



## CAPSTONE PROJECT Senior Data Scientist

# TEXT ANALYTICS

Book Fair with NLP

## My Project Background (Business Problem)

Problem		How to Solve
<b>APA</b>	<b>Event</b> pameran buku yang segera diadakan seorang pemilik toko buku perlu disukseskan	Untuk membantu mendukung acaranya, maka <b>buku jualannya akan dianalisis</b> .
<b>BERAPA</b>	<b>Skala</b> jumlah buku yang terlibat sangat banyak untuk dipersiapkan secara manual	Agar dapat menangani skala buku yang banyak, digunakan <b>machine learning untuk membantu mengatur bukunya</b> .
<b>BAGAIMANA</b>	<b>Topik</b> buku-buku yang ada perlu diatur dengan cara yang menarik calon pengunjung/pembeli	Agar layout pameran menarik calon pengunjung/pembeli, pengaturan buku <b>dibantu dengan NLP, terutama berdasarkan book summary/alur bukunya</b> .

## Method & Workflow Project

### Methods Used

Menggunakan metode machine learning, yaitu **clustering**, untuk membantu mengorganisasikan buku terutama **berdasarkan book summary/alur bukunya**, dengan **input dari analisis text analytics**.

### Workflow

#### Secara garis besar:

**Input:** Book Summary/Metadata Buku Lainnya

⇒ Analisis Dengan Proses Text Analytics

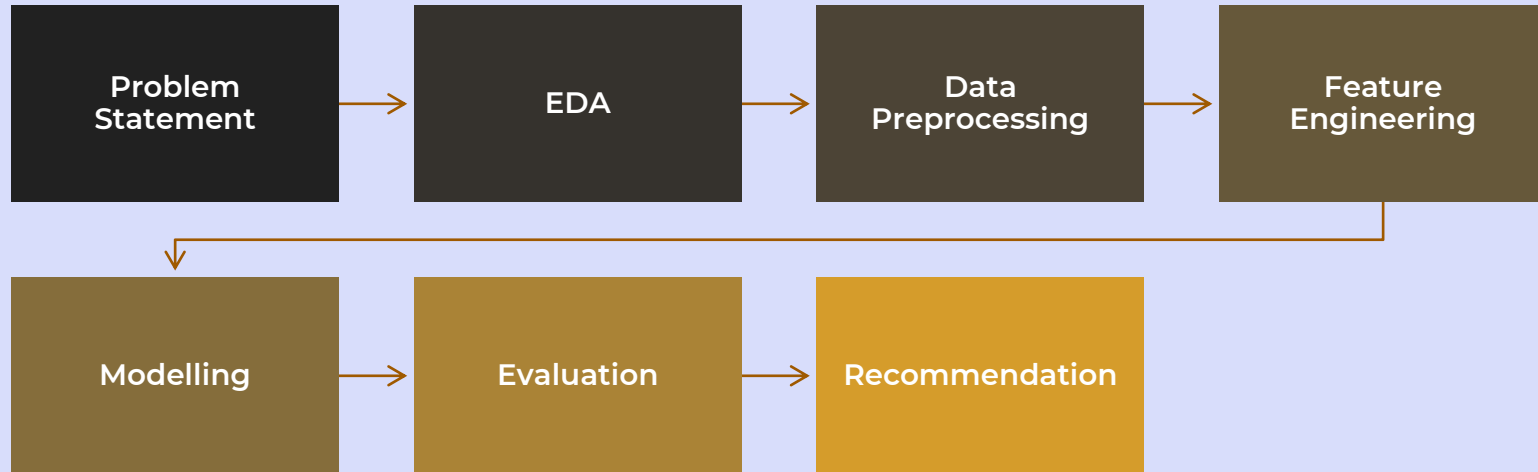
⇒ Clustering

**Output:** Insight untuk Mengorganisasikan Buku

## Demo Time!



## Flow Chart Pembuatan Project



## EDA Tentang Buku Pemilik Toko

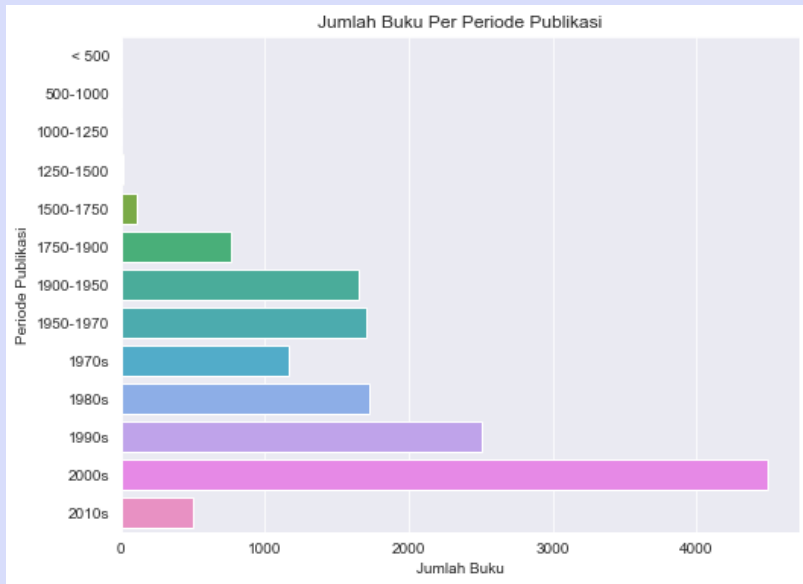
Pemilik toko buku kita memiliki berbagai macam buku!

- **~16,600** judul buku unik
- Setidaknya tersebar dalam **~230 genre** buku
- Setidaknya dari **~5,600 penulis!**\*

\* Setelah pengisian missing value



## EDA Tentang Buku Pemilik Toko



**Pemilik buku memiliki banyak stok buku antara 1970-2010\***

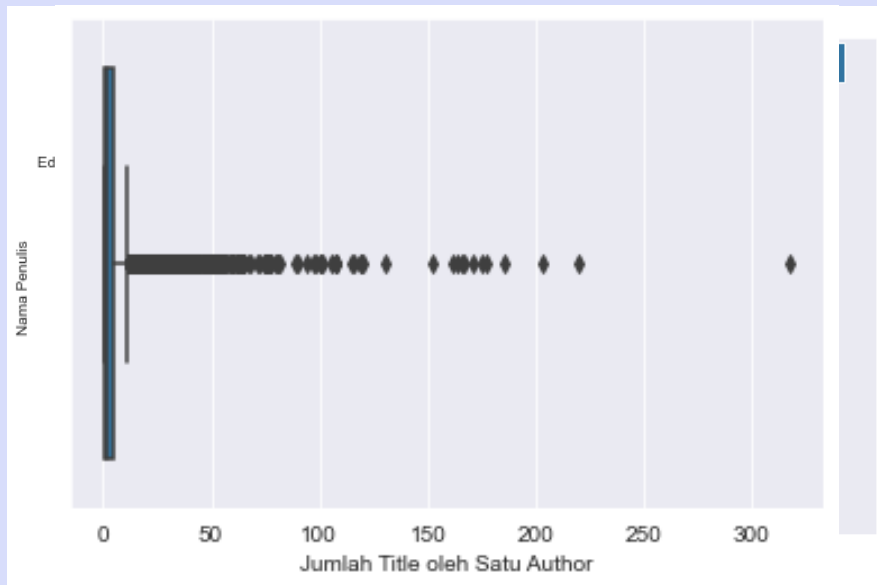
Semakin mendekati decade tahun 2000-an, semakin banyak judul buku unik yang dimiliki per dekade periode publikasi.

\* Setelah pengisian missing value

## EDA Tentang Buku Pemilik Toko

Beberapa penulis buku memiliki sangat banyak buku dalam stok toko!

Sebagian besar penulis hanya **memiliki 1-5 buku**, namun beberapa penulis memiliki hingga ratusan buku! Umumnya author **fiksi fantasi, misteri, thriller, dan sci-fi**.





