SE 953482 Natural Language Processing for SE 66/2

NLP Overview

Asst. Prof. Pree Thiengburanathum

CAMT
College of Arts, Media and Technology
Chiang Mai University

Agenda

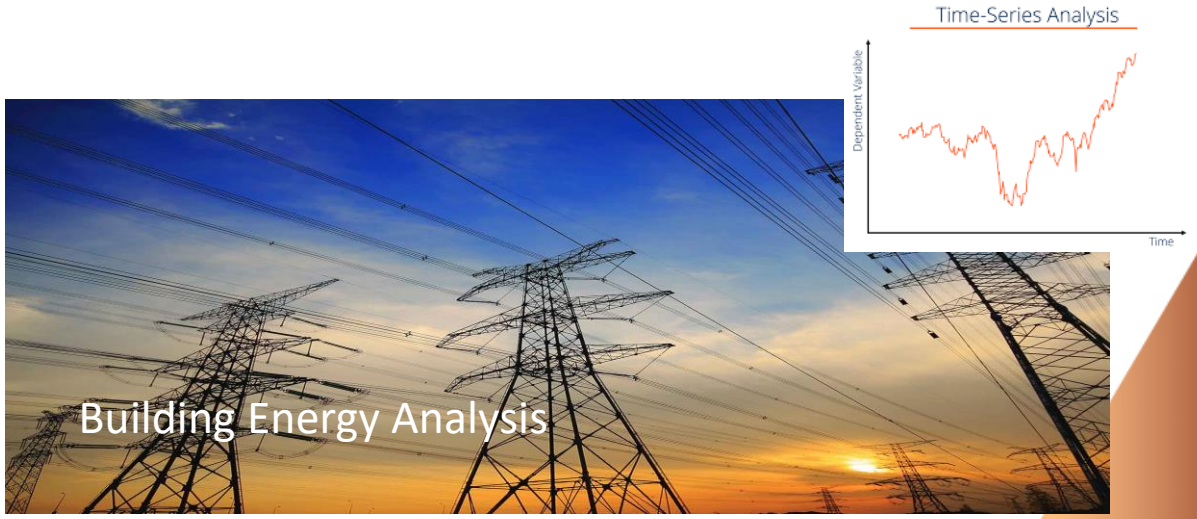
- Course outline
- Intro to unstructured data (NLP)

Education

- **Ph.D. in (Computing and Informatic) Bournemouth University,** *2013 – Mar 2018, Faculty of Science and Technology, Department of Computing and Informatics, Bournemouth University, UK.*
- **ERASMUS MUNDUS Research fellow,** *Feb 1, 2010- Dec 1, 2010, Universite De Lyon 2, France*
- **Master of Science in Computer Science**, Fall 2008, *University of Colorado at Denver, Denver, Colorado, USA.*
- **Bachelor of Science in Computer Science**, 2005, *Colorado State University, Fort Collins, Colorado, USA.*

CAMT
College of Arts, Media and Technology
Chiang Mai University

Area of interests



CAMT
College of Arts, Media and Technology
Chiang Mai University

Master students Thesis/IS

- (In progress) 2023 – Co-supervisor Trpnakorn, Zeroshot learning in Thai Sentiment Analysis
- (In progress)2023- Nuttawut, Cryptocurrency and Stock Price Prediction Using Sentiment Analysis
- (In progress 2023- SIRIYAPORN Development of Predictive Model for Cyprinid Herpes Virus 2 Infection of Goldfish
- (*In progress*) 2022 – Sukanya, An Analysis of Personal Factors Influencing the Electricity Consumption in Chiang Mai University's Dormitories during the Post-COVID-19 Recovery.
- (*In progress*) 2022 – Jukkrit, Thai Spell Correction Using Pre-train BERT for End-to-End Word Extraction based on Tesseract
- (*In progress*) 2022 – Parinya, Sentiment Analysis and visualization from online restaurant review using NLP
- (*Graduated*) 2022 – Patcharapol Y. Development of Electricity Consumption Forecasting Model for Campus-Scaled Building Using Machine Learning
- (*Graduated*) 2021 – Decheng Yang, A Comparative Study of Open-Source Crawlers Based on Robustness and Scalability Testing on E-Commerce Websites
- (*Graduated*) 2021 - Li Ye, A Real-Time Bus Arrival Time Prediction System Based on Spark Framework and Machine Learning Approaches: a case study in Chiang Mai

CAMT
College of Arts, Media and Technology
Chiang Mai University

Course outline

Course learning outcomes (CLOs): Students are able to

1 Explain the process in NLP and techniques.
2. Use the appropriate models and metrics tools for the right problem.

Course Description:

NLP overview, word tokenization and text preprocessing, text extraction methods, machine-learning models in NLP, Deep-learning models in NLP, Transformer, model evaluation and explain-ability, evaluation metrics, NLP-based systems, and case studies

Course outline (cont.)

- Pre-requisite SE 233 (953233), SE 411 (953411)

1.	NLP Overview	3
2.	Word Tokenization, Text preprocessing	6
3.	Text extraction methods	6
4.	Machine-learning models in NLP	6
5.	Deep-learning models in NLP	6
6.	Transformers	3
7.	Evaluation metrics and explain-ability	3
8.	NLP-based Systems	3
9.	Case studies and Project	9

Course outline (cont.)

- **Grading system:**
  - Midterm Examination *30%*
  - Final Examination *30%*
  - Workshops 15% (6 works)
  - Programming Assignments *25% (4-5 works)*
  - **I have right to adjust grading system based on the student performance.*
  - CMU-based (i.e., a grade A cut at 80%)

Grade policy:

- Any late assignment submissions will either be penalized (at least 50% reduction) or NOT be accepted.
- If a student is late more than 15 minutes in either lab or lecture, you will be regarded as absence.
- If a student needs to be absent with legitimate causes, please notify the lecturer or TA before the date of absence.
- The student who has come to class less than 80% <u>will NOT allow</u> to take the FINAL EXAM.
- The student who does not take the final exam gets “F” for this course. o The work that does not strictly follow the instruction is not accepted.

Course outline (cont.)

- Course communication
  - MS Team Channel
  - pree.t@cmu.ac.th, or direct message me at Team
  - Room 415-1 (T, Th)
- Programming language and tools
  - Python3
  - Anaconda framework (optional)

Textbooks

(*main*) Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems 1st Edition

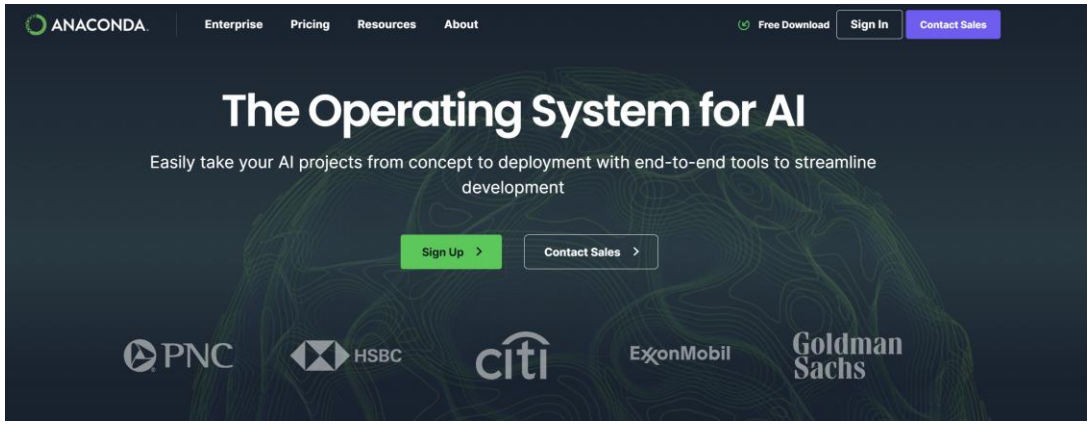
by Sowmya Vajjala (Author), Bodhisattwa Majumder (Author), Anuj Gupta (Author), Harshit Surana (Author)

(*optional*) Getting Started with Natural Language Processing

- by Ekaterina Kochmar (Author)

Anaconda/Miniconda

- Pre-install libraries
- Good package management
- Easy to install, maintain, and export
- Cross platforms
- https://www.anaconda.com/



Dev tools for the course

Kaggle

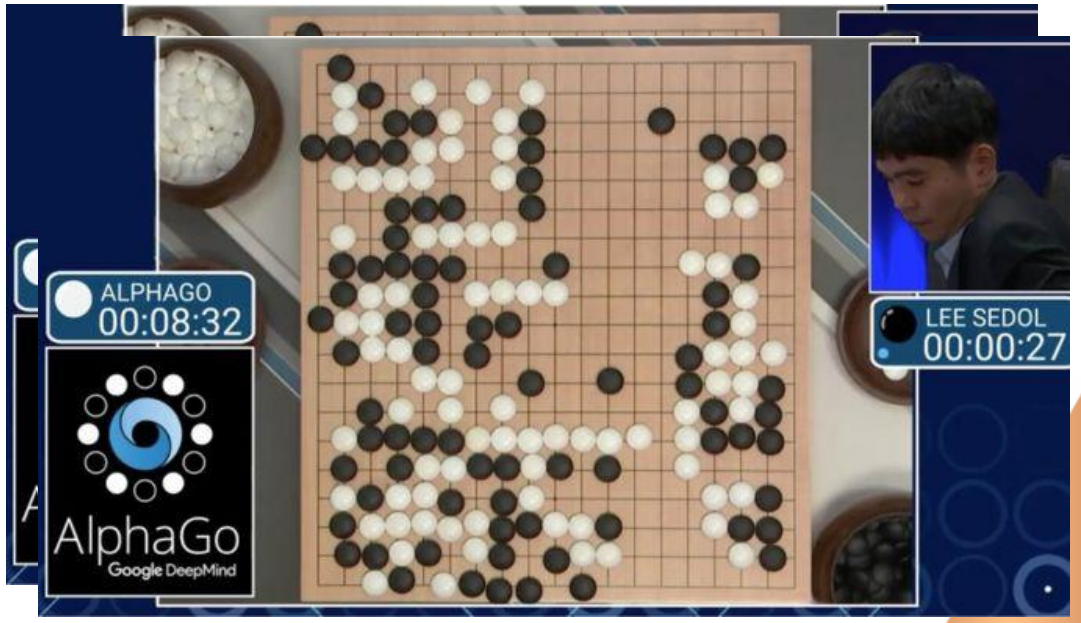
Colab

Github

Introduction (cont.)

