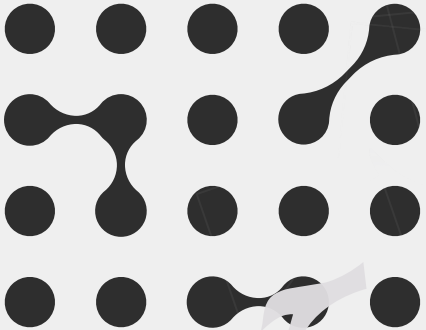
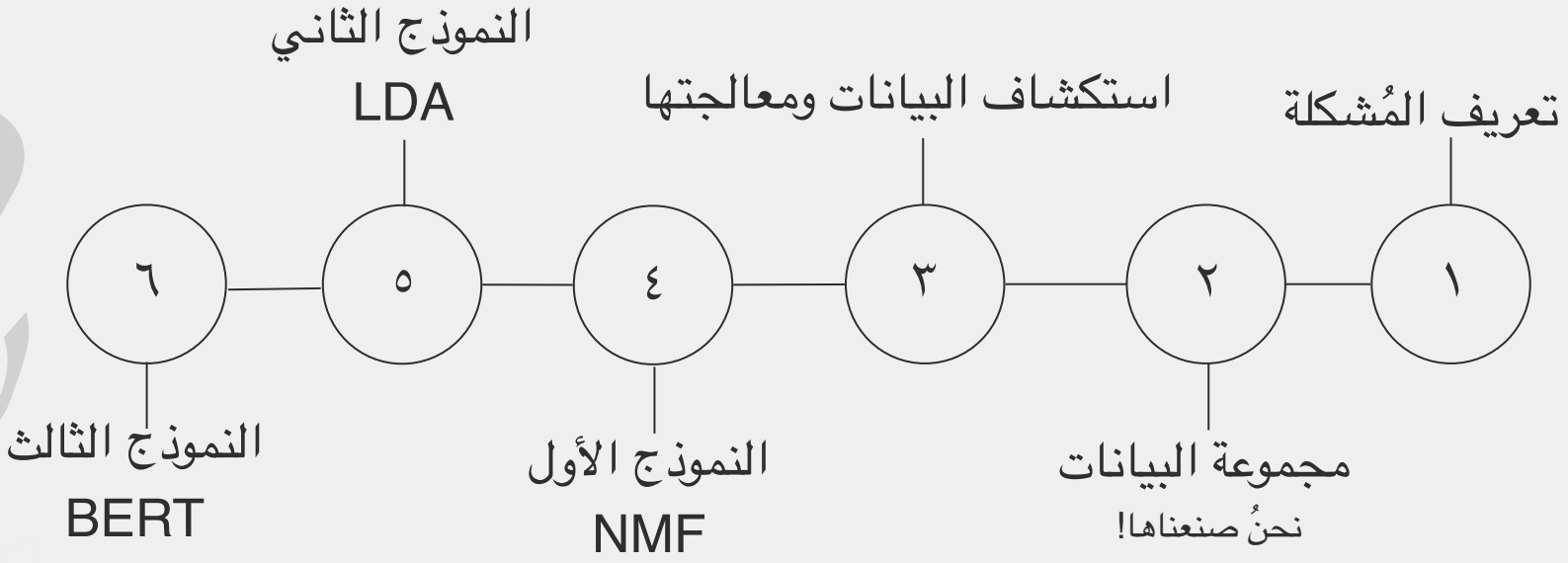
SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

مُعالجة اللغة الطبيعية

- روبوت الدردشة لموقع الإدارة العامة للمرور
  جميلة الحربي
  غالية ماهر
  غدير علي
  محمد العجمي
  محمد المالكي
  يحيى اليوبي
- بإشراف الدكتور: مجدل القحطاني



سير عمل المشروع:



تعريفُ المُشكلةِ:

- رؤية ٢٠٣٠ لمملكة ورياض يزهو بالاقتصاد.
- الوقوف الخاطئ من أكثر أسباب الازدحام شيوعا وخصوصا في وقت الذروة.

مجموعة البيانات:

- توزيع الاستبيانات.

