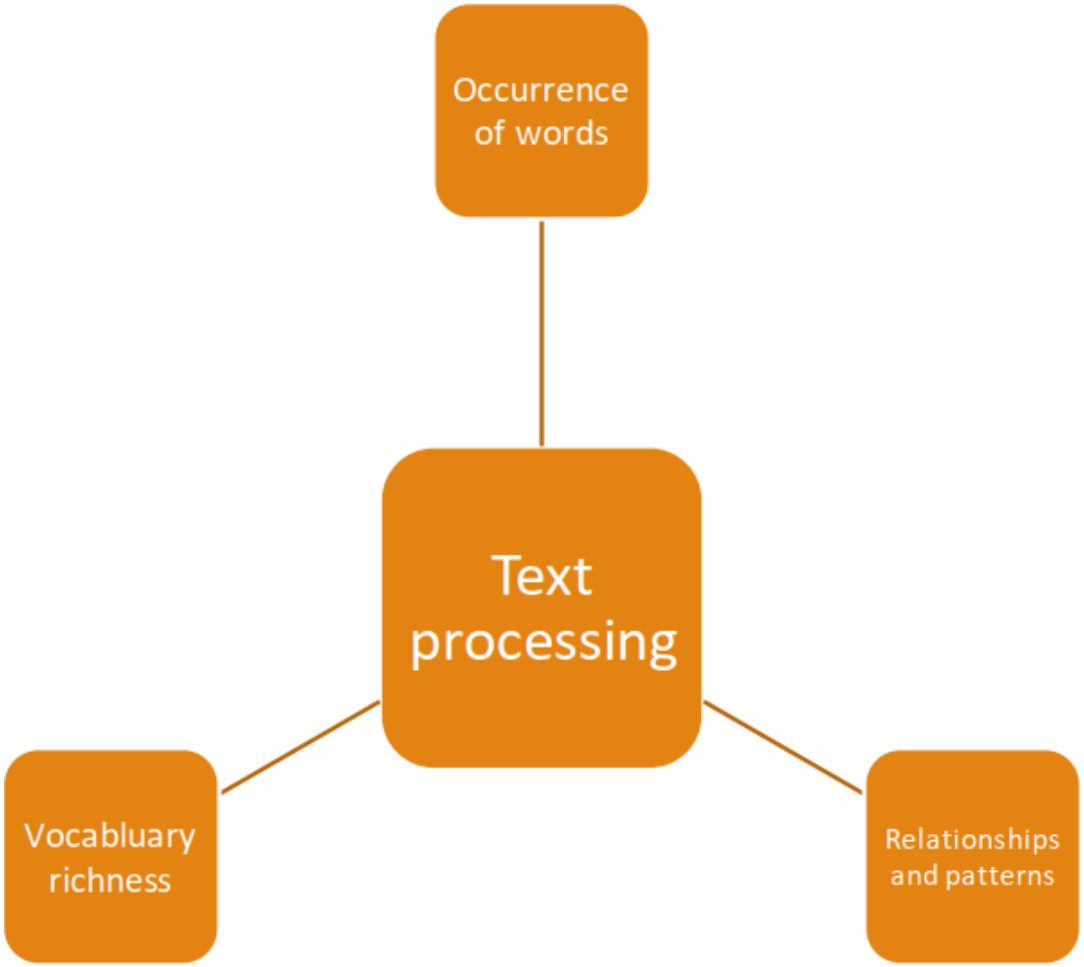
- Quá trình rút gọn các từ biến thể (hoặc đôi khi là từ phái sinh) về gốc từ, từ cơ bản hoặc gốc ngôn ngữ (còn gọi là lemma = lemmatization), đưa các từ về dạng ngữ pháp cơ bản

5. Gắn thẻ (Tagging)

- Gắn thẻ từ loại, xác định một từ thuộc loại từ nào

6. Từ đồng nghĩa (Synonyms)

- Sử dụng từ điển đồng nghĩa để nhóm các từ



Natural language processing

Biểu diễn tài liệu (Document representation)

- ***Biểu diễn tài liệu kiểu "túi từ“ (bag of words)***
  - Từ khóa
  - Tài liệu được biểu diễn bằng một tập hợp các từ khóa mô tả nội dung của tài liệu (thường được tác giả của tài liệu viết bằng tay).
  - Việc tìm kiếm tài liệu được thực hiện bằng cách cung cấp các từ khóa.