- Data continues to grow exponentially
  - Estimated to be 2.5 MTB a day
  - Grow to 40 BTB by 2020 (50 * of 2010)
- Approx. 80% of data is estimated to be unstructured/text-rich data
  - >4.5 billion web pages
  - >40 million articles (5 million in English)
  - >500 million tweets a day, 200 billion a year
  - >1.5 trillion queries on Google a year

Machine Versus Men



CAMT
College of Arts, Media and Technology
Chiang Mai University

Machine Versus Men (cont.)

- Can machine beat the best of man in what man is supposed to be the best at?
- https://www.youtube.com/watch?v=YgYSv2KSyWg
- Watson, which is called DeepQA
- Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage.

DeepQA overall architecture

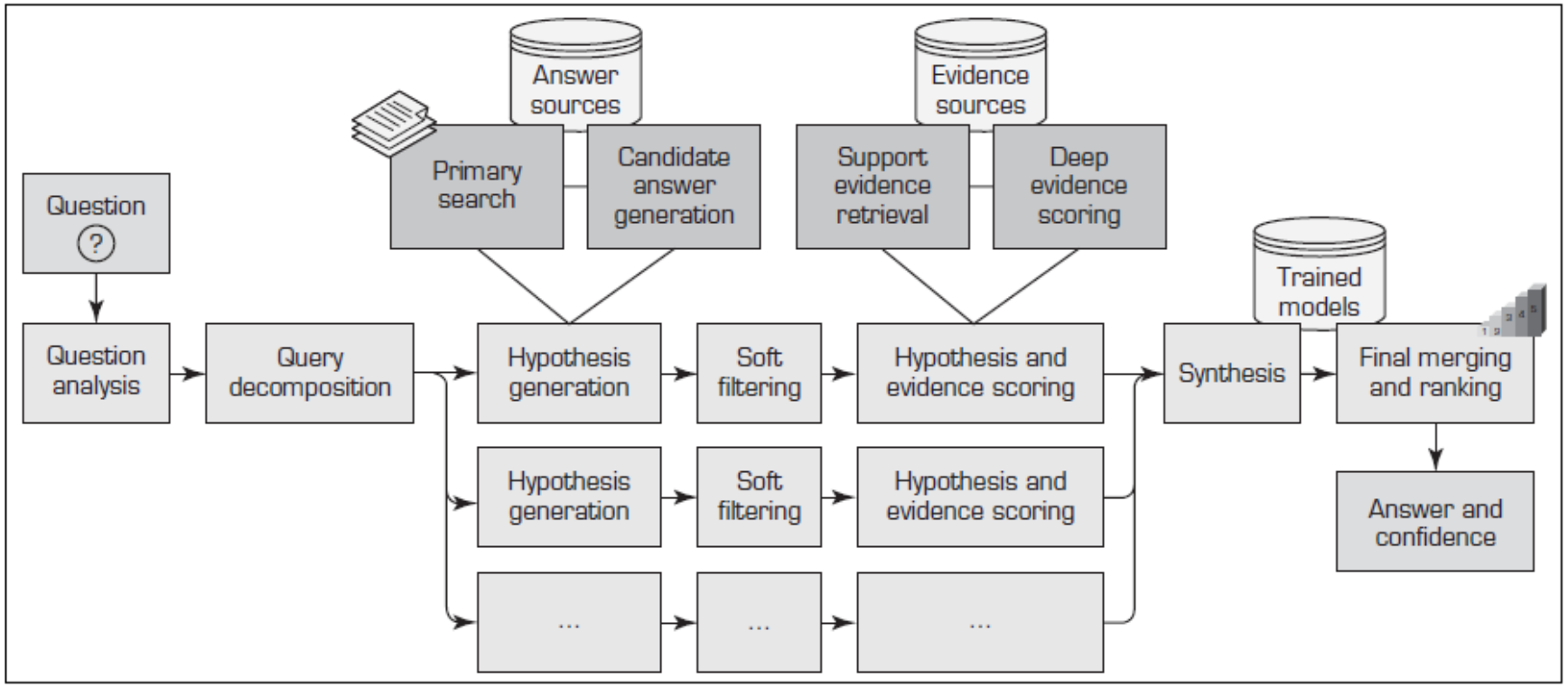


FIGURE 7.1 A High-Level Depiction of DeepQA Architecture.

Meta NLP 2022

- LaMDA – conversation robot, BERT+GPT-3
- FLORES-101 – Largest language transition dataset

~5x

Flores supports translation evaluation in 10 100 directions while the Talks data set only supports 2 162 directions

TALKS DATA SET Ye et al, 2018 [46]

Amharic	Esperanto	Japanese	Portuguese
Arabic	Estonian	Kannada	Russian
Asturian	Filipino	Kazakh	Serbian
Basque	Finnish	Lithuanian	Serbian
Belarusian	French	Macedonian	Silesian
Burmese	Galician	Malagasy	Sinhala
Catalan	Georgian	Malay	Slovak
Cebuano	Greek	Marathi	Telugu
Chinese	Gujarati	Mauritian	Thai
Czech	Haitian Creole	Creole	Turkish
Dutch	Hindi	Occitan	Vietnamese
English	Igbo	Pashto	

FLORES

Afrikaans			
Amharic			
Arabic			
Armenian			
Assamese	Gujarati	Malay	Sorani Kurdish
Asturian	Hausa	Malayalam	Spanish
Azerbaijani	Hebrew	Maltese	Swahili
Belarusian	Hindi	Māori	Swedish
Bengali	Hungarian	Marathi	Tajik
Bosnian	Icelandic	Mongolian	Tamil
Bulgarian	Igbo	Nepali	Telugu
Burmese	Indonesian	Northern Sotho	Thai
Catalan	Irish	Norwegian	Turkish
Cebuano	Italian	Nyanja	Ukrainian
Chinese Sim.	Japanese	Occitan	Umbundu
Chinese Trad.	Javanese	Oriya	Urdu
Croatian	Kabuverdianu	Oromo	Uzbek
Czech	Kamba	Pashto	Vietnamese
Danish	Kannada	Persian	Welsh
Dutch	Kazakh	Polish	Wolof
Estonian	Khmer	Portuguese	Xhosa
Filipino	Korean	Punjabi	Yoruba
Finnish	Kyrgyz	Romanian	Zulu
French	Lao	Russian	
Fula	Latvian	Serbian	
Galician	Lingala	Shona	

Meta NLP 2023

- GPT-3 4,096 and 2,049 tokens
- GPT-4 8,192 and 32,768 tokens
- Introduced “Multimodal” (e.g., can also understand image)
- Better and solve math problems
- Even more languages (with low-resources)
- Able to include reference and source of the text generated

Structured vs Unstructured data

	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Figure 1.1 An Excel table is an example of structured data.

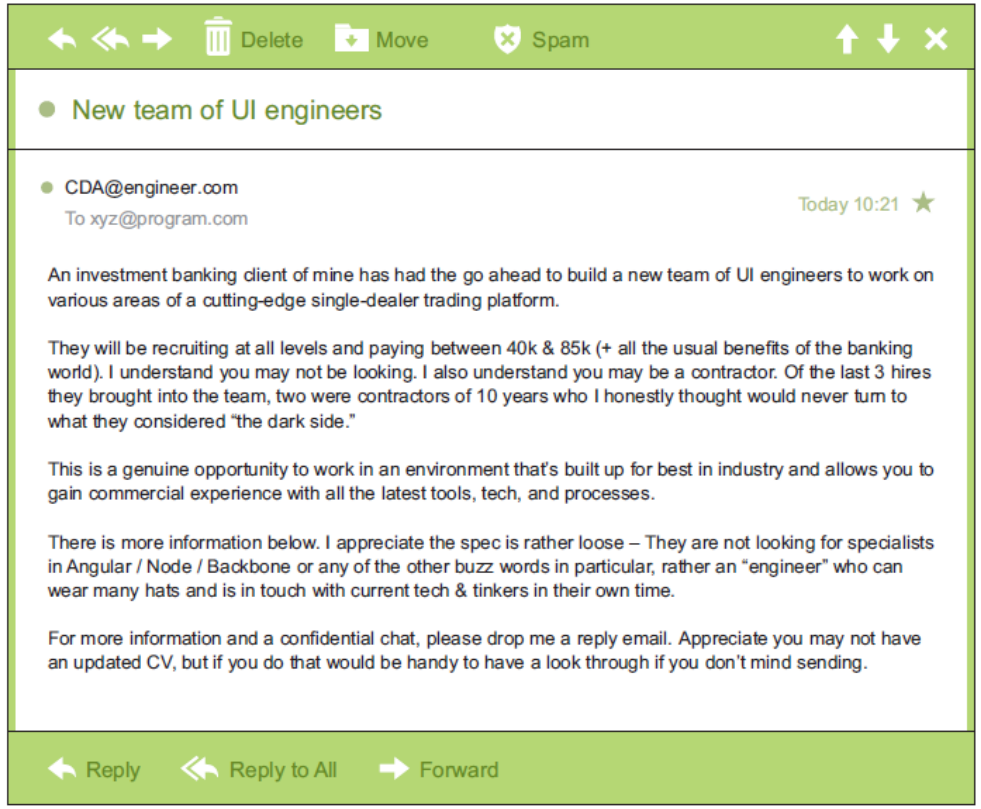


Figure 1.2 Email is simultaneously an example of unstructured data and natural language data.

