NLP – Data Cleaning

Penyusun Modul : Dennis Laorens Bawole
Editor : Rina Fitriyani

A Basic Thought Process

1. Start with a question
2. Get and clean data
3. Perform EDA
4. Apply Techniques
5. Share Insights

"If I exercise more, will I have better stamina?"

A Basic Thought Process

1. Start with a question
2. Get and clean data
3. Perform EDA
4. Apply Techniques
5. Share Insights

Name	Avg Running Time	Workout
Dennis	1 hr 5 mins	2 hrs/day
Febri	105 mins	3 hrs/day
Abay	40 mins	1 hrs/day
Dijung	35 mins	50 mins/day
Fitri	1 hr 15 mins	90 mins/day
Citra	100 mins	150 mins/day
Ade	10 mins	20 mihs/day

A Basic Thought Process

1. Start with a question
2. Get and clean data
3. Perform EDA
4. Apply Techniques
5. Share Insights

Name	Avg Running Time	Workout
Dennis	**1 hr 5 mins**	**2 hrs/day**
Febri	105 mins	**3 hrs/day**
Abay	40 mins	**1 hrs/day**
Dijung	35 mins	50 mins/day
Fitri	**1 hr 15 mins**	90 mins/day
Citra	100 mins	150 mins/day
Ade	10 mins	20 mins/day

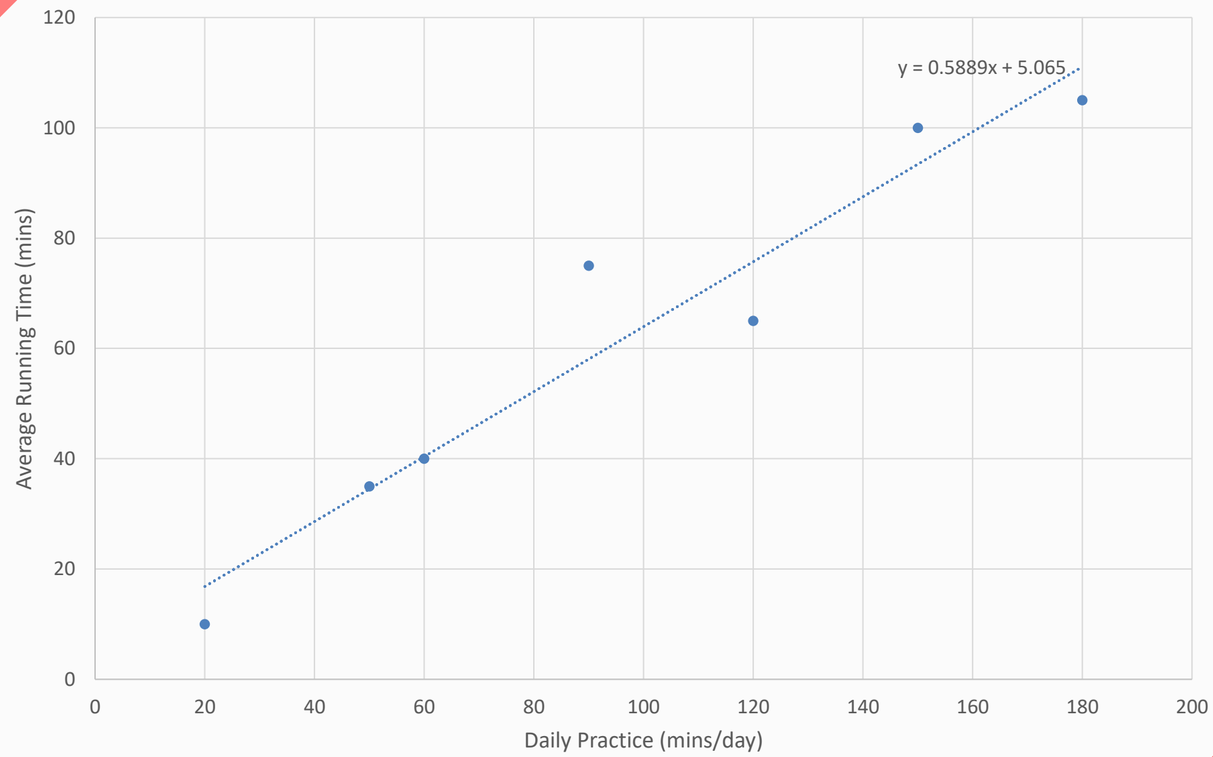
A Basic Thought Process

1. Start with a question
2. Get and clean data
3. Perform EDA
4. Apply Techniques
5. Share Insights

Name	Avg Running Time	Workout
Dennis	65 mins	120 mins/day
Febri	105 mins	180mins/day
Abay	40 mins	60 mins/day
Dijung	35 mins	50 mins/day
Fitri	75 mins	90 mins/day
Citra	100 mins	150 mins/day
Ade	10 mins	20 mihs/day

