TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

──────── \* ───────



ĐỒ ÁN

**TỐT NGHIỆP ĐẠI HỌC**

NGÀNH CÔNG NGHỆ THÔNG TIN

**XÂY DỰNG ỨNG DỤNG HỌC TIẾNG ANH LOXO**

Sinh viên thực hiện : **Nguyễn Song Hào - 20121630**

Lớp CNTT 2.04 – K57

Giáo viên hướng dẫn: ThS. **Lê Thanh Hùng**

# PHIẾU GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

1**. Thông tin về sinh viên**

Họ và tên sinh viên: **Nguyễn Song Hào**

Điện thoại liên lạc: 0975030714 Email: nguyensonghao974@gmail.com

Lớp: CNTT 2.04- K57 Hệ đào tạo: Chính quy

Đồ án tốt nghiệp được thực hiện tại: **Bộ môn công nghệ phần mềm – Viện CNTT và TT**

Thời gian làm ĐATN: Từ ngày 14 /02 /2017 đến 27 /05 /2017

2**. Mục đích nội dung của ĐATN**

Xây dựng ứng dụng học tiếng Anh trên nền tảng Android và iOS

3. **Các nhiệm vụ cụ thể của ĐATN**

* Tìm hiểu ứng dụng đa nền tảng và Ionic 2
* Tích hợp công nghệ nghiên cứu, xây dựng và phát triển ứng dụng học tiếng Anh
* Cài đặt và kiểm thử

4**. Lời cam đoan của sinh viên**

Tôi – ***Nguyễn Song Hào*** - cam kết ĐATN là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của *Ths. Lê Thanh Hùng*.

Các kết quả nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác.

|  |  |
| --- | --- |
|  | *Hà Nội, ngày 27 tháng 05 năm 2017*  Tác giả ĐATN  *Nguyễn Song Hào* |

**5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành của ĐATN và cho phép bảo vệ:**

|  |  |
| --- | --- |
|  | *Hà Nội, ngày 27 tháng 05 năm 2017*  Giáo viên hướng dẫn  *Ths Lê Thanh Hùng* |

# TÓM TẮT NỘI DUNG ĐỒ ÁN TỐT NGHIỆP

Ngày nay, công nghệ càng trở nên gần gũi với con người nhất là trong giai đoạn phát triển rực rỡ của smart phone. Mọi người đều chỉ cần một chiếc điện thoại thông minh và mạng internet là đã có thể chia sẻ và kết nối với toàn thế giới, tạo ra một khối lượng lớn dữ liệu. Nhu cầu của người dùng sẽ ngày càng yêu cầu những ứng dụng thông minh hơn, có thể hỗ trợ họ tốt hơn trong việc sử dụng và tương tác, đặc biệt là ở giai đoạn phát triển của Internet of Thing (IoT).

Trong giai đoạn phát triển của những ứng dụng thông minh và IoT, vai trò của AI và học máy là vô cùng quan trọng vì chúng sẽ mang sự thông minh, mang những khả năng gần giống con người vào phần mềm để hỗ trợ người dùng, đó sẽ là những khả năng xử lý âm thanh, ngôn ngữ, hình ảnh,…tạo ra phương pháp giao tiếp mới giữa con người với con người và con người với máy.

Đồ án này sẽ nghiên cứu về deep learning là một hệ thống các phương pháp học máy, trong thời gian gần đây đã đạt được những bước phát triển thần tốc và kỳ diệu ở lĩnh vực học máy và trí tuệ nhân tạo, đồ án sẽ ứng dụng trực tiếp phương pháp để giải quyết bài toán phân tích sắc thái tình cảm (phân tích quan điểm người nói) trong ngôn ngữ tiếng Việt.

Ngoài ra đồ án này sẽ giới thiệu bộ công cụ fastText của Facebook, một bộ công cụ hữu dụng cho việc phân loại và xử lý văn bản.