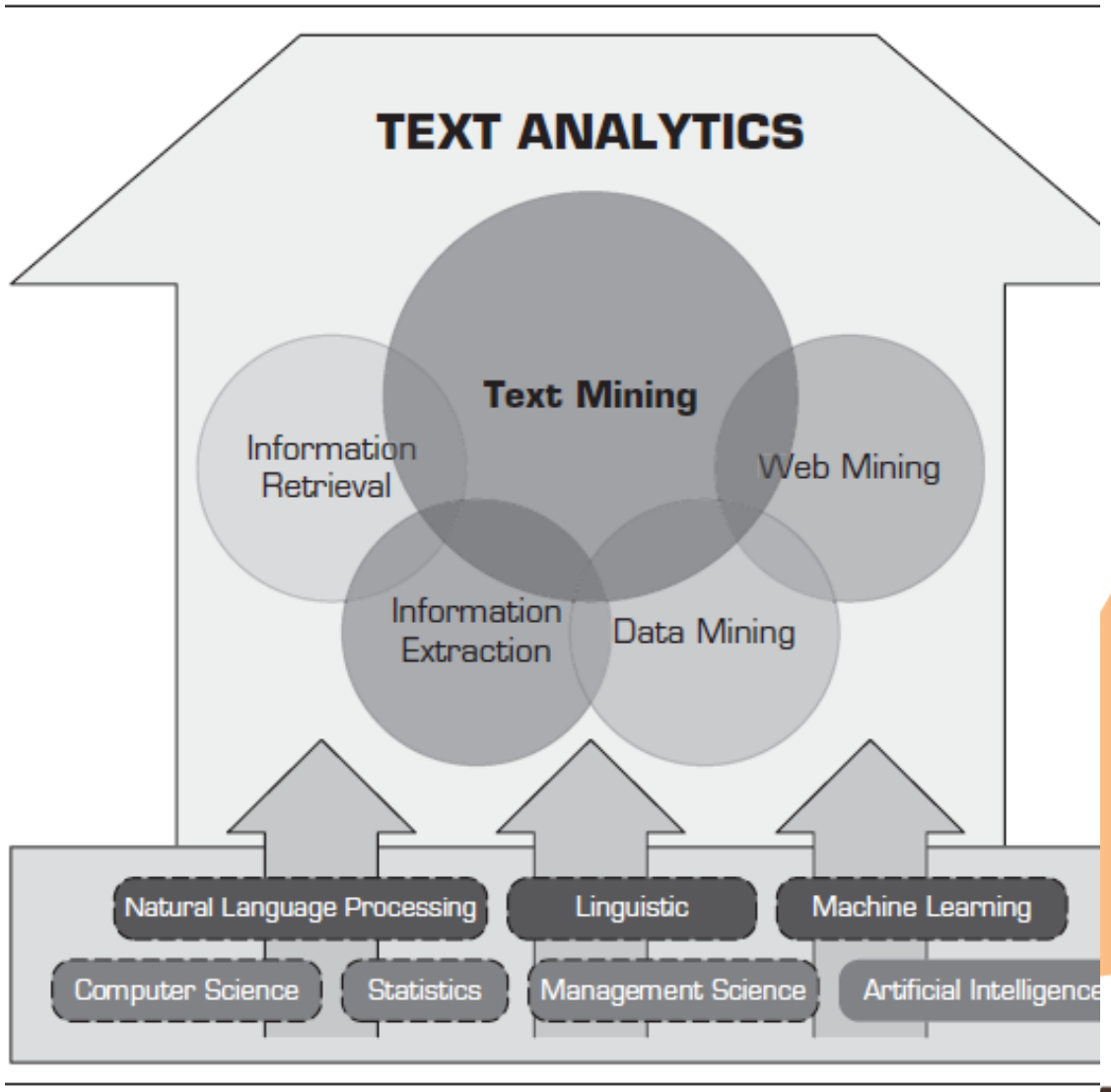
Images source: Course

Text Analytics

- The vast majority of business data is stored in text documents that are virtually unstructured.
- Text analytics is a broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms) as well as information extraction, data mining, and Web mining,
- Whereas text mining is primarily focused on discovering new and useful knowledge from the textual data sources.

Text Analytics (cont.)

- Text Mining is a derivative of Data Mining.
- Sentimental Analysis is a derivative of Text Mining.



Text Mining

- AKA. Text data mining/knowledge discovery in textual database
- Large amount of unstructured data
- Word, PDF, XML, etc.

- Benefit domains:
  - In the areas where very large amount of textual data are being generated.
  - Could you give some example?

Text Mining (cont.)

- Law – court orders
- Academic research – research article
- Finance – quarterly report
- Medicine – discharge summaries
- Biology- molecular interactions
- Marketing – customer comments
- Social Science – Web board, Twitter , etc.
- Technology – e-mail platforms? Is Gmail the smartest?

Text Mining applications

- **Information extraction** – identify the key phrases and relationship within text
- **Topic tracking** – predict/recommend other document of interest to user.
- **Summarization** – summarizing a document to save time of the reader.
- **Categorization** – identify the main themes of a document and put to the right themes
- **Clustering** – group similar documents without having a predefined set of categories
- **Concept linking** – connects related documents by identify their shared concepts.
- **Question answering** – find the best answer to a given question

Text Mining applications (cont.)

- **Intention mining/recognition/detection** – discover user intention based on comments, reviews, tweets, blogs
- **Concept mining** – extract idea and concept from large static social media
- **Sentiment Analysis** – categorize text to sentiment polarity (pos, neg, neu)
- Topic modeling – unconver the topical structure of a large collection of docs.

NLP applies in Software Engineering

- **Code Generation and Understanding:** NLP techniques can be used to convert natural language commands into code and to help developers understand complex codebases.
- **Automated Documentation:** NLP can be used to generate and maintain technical documentation based on code changes.
- **Bug Tracking and Analysis:** It can assist in categorizing and prioritizing bugs by analyzing bug reports.
- **Customer Support:** NLP can power chatbots and support systems that interact with users to solve technical problems.
- **Code Reviews:** NLP can automate some aspects of code reviews by summarizing changes and identifying potential issues.

Level of difficulty

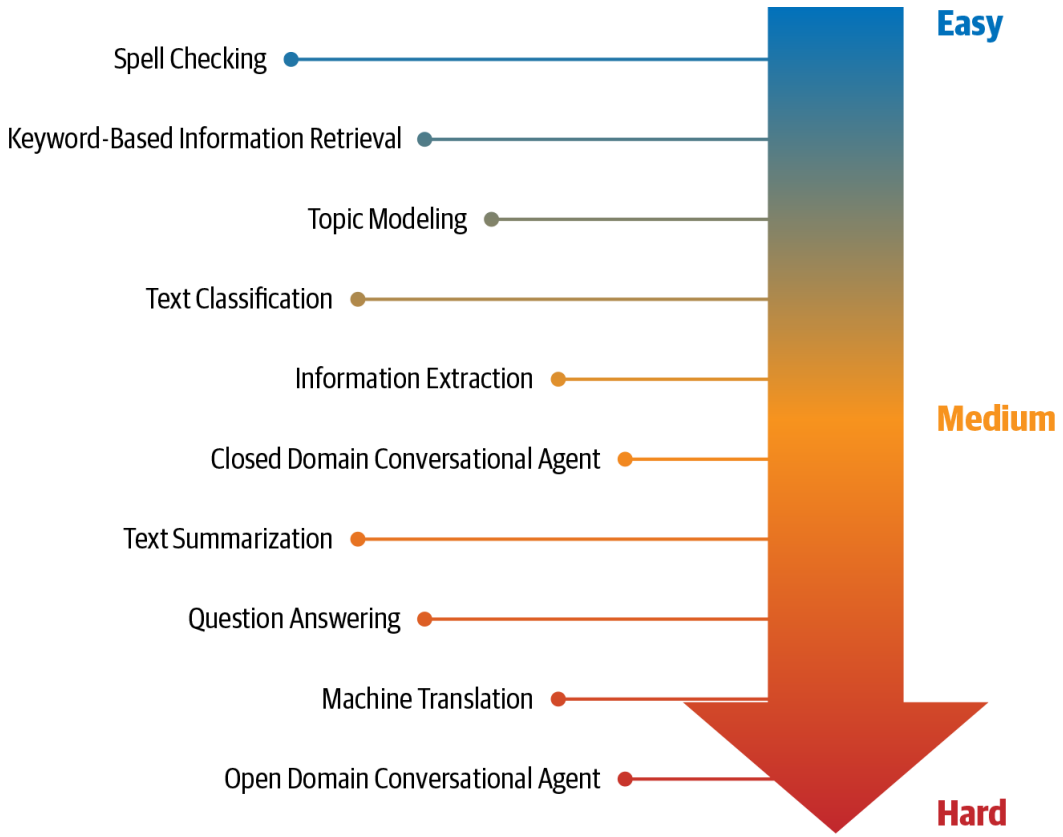


Figure 1-2. NLP tasks organized according to their relative difficulty

Intro Natural Language Processing (NLP)

- Natural language vs Programming language
  - NL - human share information with human
  - PL – human tells machines what to do
- NLP – Machine can now process natural language (i.e., interpreter)

Intro NLP (cont.)

- Subfield of AI and Computation Intelligent/linguistics.
- Try to understanding the natural human language.
- Moving forward to syntax-drive (word counting) to true understanding and processing of NLP
- Considering grammatical, semantic constraint and context.

NLP practical applications

- **Editing** – spelling, grammar, style
- **Dialog** – Chatbot, assistant, scheduling
- **Email** – spam filter, classification, prioritization
- **Text mining** – Summarization, knowledge extraction
- **News** – event detection, fact checking, fake news detection
- **Attribution** – plagiarism detection, literacy forensics,
- **Creative writing** – Movie scripts, poetry, song lyrics.
- **Search**- web, documents, autocomplete
- **Chatbot** – Ambiguous commands, Q/A, scheduling

Intro Natural Language Processing (NLP)

- Bag-of-words (classical method)
  - Text, sentences, paragraph, or document -> words
  - Classification model <spam/legitimate>
  - One bag is filled with words found in spam messages (Viagra, stock, buy)
  - Another bag is filed with words related to user's friend or workplace.
- Human do not use words without some order or structure
  - Semantic and syntactic structure
- Text mining need to look for ways beyond the bag-of-words.

