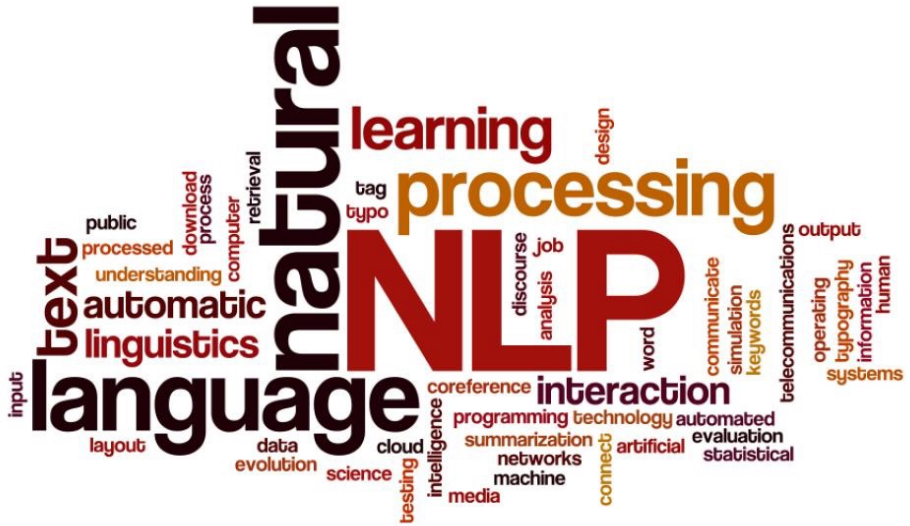
Introduction to NLP

CSX4210/INX4210 Natural Language Processing and Social Interaction

Vincent Mary School of Science and Technology

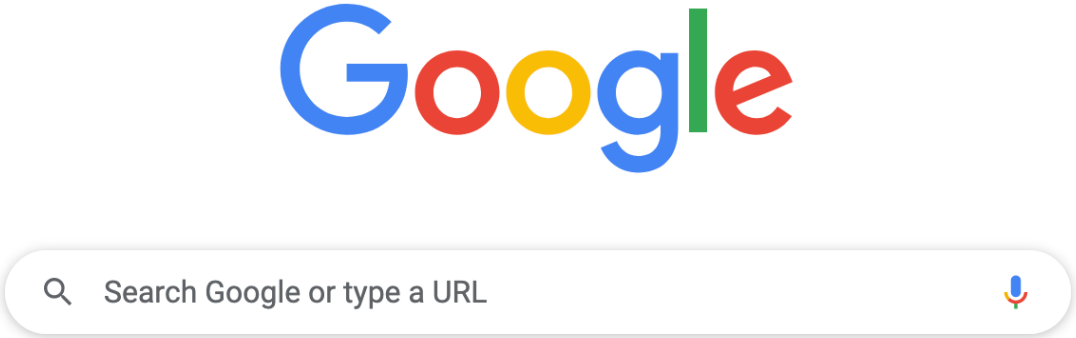
Assumption University



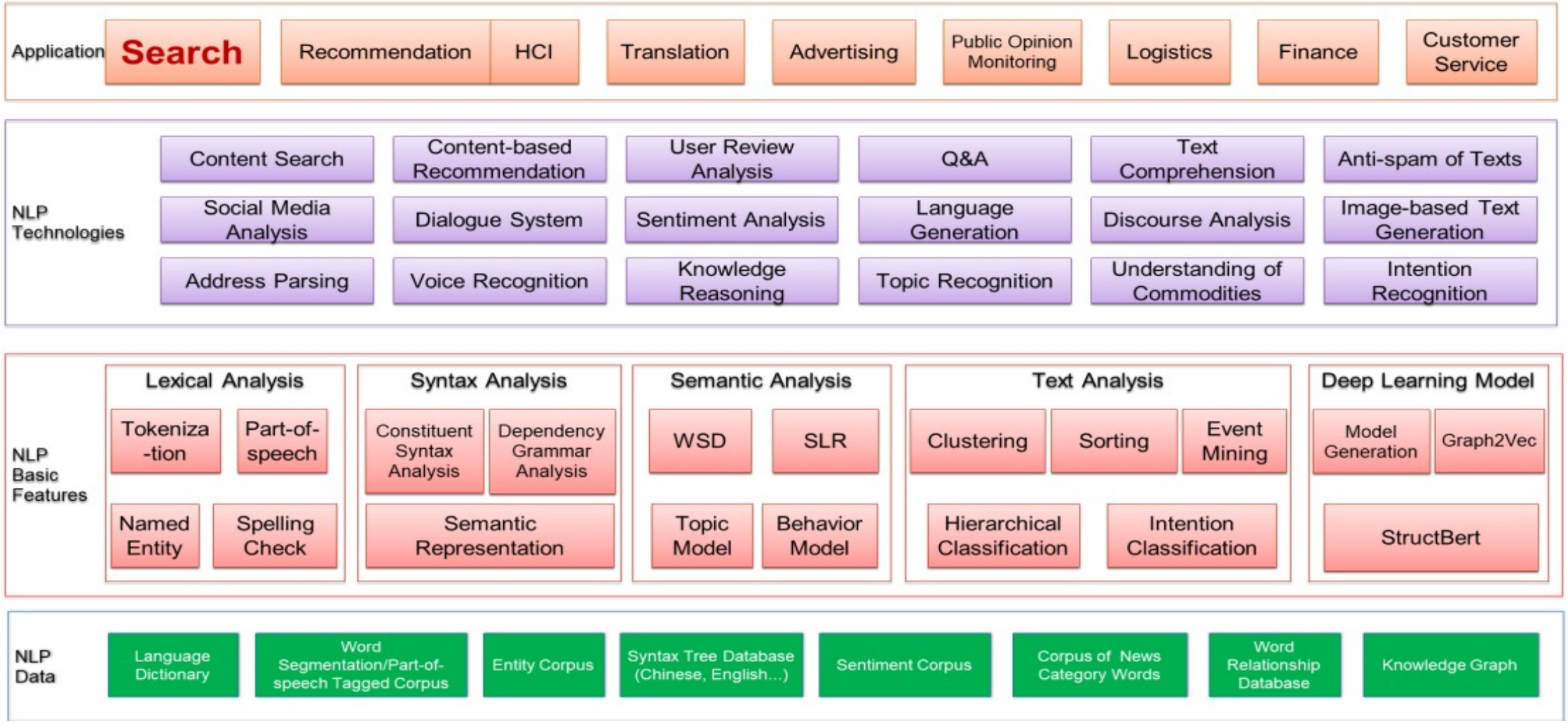
What is NLP?