Đồ án được chia thành 4 chương:

Chương 1: Tổng quan về phân tích quan điểm người dùng

Chương 2: Cơ sở lý thuyết của đồ án

Chương 3: Xây dựng mô hình phân tích sắc thái

Chương 4: Cài đặt mô hình và đánh giá kết quả

Chương 5: Áp dụng mô hình huấn luyện cho bài toán cụ thể

# LỜI CẢM ƠN

Lời đầu tiên, em xin được gửi lời cảm ơn chân thành đến các thầy cô trong trường Đại Học Bách Khoa Hà Nội cũng như các thầy cô giáo trong Viện Công Nghệ Thông Tin và Truyền Thông đã truyền dạy cho em những kiến thức và kinh nghiệm quý báu trong suốt quá trình học tập và tu dưỡng trong 5 năm qua.

Em xin được gửi lời cảm ơn chân thành tới thầy Ths. Lê Thanh Hùng - Giảng viên bộ môn công nghệ phần mềm, Viện Công Nghệ Thông tin và Truyền Thông đã giúp đỡ, hướng dẫn và chỉ dạy em trong quá trình học tập nghiên cứu với những lời khuyên, bài học bổ ích.

Tôi cũng muốn gửi lời cảm ơn trân thành đến tập thể lớp CNTT 2.04 – K57, là một tập thể đoàn kết, cùng nhau phát triển, tạo điều kiện cho các thành viên trong lớp học hỏi, cùng nhau trau dồi thêm những kiến thức mới.

Cuối cùng, em xin được gửi lời cảm ơn chân thành tới gia đình, bạn bè đã quan tâm, động viên, đóng góp ý kiến giúp đỡ em trong quá trình học tập, nghiên cứu và hoàn thành đồ án tốt nghiệp.

Tuy nhiên, do thời gian và trình độ có hạn nên đồ án không thể tránh khỏi những thiếu sót. Chính vì vậy, em rất mong có được sự góp ý từ các thầy cô giáo và toàn thể các bạn.

Em xin chân thành cảm ơn!

Hà Nội - Ngày 27 Tháng 05 Năm 2017

Nguyễn Song Hào

CNTT 2.04 – K57

Mục lục

[PHIẾU GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP 2](#_Toc470248900)

[TÓM TẮT NỘI DUNG ĐỒ ÁN TỐT NGHIỆP 4](#_Toc470248901)

[LỜI CẢM ƠN 5](#_Toc470248902)

[DANH SÁCH CÁC HÌNH ẢNH 10](#_Toc470248903)

[DANH SÁCH CÁC BẢNG 11](#_Toc470248904)

[DANH MỤC THUẬT NGỮ 12](#_Toc470248905)

[CHƯƠNG 1 : TỔNG QUAN VỀ PHÂN TÍCH QUAN ĐIỂM NGƯỜI DÙNG 13](#_Toc470248906)

[1.1 Nhu cầu hiện tại của bài toán 13](#_Toc470248907)

[1.2 Các giải pháp hiện tại cho bài toán 14](#_Toc470248908)

[*1.2.1* *Các ứng dụng phân tích sắc thái văn bản trên thế giới* 14](#_Toc470248909)

[*1.2.1.1* *Sản phẩm Sysomos* 14](#_Toc470248910)

[*1.2.1.2* *Sản phẩm Visible* 16](#_Toc470248911)

[*1.2.1.3* *Sản phẩm Oracle Social Cloud* 17](#_Toc470248912)

[*1.2.1.4* *Đánh giá xếp hạng tổng quan Top 9 sản phẩm.* 19](#_Toc470248913)

[*1.2.2* *Các ứng dụng phân tích sắc thái văn bản tại Việt Nam* 20](#_Toc470248914)

[*1.2.2.1* *Social heat* 20](#_Toc470248915)

[*1.2.2.2* *SMCC* 21](#_Toc470248916)

[1.3 Đề xuất giải pháp và mục tiêu của đồ án 22](#_Toc470248917)

[*1.3.1* *Đề xuất giải pháp cho đồ án* 22](#_Toc470248918)

[*1.3.2* *Mục tiêu của đồ án* 23](#_Toc470248919)

[CHƯƠNG 2 : CƠ SỞ LÝ THUYẾT CỦA ĐỒ ÁN 25](#_Toc470248920)