Challenge in NLP (non-practical)

- **Part of speech tagging (POS-tagging)**- identify Adverb verb, noun in the sentence.
- **Text segmentation** - Chinese/Thai/Other languages.
- **Word sense disambiguation** – a word may has more than one meaning.
- **Syntactic ambiguity** – grammar is ambiguous
- **Imperfect or irregular input** – typos , grammar errors

Text mining terminologies

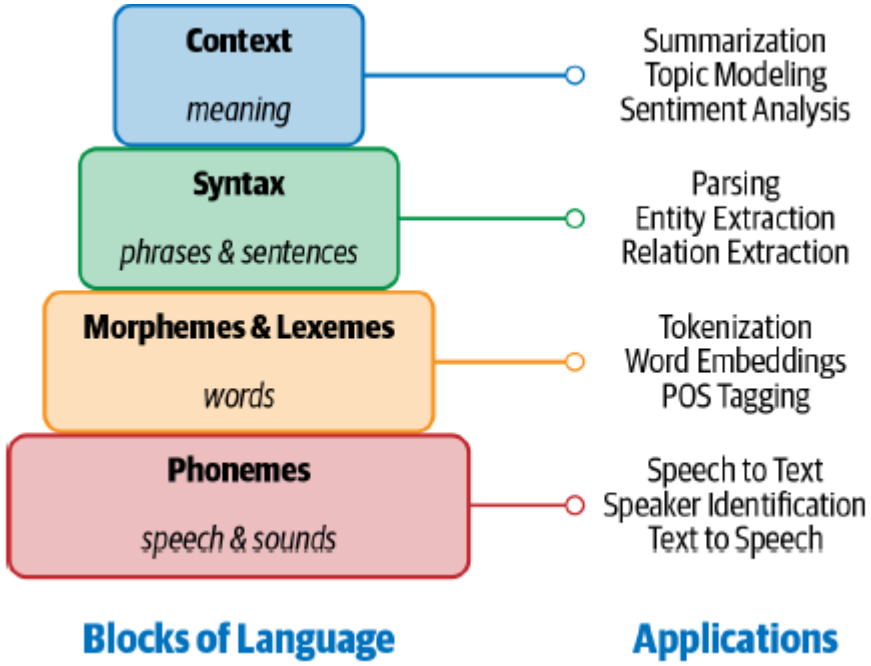
- ***Unstructured data (versus structured data).*** *– human readable*
- ***Corpus*** *= dataset*
- ***Terms*** *– single word or multiword pharse from corpus*
- ***Concepts*** *– features generated from documents.*
- ***Stemming*** *– process of reducing inflected word to their root*
- ***Stop words-*** *words that are filtered out after processed.*
- ***Synonyms and polysemes*** – syntactically different/identical words
- ***Tokenizing*** *– block of text*
- ***Word frequency*** *- #time that word occur in the document*

Text mining terminologies (cont.)

- ***Morphology*** *– form and formation of words in a language*
- ***Word frequency*** *– the number of times a word is found*

Language

- “Language is a structured system of communication that involves complex combinations of its constituent components, such as characters, words, sentences, etc.”



Language (cont.)

- Phonemes – smallest unit of sound in language
  - English has 44 phonemes (single letter or combo)
  - Useful in apps like speech reg, speech-to-text/text-to-speech
- Morphemes – smallest unit of language that has meaning
  - Combination of phonemes
  - Cats = Cat + s
  - Unbreakable = Un + break + able

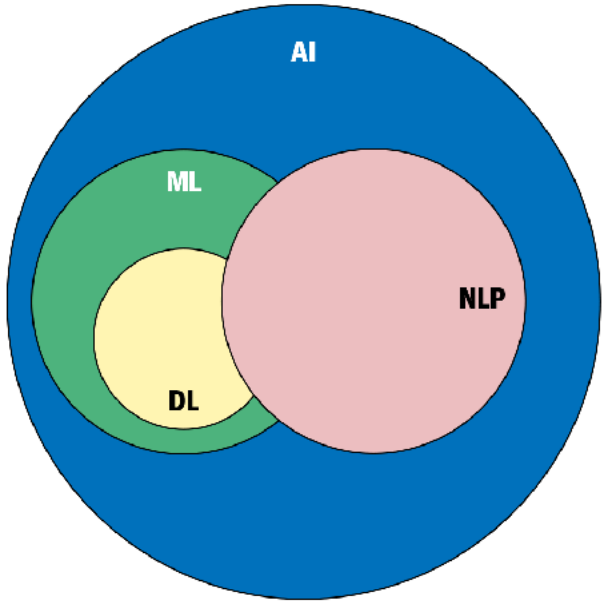
Language (cont.)

- Syntax – a set of rules to construct grammatically correct sentences
  - *Runs she fast*.
  - *She runs fast*.

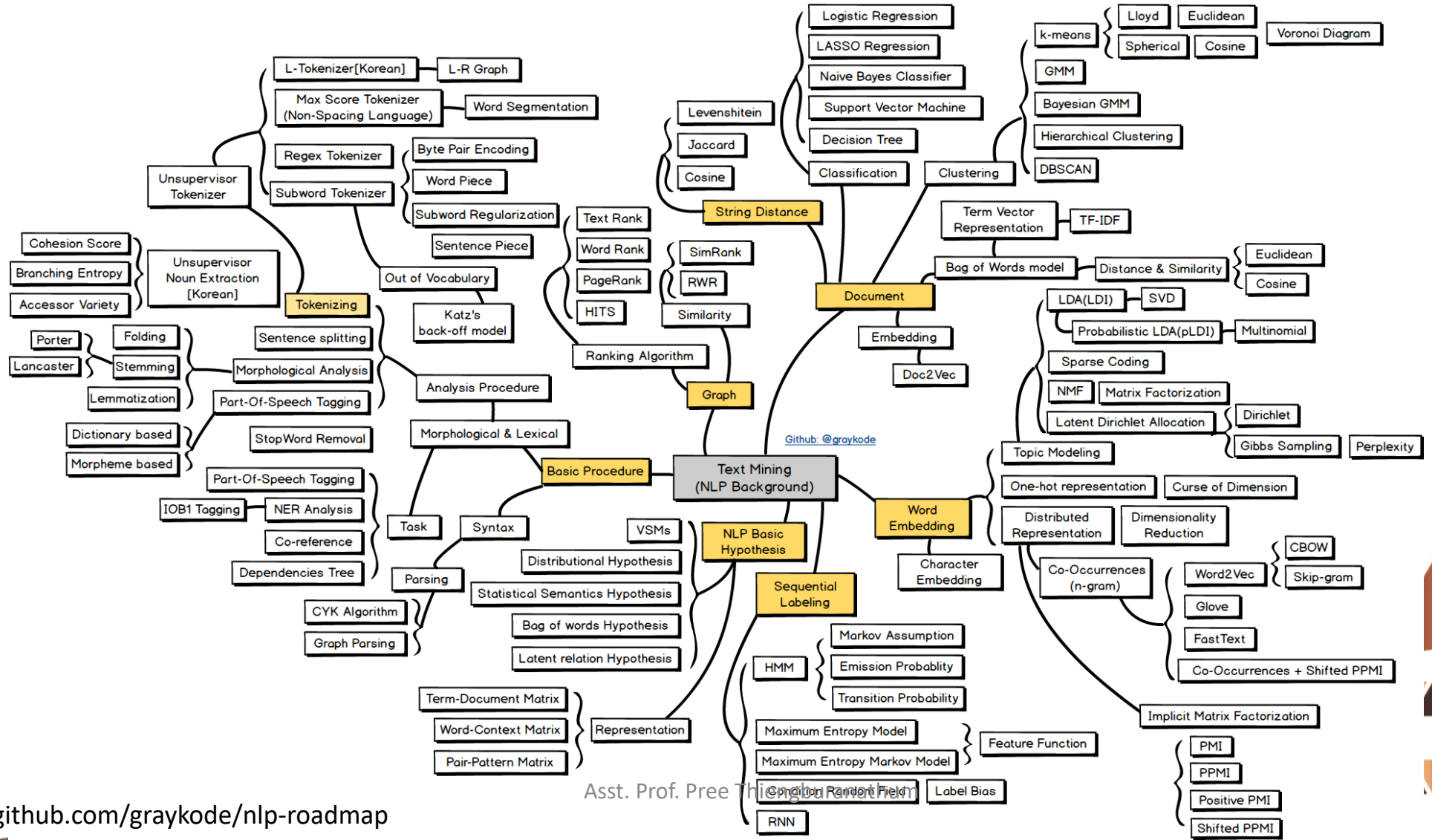
Why NLP is challenges?

- The **ambiguity** and **creativity** of human language
- Ambiguity - uncertainty of meaning. Most human languages are inherently ambiguous
  - “I made her duck.”
  - “Call me a taxi.”
  - “The teacher said the test would be difficult tomorrow.”
- Creativity – language is not a rule-based driven.
  - Various styles, dialects
  - Poem is a great example.

NLP related fields of studies.