Natural Language Processing (NLP) is the ability to receive language information, interpret it, complete a task based on the language information and produce appropriate human language. [John Medicine, 2020]

How is NLP used in these applications?



Capabilities of NLP

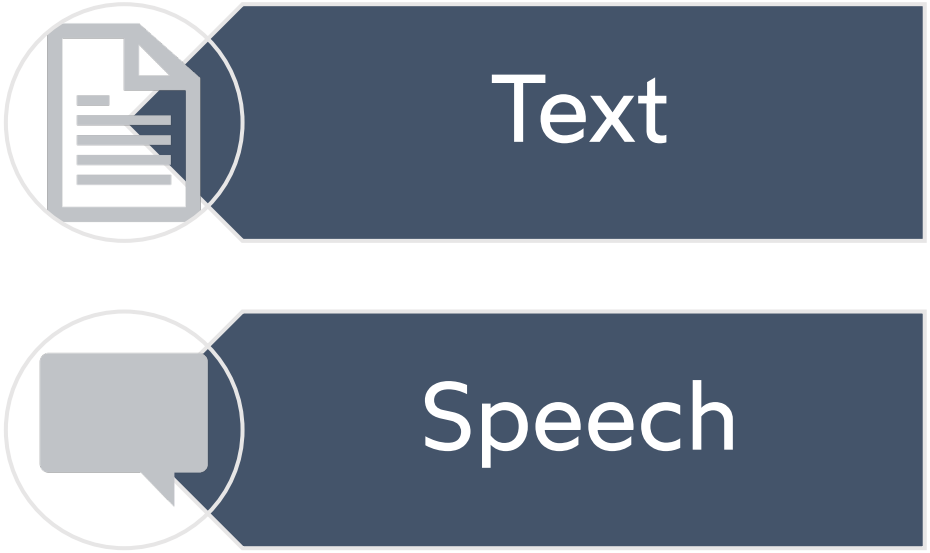


Areas of

NLP

Formats of

Inputs



Challenges in NLP

- Natural language is ambiguous.
- Some words have different meaning in different context (Semantic Ambiguity).

Lexical Ambiguity

John went to the bank.

Financial institution?

Bank of the river?

- Syntactic Ambiguity

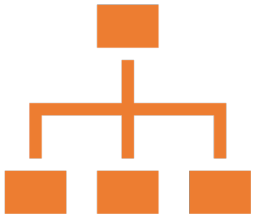
Put the box on the table in the kitchen

Is the box already on the table, and to be put in the kitchen

Is the box to be put on the table which is in the kitchen?

Analysis in

NLP



Syntactic Analysis

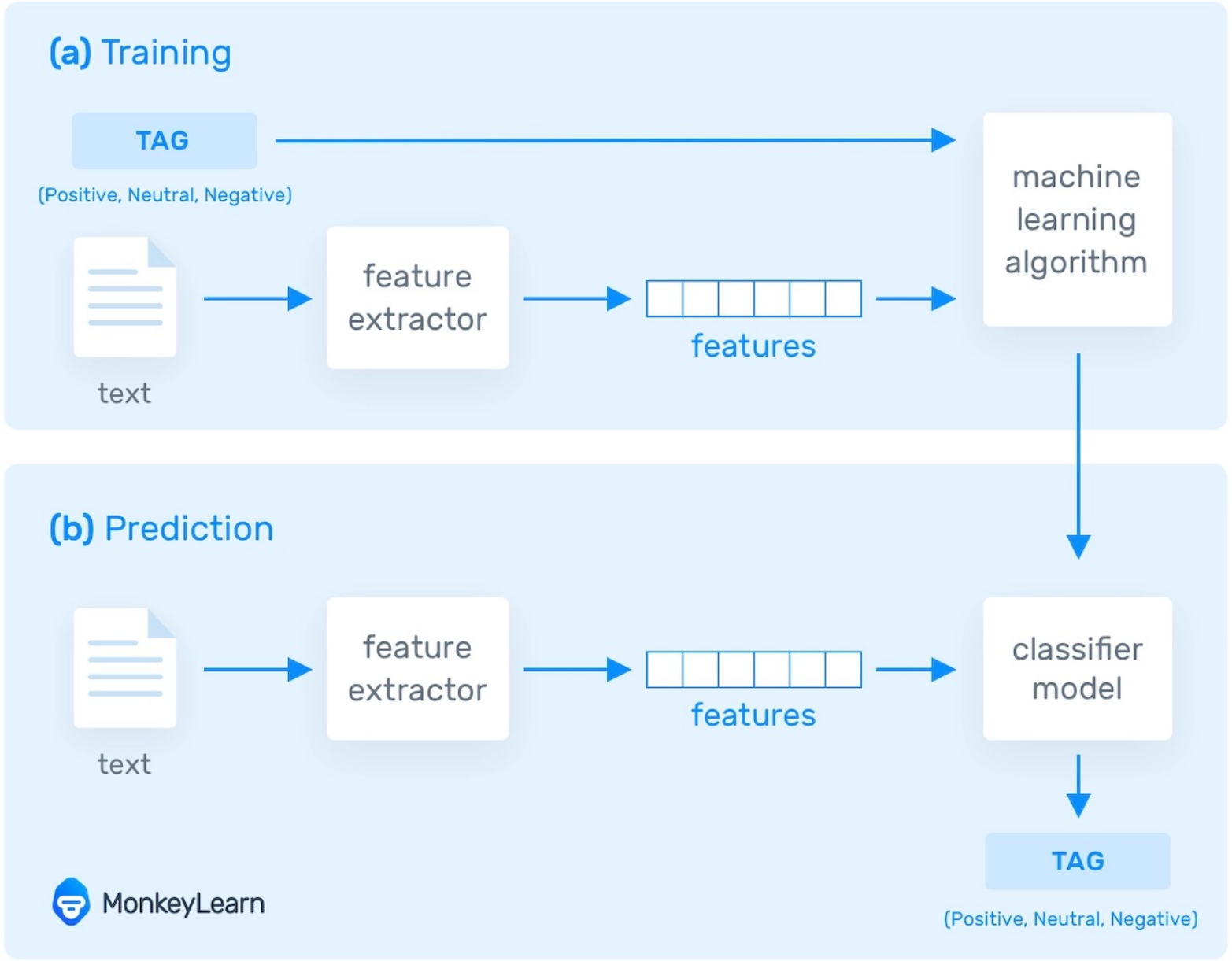
Structure

Semantic Analysis

Meaning

NLP
Pipeline

ML Process in NLP



Preparing/cleaning the data in such a way that it can be fed to the model.

