# Giới thiệu Gate Tool

## Khái quát về Gate

* Là một Natural Processing Language (NPL) tookit được phát triển tại đại học Sheffield, nước Anh kể từ năm 1995.
* Là mã nguồn mở Java và được phát hành miễn phí <http://gate.ac.uk>.
* Có sự đóng góp của nhiều cộng đồng phát triển, người dùng, nhà giáo dục, sinh viên và các chuyên gia trong nhiều lĩnh vực khác nhau.
* Nhiệm vụ chính là xử lý ngôn ngữ bao gồm giọng nói, nghiên cứu ung thư, hỗ trợ quyết định, khai thác thông tin web, chú thích ngữ nghĩa …
* Có trên Maven được sử dụng như API, Servlet cho việc phát triển web

## Các tiện ích trong Gate

* Thành phần xử lý ngông ngữ: Parser, machine learning tools, stemmers, IR tool, IE components cho nhiều ngôn ngữ khác nhau như Anh, Đức, Pháp…
* Tool Trực quang hóa, biểu diễn và thao tác văn bản(chú tyhích, ontologies, parse tree..)
* Tool trích xuất nhiều thông tin khác nhau: khám phá những thông tin chưa biết trước đó
* Tool Công cụ đánh giá và tính điểm chuẩn
* **Các thuật toán được sử dụng trong tool là hoàn toàn tách biệt, nó được phát triển một cách độc lập từ nhiều người dùng có chuyên môn khác nhau**

## Lợi ích của Gate

* Tiết kiệm thời gian quản lý văn bản và dữ liệu từ nhiều nguồn khác nhau
* Tìm link ẩn thông qua các lượng lớn thông tin đã định dạng
* Thu thập thông tin và trích xuất các nhân tố mới

## Tải và cài đặt Gate

* Tải Gate tại <http://gate.ac.uk/download/>
* Yêu cầu máy phải cài đặt Java JDK phiên bản tối thiểu là 7 cho Gate ver8, để biết them chi tiết truy cập <https://gate.ac.uk/sale/tao/splitch2.html#x5-170002.1>
* Tuy Gate là một tool nhưng chúng ta không cần cài đặt, sau khi tải về chỉ việc giải nén và Double-Click file Gate.exe là có thể sử dụng được

# Các ứng dụng xử lý trong Gate

## ANNIE

* ANNIE là 1 plugin của Gate, nằm trong IE system (sử dụng các kỹ thuật phân tích để tìm các interesting patterns), được phát triển bởi đội ngũ của Gate.
* Sử dụng ngôn ngữ quy tắc hành động kiểu mẫu hữu hạn JAPE (Java Annotation Patterns Engine) để xử lý các từ vựng.
* Ứng dụng ANNIE bao gồm 1 tập PRs (Processing Resources) như là:
  + English Tokeniser: Tách văn bảng thành các từ có nghĩa dựa vào khoản trắng.
  + Sentence Splitter: Phân văn bản thành các câu dựa vào dấu chấm, \t, \r.
  + POS tagger: Gán nhãn loại từ cho từng vựng dựa vào JAPE
  + Gazetteer: Xây dựng wordlist để thực hiện Named Entity. Kết quả là tập Named Entity được đánh dấu Name entity tagger:
  + Orthographic corefernce: Bổ sung tên cho thực thể được tìm thấy bởi NE transducers

## Các thư viện chuyên xử lý ngôn ngữ khác

* Gate là một tool mà tập hợp nhiều bộ thư viện xử lý ngôn ngữ khác nhau. Do đó, người dùng có thể cùng lúc tạo nhiều ứng dụng với các bộ thư viện ANNIE, OpenNPL, Stanford\_CoreNLP … để so sánh hiệu năng và tính đúng đắng đối với bộ dữ liệu chúng ta cần xử lý.
* Về cơ bản, các bộ thự viện đều có các class để sử lý Tokenizer, Sentence Splitter, POS tagger. Nhưng do nhiều người đóng góp củng như văn hóa dùng ngôn ngữ mà kết quả đạt được của các thuật toán là khác nhau.

## Một số Tagger và thuật toán khác

* Ngoài một số thư viện chuyên sử lý ngôn ngữ, Gate còn tiếp nhận sự đóng góp từ cộng đồng một số thư để phụ vụ việc phân số một số dữ liệu nhất định như:
  + Các Tagger: Tagger\_MetaMap, Tagger\_Date, Tagger\_Number, Tagger\_Chemistry…
  + Các thư viện: Machine Learning, Wordnet…

# Tạo ứng dụng

## Dữ liệu đầu vào của Gate