[2.1 Sentiment Analysis 25](#_Toc470248921)

[*2.1.1* *Xử lý ngôn ngữ tự nhiên* 25](#_Toc470248922)

[*2.1.2* *Các hướng tiếp cận bài toán truyền thống* 28](#_Toc470248923)

[*2.1.3* *Bước đột phá mới cho bài toán khi mạng neural phát triển.* 29](#_Toc470248924)

[2.2 Machine Learning 30](#_Toc470248925)

[*2.2.1* *Thuật toán học máy* 31](#_Toc470248926)

[*2.2.2* *Các phương pháp học máy* 32](#_Toc470248927)

[*2.2.2.1* *Học có giám sát* 33](#_Toc470248928)

[*2.2.2.2* *Học không giám sát* 34](#_Toc470248929)

[*2.2.2.3* *Học bán giám sát* 35](#_Toc470248930)

[*2.2.2.4* *Học tăng cường* 35](#_Toc470248931)

[2.3 Deep Learning 36](#_Toc470248932)

[*2.3.1* *Neural Network* 37](#_Toc470248933)

[*2.3.1.1* *Kiến trúc tổng quan của một mạng neural* 38](#_Toc470248934)

[*2.3.1.2* *Quá trình xử lý thông tin của một mạng neural* 39](#_Toc470248935)

[*2.3.1.3* *Quá trình học của một mạng neural* 41](#_Toc470248936)

[2.3.1.3.1 Nguyên tắc huấn luyện 42](#_Toc470248937)

[2.3.1.3.2 Lan truyền thẳng và lan truyền ngược 42](#_Toc470248938)

[*2.3.2* *Mạng neural nhân chập (Convolution neural network)* 44](#_Toc470248939)

[*2.3.2.1* *Convolution – tích chập* 45](#_Toc470248940)

[*2.3.2.2* *Kiến trúc của một mạng CNN* 45](#_Toc470248941)

[CHƯƠNG 3 : XÂY DỰNG MÔ HÌNH PHÂN TÍCH SẮC THÁI 48](#_Toc470248942)

[3.1 Word Embedding 49](#_Toc470248943)

[*3.1.2* *Huấn luyện word vectors* 51](#_Toc470248944)

[*3.1.2.1* *Neural Network Language Models* 51](#_Toc470248945)

[*3.1.2.2* *Mô hình Continuos Bag-of-words (CBOW)* 54](#_Toc470248946)

[*3.1.2.3* *Mô hình Continuos Skip-gram* 56](#_Toc470248947)

[*3.1.2* *Bổ sung word vector với thông tin của các subword* 57](#_Toc470248948)

[*3.1.2.1* *Biểu diễn hình thái học của từ* 58](#_Toc470248949)

[*3.1.2.2* *Các đặc trưng mức ký tự cho NLP* 58](#_Toc470248950)

[*3.1.2.3* *Mô hình subword* 59](#_Toc470248951)

[*3.1.2.4* *Từ điển n-gram ký tự* 59](#_Toc470248952)

[3.2 Sử dụng mạng neural CNN cho bài toán phân loại văn bản 60](#_Toc470248953)

[*3.2.1* *Kiến trúc mô hình* 61](#_Toc470248954)

[*3.2.2* *Chuẩn hóa* 62](#_Toc470248955)

[*3.2.3* *Các siêu tham số và quá trình huấn huyện* 62](#_Toc470248956)

[3.3 Ứng dụng thư viện fastText cho bài toán phân tích sắc thái. 63](#_Toc470248957)

[*3.3.1* *fastText* 64](#_Toc470248958)

[*3.3.2* *Phân loại văn bản* 64](#_Toc470248959)

[CHƯƠNG 4 : CÀI ĐẶT MÔ HÌNH VÀ ĐÁNH GIÁ KẾT QUẢ 67](#_Toc470248960)

[4.1 Xây dựng bộ dữ liệu 67](#_Toc470248961)

[*4.1.1* *Dữ liệu học mô hình ngôn ngữ* 67](#_Toc470248962)