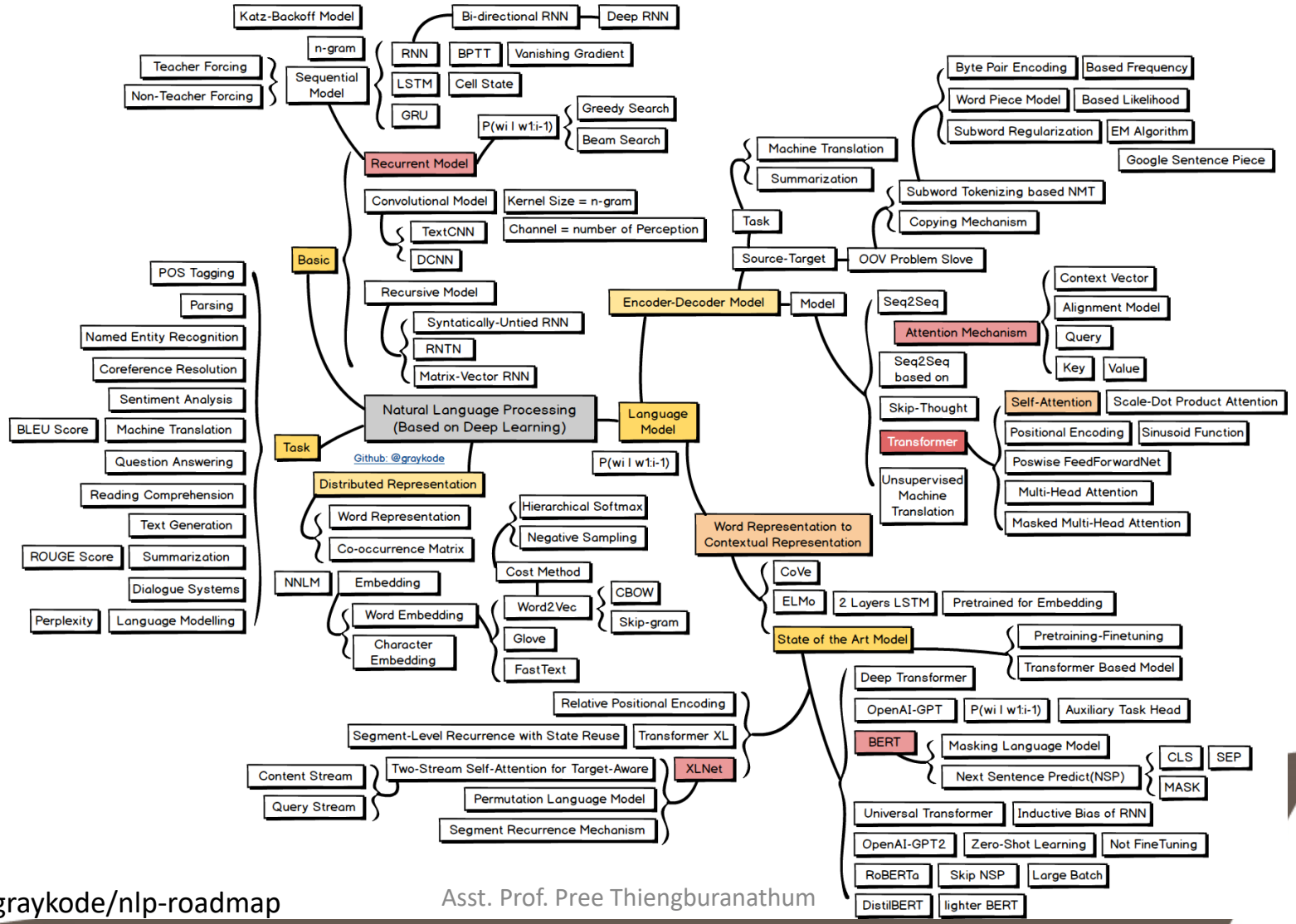


NLP road map



Source:https://github.com/graykode/nlp-roadmap

NLP road map (cont.)



Source: https://github.com/graykode/nlp-roadmap

Case study 2021 / NLP application in Action

- AI chat bot with Stress detection
- Read more at:
  - https://arxiv.org/abs/1911.00133

Conversational Agents

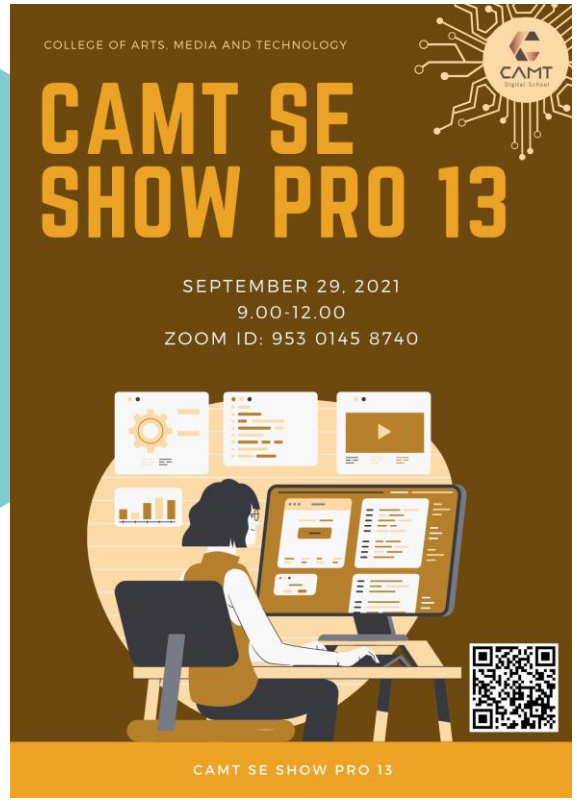
- AKA. Dialogue System, Dialogue Agents, Chatbots
- Personal Assistants on phones or other platforms
  - Alexa, SIRI, Google Assistant, Cortana
- Playing music, setting times and clock
- Chatting for fun
- Booking, scheduling reservation
- Clinical uses for mental heath (like my project)

Conversational Agents (cont.)

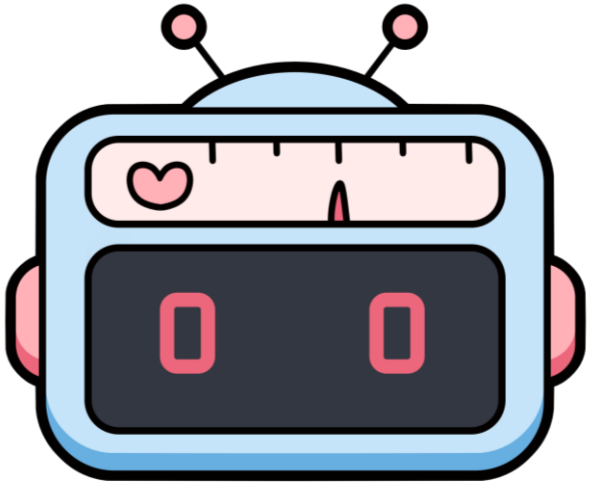
- Chatbots
  - Mimic informal human chatting
  - For fun, even for therapy
- Task-based
  - Personal assistant
  - Book seat in restaurant, movie theater, flights

Chatbot Arch.

- Rule-based
  - Pattern-action rules (ELIZA)
- Corpus-based(data-driven)
  - Information Retrieval
  - Model-based (My chatbot)



AI Chatbot with Stress Detection



Group name: *FTW_SD*
Presenter: *Pakin Kampeera (612115005)*
Aoxue Gui (612115501)
Project Advisor: *Dr. Pree Thiengburanathum*

CAMT
College of Arts, Media and Technology
Chiang Mai University

Background

- **Factors that affect people's active medical treatment:**
  - **Cost**
  - **Inconvenience of making an appointment**
  - **Obstacles to confiding in strangers**

- **Stress hurts human health.** (Hathaway, 2012)

- **The medical expenses invested to alleviate people's psychological problems are far from enough worldwide.** (United Nations, 2019)

- **The pandemic has caused a surge in the number of people calling the mental health hotline.** (CDC, 2020)

- **Most people have stress problems.** (Mental Health Foundation, 2018; APA, 2020; United Nations, 2019; KFF, 2020)

Motivation

we hope to develop a "self-help" tool to let limited psychological counselling services help those who need it more

Aim

Develop a chatbot to detect the user stress during communication and a dashboard to visually shows the statistical data.

Feature Overview

05.

Data cleaning, Data preprocessing, Data analysis

- Analyze sentences.

06.

Report generate

- View own stress at last 6 months informatio n in line chart.

07.

Notification

- View notification message.

Feature Overview

05.

Data cleaning, Data preprocessing, Data analysis

- Analyze sentences.

06.

Report generate

- Generate intuitive report to see stress change during the last six months.

07.

Notification

- Send notification message.

CAMT
College of Arts, Media and Technology
Chiang Mai University

Feature Overview

01.

Authentication

- Register
- Login
- Logout
- Reset password

02.

Dashboard

- View statistic data
- View sentences table
- Export table to CSV

03.

User management

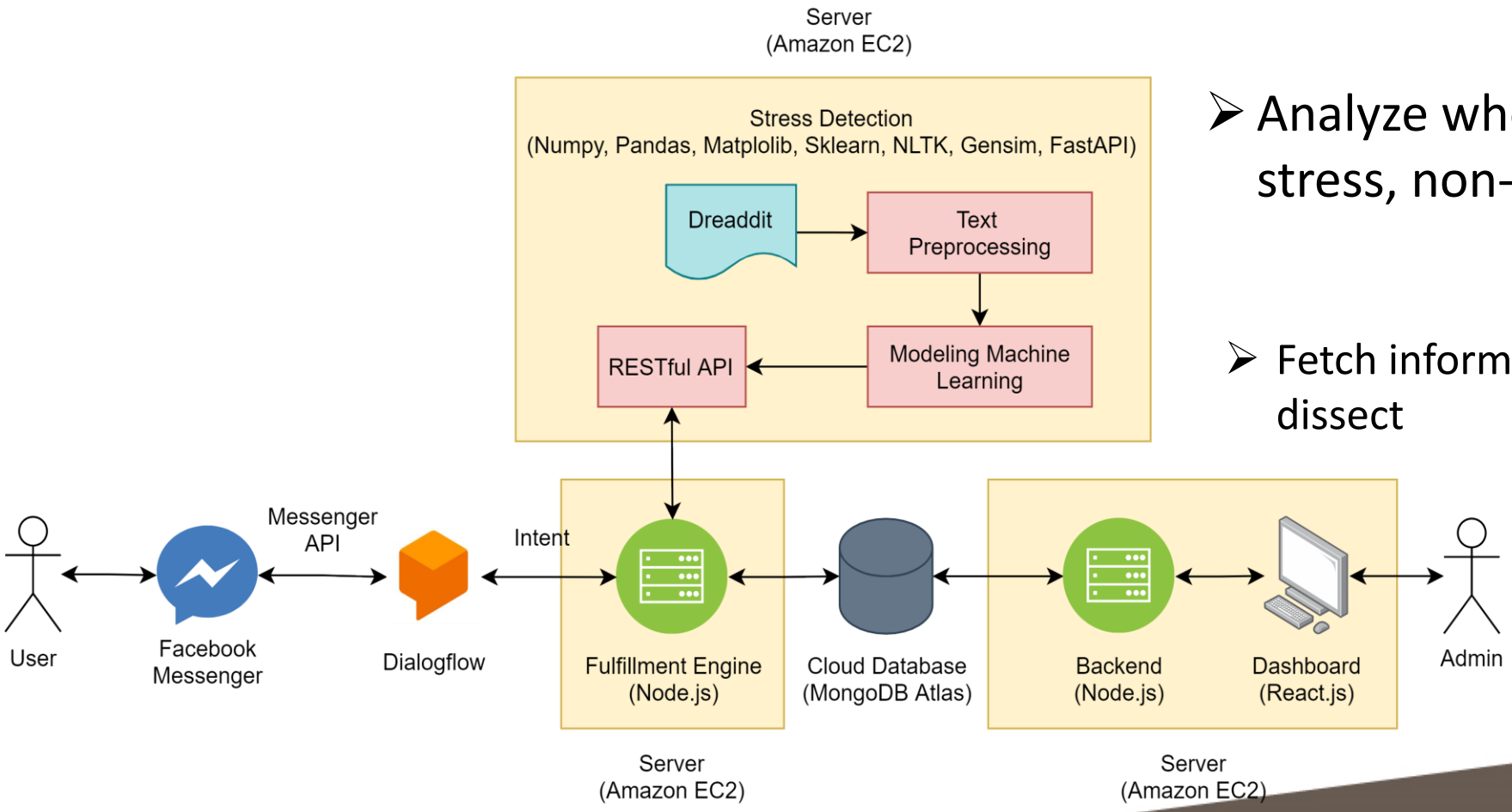
- View account information
- Change user

04.