Pre-Processing

Pre-Processing

- Tokenization
- Stop Words Removal
- Normalization
  - Stemming
  - Lemmatization
- Spelling Correction

Tokenization

- The process of separating streams of text into smaller units (which can be sentences, tokens or words).
- Sentence tokenization breaks long text into sentences.
- Word tokenization splits text in a sentence into words.
- Language Issues:
  - No word boundaries (no space between words) e.g. Chinese, Japanese, Thai
  - No sentence boundaries (no period at the end of the sentence) e.g. Thai
- Challenges:
  - Handling words that are often appearing together.
  - Artificial Intelligence, Computer Science, Machine Learning

Example: Tokenization

- I love to travel.

I | love | to | travel

- Artificial Intelligence can make it better.

Artificial | Intelligence | can | make | it | better

Artificial Intelligence | can | make | it | better

Stop Words Removal

- Stop words are commonly used words in the language.
  - a, an, the, in, on, and etc.
- They are normally considered unimportant; hence, usually are removed so the algorithm can focus on the more important ones.
- The set of stop words to use and the decision whether to remove them depend on applications.

i, in, on, my, to

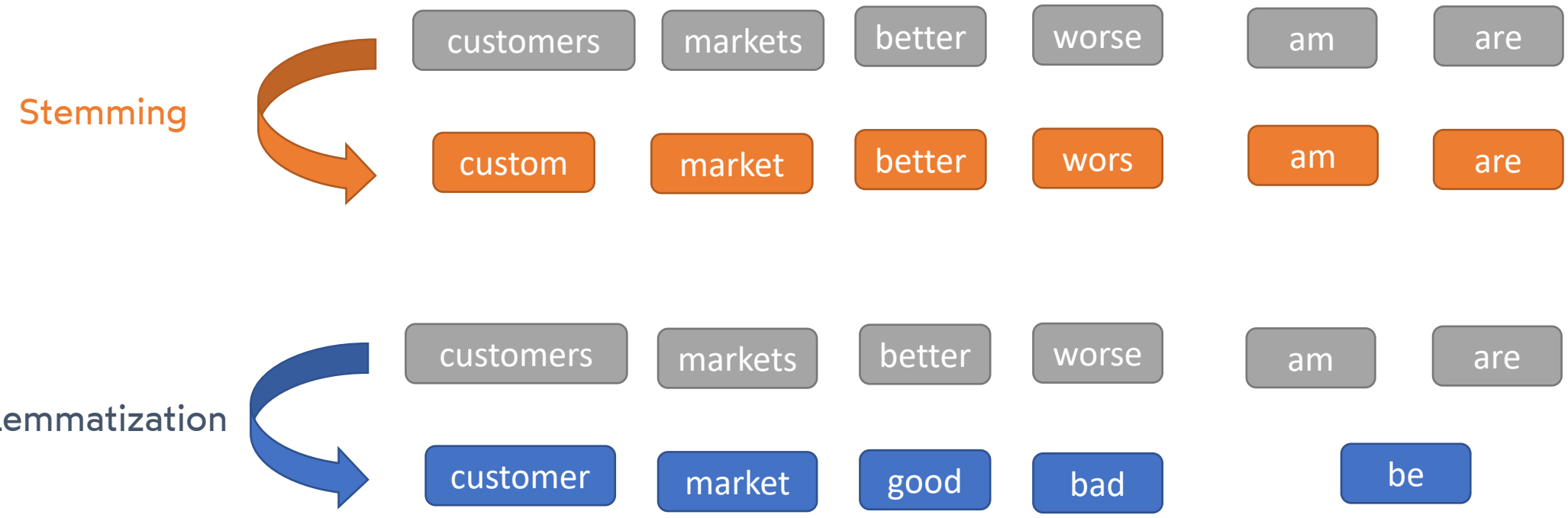
I need help on how to adopt AI in my business.

need help how adopt AI business

Normalization

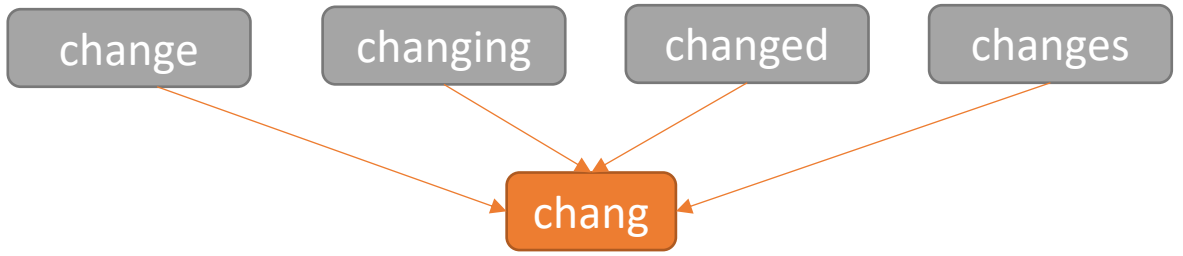
- Both lemmatization and stemming reduce the inflected word to its root form.
- Stemming
  - is based on heuristic algorithm that removes the end of words.
  - may not give out the actual language words.
- Lemmatization
  - uses morphological analysis of words.
  - returns the dictionary form of the words (*aka* lemma).