[*4.1.2* *Dữ liệu học mô hình phân tích sắc thái* 68](#_Toc470248963)

[4.2 Cài đặt mô hình word embedding 69](#_Toc470248964)

[4.3 Cài đặt các mô hình 70](#_Toc470248965)

[*4.3.1* *Phương pháp sử dụng CNN* 70](#_Toc470248966)

[*4.3.2* *Phương pháp sử dụng thư viện fastText* 71](#_Toc470248967)

[4.4 So sánh và đánh giá kết quả. 75](#_Toc470248968)

[CHƯƠNG 5 : ÁP DỤNG MÔ HÌNH HUẤN LUYỆN CHO BÀI TOÁN CỤ THỂ 79](#_Toc470248969)

[5.1 Bài toán thực tế và giải pháp đề xuất 79](#_Toc470248970)

[*5.1.1* *Bài toán thực tế* 79](#_Toc470248971)

[*5.1.2* *Đề xuất giải pháp* 80](#_Toc470248972)

[5.2 Kiển trúc hệ thống 82](#_Toc470248973)

[*5.2.1* *Webhook server* 82](#_Toc470248974)

[*5.2.2* *NLP server* 83](#_Toc470248975)

[*5.2.3* *Elasticsearch server* 85](#_Toc470248976)

[*5.2.4* *Client* 87](#_Toc470248977)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 90](#_Toc470248978)

[TÀI LIỆU THAM KHẢO 92](#_Toc470248979)

[INDEX 94](#_Toc470248980)

# DANH SÁCH CÁC HÌNH ẢNH

Hình 1.1: Giao diện hệ thống Sysomos 15

Hình 1.2: Giao diện hệ thống Visible 16

Hình 1.2: Kiến trúc của sản phẩm Oracle Social Cloud 18

Hình 1.3: Giao diện của sản phẩm Social heat 21

Hình 1.4: Mô hình các chức năng của hệ thống SMCC 22

Hình 1.5: So sánh đánh giá fastText và các mô hình mạng Neural 23

Hình 2.1: Các phương pháp giải quyết bài tóan phân tích sắc thái 29

Hình 2.2: Kiến trúc tổng quan của mạng neural 38

Hình 2.3: Quá trình xử lý thông tin của mạng neural 39

Hình 2.4: Cách tính tổng trọng số ở mỗi tầng ẩn 40

Hình 2.5: Hàm kích hoạt 41

Hình 2.6: Lan truyền ngược 44

Hình 2.7: Phép nhân chập 2 ma trận 45

Hình 2.8: Kiến trúc mạng CNN 46

Hình 3.2: Biểu diễn sự liên quan giữa các vector từ 51

Hình 3.3: Kiến trúc của mô hình ngôn ngữ mạng neural 53

Hình 3.4: Mô hình CBOW 55

Hình 3.5: Mô hình Skip-gram 56

Hình 3.6: Kiến trúc mạng CNN cho bài toán phân tích sắc thái 61

Hình 3.7: Mô hình phân loại văn bản của fastText 65

Hình 5.1: Vấn đề dừng hoạt động của các fanpage 80

Hình 5.2: Kiển trúc hệ thống 82

Hình 5.3: Dữ liệu trả về webhook server từ facebook 83

Hình 5.4: Giao diện đăng nhập qua facebook 87

Hình 5.5: Giao diện hiển thị các post của fanpage 88

Hình 5.6: Giao diện hiển thị comment của post 88

Hình 5.7: Giao diện thống kê các người dùng trong fanpage 89

# DANH SÁCH CÁC BẢNG

[Bảng 1.1: So sánh độ giám sát top 9 sản phẩm trên thế giới 19](#_Toc470123458)

[Bảng 1.2: So sánh độ phân tích top 9 sản phẩm trên thế giới 19](#_Toc470123459)

[Bảng 4.1: Danh sách tham số huấn luyện mô hình word-embedding 70](#_Toc470123460)

[Bảng 4.2: Tham số khởi tạo của mô hình CNN 71](#_Toc470123461)

