Introduction to Natural Language Processing (NLP)

Presented By Wing Chan

Agenda

What is NLP?

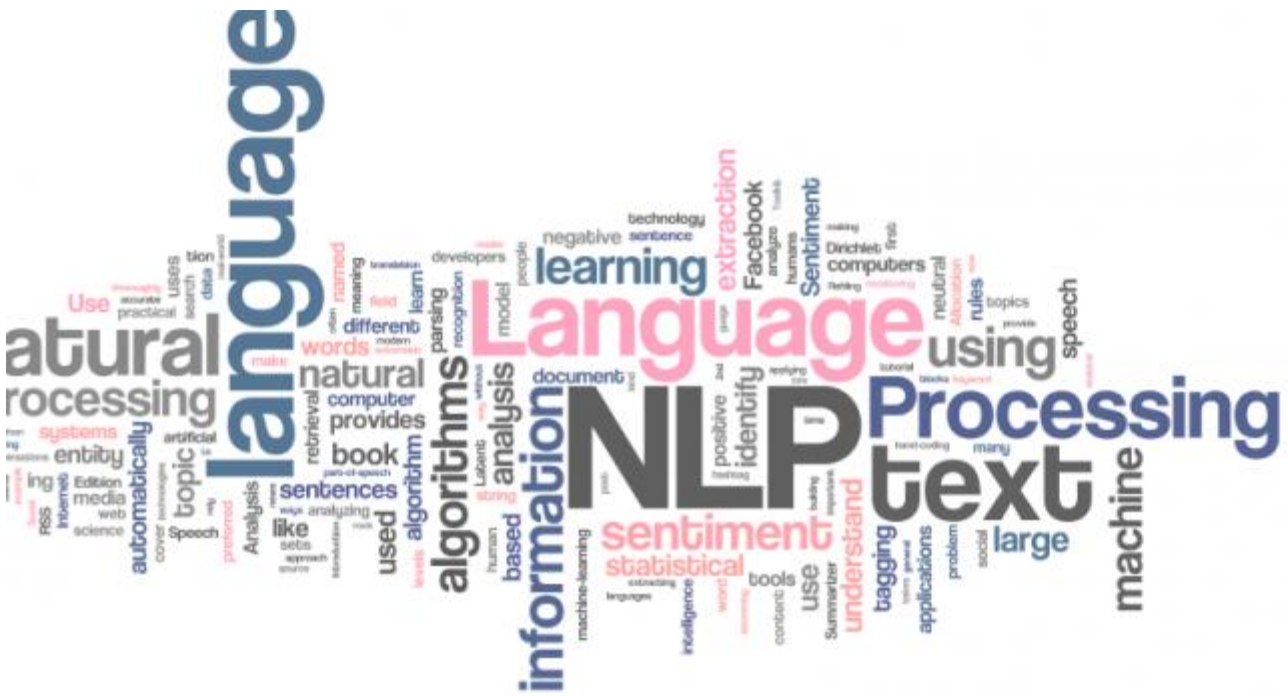
How NLP works?