Example: Stemming vs. Lemmatization

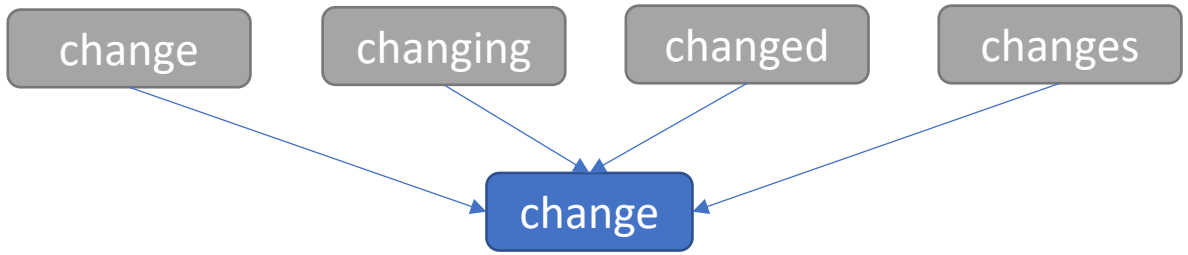


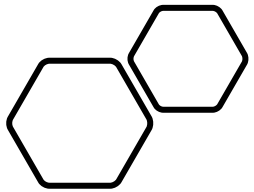
Example: Stemming vs. Lemmatization

Stemming



Lemmatization





NLP Subtasks

Part-of-Speech Tagging

- *aka* POS Tagging
- The process of assigning a part-of-speech or lexical class marker to each word in a collection.
- Traditional Parts of Speech
  - noun, verb, adjective, preposition, adverb, article, interjection, pronoun, conjunction, etc.

The	suite	was	very	good	with	comfortable	furniture
DET	N	V	ADV	ADJ	P	ADJ	N

Named Entity Recognition

- The process of detecting named entities in the given text.
- Useful in co-reference resolution.
- Examples of named entities:
  - Person's name e.g. James Bond, Joe Biden
  - Location e.g. Bangkok, New York City
  - Organization name e.g. Assumption University, Bank of America

Joe Biden is the current president of the United States of America.

Joe Biden is the current president of the United States of America.

Person

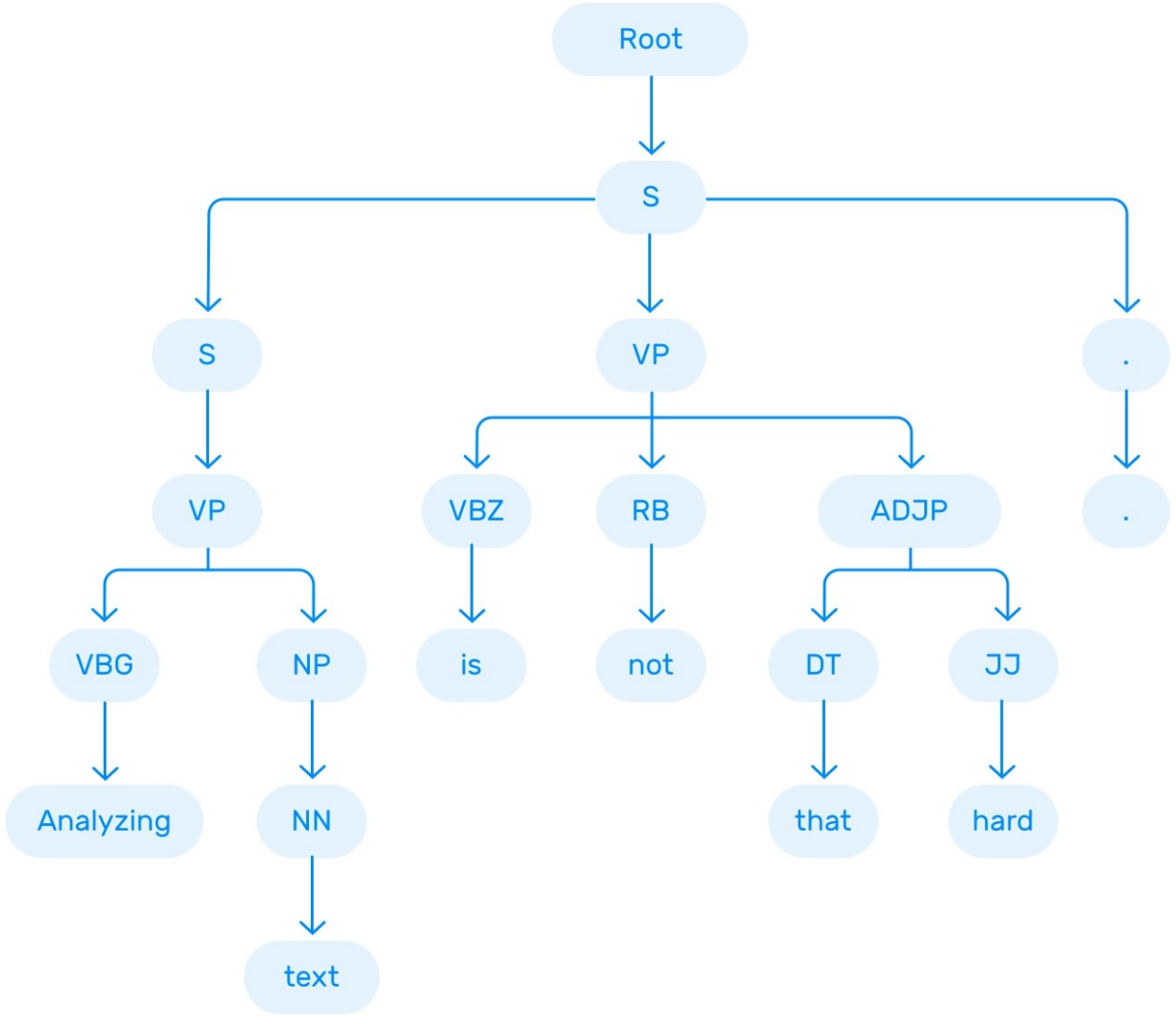
Country

Parsing

Understanding the structure of the text, check if it conforms to the grammar and form a parse tree.