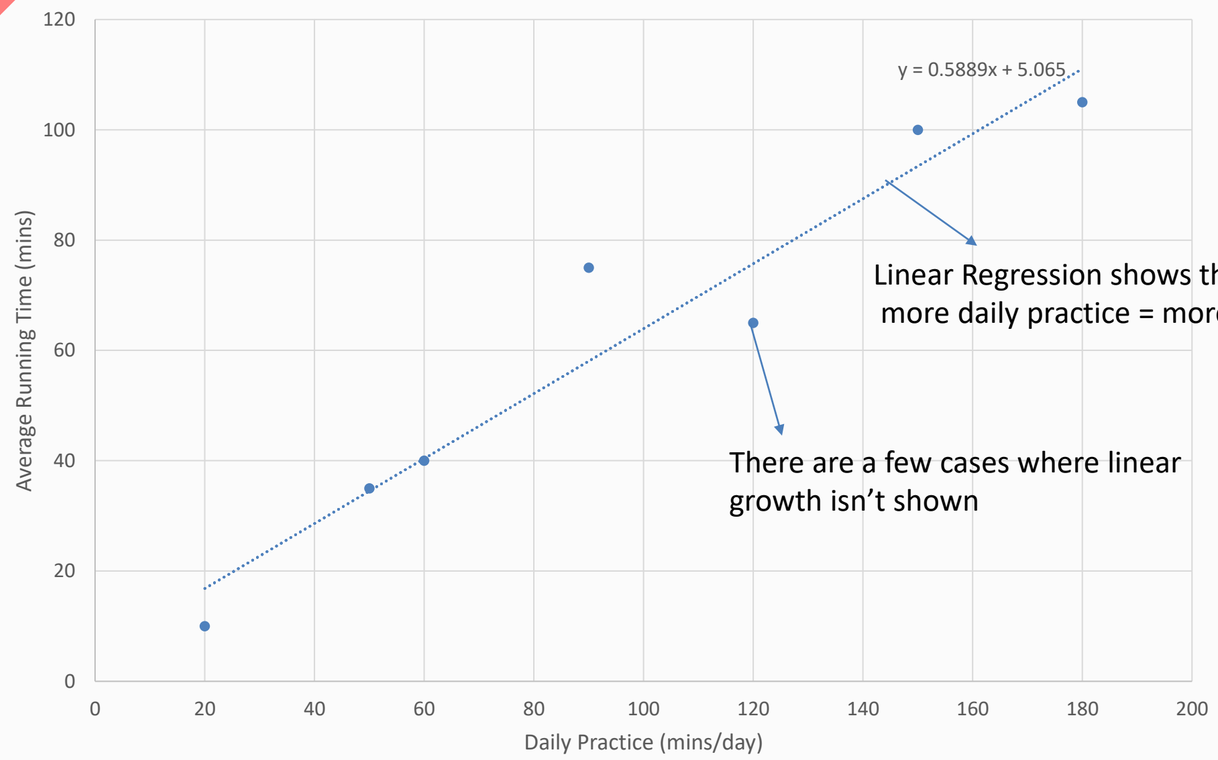
A Basic Thought Process

1. Start with a question
2. Get and clean data
3. Perform EDA
4. Apply Techniques
5. Share Insights



A Basic Thought Process

1. Start with a question
2. Get and clean data
3. Perform EDA
4. Apply Techniques
5. Share Insights



A Basic Thought Process

1. Start with a question
2. Get and clean data
3. Perform EDA
4. Apply Techniques
5. Share Insights

"If I exercise more, will I have better stamina?"

The data shows that there is a **positive correlation** between the daily practice time and the average running time before a person needs rest