- سحب البيانات من المقالات و أخرى قمنا بكتابتها وجمعها في ملف Excel

	text
0	هناك شخص واقف خلف سيارتي
1	هناك شخص متوقف خلف سيارتي
2	في احد موقف ورا سيارتي
3	فيه واحد موقف ورا سيارتي
4	فيه واحد واقف وراي
...	...
607	ما هو مبلغ مخالفاتي
608	وجدت مخالفات جديدة وأودّ الاستفسار عن وقت حدوثها
609	هل علي مخالفات
610	ماهي فواتيري
611	هل لي ان اعرف نوع المخالفة

612 rows × 1 columns

استكشاف البيانات ومعالجتها:

- حذف جميع الأرقام وعلامات الترقيم.
- انشاء قاموس كلمات الإيقاف "حروف الجر والضمائر" ثم إزالتها جميعًا من مجموعة البيانات.
- استبدل كلمة "وراي" بكلمة "خلفي".
- تحويل مايلي :

كل "أ ، إ ، آ" إلى "ا"
كل "ى" إلى "ي"

استكشاف البيانات ومعالجتها:

- اقتصاص الأحرف الزائدة من الكلمة Stemming > farasapy
- جزء الكلمة في الجملة Part of the speech <

CAMeL-Lab/bert-base-arabic-camelbert-mix-pos-glf

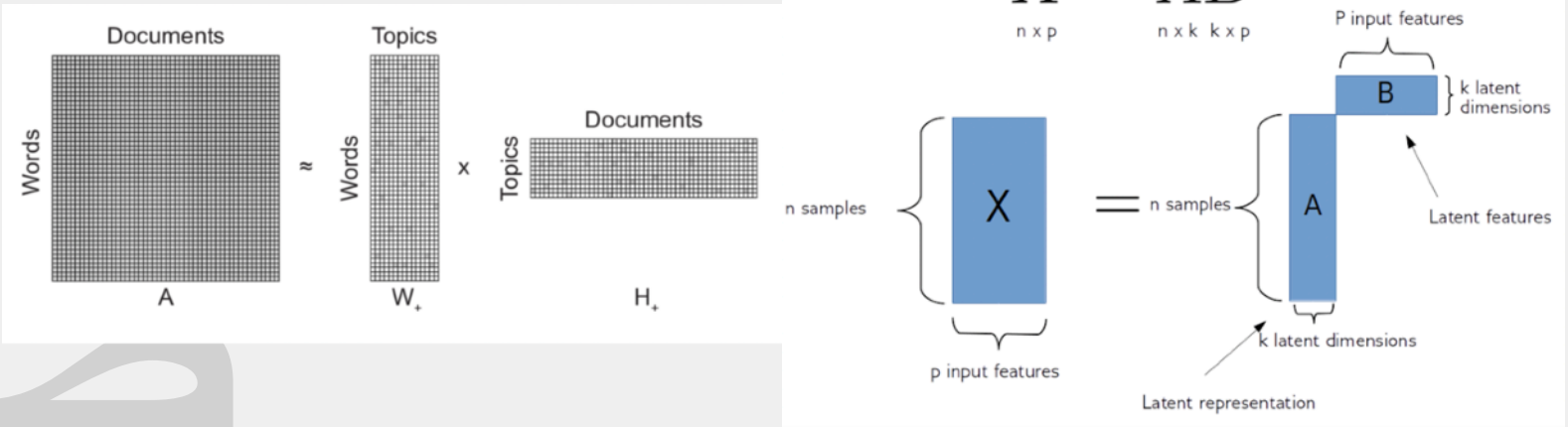
لقيت سيارتي مصدومه > لقيت فعل

٤ سيناريوهات:

- شخص قفل على سيارتي!
- لقيت طفل لوحده بالسيارة
- ابغى استفسر عن مخالفاتي
- صُدمت سيارتي

النموذج الأول NMF

(Non-negative Matrix Factorization)