## EDA Tentang Buku Pemilik Toko

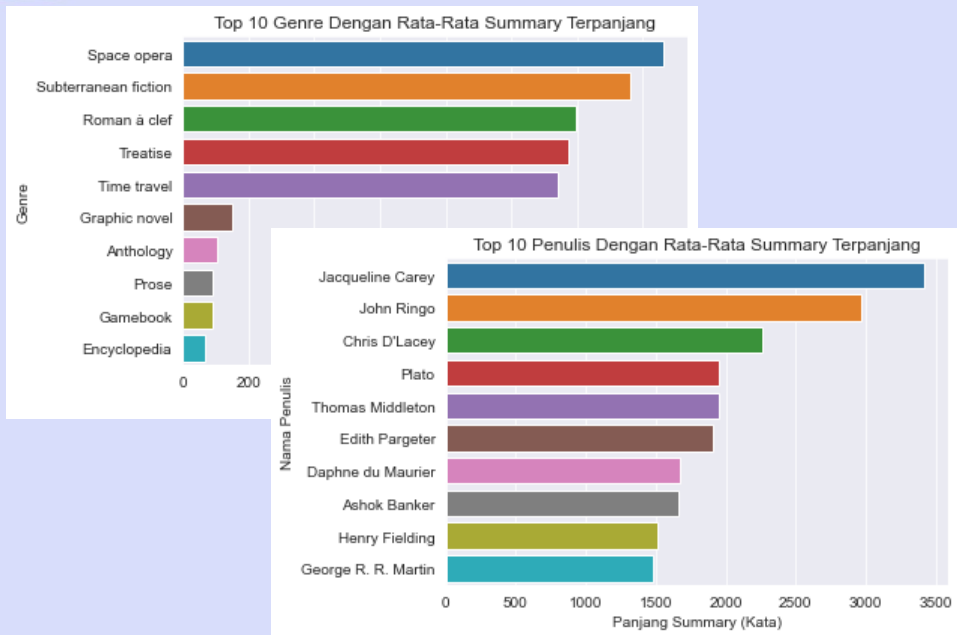
### Panjang summary buku cukup bervariasi!

- Rata-rata sepanjang **1-2 halaman\***
- Ada buku-buku dengan **summary belasan halaman!**
- Ada juga buku-buku dengan **summary < 8 kata\***

\* Asumsi 1 halaman 450-500 kata



## EDA Tentang Buku Pemilik Toko



**Ada pola dalam panjang book summary...**

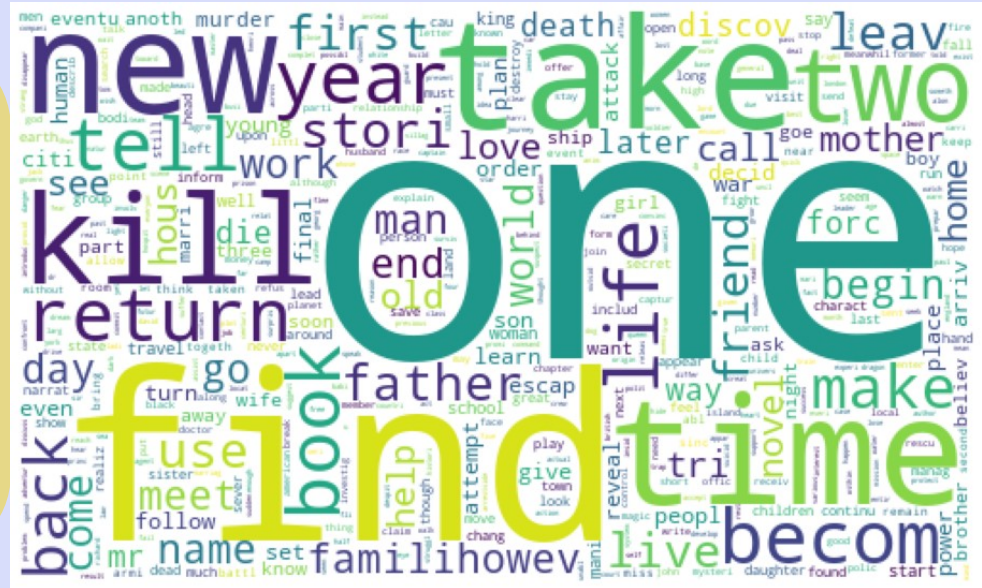
Penulis dan genre dengan book summary panjang umumnya memang berkaitan dengan buku yang plotnya rumit hingga membutuhkan summary panjang, dan sebaliknya juga untuk genre.