Facebook messenger Chatbot

- Search
- Communic ate with Chatbot

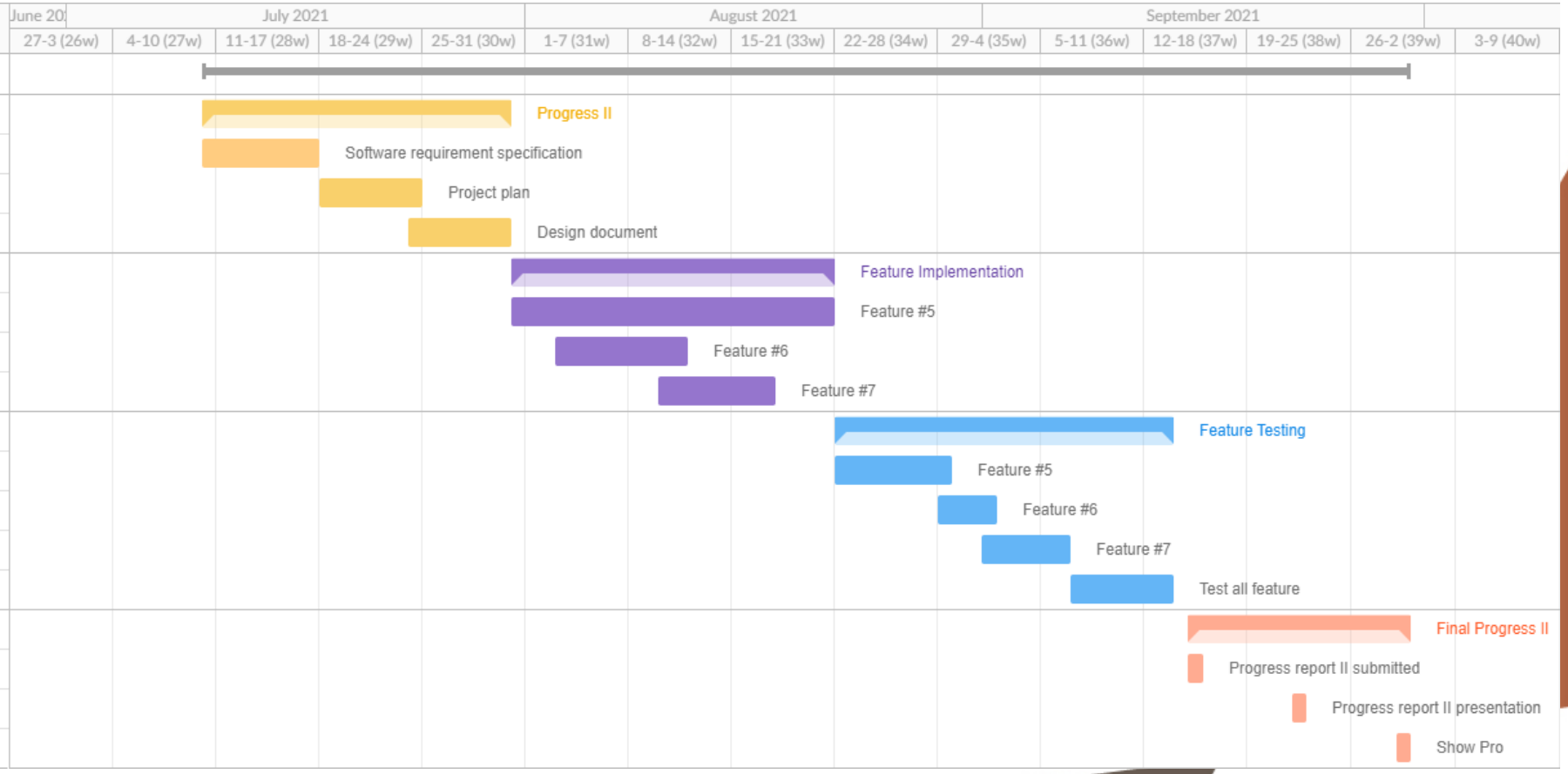
Product Perspective



- Analyze whether the sentences are stress, non-stress and cannot tell
- Fetch information from database to dissect

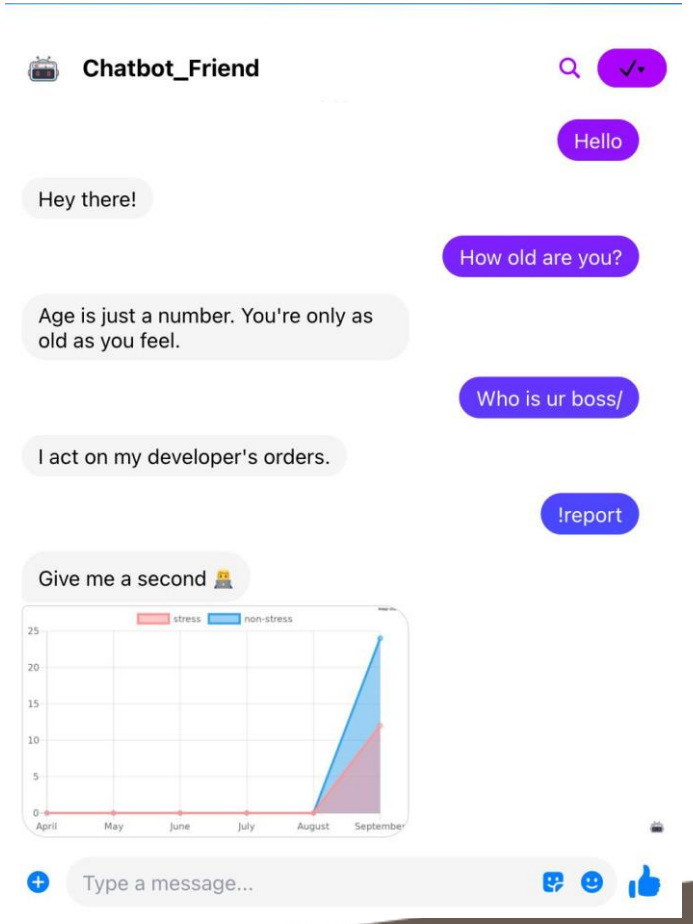
Milestone

	Task name	Start date	End date
		10/07/2021	**29/09/2021**
1	**Progress II**	**10/07/2021**	**30/07/2021**
1.1	Software requirement specification	10/07/2021	17/07/2021
1.2	Project plan	18/07/2021	24/07/2021
1.3	Design document	24/07/2021	30/07/2021
2	**Feature Implementation**	**31/07/2021**	**21/08/2021**
2.1	Feature #5	31/07/2021	21/08/2021
2.2	Feature #6	03/08/2021	11/08/2021
2.3	Feature #7	10/08/2021	17/08/2021
3	**Feature Testing**	**22/08/2021**	**13/09/2021**
3.1	Feature #5	22/08/2021	29/08/2021
3.2	Feature #6	29/08/2021	01/09/2021
3.3	Feature #7	01/09/2021	06/09/2021
3.4	Test all feature	07/09/2021	13/09/2021
4	**Final Progress II**	**15/09/2021**	**29/09/2021**
4.1	Progress report II submitted	15/09/2021	15/09/2021
4.2	Progress report II presentation	22/09/2021	22/09/2021
4.3	Show Pro	29/09/2021	29/09/2021