النموذج الأول NMF

(Non-negative Matrix Factorization)

```
[44] for i in range(0,len(topics_NMF)):
       print(topics_NMF[i])
```

```
['اطفال', 'خط', 'مركب', 'اولاد', 'أحد', 'سيارة', 'وجد', 'شخص', 'طريق', 'مقفل']
['جديد', 'استفسار', 'عدد', 'تفصيل', 'مجموع', 'أظهر', 'مرصوده', 'كم', 'مبلغ', 'مخالفة']
['اهل', 'ناسيين', 'نايم', 'نائم', 'وجد', 'مغلق', 'محجوز', 'وحد', 'سيارة', 'طفل']
['غلط', 'خلف', 'واحد', 'سيارة', 'هرب', 'هي', 'واقف', 'صدم', 'حك', 'أحد']
['وجد', 'نائم', 'محجوزين', 'داخل', 'مغلق', 'وحد', 'مركب', 'سيارة', 'محتجز', 'أطفال']
['غلط', 'شكل', 'خاطئ', 'زجاج', 'موقف', 'كسر', 'متوقف', 'سيارة', 'خلف', 'شخص']
['خدش', 'طافي', 'صدمة', 'عدد', 'كم', 'عند', 'ماسبب', 'أعرف', 'نوع', 'مخالف']
```

```
[92] cm = CoherenceModel(topics=topics_NMF, texts=texts, corpus=corpus, dictionary=id2word, coherence='c_v')
     coherence_nmf = cm.get_coherence()
     print('\nCoherence Score: ', coherence_nmf)
```

```
Coherence Score:  0.3768295547960176
```

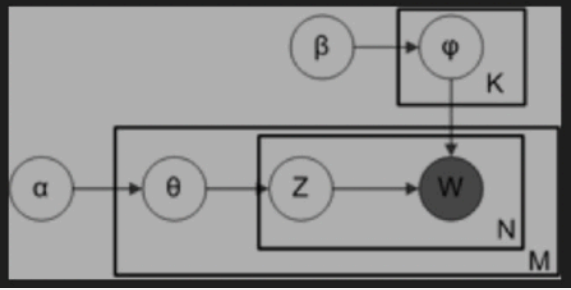
النموذج الثاني LDA
(Latent Dirichlet Allocation)

θ_i is the topic distribution for document *i*

φ_k is the word distribution for topic *k*

z_{ij} is the topic for the *j*-th word in document *i*

w_{ij} is the specific word.



النموذج الثاني LDA

(Latent Dirichlet Allocation)

```
[38] lda.print_topics()
```

```
[(0,
  '0.205*"شخص"*0.029 + "غلط"*0.032 + "واحد"*0.037 + "هرب"*0.037 + "مقفل"*0.040 + "مخالفة"*0.049 + "واقف"*0.057 + "صدم"*0.066 + "أحد"*0.074 + "سيارة"'),
 (1,
  '0.136*"أحد"*0.041 + "شخص"*0.046 + "خروج"*0.051 + "لا"*0.051 + "أستطيع"*0.051 + "متوقف"*0.061 + "مخالفة"*0.071 + "موقف"*0.071 + "خلف"*0.086 + "سيارة"'),
 (2,
  '0.242*"نايم"*0.023 + "متوقف"*0.026 + "حك"*0.028 + "ناس"*0.034 + "هرب"*0.034 + "واحد"*0.034 + "شخص"*0.051 + "مقفل"*0.057 + "طفل"*0.082 + "سيارة"'),
 (3,
  '0.164*"مقفل"*0.020 + "فاتورة"*0.046 + "مبلغ"*0.046 + "خاطئ"*0.047 + "متوقف"*0.047 + "تكل"*0.047 + "شخص"*0.047 + "سيارة"*0.052 + "مخالف"*0.101 + "مخالفة"'),
 (4,
  '0.225*"هي"*0.022 + "واقف"*0.022 + "وجد"*0.028 + "طريق"*0.033 + "طفل"*0.051 + "أطفال"*0.053 + "وجد"*0.057 + "شخص"*0.067 + "مقفل"*0.100 + "سيارة"'),
 (5,
  '0.156*"كم"*0.028 + "مغلق"*0.031 + "شخص"*0.031 + "مقفل"*0.031 + "وجد"*0.034 + "مخالفة"*0.040 + "مخالف"*0.046 + "وجد"*0.052 + "أطفال"*0.077 + "سيارة"'),
 (6,
  '0.171*"محتجز"*0.026 + "مقفل"*0.028 + "زجاج"*0.034 + "حك"*0.034 + "أطفال"*0.042 + "أحد"*0.044 + "وجد"*0.044 + "شخص"*0.044 + "طريق"*0.050 + "سيارة"')]
```