## EDA Tentang Buku Pemilik Toko

**Kata-kata yang paling umum cukup penuh adventure dan kekerasan**

Terdapat kata-kata **seputar keluarga/hubungan** atau **angka** pula, tetapi kata-kata paling umum adalah kata-kata *prompt* petualangan seperti **take, return, find, time, leave**.\*

\* See: Monomyth. Kata-kata telah distemming.



## EDA Tentang Buku Pemilik Toko

Kata-kata yang paling umum cukup penuh adventure dan kekerasan

Kata-kata **per genre** umumnya juga **didominasi kata-kata populer** ini, **dengan beberapa kata-kata spesifik** seperti *vampire*, *demon*, *night*, dan entah kenapa nama Anita.\*

\* Kata-kata telah distemming.

Kata-kata yang paling umum cukup penuh adventure dan kekerasan

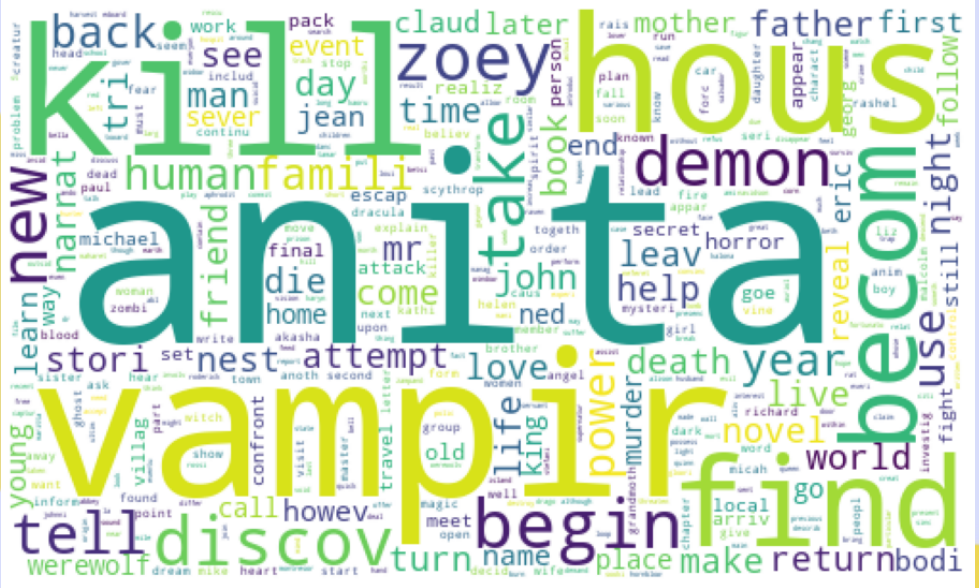
Kata-kata **per genre** umumnya juga **didominasi kata-kata populer** ini, **dengan beberapa kata-kata spesifik** seperti *vampire*, *demon*, *night*, dan entah kenapa nama Anita.\*

\* Kata-kata telah distemming.

Kata-kata yang paling umum cukup penuh adventure dan kekerasan

Kata-kata **per genre** umumnya juga **didominasi kata-kata populer** ini, **dengan beberapa kata-kata spesifik** seperti *vampire*, *demon*, *night*, dan entah kenapa nama Anita.\*

\* Kata-kata telah distemming.