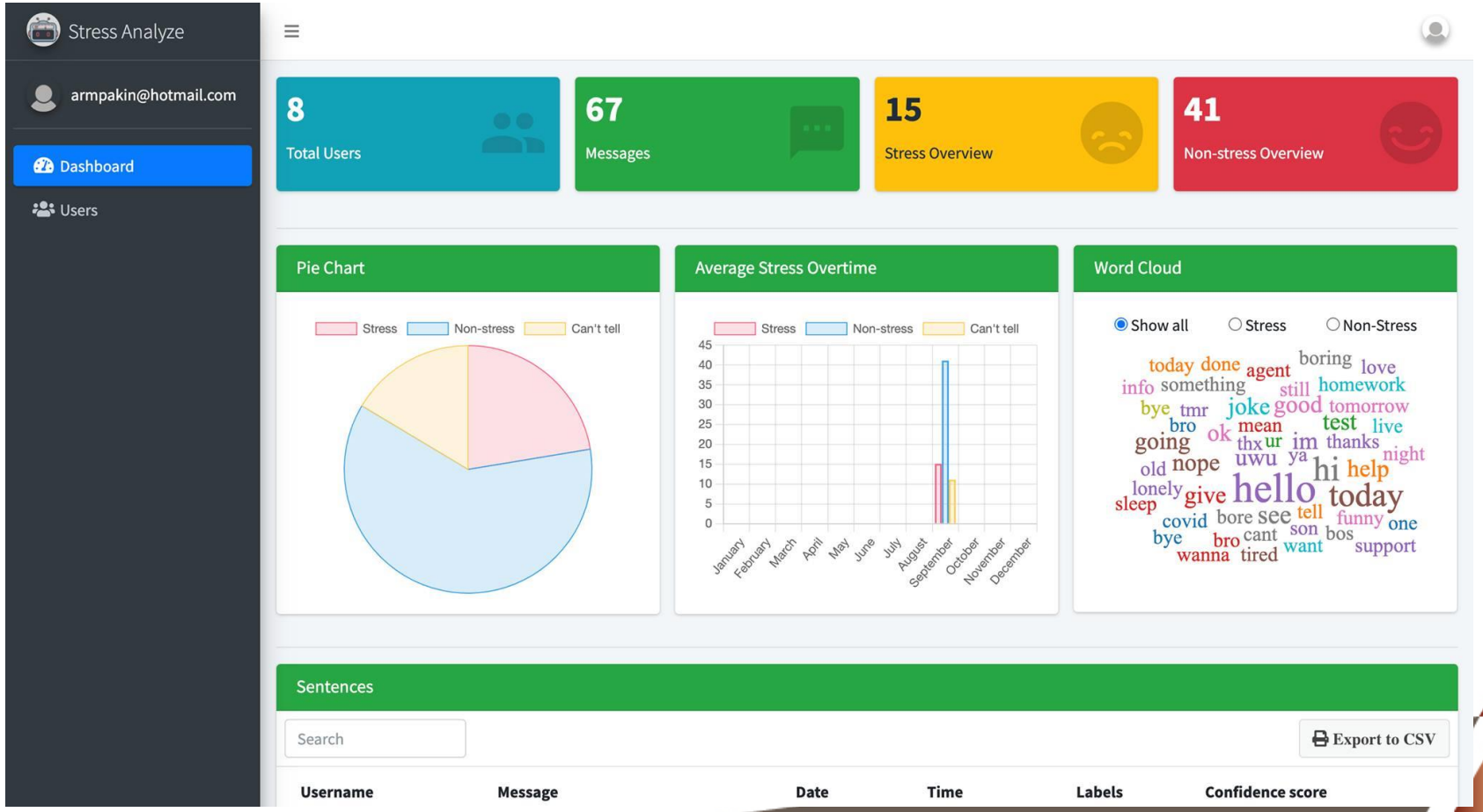


CAMT
College of Arts, Media and Technology
Chiang Mai University

User interface



User interface



CAMT
College of Arts, Media and Technology
Chiang Mai University

User interface

Sentences

Search

Export to CSV

Username	Message	Date	Time	Labels	Confidence score
Pathomsakul Supamanee	UwU mean uwu	9/26/2021	11:37:08 PM	stress	0.5428668336031935
Pathomsakul Supamanee	UwU	9/26/2021	11:36:57 PM	can't tell	-
Pathomsakul Supamanee	r u winning?	9/26/2021	11:36:49 PM	can't tell	-
Pathomsakul Supamanee	hello my son	9/26/2021	11:36:42 PM	non-stress	0.13118900402470252
Pakin Kampeera	see you tmr bye	9/26/2021	9:53:01 PM	stress	0.5542619158800078
Pakin Kampeera	no bro, its still covid 19	9/26/2021	9:52:52 PM	stress	0.9356098723333347
Pakin Kampeera	today is so boring	9/26/2021	9:52:34 PM	non-stress	0.42549680705190424
Pakin Kampeera	How is it going today?	9/26/2021	9:52:22 PM	non-stress	0.42549680705190424
Pakin Kampeera	Hello	9/26/2021	9:52:08 PM	non-stress	0.060582713902607
Pakin Kampeera	see you tomorrow	9/26/2021	9:47:22 PM	stress	0.6573186070424237

10

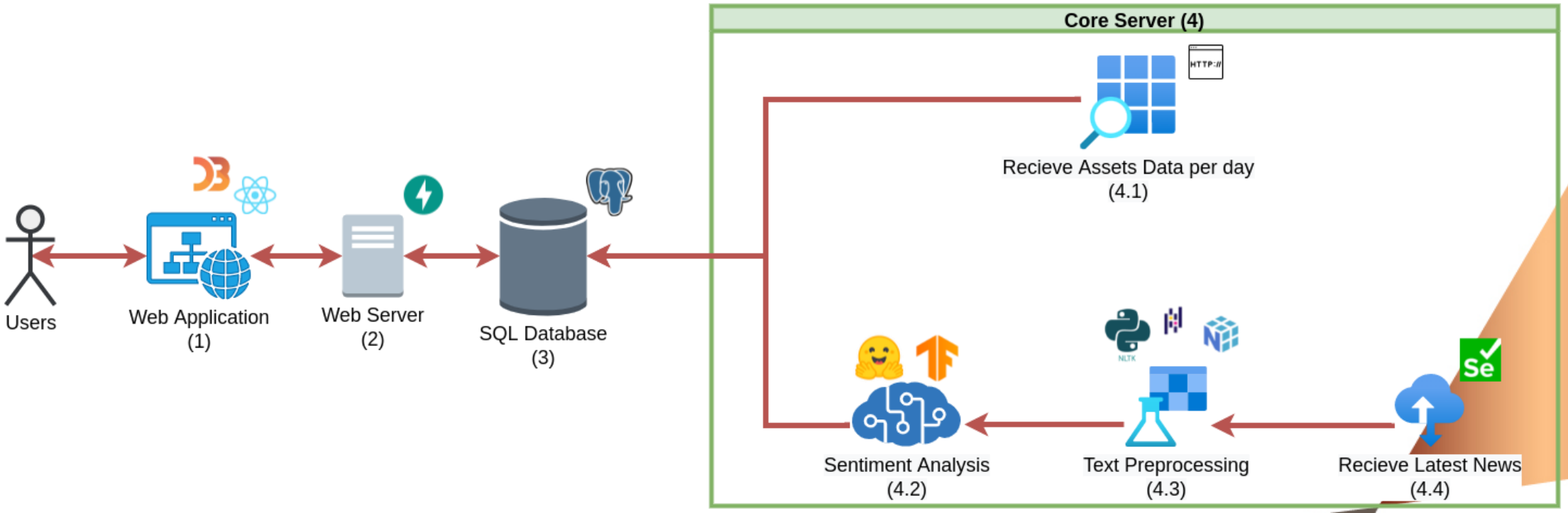
<< < 3 4 5 6 7 >

Version 0.1.0

CAMT
College of Arts, Media and Technology
Chiang Mai University

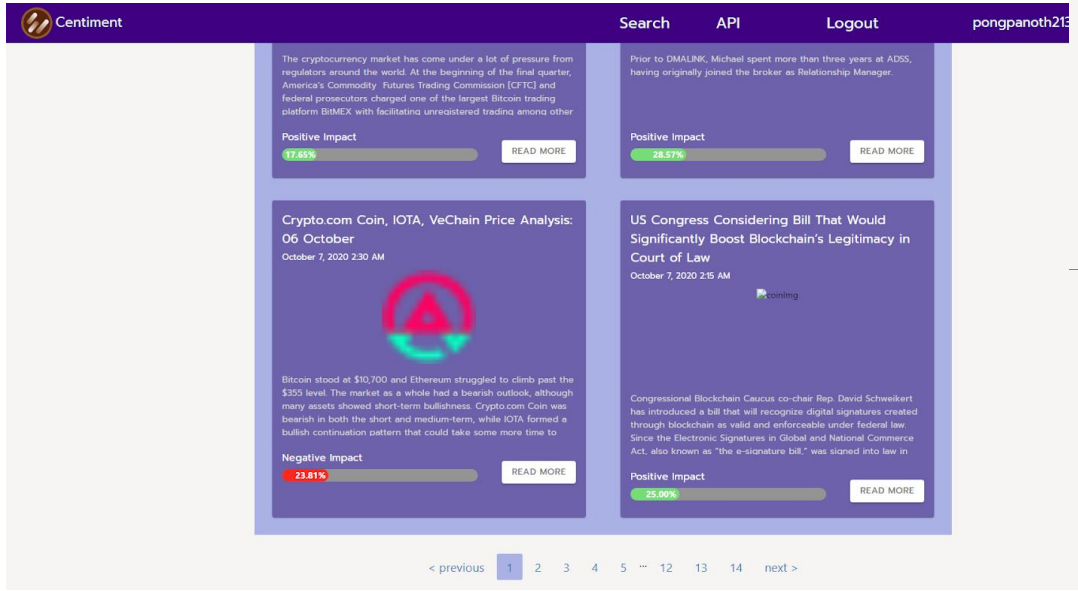
Cryptocurrency News Sentiment Analysis

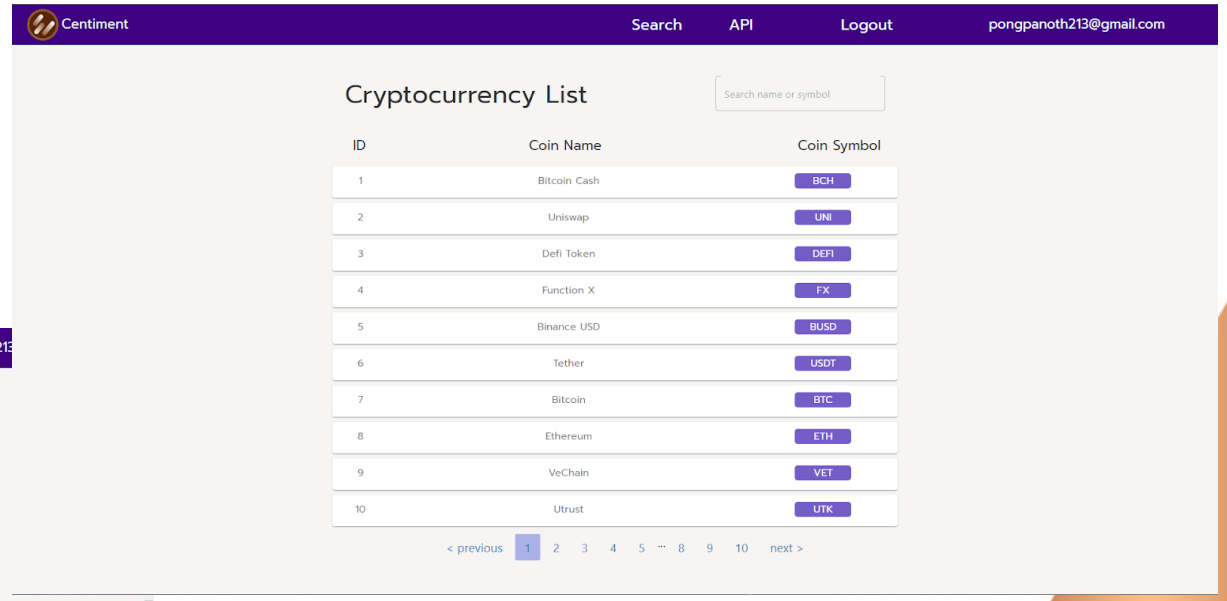
Thanatorn Kanthala 622115014
Pongpanoth Panya 622115024