NLP Applications

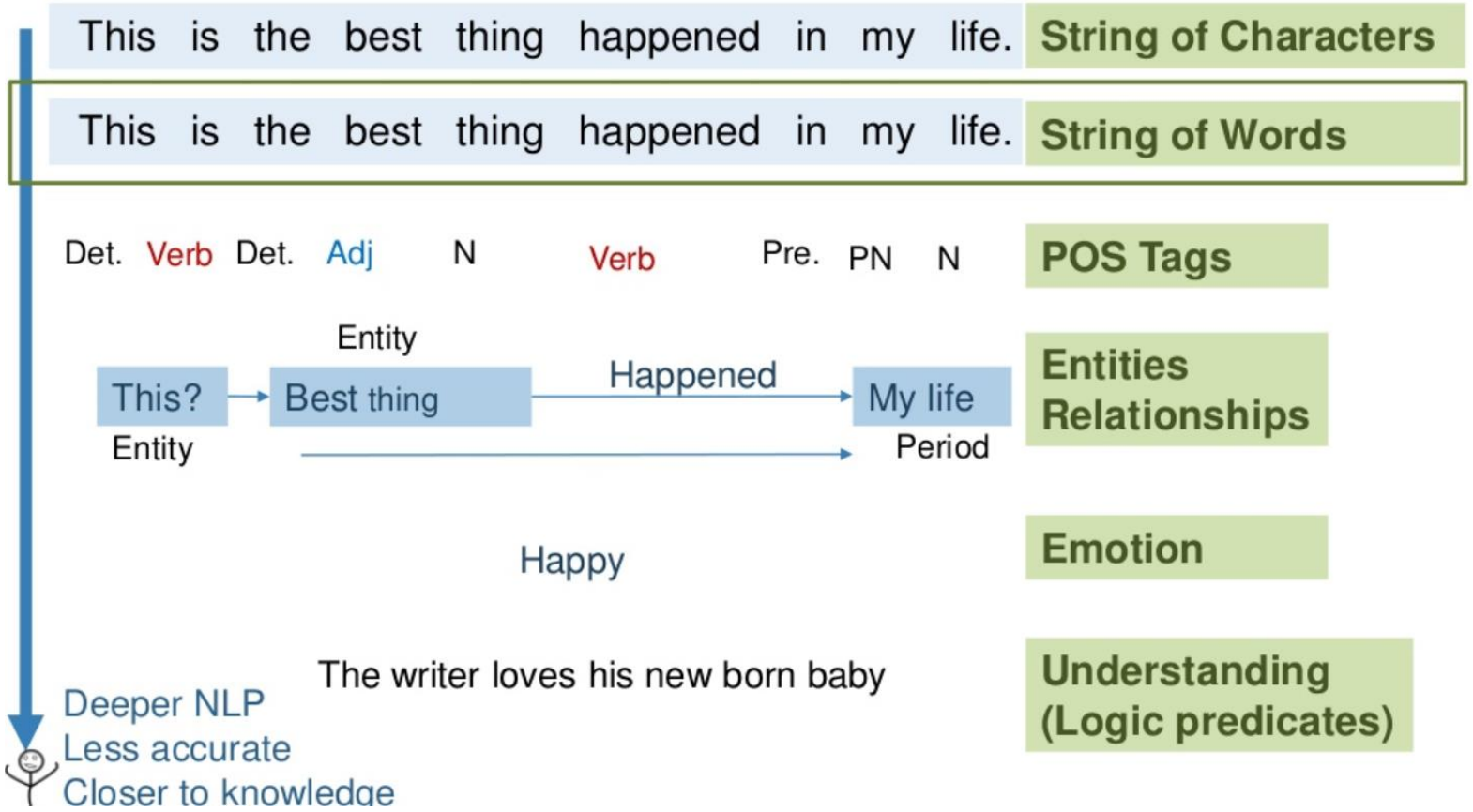
Project Demo – Switch

What's NLP?

- Stands for Natural Language Processing.
- A process that try to understand or generate pieces of human language. NLP often uses AI techniques such as Machine Learning.
- To convert unstructured data such as text documents and voice data from human to structured and meaningful knowledges. This is an important part of Text mining / Text Analysis.



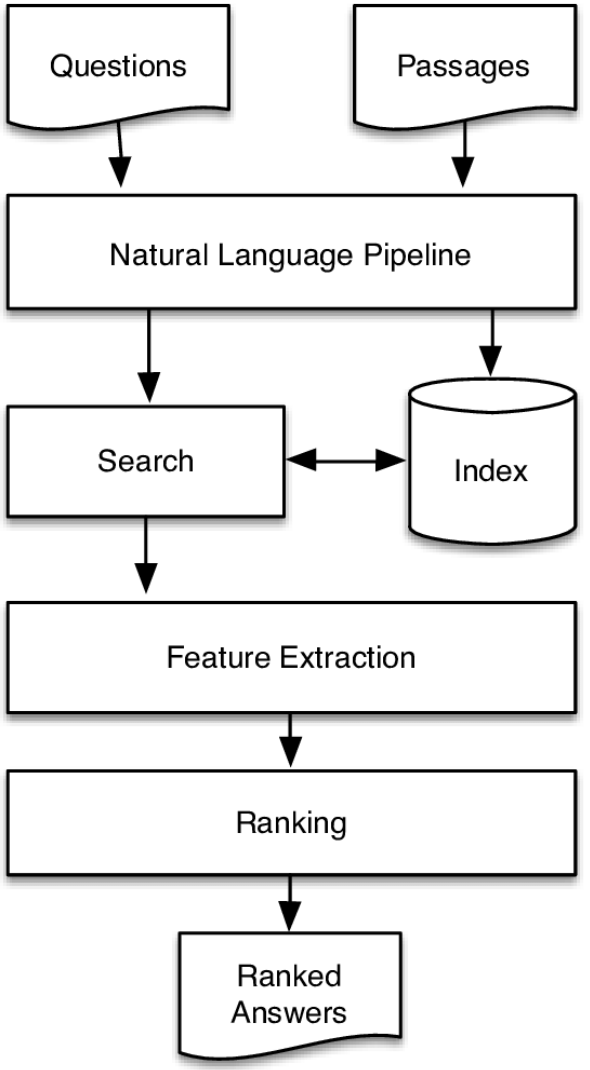
Basic Concepts in NLP



How NLP Works

Key features in NLP pipeline can do:

- Entity recognition - extracting entities from text documents
- Topics analysis - extracting features and topics
- Sentiment analysis - Analyzing text for positive and negative feelings
- Classification - Classifying documents



List of common NLP Algorithms

Algorithm	What It Does
Navie Bayes	Classifies data based on the probability of something happening from multiple data points.
K nearest neighbor (KNN)	Classifies data by a majority vote of its neighbors, with the data being assigned to the class most common among its k nearest neighbors.
Naïve Bayes - Multi Class	Assigns multiple classifications based on the probability of something happening from multiple data points.
SVM - Multi Class	Attempts to find a clear separator (hyperplane) between multiple paired classifications of data.
K-Means	Clusters groups of similar textual data in clusters to extract insights from unstructured data.