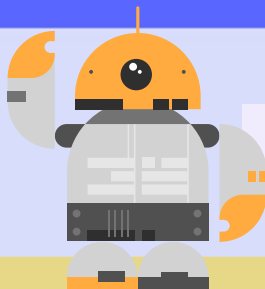
# Preprocessing

## Preprocessing General

- Koreksi **tipe data**
- Penghapusan **baris-baris yang duplikat** karena duplikasi genre
- Pengecekan **missing value**
- Pengisian beberapa **missing value dari Wikidata** (author, publication date)
- **Koreksi ID genre misteri**

## Preprocessing NLP

- **Lowercasing**
- Pembersihan **pattern/symbol noise**
- **Tokenisasi**
- Pembersihan **stopwords** (dengan NLTK stopwords serta kata-kata angka)
- **Lemmatisasi** dan **stemming** (dengan WordNet dan Snowball Stemmer)




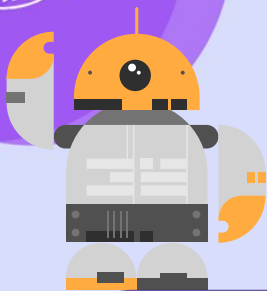
# Feature Engineering

## Doc2Vec

- Dipilih karena dapat menghasilkan vektor langsung per dokumen dengan jumlah dimensi ekonomis
- Menggunakan Gensim
- Dirasa lebih baik dibandingkan merata-ratakan Word2Vec untuk mendapatkan vektor
- Alternatif: TF-IDF, namun tidak dipakai karena sudah dicoba dan performa kurang optimal saat clustering

## Parameter

- 
- `vector_size = 250`
  - `min_df = 3`
  - `epochs = 30`



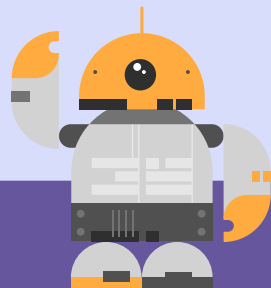
## Parameter

- Daftar entity:  
['NORP', 'ORG', 'GPE',  
'LOC', 'EVENT']
- Minimum dokumen dengan  
entity = 10

# Feature Engineering

## NER Event/Lokasi/Organisasi

- NER diaplikasikan pada summary buku karena *setting* buku bisa sangat berdampak pada alur buku!
- NER dilakukan dengan spaCy karena pertimbangan waktu, lalu nama orang yang salah deteksi dikoreksi dengan Stanford NER (Stanza)
- Hasil NER kemudian diproses dengan CountVectorizer() dan dipilih yang umum saja



## Feature Engineering

### Fitur-Fitur Lainnya

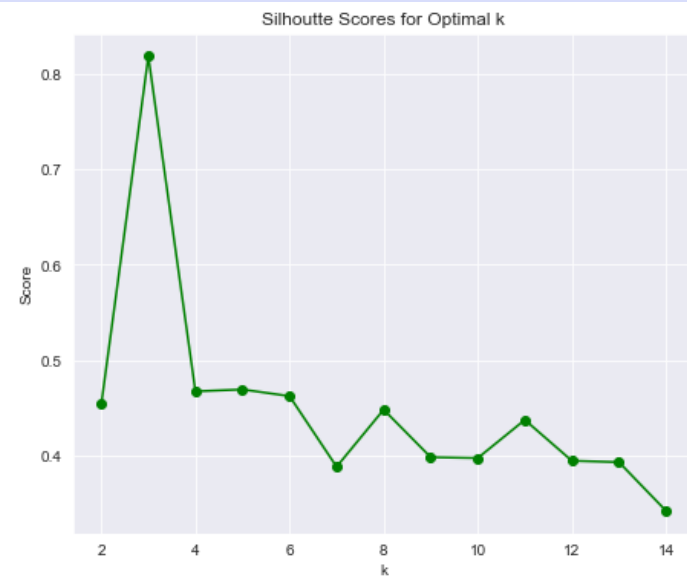
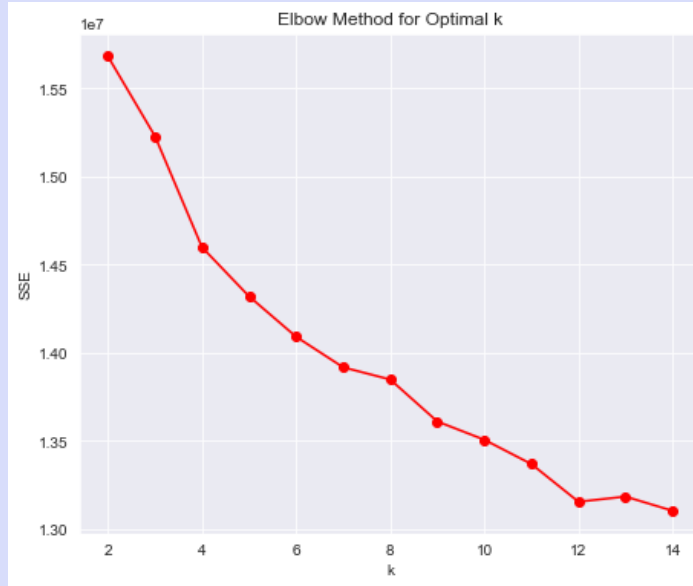
- **Periode publikasi** dibuat dari tahun publikasi yang telah diimputasi dan dilakukan binning yang sesuai, dan periode publikasi diyakini dapat mempengaruhi gaya atau alur populer zaman tersebut
- **Panjang summary** dibuat dari jumlah kata dalam summary tiap buku, yang saat EDA dapat membedakan tipe buku/genre dan penulis