CAMT
College of Arts, Media and Technology
Chiang Mai University

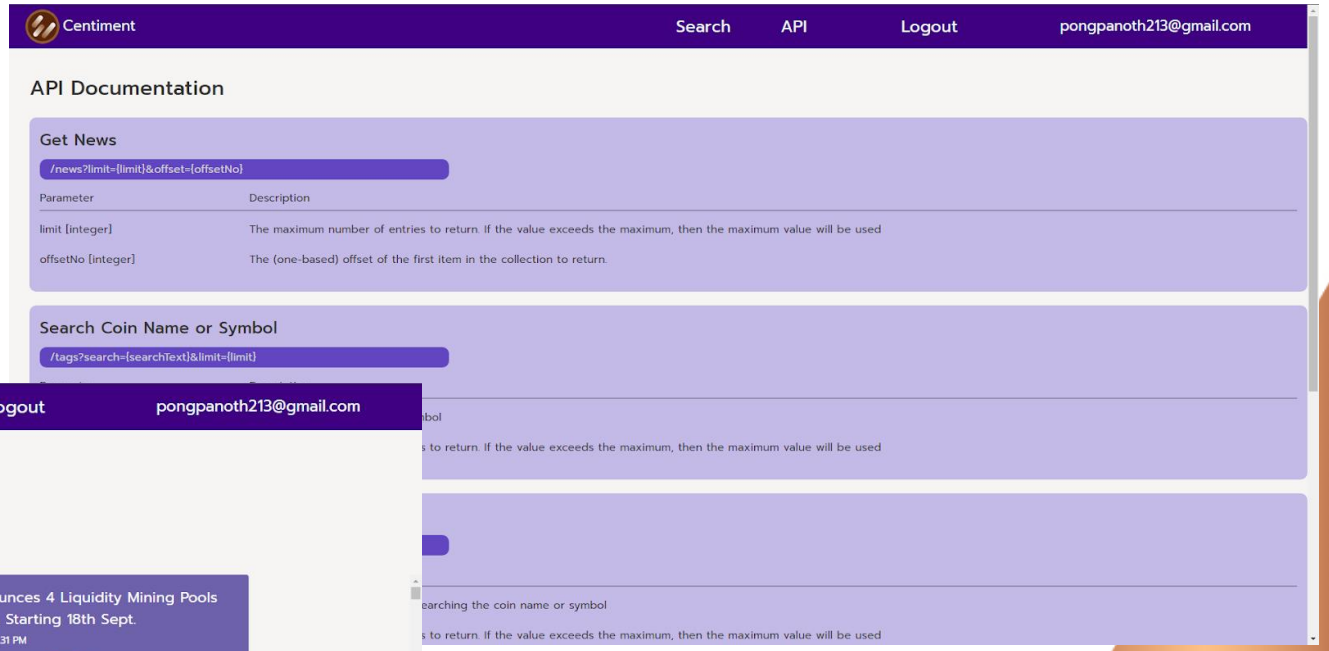
UI

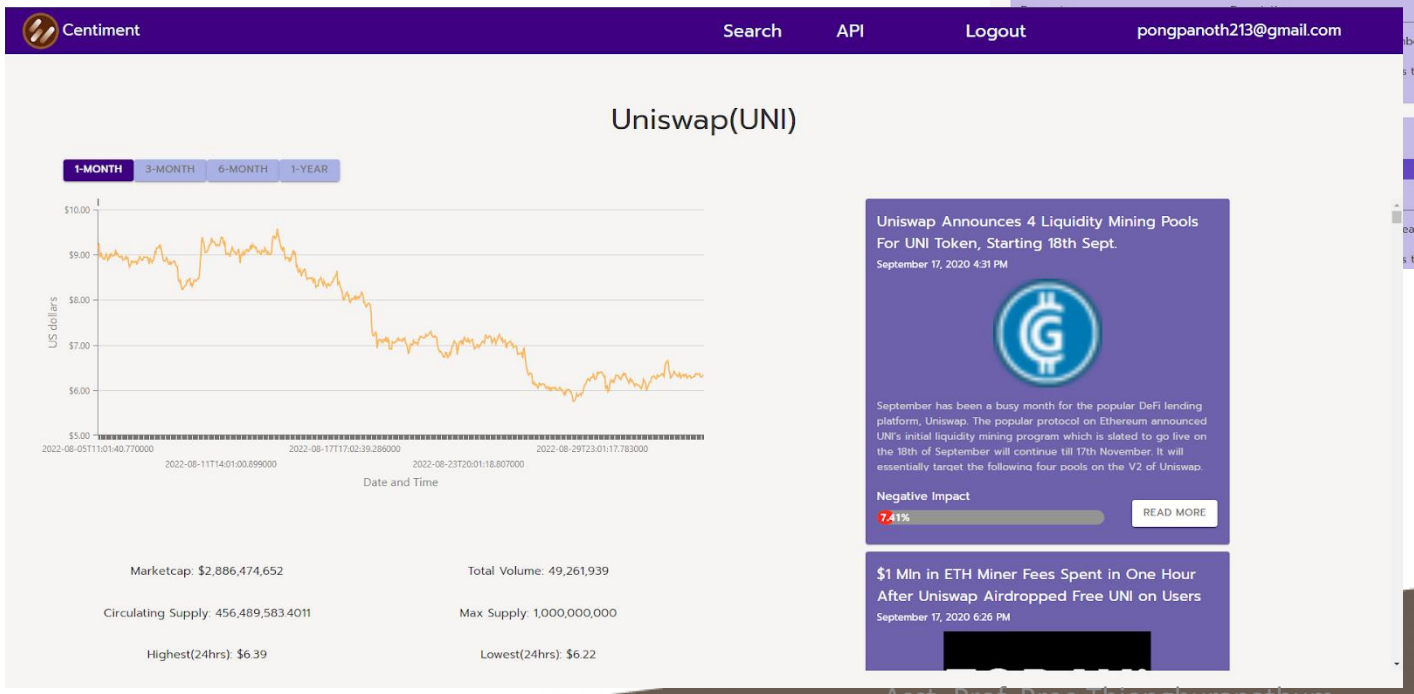




CAMT
College of Arts, Media and Technology
Chiang Mai University

UI


Centiment
Search
API
Logout
pongpanoth213@gmail.com
API Documentation
Get News
/news?limit={limit}&offset={offsetNo}
Parameter
Description
limit [integer]
The maximum number of entries to return. If the value exceeds the maximum, then the maximum value will be used
offsetNo [integer]
The (one-based) offset of the first item in the collection to return.
Search Coin Name or Symbol
/tags?search={searchText}&limit={limit}


Centiment
Search
API
Logout
pongpanoth213@gmail.com
Uniswap(UNI)
1-MONTH
3-MONTH
6-MONTH
1-YEAR
US dollars
Date and Time
Uniswap Announces 4 Liquidity Mining Pools For UNI Token, Starting 18th Sept.
September 17, 2020 4:31 PM
September has been a busy month for the popular DeFi lending platform, Uniswap. The popular protocol on Ethereum announced UNI's initial liquidity mining program which is slated to go live on the 18th of September will continue till 17th November. It will essentially target the following four pools on the V2 of Uniswap.
Negative Impact
7.41%
READ MORE
$1 Mln in ETH Miner Fees Spent in One Hour After Uniswap Airdropped Free UNI on Users
September 17, 2020 6:26 PM
Marketcap: $2,886,474,652
Total Volume: 49,261,939
Circulating Supply: 456,489,583.4011
Max Supply: 1,000,000,000
Highest(24hrs): $6.39
Lowest(24hrs): $6.22

GPT-Baker

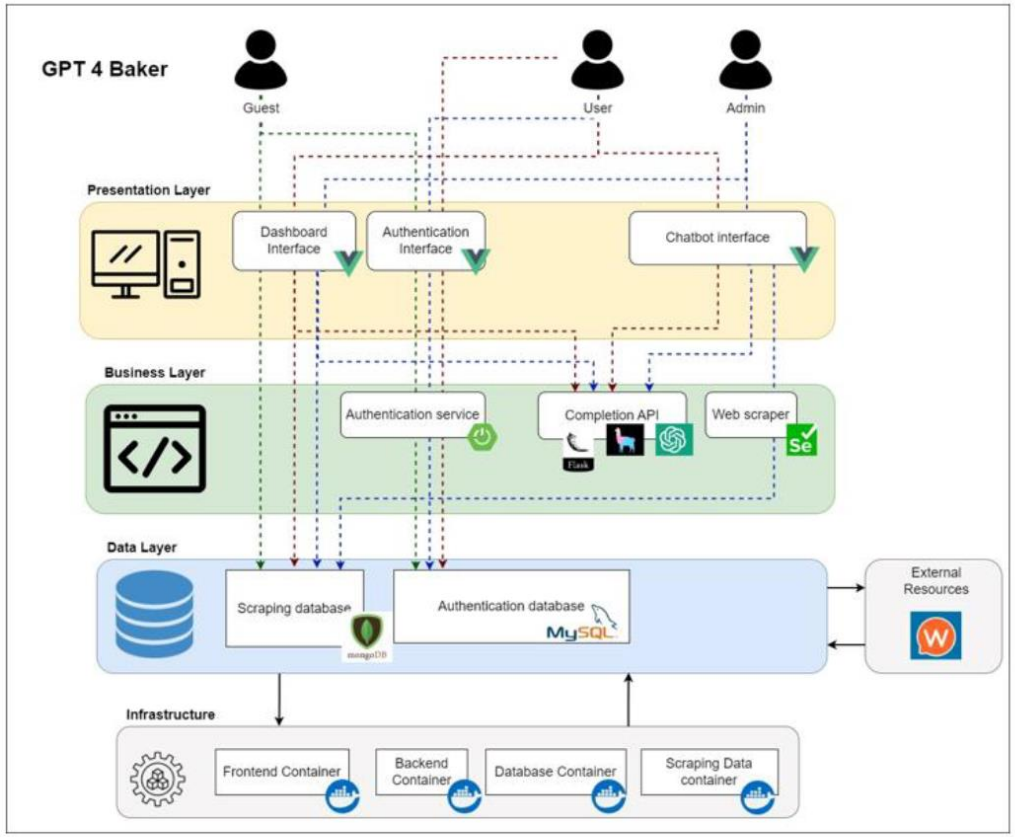


Figure 1: GPT 4 Baker System Architecture

Class activity (if we have time)

- https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

- Answer the following question:
- How many rows/sample?
- What is the longest sample?
- How many word?
- What is the average word length?

References

- Lane, H., Hapke, H., & Howard, C. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python* (1st edition). Manning.
- Turban, E., Delen, D., & Sharda, R. (n.d.). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*.
- Turcan, E., & McKeown, K. (2019). Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. *ArXiv:1911.00133 [Cs]*. http://arxiv.org/abs/1911.00133
- **Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems** 1st Edition