NLP Applications

- Spellcheckers or Spam filters
- Recommendation system, i.e. Netflix
- Voice assistants, I.e. Alexa, Siri
- Online search, i.e. Google
- Language Translation, i.e. Google translate

Project Demo - Switch

- Switch is a comprehensive Jobs search engine using content from Twitter. It allows Job seekers to search relevant job posting, submitted by Twitter users. We collect all tweets related to job posting using Twitter's API and process them with Natural Language Processing (NLP) techniques to return relevant job results to our users.

- Website: http://switch-ui.s3-website.us-east-2.amazonaws.com

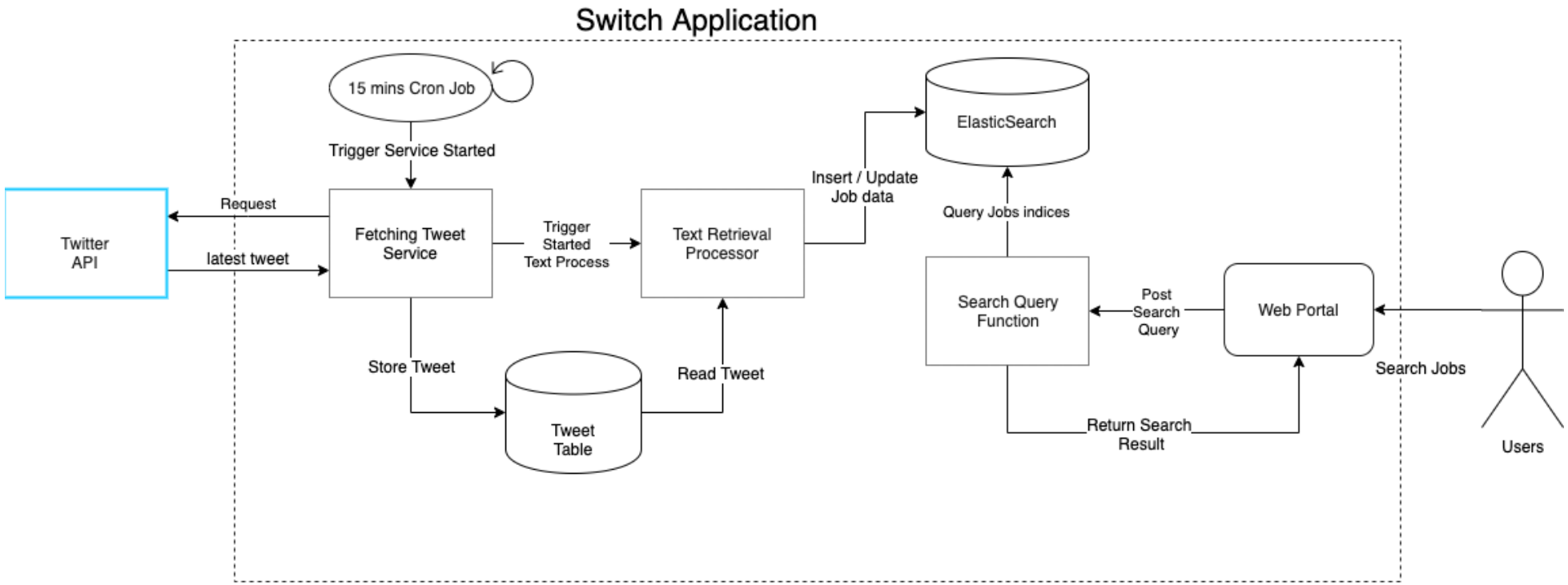
Technologies we used

- Twitter APIs - To scrape data from data source Twitter we used Twitter API
- Amazon Web Service (AWS) - We used AWS to host our services. We leveraged many AWS services for this project includes Lambda functions, DynamoDB, Elastic Search Engine, S3 and CloudWatch Events.
- Python Libraries used are Tweepy, Spacy, Pandas, NumPy, re, json, requests
- Web Development Framework– For our end user Web portal we used React as our platform of development.
- Other Tools – Google Colabs (Jupyter notebook), Trello (Kanban board).

NLP techniques we used

- Named Entity Recognition – is an information extraction technique to extract unstructured text (i.e. Tweet messages) into pre-defined categories information such as organization name and geographic location.
- Naive Bayes Classification – is a simple text retrieval technique for constructing a classifier based on applying Bayes' algorithm with independence features and labeled training data. We used it to identify and discard irrelevant tweets in this project. This is also a type of supervised learning methods.

Architecture Diagram



Lesson learned

- Data exploration is important – be sure to understand your data by sampling them first.
  - The default tweet message was 140 chars. We need to get the full text message (256 chars) instead.
  - Missing valuable field values, i.e. Screen name (@interstate_batteries)
  - Collecting good training data is hard.
- Understand your platform limitation
  - AWS Lambda deployment package limitation - 50 MB (zipped for direct upload).

References

- Switch Web Portal: http://switch-ui.s3-website.us-east-2.amazonaws.com/
- Switch Source Codes and Documentation: https://lab.textdata.org/wingkc2/switch
- Presentation Video: https://youtu.be/loMG4rdGXkg