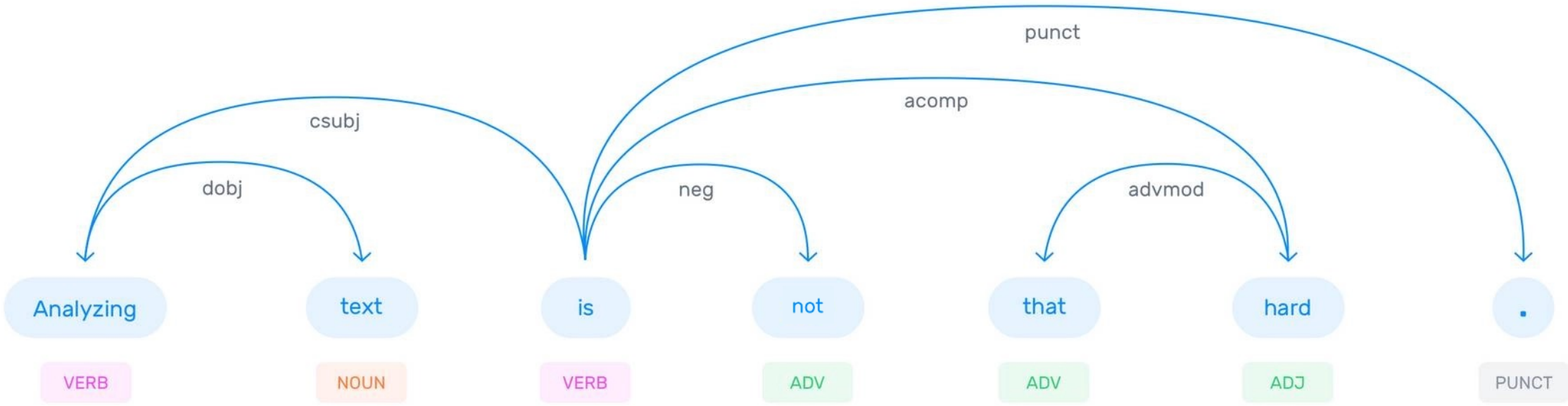
Constituency Parsing

Analyze the syntactic structure of a sentence (phrase structure).



Dependency Parsing

Analyze how words in a sentence are related.



Word Sense Disambiguation

- *aka* WSD
- The process of determining which sense of a word is meant in a sentence.

frequency

I can hear bass sound.

He likes to eat grilled bass.

fish

NLP
Tasks

Topic Modeling

Text Summarization

Topic Classification

Sentiment Analysis

Aspect-Based Sentiment Analysis

Emotion Analysis

Machine Translation

Speech Recognition (Speech-to-Text)

Speech Synthesis (Text-to-Speech)

and many more...

Topic Modeling

- Discover hidden topics or themes from a collection of documents

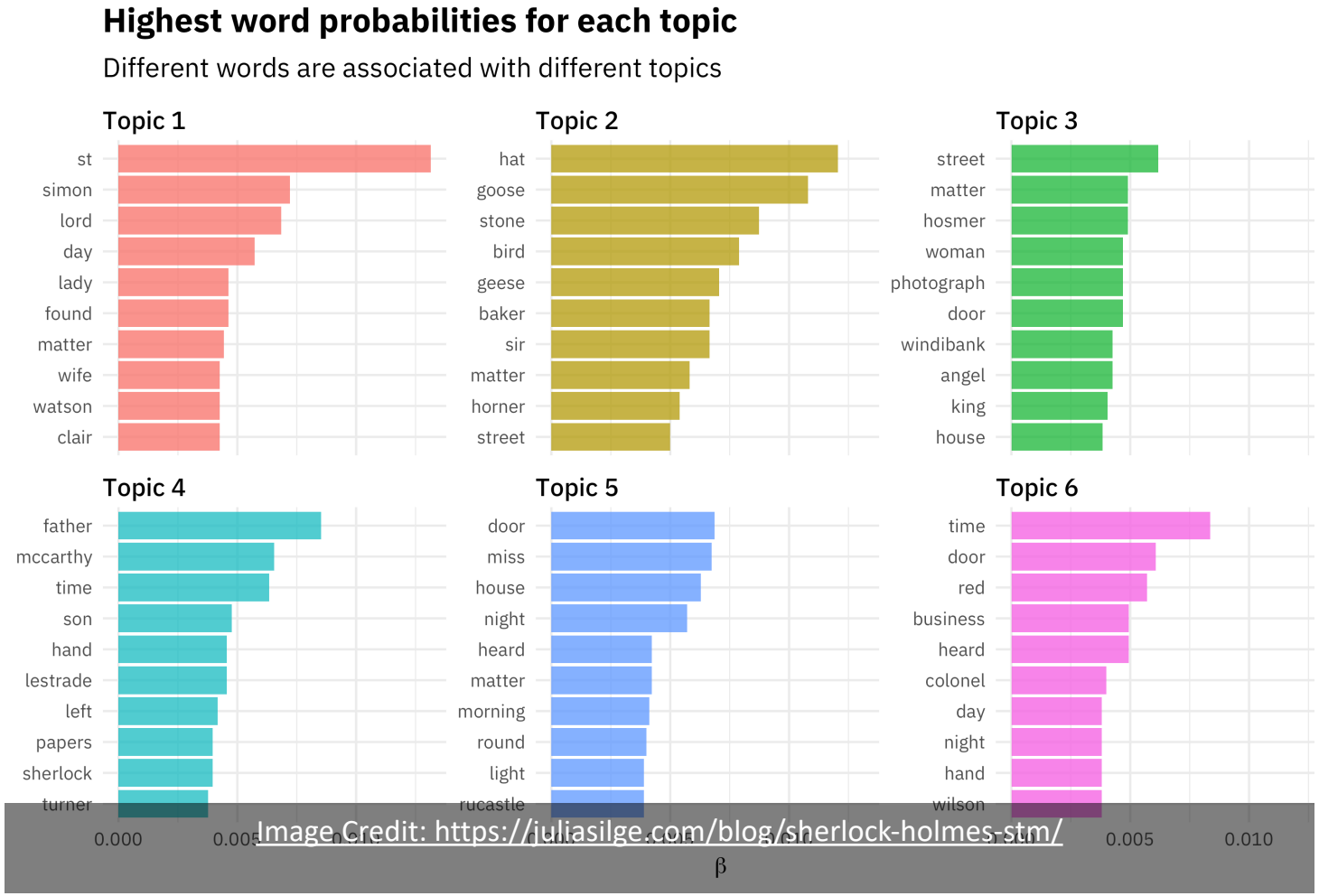


Image Credit: https://juliasilge.com/blog/sherlock-holmes-stm/

Text Summarization

- Summarize the given text.