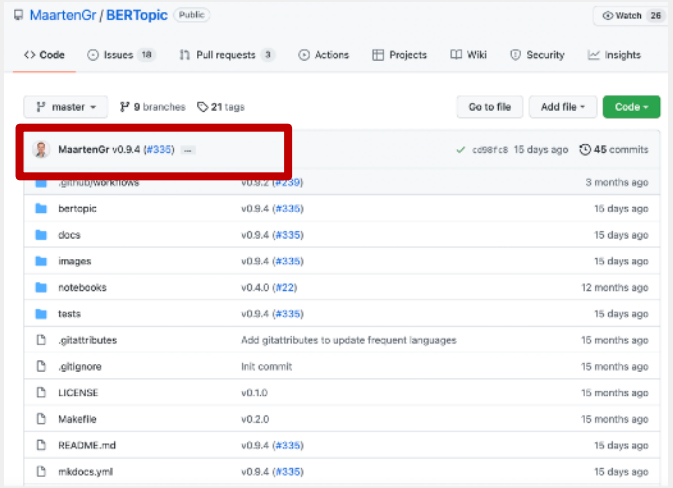
```
coherence_model_lda = CoherenceModel(model=lda, texts=texts, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)
```

```
Coherence Score:  0.584442132205764
```

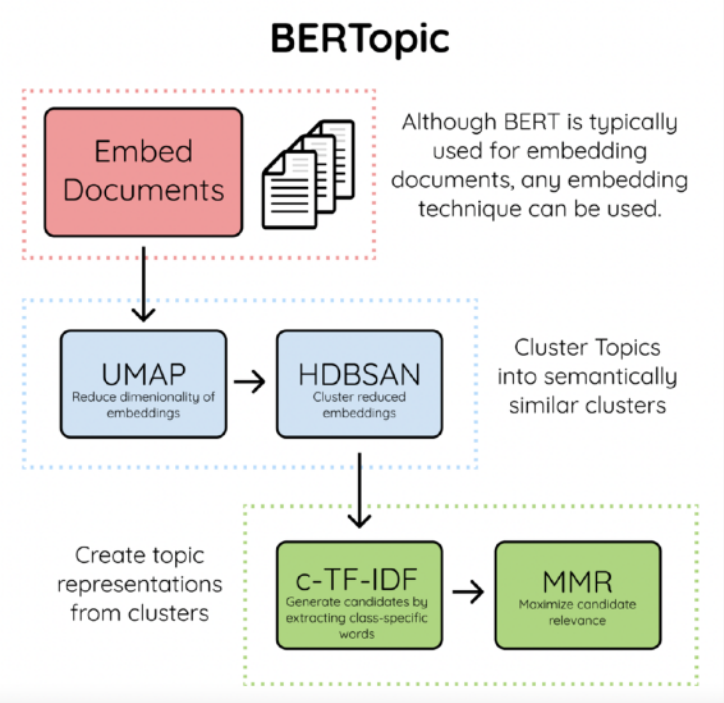
النموذج الثالث BERTopic

- النموذج مُتدرب على بليون ونصف كلمة وعدة مقالات

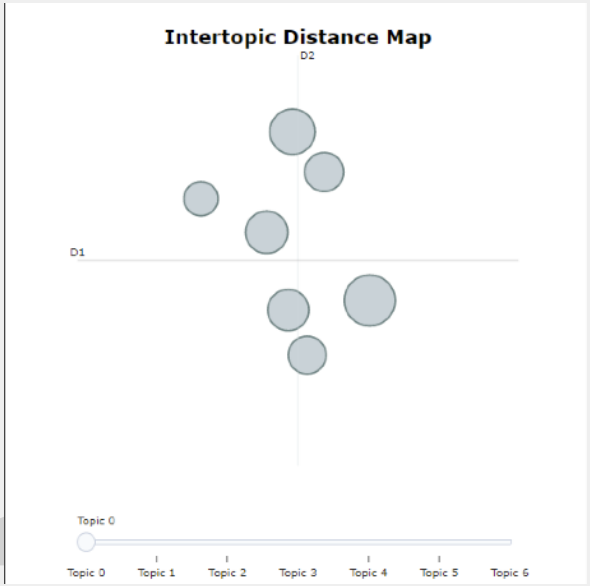
Model	HuggingFace Model Name	Size (MB/Params)	Pre-Segmentation	DataSet (Sentences/Size/nWords)
AraBERTv0.2-base	bert-base-arabertv02	543MB / 136M	No	200M / 77GB / 8.6B
AraBERTv0.2-large	bert-large-arabertv02	1.38G 371M	No	200M / 77GB / 8.6B
AraBERTv2-base	bert-base-arabertv2	543MB 136M	Yes	200M / 77GB / 8.6B
AraBERTv2-large	bert-large-arabertv2	1.38G 371M	Yes	200M / 77GB / 8.6B
AraBERTv0.1-base	bert-base-arabertv01	543MB 136M	No	77M / 23GB / 2.7B
AraBERTv1-base	bert-base-arabert	543MB 136M	Yes	77M / 23GB / 2.7B

