Introduction to Natural Language Processing (NLP)

Presented By Wing Chan

Agenda

What is NLP?

How NLP works?

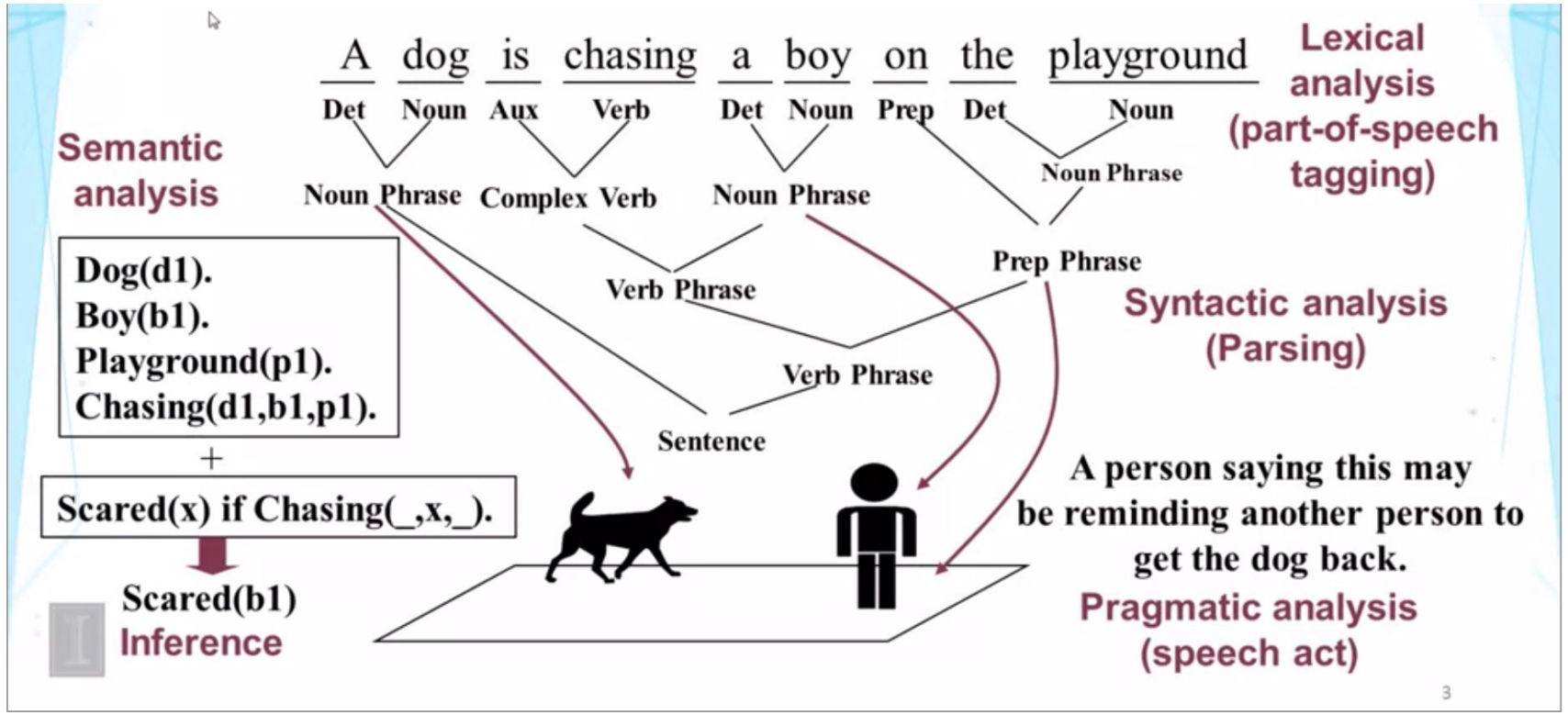
NLP Applications

Project Demo – Switch

What's NLP?

- NLP stands for Natural Language Processing.
- NLP is automated programs that try to understand or generate pieces of human language. NLP often uses AI techniques.
- Mainly process text documents and voice data from human.