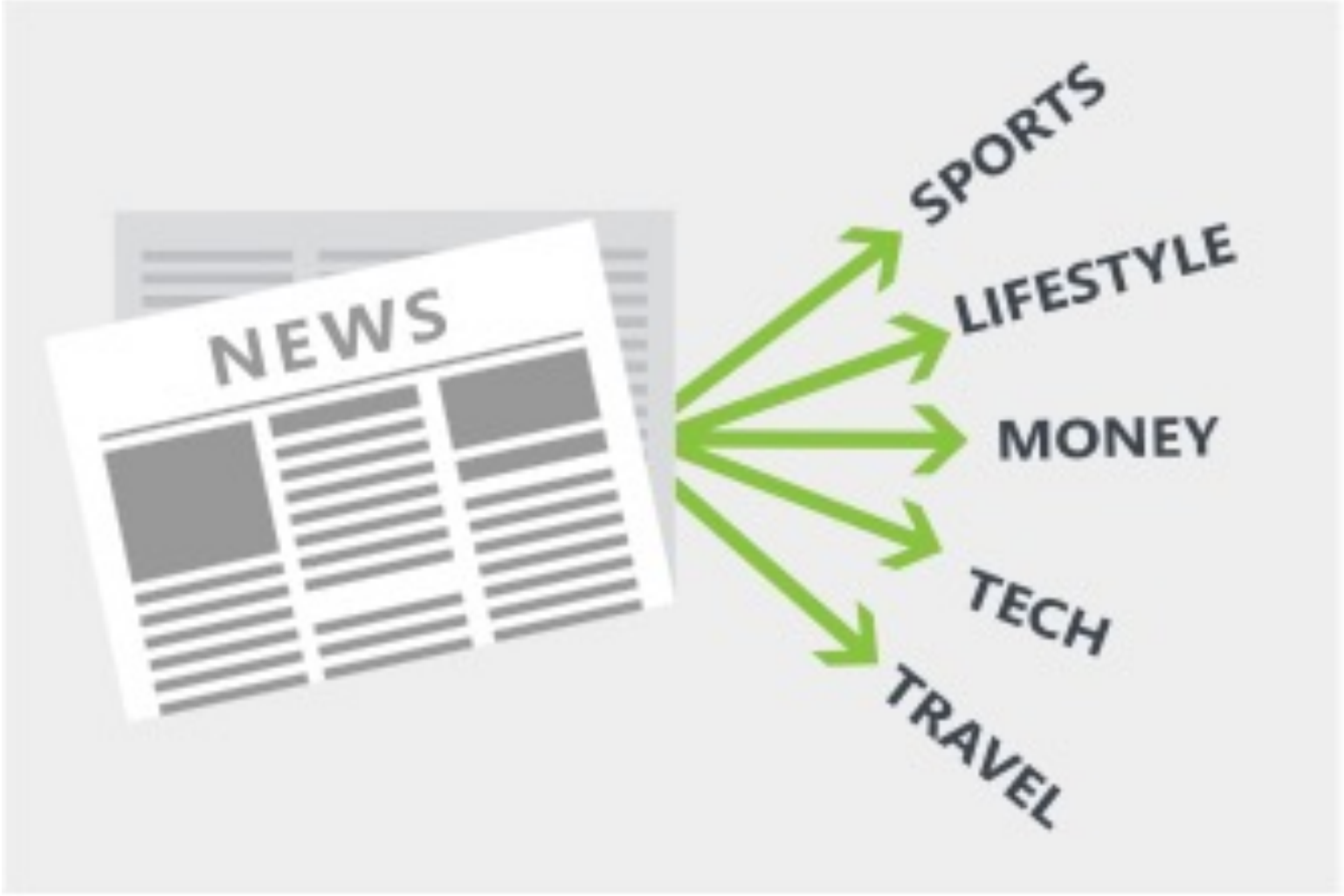
Summarize this for a second-grade student:

Jupiter is the fifth planet from the Sun and the largest in the Solar System. It is a gas giant with a mass one-thousandth that of the Sun, but two-and-a-half times that of all the other planets in the Solar System combined. Jupiter is one of the brightest objects visible to the naked eye in the night sky, and has been known to ancient civilizations since before recorded history. It is named after the Roman god Jupiter.[19] When viewed from Earth, Jupiter can be bright enough for its reflected light to cast visible shadows,[20] and is on average the third-brightest natural object in the night sky after the Moon and Venus.

Jupiter is the fifth planet from the Sun and the biggest in our Solar System. It is so bright that you can see it in the night sky. It was named after the Roman god Jupiter and it is the third brightest thing in the sky after the Moon and Venus.

Topic Classification

- Classify a document or article based on a predefined set of topics.



Sentiment Analysis

- Determine general feelings e.g. positive or negative.

Aspect-Based Sentiment Analysis

- Identify the fine-grained polarity or sentiment towards an aspect.

Very friendly and helpful staff. Amazing and very quiet location directly at the beach.

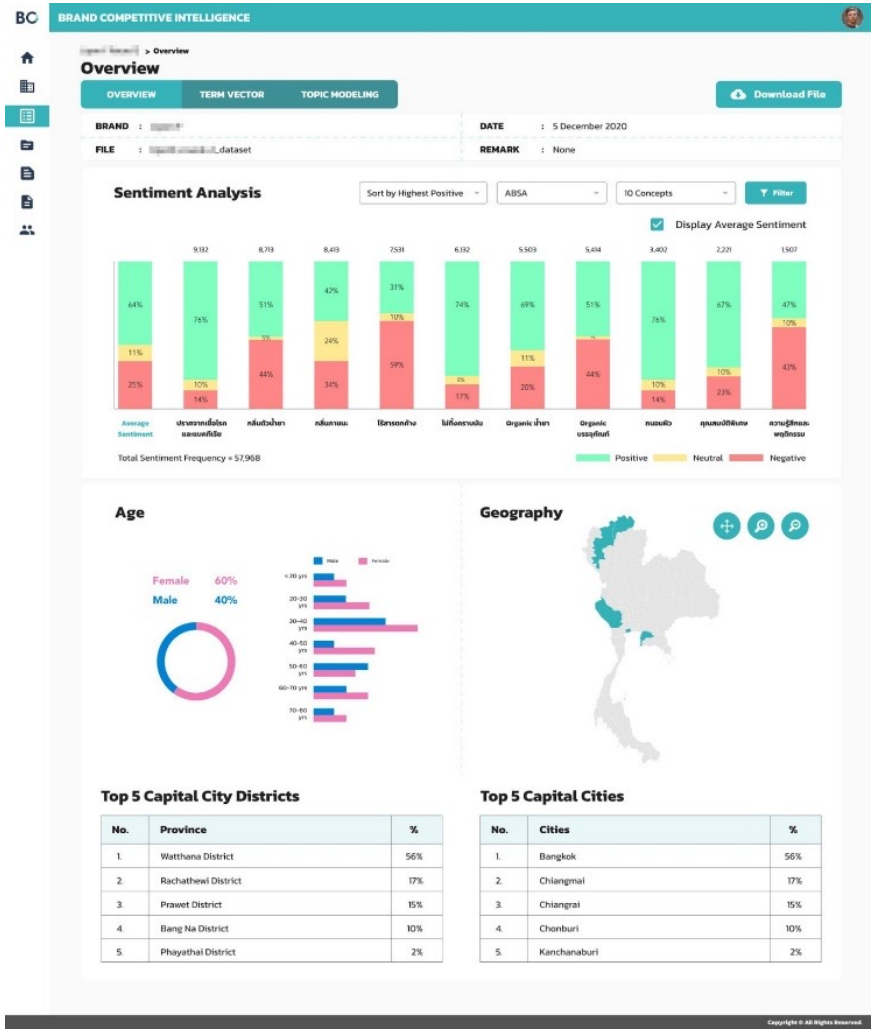
The grounds were beautiful but the price of the townhouse was a little pricey considering what you got