[Bảng 4.3: Danh sách tham số huấn luyện mô hình phân tích sắc thái 72](#_Toc470123462)

[Bảng 4.5: Giá trị tham số mô hình fastText 1 73](#_Toc470123463)

[Bảng 4.6: Giá trị tham số mô hình fastText 2 74](#_Toc470123464)

[Bảng 4.7: Giá trị tham số mô hình fastText 3 75](#_Toc470123465)

[Bảng 4.8: So sánh độ chính xác giữa các mô hình 76](#_Toc470123466)

[Bảng 4.9: So sánh thời gian huấn luyện giữa các mô hình 76](#_Toc470123467)

[Bảng 4.10: So sánh độ chính xác giữa các mô hình theo [2] 77](#_Toc470123468)

[Bảng 4.11: So sánh thời gian huấn luyện giữa các mô hình theo [2] 77](#_Toc470123469)

[Bảng 5.1: Giá trị tham số huấn luyện mô hình chính trị 84](#_Toc470123470)

[Bảng 5.2: Cấu trúc dữ liệu trên elasticsearch 86](#_Toc470123471)

# DANH MỤC THUẬT NGỮ

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Dạng đầy đủ** | **Giải nghĩa** |
| 1 | SO | Sentiment Orientation | Chỉ số thể hiện quan điểm của văn bản |
| 2 | NLP | Nature Language Processing | Xử lý ngôn ngữ tự nhiên |
| 3 | SVM | Support Machine Vector | Phương pháp học máy dựa vào vector hỗ trợ |
| 4 | IoT | Internet of thing | Mạng lưới vạn vật kết nối Internet hoặc là Mạng lưới thiết bị kết nối Internet |
| 5 | Deep Learning | Deep Learning | Học sâu |
| 6 | Machine Learning | Machine Learning | Học máy |
| 7 | AI | Artificial intelligence | Trí tuệ nhân tạo |
| 8 | ANN | Artificial Neural Network | Mạng nơ-ron nhân tạo |
| 9 | CNN | Convolution neural network | Mạng nơ –ron nhân chập |
| 10 | RNN | Recurrent neural network | Mạng nơ-ron hồi quy |
| 11 | SA | Sentiment analysis | Phân tích sắc thái tình cảm |
| 12 | API | Application Programming Interface | Giao diện lập trình ứng dụng |

# CHƯƠNG 1 : TỔNG QUAN VỀ PHÂN TÍCH QUAN ĐIỂM NGƯỜI DÙNG

|  |
| --- |
| **Nội dung chương này trình bày những vấn đề sau:**   * Nhu cầu phân tích sắc thái văn bản * Các giải pháp hiện tại * Đề xuất giải pháp và mục tiêu của đồ án |

## Nhu cầu hiện tại của bài toán

Với sự phát triển mạnh của công nghệ và đặc biệt là các ứng dụng smart phone cùng với mạng internet đã tạo ra một cuộc bùng nổ dữ liệu trên thế giới. Mỗi ngày hiện nay chúng ta đang tạo ra 2.5 quintillion bytes dữ liêu, 90% dữ liệu hiện nay trên thế giới được xuất hiện trong 2 năm trở lại đây. Một lượng lớn thông tin sẽ được tạo ra mỗi ngày, trong khi con người mỗi ngày chỉ có 24 tiếng để làm việc và đáp ứng những nhu cầu của mình, họ phải cần đến một sự hỗ trợ lớn đến từ máy móc. Với xu hướng phát triển của dữ liệu và các thiết bị thông minh kết nối internet, thì việc tạo ra cho máy móc một trí thông minh sẽ tạo lên mộ sự hỗ trợ không hề nhỏ cho con người trong công việc và cuộc sống hàng ngày.

Một tương lai không xa, mọi thiết bị sẽ đều có kết nối internet và có thể điều khiển bằng smartphone trở lên phổ biến và sẽ đòi hỏi khả năng chúng có thể làm việc và giao tiếp như con người để sự tương tác giữa người và máy trở nên dễ dàng hơn. Và nhu cầu tiếp cận thông tin một cách nhanh chóng và chính xác là cần thiết.