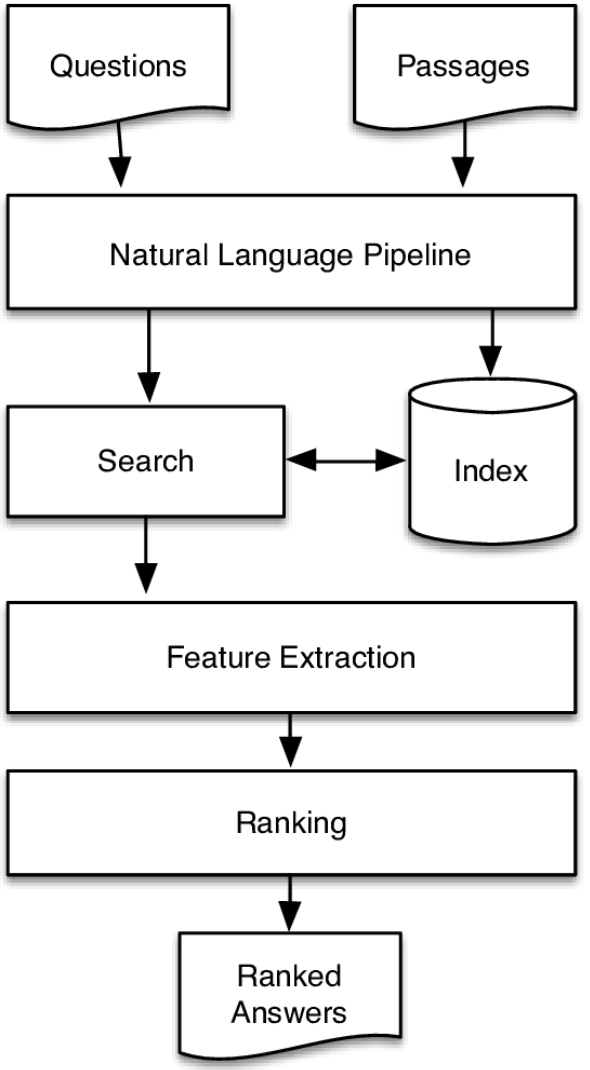
Basic Concepts in NLP



How NLP Works

NLP Pipeline can do:

- Extracting information from text documents (Entity recognition)
- Extracting features and topics (Topics analysis)
- Analyzing text for positive and negative feelings (sentiment analysis)
- Classifying documents (Classification)



NLP Applications

Spellcheckers:

Suggested words

Alexa, Siri and other voice assistants

Online search

Google translate and other machine translation apps

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

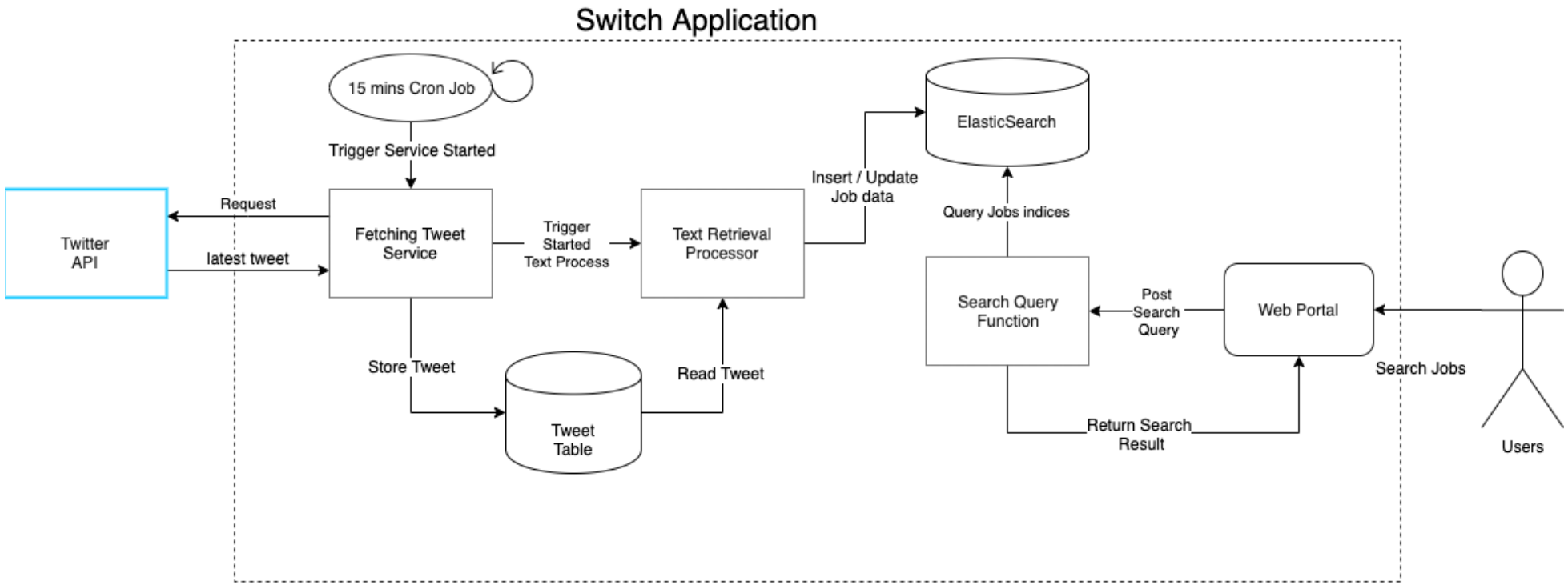
Technologies we used

- 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
- React – 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 – To extract the organization name and geographic location information from Tweets.
- Naive Bayes Classification - To identify and discard irrelevant tweets.

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