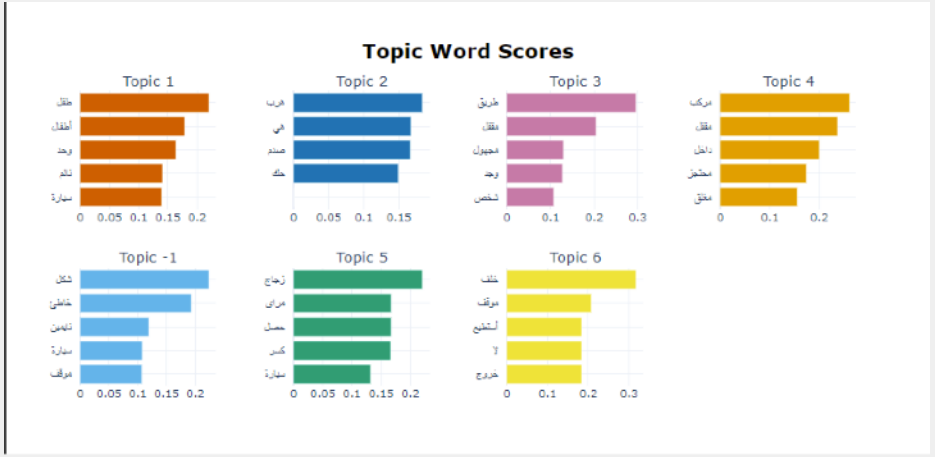


النموذج الثالث BERTopic



النموذج الثالث BERTopic





النموذج الثالث BERTopic

```
[ ]  my_model.topic_names

     {0: 'مخالفات',
      1: 'طفل محتجز',
      2: 'اصتدام',
      3: 'طريق مقفل',
      4: 'طفل محتجز',
      5: 'اصتدام',
      6: 'طريق مقفل'}
```

```
[ ]  cosine_similarity((scores_topic3, scores_topic6)).round()

     array([[1., 1.],
            [1., 1.]])
```

النموذج الثالث BERTopic

```
[17] # Load model

     my_model = BERTopic.load("/content/drive/MyDrive/NLP/my_model4")
```

```
Downloading: 100% 381/381 [00:00<00:00, 5.96kB/s]
Downloading: 100% 805k/805k [00:01<00:00, 1.08MB/s]
Downloading: 100% 2.52M/2.52M [00:01<00:00, 2.31MB/s]
Downloading: 100% 112/112 [00:00<00:00, 975B/s]
Downloading: 100% 384/384 [00:00<00:00, 9.87kB/s]
```

```
coherence_model_lda = CoherenceModel(model=my_model, texts=texts, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)
```

```
Coherence Score:  1.0
```

النتائج:

النموذج	مقياس الاتساق
Non-negative matrix frequancy (NMF)	0.376
Latent dirichlet allocation (LDA)	0.584
BERTopic model	1.000

العمل المستقبلي :

العمل المستقبلي لهذا المشروع سيتضح بعد أسبوعين في المشروع النهائي ان شاء الله...

نلقاكم على خير☺

شكرا لحسن استماعكم

بكِ تاجُ فخري وانطلاقُ لساني
و مرورُ أيامي ودفـءُ مكاني
لغة الجدودِ ودربُنا نحوَ العُلا
و تناغمُ الياقوتِ والمَرجانِ