TRƯỜNG ĐẠI HỌC KHOA HỌC
KHOA CÔNG NGHỆ THÔNG TIN

BÀI BÁO CÁO

Đề tài: Trích xuất concepts (classes) trong corpus Du lịch. Bằng phương pháp thống kê truyền thống.

Mã học phần: **TIN4633**

Học phần**: Xử lý ngôn ngữ tự nhiên**

Nhóm sinh viên thực hiện:

19T1021253 – Huỳnh Trọng Thiện (NT)

18T1021215 – Võ Chí Nhân

19T1021118 – Đinh Bộ Lĩnh

Huế, tháng 10 năm 2021

1. Các khái niệm

- **Corpus:** kho ngữ liệu, được hiểu đơn giản như là nguồn dữ liệu gồm tập hợp các văn bản thuộc về một lĩnh vực.
- **Domain:** lĩnh vực.
- **Concept:** trong trường hợp này được hiểu như là đối tượng, từ đặc trưng, từ khoá của một văn bản. Trong bài toán này là các danh từ, không bao gồm tên riêng.

- **NN**: Nhãn được gán cho danh từ số ít. *Ví dụ: 'desk'*
- **NNS**: Nhãn được gán cho danh từ số nhiều. *Ví dụ: 'desks'*
- **NLTK:** là bộ công cụ xử lý ngôn ngữ tự nhiên, nó chứa các thư viện xử lý ngôn ngữ. Phục vụ trong các chương trình Python với các bài toán xử lý ngôn ngữ tự nhiên như mã hoá, phân tích cú pháp, gán nhãn, lập luận ngữ nghĩa….

2. Bài toán

Trích xuất concepts (classes) trong corpus Du lịch. Bằng phương pháp thống kê truyền thống.

Input: Corpus thuộc Domain Du lịch *(1800 file *.txt)*

Output: Concepts trong Copurs *(file .csv chứa các concept với tần suất xuất hiện tương ứng)*

3. Ý tưởng thực hiện

Bước 1: Tách biệt các câu trong mỗi đoạn văn thành các từ riêng biệt.

Bước 2: Gán nhãn từng từ.

Bước 3: Thu góp các từ có nhãn 'NN' hoặc 'NNS'. Lưu ý, nếu từ đó có nhãn 'NNS' thì cần chuẩn hoá hình thái từ.

Bước 4: Đếm tần suất xuất hiện của các từ vừa thu góp được.

4. Thuật toán và cài đặt

- Thuật toán:

Bước 1: Đọc file trong Corpus. Nếu đã hết file, sang bước 8.

Bước 2: Đọc câu trong file. Nếu không còn câu, quay lại Bước 1.

Bước 3: Đọc từ trong câu. Nếu đã hết từ trong câu, quay lại Bước 2.

Bước 4: Gán nhãn cho từ vừa đọc được.

Bước 5:

- Nếu là nhãn **'NNS'**, sang Bước 6.
- Nếu là nhãn **'NNS'**, sang Bước 7.

Bước 6: Lưu từ vào file **word.txt** (chứa các concept). Quay lại Bước 3.

Bước 7: Chuẩn hoá hình thái của từ nhờ hàm WordNetLemmatizer.*lemmatize()* Sang bước 6.

Bước 8: Duyệt file **word.txt** và tiến hành đếm tần suất xuất hiện của các concept. Kết quả lưu vào file **outPut.csv** *sau đó*, **Kết thúc./.**

- **Cài đặt:**

Phần cài đặt được mô tả ở link git hub:

5. Phương pháp khác (Yake)

5.1 Giới thiệu

Yake là một phương pháp trích xuất từ khoá tự động không giám sát, không phụ thuộc vào từ điển, chỉ thực thi trên từng tài liệu riêng lẻ.

5.2 Mô tả thuật toán

Bước 1: Bao gồm các bước tiền xử lý dữ liệu.

Bước 2: Xác định các thuật ngữ có khả năng là ứng viên cho concept.

Bước 3: Tính toán, đánh giá điểm cho các ứng viên *(ở đây nhóm em rất cần sự chia sẻ của cô và các bạn, để có cơ hội hiểu hơn về điểm tính toán này)*

Bước 4: Xử lý dữ liệu trùng lặp và xếp hạng. Với các ứng viên có điểm càng thấp thì sẽ càng có liên quan.

6. Đánh giá giữa hai phương pháp

Nội dung	Phương pháp truyền thống	Phương pháp Yake
Đối tượng đầu vào	Kho ngữ liệu	Một văn bản độc lập
Gán nhãn	Gán nhãn trên từng từ, dựa vào từ điển, không dựa vào ngữ cảnh	Gán nhãn trên cụm từ có sử dụng mô hình n-gam. Không dựa vào từ điển, có dựa vào ngữ cảnh
Đánh giá	Dựa vào tần suất xuất hiện của từng từ được xem là concept của văn bản	Đánh giá theo các chỉ số riêng của thuật toán dựa trên các đặc điểm của các concept, không phải dựa vào tần suất xuất hiện
Đầu ra	Concept là từng từ	Concept là từng từ hoặc cụm từ

Những thông tin trên được xây dựng trên sự tìm hiểu và trao đổi giữa các thành viên trong nhóm, thật khó để tránh khỏi sai, thiếu xót. Rất mong cô và các bạn bỏ quá và đóng góp thêm ý kiến để bài báo cáo của nhóm được hoàn thiện hơn. Nhóm chúng em xin chân thành cảm ơn!