### Yang Tidak Masuk

- **Author** (sulit diimputasi)
- **Genre** (belum diimputasi dan missing value signifikan)



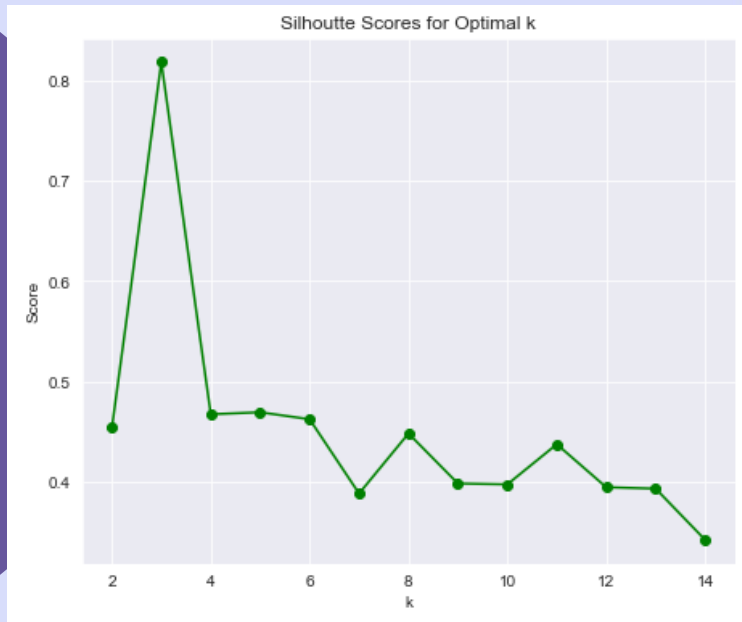
# Modelling (clustering)



# Modelling (clustering)

## Dipilih K-Means dengan $k = 6$

- Secara silhouette score, kandidatnya adalah  $k = 4$  hingga 6, sementara secara elbow method agak ambigu
- Mempertimbangkan jumlah kluster yang diperlukan untuk pameran dan hasil dari beberapa kali random state, maka angka 6 dipilih



## Result of Project and Recommendations

### To Be Continued...

- PPT dipublikasi hingga *step modelling* selesai
- Bagian selanjutnya *under renovation* karena ada yang mau ditingkatkan
- Ditunggu pada *update* selanjutnya!

## Learning Takeaways

What learning do I get from working on this project?

**Mencoba banyak metode baru** di luar kelas pada bidang yang belum terlalu saya kenal sebelumnya (**NLP**)

What other learning do I feel throughout my journey in #SDSNarasioData?

Belajar untuk **siap dengan semua kemungkinan** dan belajar **membaca kode** yang **kompleks**

A vibrant, stylized illustration of a collaborative work environment. Three people are seated around a table, working on laptops. A fourth person stands, holding a yellow sticky note, and pointing towards a large screen. The screen displays a lightbulb icon, a bar chart, and various documents. The background is a deep blue with abstract shapes, including a large yellow circle on the left, a purple circle at the top, and a gear icon. A small potted plant sits on the floor near the table.

**THANK YOU!**