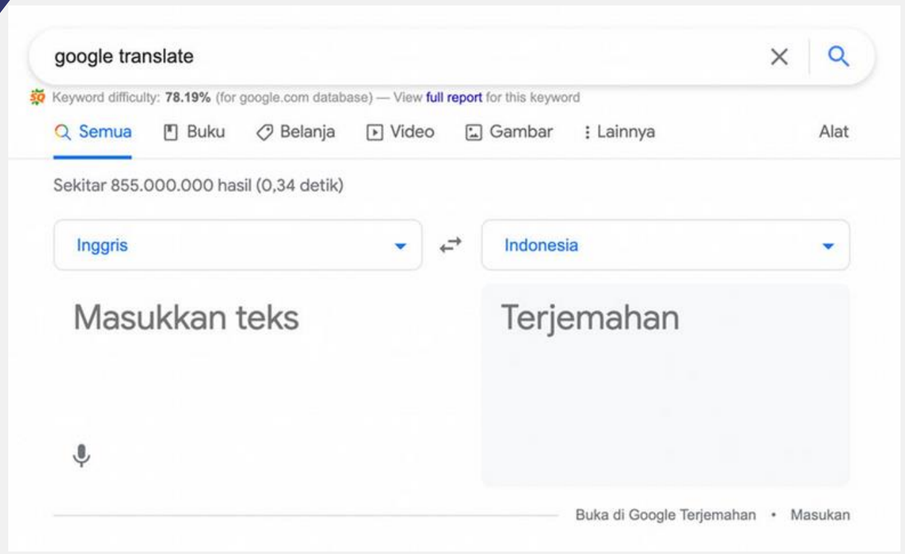
There will be cases where the data doesn't match up with the trendline which shows that a small population can be unique and different

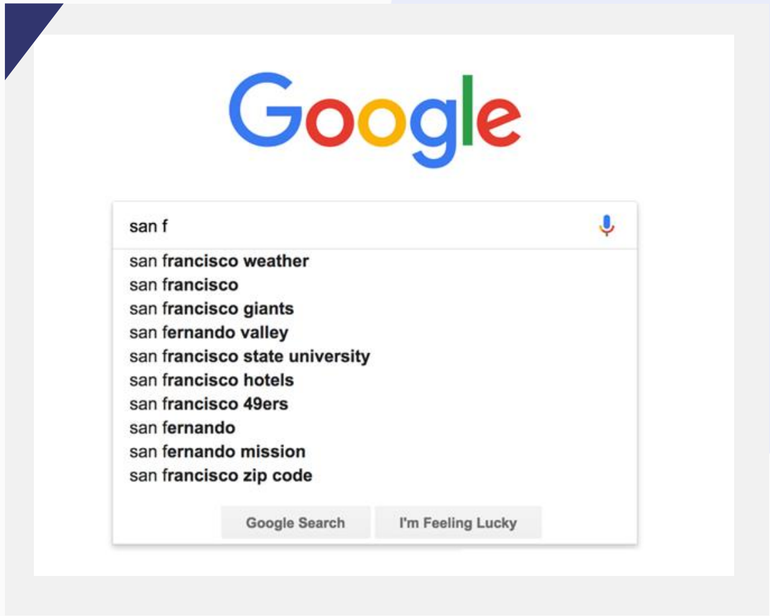
What's our Question?

For NLP, a "word" based data question should be asked, the running problem isn't a case where NLP can be used for, so what are some good questions to be asked?