Ví dụ:

Một bài báo về "biến đổi khí hậu" có thể được gán các từ khóa: "nhiệt độ", "khí thải", "băng tan", "nóng lên toàn cầu"

Khi tìm kiếm với từ khóa "nhiệt độ", bài báo này sẽ được hiển thị

Biểu diễn tài liệu (Document representation)

Biểu diễn tài liệu trong không gian vector

Ma trận tần suất từ

Tài liệu văn bản được biểu diễn bằng vector tần suất xuất hiện của các từ khóa, và tất cả các vector được tập hợp trong Ma trận Tần suất Từ.

Ví dụ:

Giả sử có 3 từ khóa: "nhiệt độ", "khí thải", "băng tan“

Một tài liệu có thể được biểu diễn bằng vector: [5, 3, 2]

"nhiệt độ" xuất hiện 5 lần"khí thải" xuất hiện 3 lần"băng tan" xuất hiện 2 lầ

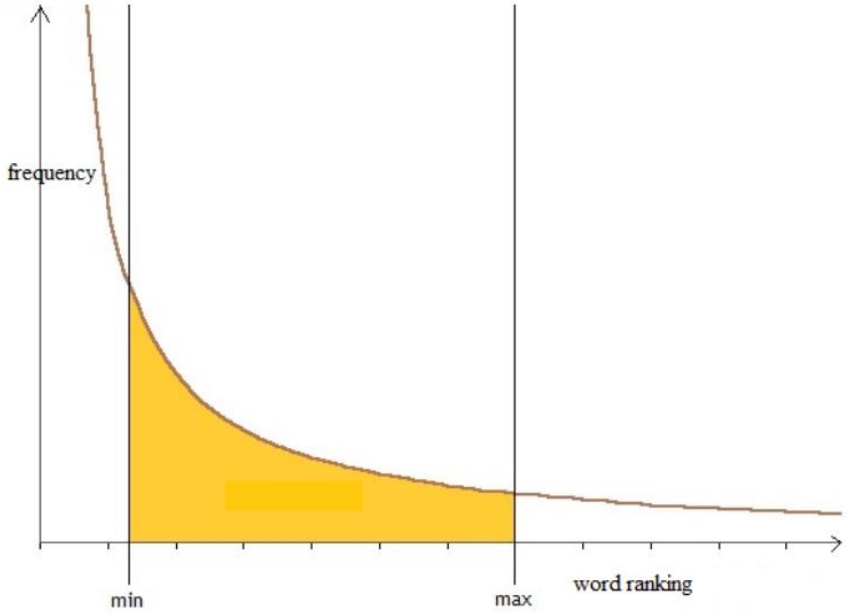
Doc 1	Coca-Cola announced earnings on Tuesday, Jan 12, 2017
Doc 2	Coca-Cola's profits are down as of 12/01/2017

Parsed term	ID	D1	D2
Coca-cola	1	1	1
+announce	2	1	0
+earnings	3	1	0
on	4	1	0
Tuesday	5	1	0
+profit	6	0	1
down	7	0	1
as of	8	0	1
‘s	9	0	1
+be	10	0	1
+[12/01/2017]	11	1	1

Tần suất từ

Thống kê tần suất từ từ bộ sưu tập 46 nghìn bài báo với 19 triệu từ được xác định

Word	Frequency	Word	Frequency	Word	Frequency
the	1,130,021	is	152,483	with	101,210
of	547,311	said	148,302	from	96,900
to	516,635	it	134,323	he	94,585
a	464,736	on	121,173	million	93,515
in	390,819	by	118,863	year	90,104
and	387,703	as	109,135	its	86,774
that	204,351	at	101,779	be	85,588
for	199,340	mr	101,679	was	83,398



Zipf’s law

Định nghĩa cơ bản về tần suất và thứ hạng:

Tần suất của từ trong văn bản tỷ lệ nghịch với thứ hạng của nó (vị trí trong bảng xếp hạng)

Tần suất là số lần từ đó xuất hiện trong văn bản

Zipf’s law

Quy luật xếp hạng:

- Từ phổ biến nhất có hạng = 1
- Từ phổ biến thứ hai có hạng = 2
- Từ tiếp theo có hạng = 3
- Và cứ tiếp tục như vậy...