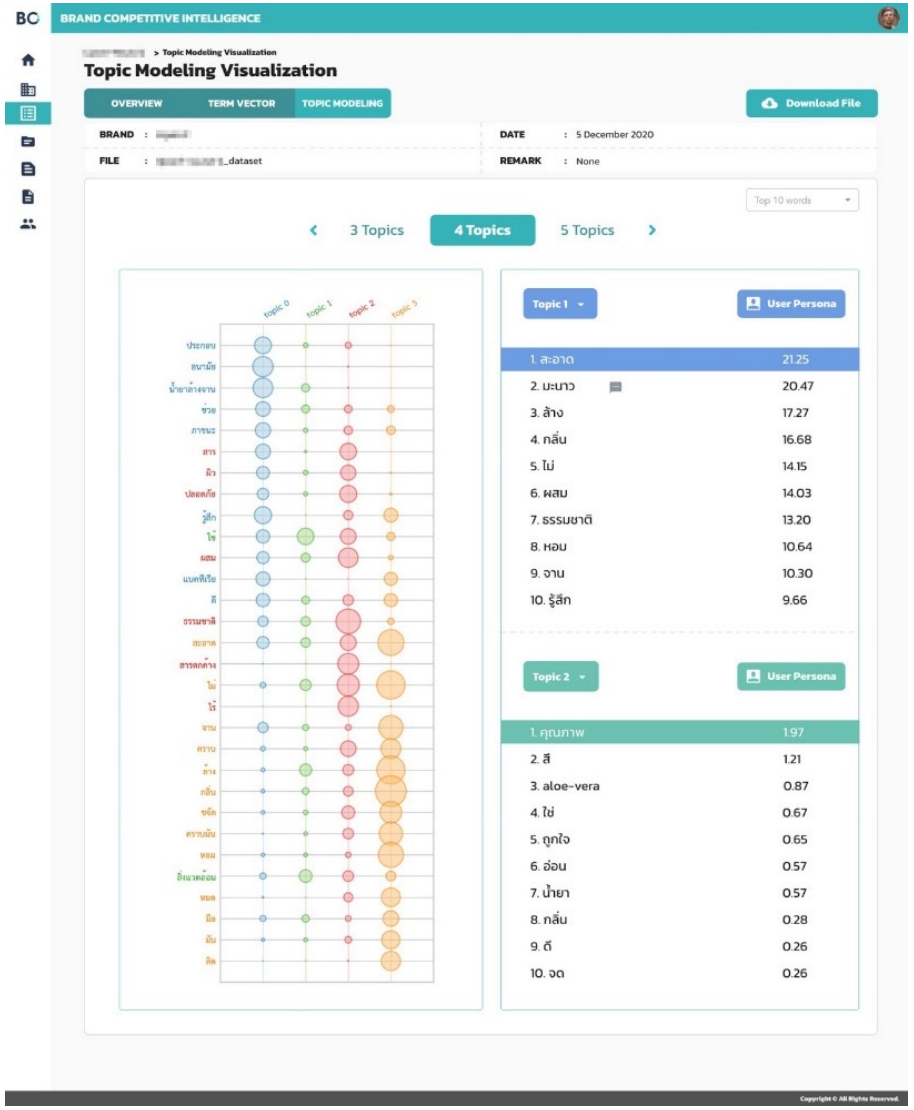
Emotion Analysis

- Determine emotions conveyed in the given data (text or speech).