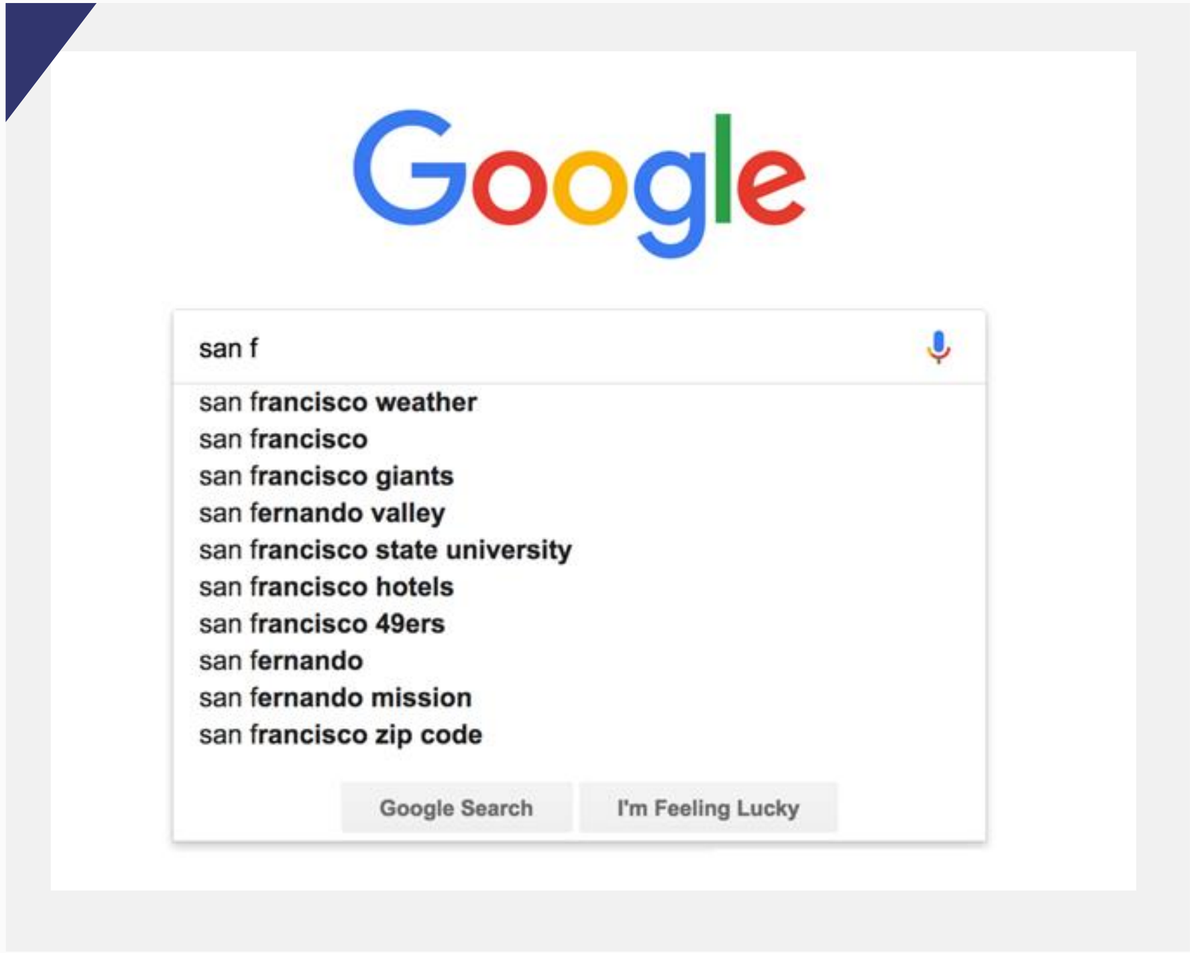
What’s our Question?

Google Assistant





What’s our Question?



What is a good movie like Spiderman?

What is a similar movie like Spiderman: No Way Home?

Data Gathering + Cleaning

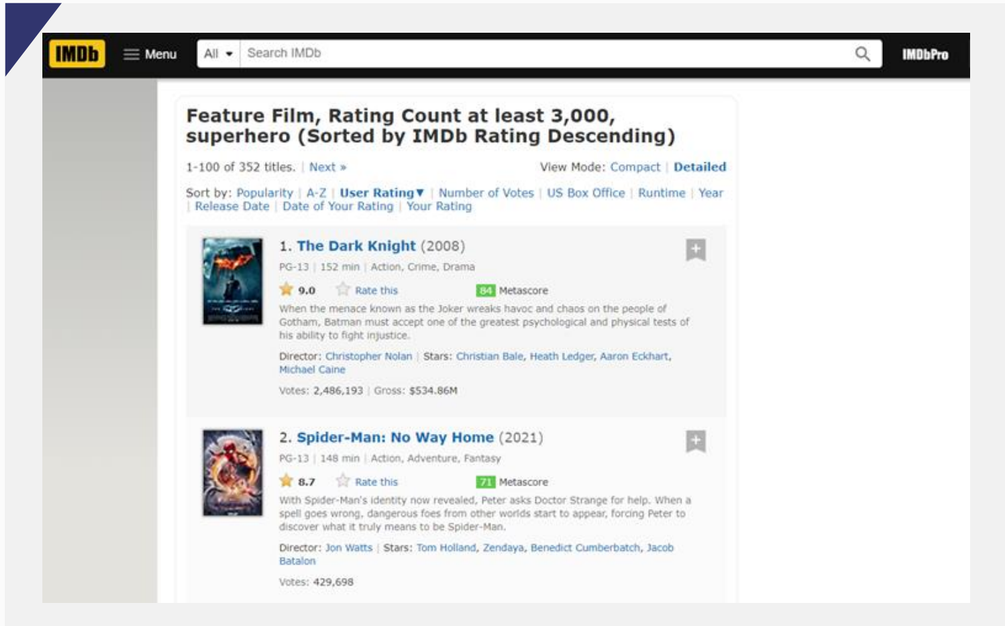
- Input : the question “What is a similar movie like spiderman”
- Output : clean and organized text-based data

