# Kế hoạch thực hiện

# Bảng phân công công việc và đánh giá

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MSSV** | **Thành viên** | **Danh sách công việc** | **Mức độ hoàn thành** | **Tổng hợp đánh giá chéo** |
| 16520724 | Bùi Nguyên Mão |  |  |  |
| 2 | Cao Văn Nam |  |  |  |
| 3 | Thừa Lê Thanh Chương |  |  |  |
| 4 | Bùi Văn Thuận |  |  |  |

# Nội dung

## Giới thiệu đề tài:

Công nghệ thông tin hiện nay đã mở ra một thời kỳ phát triển mạnh mẽ của xã hội tri thức, nơi mà tất cả mọi người đều có thể tiếp cận dễ dàng với tin tức bất kể không, thời gian. Tuy nhiên, mặt trái của nó đó là sự phát tán các tin tức nhảm, sai lệch vẫn luôn hiện diện như một căn bệnh từng bước đẩy con người vào bước đường lạc lối. Với số lượng tin bài được xuất hiện mỗi giây, rất khó có một tổ chức hay các nhân nào có thể kiểm tra, rà soát được sự đúng đắn hay nội dung vì vậy đây là một vấn đề nhức nhối mà cần phải có biện pháp để giải quyết.

Sự phát triển của machine learning và natural language processing như một vị cứu tinh cho đứa con chiên đang lạc lối. Với những công nghệ này, con người có thể giải quyết được một phần vấn đề trên.

Đề tài này chỉ là một bước tìm hiểu nhỏ nhưng sẽ mang lại một cái nhìn rõ hơn cho mọi người về cách mà NLP đã và đang tiếp tục hoạt động, phát triển để góp phần xây dựng một môi trường dữ liệu trong sạch.

*Bài toán đặt ra:*

kenh12 sau khi nhận đươc nhiều báo cáo vì số lượng tin bài giả đang gia tăng nghiêm trọng trên mạng xã hội của mình đã quyết đinh phải kiểm duyệt tất cả tin tức bài đăng, nhưng do số lượng tin bài nhận được mỗi ngày quá nhiều, bộ phận kiểm tra gồm 100 nhân viên làm ngày làm đêm vẫn không thể hoàn thành được nhiệm vụ. Một cuộc họp khẩn giữa 12 giờ đêm đã được diễn ra, cuối cùng cũng đưa ra một phương pháp có thể giải quyết được vấn đề trên.

*Giải pháp:*

* Sử dụng máy tính để làm việc kiểm duyệt thay cho con người.
* Xây dựng bộ dữ liệu lớn để máy phân loại được các bài viết theo tiêu chí các bài sử dụng các từ tương đồng nhau thường có mục đích như nhau.
* Để tối ưu chi phí và nguồn lực hiện có, tạm thời chỉ kiểm tra với tiêu đề của bài viết.

## Dữ liệu - Phân tích - Hướng giải quyết:

### Bộ dữ liệu

Đề tài sử dụng bộ dữ liệu tiêu đề tin tức ở nguồn: [*https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection*](https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection)

Thông tin bộ dữ liệu:

* Dữ liệu được thu thập ở 2 trang: *theonion* chuyên viết các bài châm biếm, mỉa mai và *huffingtonpost* chuyên viết tin chính thống.
* Gồm 26709 dòng dữ liệu kiểu JSON, có các thông tin: link bài viết, tiêu đề và nhãn (có phải tin bài mỉa mai)
* Các thông tin tiêu đề đều bằng tiếng Anh.

### Hướng giải quyết

Bài toán trở thành một bài toán phân lớp dữ liệu ngôn ngữ tự nhiên. Máy tính chỉ có thể xử lý với các con số, nên cần có bước đầu tiên là bước chuyển dữ liệu chữ sang số. Có một số kỹ thuật để giải quyết vấn đề này nhưng ở đây, chúng ta sẽ chọn cách đơn giản là Bag of Words. Kỹ thuật này sẽ xây dựng một bộ dữ liệu tất cả các từ được sử dụng rồi từ đó sẽ tạo ra các vector với số chiều là số lượng từ tương ứng với mỗi chuỗi đầu vào.

Tuy nhiên, trước đó dữ liệu văn bản cần một bước làm sạch với các quá trình chuẩn hóa từ và loại bỏ các từ không cần thiết. Trong phần lớn các ngôn ngữ, có một số từ được thêm vào câu với mục đích tăng sắc thái biểu đạt hoặc cảm xúc, những từ này cần được loại bỏ để giảm số lượng chiều của vector đầu vào, tăng hiệu quả mô hình.