NLP

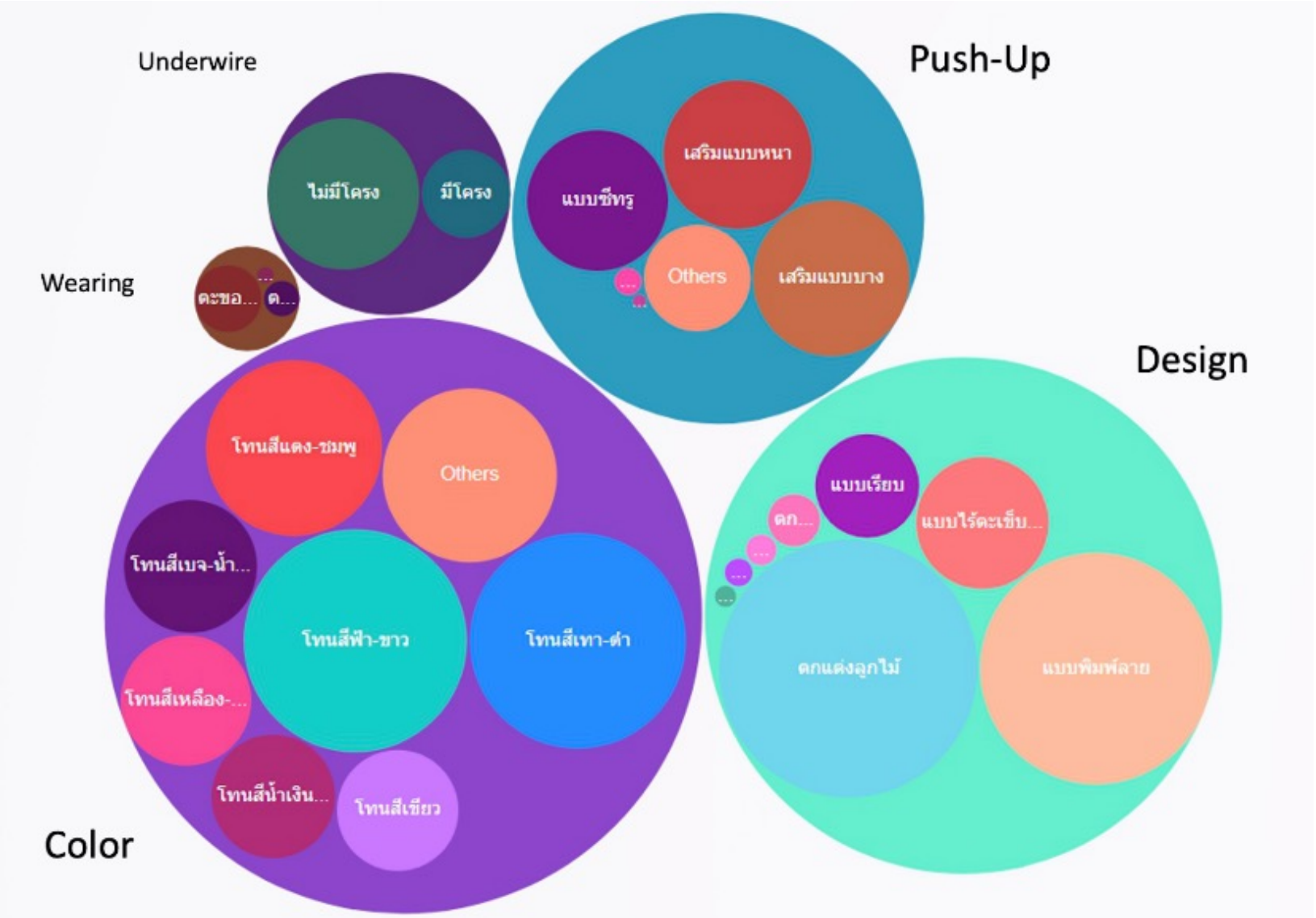
for Market Research





NLP

for Market Research



New Kid in Town

ChatGPT

- Chat Generative Pre-Trained Transformer
- A chatbot by OpenAI launched in November 2022

- Try it out at chat.openai.com

Conversational AI

ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

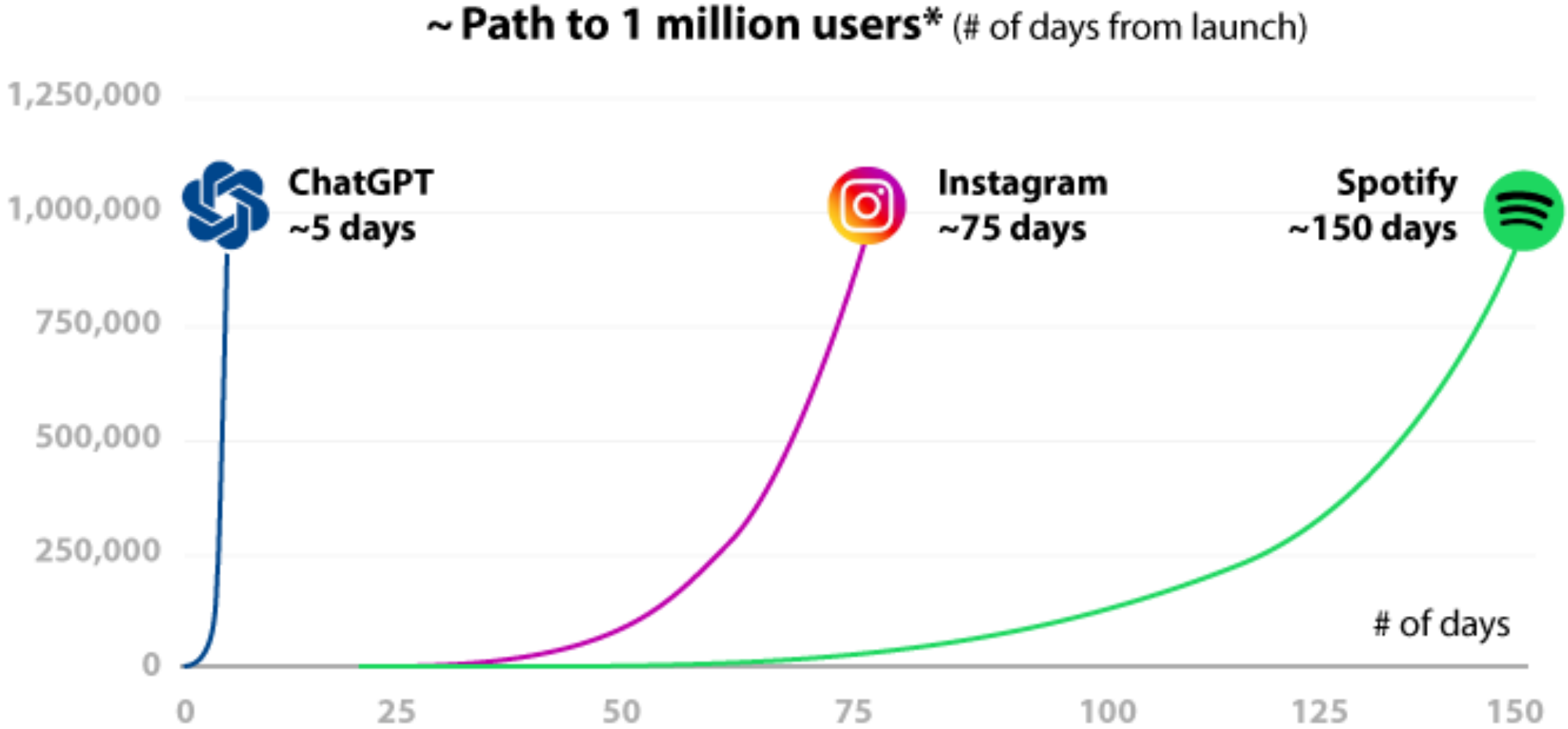
Trained to decline inappropriate requests

Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021



Sources: Google, Subredditstats, Media Reports