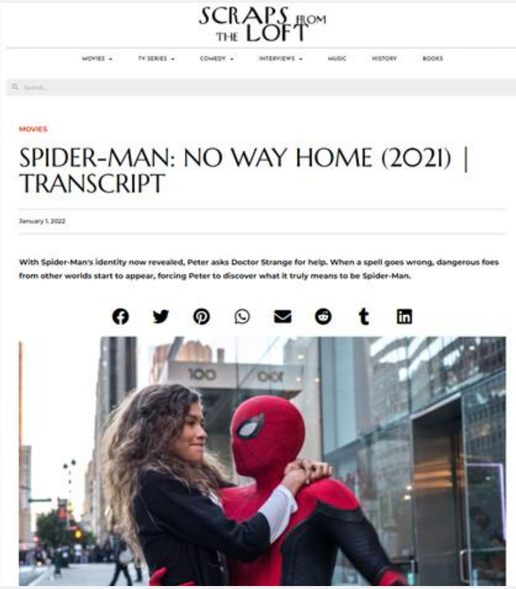
Movie	Organization	Transcript
Spiderman: No Way Home	Marvel	We come to you now with revelations about last week’s attack in London.....
Sang Chi: The Legend of The Ten Rings	Marvel	Shang-Chi is a young man who is in denial about his vocation and his magnificent warrior destiny.....
Joker	DC	Arthur is invited to appear on Murray’s show due to the popularity of his stand-up routine clips.....

What type of data should we get?

Where does the data come from?

How much data do we need?





1. Question | 2. Data | 3. Perform EDA | 4. Techniques | 5. Insight

Limiting our Data

Marvel

DC

1. Question | 2. Data | 3. Perform EDA | 4. Techniques | 5. Insight

Data Gathering

Web Scraping

- Request
  - Formal: Make HTTP requests
  - Simple: Get info from a Website

- Beautiful Soup
  - Formal: Parse HTML documents
  - Simple: Extract parts of a website

Saving Python Data

- Pickle
  - Formal: Serialize Python objects
  - Simple: Save data for later

Data Gathering

Format #1: Corpus

- Pandas: Python library for data analysis
- Dataframe: A pandas object, essentially a table

Corpus= Collection of texts

Movie	Transcript
Spiderman: No Way Home	We come to you now with revelations about last week's attack in London.....
Sang Chi: The Legend of The Ten Rings	Shang-Chi is a young man who is in denial about his vocation and his magnificent warrior destiny.....
Joker	Arthur is invited to appear on Murray's show due to the popularity of his stand-up routine clips.....

Data Cleaning

I am going to play at Gelora Bung Karno stadium on February 4th

Common data cleaning steps:

- Remove punctuation
- Lowercase all the letters
- Remove numbers

Data Cleaning

i am going to play at gelora bung karno stadium on february

How does a computer read these words? Tokenization must be done and be put into a matrix

Data Cleaning

i am going to play at gelora bung karno stadium on february

How does a computer read these words? Tokenization must be done and be put into a matrix

i	am	going	to	play	at
gelora	bung	karno	stadium	on	february

Data Cleaning

i	am	going	to	play	at
gelora	bung	karno	stadium	on	february

You can remove the unimportant words and it ends up with

Data Cleaning

i	am	going	to	play	at
gelora	bung	karno	stadium	on	february

You can remove the unimportant words and it ends up with

i	going	play	stadium	february

The data is stored in a matrix to be used flexibly

Data Cleaning

Movie					
Spiderman					
Sang Chi					

Data Cleaning

Movie	What	Where	Oscorp	Atom	Multiverse
Spiderman					
Sang Chi					

Data Cleaning

Movie	What	Where	Oscorp	Atom	Multiverse
Spiderman	12	19	5	4	8
Sang Chi	10	5	0	0	0

Data Cleaning

Movie	What	Where	Oscorp	Atom	Multiverse
Spiderman	12	19	5	4	8
Sang Chi	10	5	0	0	0

Document term matrix:

- Each row is a different document (or a different transcript)
- Each column is a different term (usually words)
- The values are word counts

Data Cleaning

Document term matrix

Movie	What	Where	Oscorp	Atom	Multiverse
Spiderman	12	9	5	4	8
Sang Chi	10	5	0	0	0

Corpus

Movie	Transcript
Spiderman: No Way Home	We come to you now with revelations about last week's attack in London.....
Sang Chi: The Legend of The Ten Rings	Shang-Chi is a young man who is in denial about his vocation and his magnificent warrior destiny.....

Summary

Input

What are similar movies like "Spiderman: No Way Home?"

Data Cleaning

Put data into a structured format for analysis

Data Gathering

Determining the scope and limits of the project and gathering data from the internet

Output

A corpus and document-term matrix

Google Colab Code