Sau khi đã chuyển dữ liệu sang dạng số, các vector ở đây là vector thưa có chiều rất lớn, vì vậy để tăng hiệu quả quá trình học và giảm thời gian chạy cần có biện pháp giảm số chiều dữ liệu nhưng vẫn phải cố gắng giữ được tương quan giữa các dữ liệu với nhau.

Sau khi xong phần chuẩn bị dữ liệu và tiền xử lý, cần lựa chọn thuật toán phù hợp. Với dạng bài toán binary classification, có khá nhiều phương pháp để xử lý như regression, knn, decision tree, svm,… Ở đây, chúng ta dùng naive bayes.

## Tiền xử lý

Đây là giai đoạn cực kỳ quan trọng của quá trình học máy, thường chiếm một phần lớn thời gian của cả quá trình. Với bài toán trên, sẽ có những bước tiền xử lý sau:

### Chuẩn hóa dữ liệu và tách từ:

Dữ liệu hiện tại đã được chuẩn hóa về nội dung, cấu trúc, không có trùng lặp, vô lý,…

Tùy thuộc vào loại dữ liệu, loại ngôn ngữ và mục đích sử dụng, có một số cách để chia nhỏ nội dung với các đặc điểm sau:

* Tùy theo loại ngôn ngữ mà cấu trúc của phần con được chia ra cho phù hợp nghĩa, ví dụ trong tiếng Anh, mỗi câu gồm nhiều từ, một từ thường cách nhau theo khoảng trắng, nhưng trong tiếng Việt, câu được tạo thành từ nhiều tiếng, một từ có thể phải được kết hợp từ nhiều tiếng nên việc tách câu theo khoảng trằng là không đảm bảo nội dung. Bên cạnh đó, tùy thuộc yêu cầu bài toán mà phải tách từ hoặc cụm từ, cụm cấu trúc (chủ ngữ, vị ngữ),…
* Có nhiều kiểu tách từ trong tiếng Anh, tuy nhiên khó có cách nào có thể tách trọn vẹn, hoàn chỉnh nội dung của các câu bởi số lượng từ vựng và cụm từ lớn, các dấu câu mang các ý nghĩa khác nhau tùy thuộc cách sử dụng,

Với dữ liệu ở đây, cách đơn giản nhất được sử dụng là loại bỏ tất cả dấu và ký tự đặc biệt tách từ đơn theo khoảng trắng.

### Hoàn nguyên từ:

Với một số ngôn ngữ trong đó có tiếng Anh, các từ trong quá trình sử dụng có sự biến đổi thành các dạng khác nhau nhưng phần lớn tạo ra ý nghĩa khá là gần nhau.

ví dụ: learn → learns → learning

Trong máy học với công nghệ hiện tại, các sự biến đổi này là không cần thiết, nó chỉ làm tăng thêm độ phức tạp mà máy cần phải xử lý vì vậy cần có một bước để đưa các từ về với dạng gốc ban đầu.

Có 2 phương pháp đưa từ về dạng gốc:

* stemming là một cách hoàn nguyên từ theo kiểu đánh đồng, các từ với các kết cấu giống nhau sẽ được đưa về dạng giống nhau phần lớn theo dạng bỏ bớt ký tự. ví dụ: learning → learn, machines→ machin, studies → studi
* lemmatizing là một phiên bản nâng cấp của stemming khi các từ tạo ra phải qua quá trình kiểm tra lại với từ điển để đảm bảo sự chính xác. Tất nhiên cái giá phải trả là thời gian và quá trình xử lý. ví dụ: machines → machine, studies → study.

Cả 2 phương pháp đều có thể được sử dụng tùy theo mục đích và yêu cầu. Ở bài toán này, chúng tôi chọn lemmatizing.

Do lemmatizing yêu cầu một bộ từ điển để xác thực từ được tạo ra nên chúng tôi sử dụng bộ wordnet của thư viện nltk.

### Loại bỏ stop words

Trong các câu văn, có những từ chứa ít dữ liệu cho quá trình xử lý phân lớp sẽ gây nhiễu rất lớn đối với mô hình BoW.

Các stop words trong tiếng anh thường là giới từ, từ nối,... như *the, there, an, anyway, hereby,...*

Ví dụ trong tiếng Anh:

*I hope that, when I've built up my savings, I'll be able to travel to Mexico.*

Câu này nếu được rút gọn lại với các từ: hope, built, up, savings, able, travel, Mexico. thì cơ bản nghĩa vẫn sẽ được dữ lại. Quá trình học máy chỉ cần có thế thôi.

### Tạo vector từ dữ liệu:

## Huấn luyện với naive bayes

### Lý do chọn naive bayes

### Hiện thực với python

## Kết quả, đánh giá

### Phân tích kết quả

### Kết luận

## Hướng phát triển

# Tài liệu tham khảo