Zipf's law

Phân tích về từ loại thường gặp:

- Các từ xuất hiện vài trăm lần chủ yếu là:
  - Đại từ
  - Giới từ
  - Các từ loại khác hầu như không mang thông tin ngữ nghĩa

Về bảng xếp hạng:

- Là danh sách các từ được sắp xếp giảm dần theo số lần xuất hiện trong văn bản
- Thể hiện mối quan hệ giữa tần suất và thứ hạng của từ

Định luật Lotka (1926) - Luật về năng suất khoa học (phân bố trích dẫn bài báo khoa học)**

- Một phần đáng kể các công bố khoa học là công trình của một số nhỏ các nhà khoa học có hiệu suất đặc biệt cao

- Ví dụ: 20% nhà khoa học có thể tạo ra 80% số công bố

" (một số ít đối tượng chiếm phần lớn tác động)

Định luật Pareto (1951) - Phân bố thu nhập dân số (nguyên tắc 80/20)**

- 20% thành viên giàu nhất trong xã hội (nhóm một phần năm trên) tạo ra 80% thu nhập

- Ví dụ: Trong một công ty, 20% khách hàng có thể tạo ra 80% doanh thu

3. Định luật Gibrat (1931) - Phân bố quy mô thành phố

- Thành phố càng nhỏ, số lượng càng nhiều
- Ví dụ:
  + Thành phố lớn (>1 triệu dân): rất ít
  + Thành phố trung bình: nhiều hơn
  + Thành phố nhỏ: rất nhiều

Đặc điểm chung:

- Tất cả đều tuân theo mô hình phân bố không đồng đều
- Có mối quan hệ tỷ lệ nghịch giữa quy mô và số lượng
- Thường tuân theo quy tắc "thiểu số chiếm đa số" (một số ít đối tượng chiếm phần lớn tác động)

Stoplist (danh sách từ dừng)

Stoplist (Danh sách từ dừng): các từ có nghĩa không đáng kể được tập hợp dưới dạng bảng

Mục đích: cho phép bỏ qua các từ cụ thể, giúp tăng tốc quá trình phân tích văn bản và mang lại kết quả tốt hơn

TERM
a
about
according
accordingly
actually
after
afterwards
against
ain
almost
along
already
also
although
am
among
amongst
an

Ví dụ stoplist trong tiếng Việt:

Ví dụ áp dụng:

- Câu gốc: "Tôi đã đến trường và tôi gặp bạn của tôi ở đó"
- Sau khi loại bỏ stopwords: "đến trường gặp bạn" -> Câu ngắn gọn hơn, giữ lại các từ có ý nghĩa chính

1. **Từ nối**:
   - và, hay, hoặc, nhưng, vì, bởi vì
   - với, cùng, của
2. **Đại từ**:
   - tôi, bạn, anh, chị, nó
   - này, kia, ấy, đó
3. **Trợ từ**:
   - đã, đang, sẽ
   - rất, lắm, quá
4. **Mạo từ**:
   - những, các, một, mỗi

Stemming (Rút gọn từ)

Root	Term
reach	reaches, reached, reaching
big	bigger, biggest
aller	vais, vas, va, allons, allez, vont

Stemming (Rút gọn từ)

Stemming (Rút gọn từ) là quá trình tạo ra các biến thể hình thái của một từ gốc/từ cơ bản

Mục tiêu của cả stemming và lemmatization là giảm các dạng biến đổi và đôi khi là các dạng phái sinh của một từ về một dạng cơ bản chung.

Có hai lỗi chính trong stemming - rút gọn quá mức và rút gọn thiếu:

- Rút gọn quá mức (Over-stemming) xảy ra khi hai từ được rút gọn về cùng một gốc (nhưng thực tế chúng có gốc khác nhau)
- Rút gọn thiếu (Under-stemming) xảy ra khi hai từ được rút gọn thành các gốc khác nhau (nhưng thực tế chúng có cùng gốc)

Stemming (Rút gọn từ)

Stemming bình thường:
"học tập" -> "học"
"học hành" -> "học"
Over-stemming (rút gọn quá mức):
"university" -> "univers"
"universe" -> "univers" (Hai từ khác nghĩa bị rút gọn về cùng gốc)
Under-stemming (rút gọn thiếu):
"connect" -> "connect"
"connected" -> "connect"
"connection" -> "connection"
(Các từ cùng gốc nhưng không được rút gọn về cùng một dạng)
Trong tiếng Việt:
Rút gọn đúng: "viết", "viết lách" -> "viết"
Over-stemming: "sinh viên", "sinh học" -> "sinh"
Under-stemming: "học", "học tập", "học hành" -> không rút gọn về "học"

Từ đồng nghĩa (Synonyms)

Các từ đồng nghĩa không có chung một dạng biến đổi ngữ pháp, nhưng chúng mang cùng một thông tin.

- Ví dụ, từ "dạy" có các từ đồng nghĩa:
- giảng dạy, giáo dục, đào tạo, huấn luyện, thuyết giảng

Các đặc điểm chính:

- Các từ này mang nghĩa tương tự nhau
- Chúng không phải là các dạng biến đổi của cùng một từ gốc
- Có thể thay thế cho nhau trong nhiều ngữ cảnh
- Giúp mở rộng phạm vi tìm kiếm và phân tích văn bản