Natural language Processing - Chatbot

Abstract

With the increase in population density in the Kingdom of Saudi Arabia especially in Riyadh city which is considered the capital of Saudi and the largest Arab cities in term of area also the fastest expanding cities in the world.

With all these reasons, the crowding increases, especially in the morning and evening peak times.
One of the problems that facing the city is the improper parking which consider the most common causes of traffic congestion and disruption not only in Riyadh but in the whole Kingdom. Therefore, this behavior is considered a traffic violation in the Saudi Traffic Law, It is called an “Illegal parking violation”.

Our goal in this project is to serve The General Saudi Department of Traffic from a side and the civil from the other side by building a model in chatbot that can answer all the questions, receive the inquiries and objections that are being asked by the civilian.

For more explanation, this project is based primarily on Natural Language Processing NLP.
The concept is about analyzing the text that sent by the end-user (Input) then the model will find the best answer to replay within an active human-based conversation (Output).

Data Collection

The dataset that we used comes from:

1. Surveys we spread to collect as much as we can of sentences and words that related to our project.
2. Web scraping of articles that covers the 3 basic scenarios that meets our project expectations:
   - Inquiry about a violation.
   - Reporting a violation such as: having a lonely child in the car, My car has been scrached.
   - Wrong parking caused line traffic.

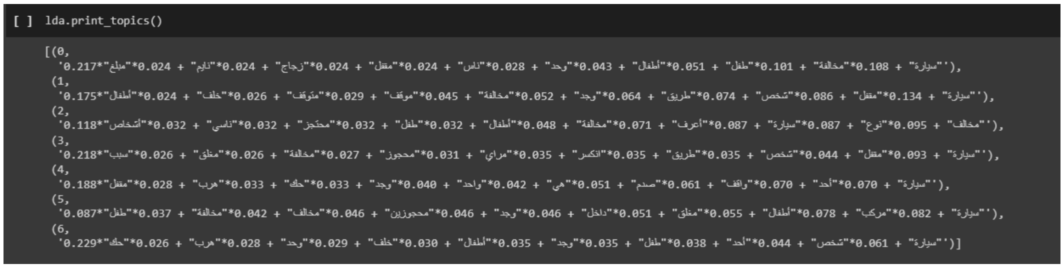
Pre-Processing

- Remove number and punctuation.
- Create stop word dictionary “prepositions and pronouns” and then remove them all from the dataset.
- Replace the word “وراي” by the word “خلفي”.
- Normalize the following from Camel tool:
  - All "أ، إ ، آ" with “ ا ”
  - All “ ه ” to “ ة ”
  - All “ ى ” to “ ي ”
- Stemming from Farasah library

- Part of speech (P.O.S) using Camel tool `model CAMeL-Lab/bert-base-arabic-camelbert-mix-pos-glf`
  لقيت سيارتي مصدومه the tool could figure out that the word لقيت is verb

Models

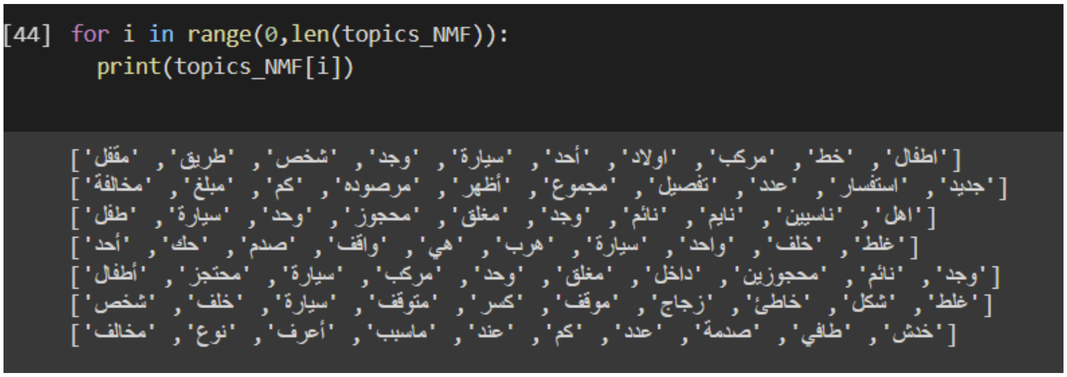
First Model:
LDA (Latent Dirichlet Allocation)
LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.
We apply the model and then print all the results. however, it wasn't that good cause there was a coherence and this method isn't applicable on Arabic.



```
[ ]  lda.print_topics()

    [(0,
      '0.217*"سيارة" + 0.108*"مخالفة" + 0.101*"طفل" + 0.051*"أطفال" + 0.043*"وحد" + 0.028*"ناس" + 0.024*"مقفل" + 0.024*"زجاج" + 0.024*"نايم" + 0.024*"مبلغ"'),
     (1,
      '0.175*"سيارة" + 0.134*"مقفل" + 0.086*"شخص" + 0.074*"طريق" + 0.064*"وجد" + 0.052*"مخالفة" + 0.045*"موقف" + 0.029*"متوقف" + 0.026*"خلف" + 0.024*"أطفال"'),
     (2,
      '0.118*"مخالف" + 0.095*"نوع" + 0.087*"سيارة" + 0.087*"أعرف" + 0.071*"مخالفة" + 0.048*"أطفال" + 0.032*"طفل" + 0.032*"محتجز" + 0.032*"ناسي" + 0.032*"أشخاص"'),
     (3,
      '0.218*"سيارة" + 0.093*"مقفل" + 0.044*"شخص" + 0.035*"طريق" + 0.035*"انكسر" + 0.035*"مراي" + 0.031*"محجوز" + 0.027*"مخالفة" + 0.026*"مغلق" + 0.026*"سبب"'),
     (4,
      '0.188*"سيارة" + 0.070*"أحد" + 0.070*"واقف" + 0.061*"صدم" + 0.051*"هي" + 0.042*"واحد" + 0.040*"وجد" + 0.033*"حك" + 0.033*"هرب" + 0.028*"مقفل"'),
     (5,
      '0.087*"سيارة" + 0.082*"مركب" + 0.078*"أطفال" + 0.055*"مغلق" + 0.051*"داخل" + 0.046*"وجد" + 0.046*"محجوزين" + 0.046*"مخالف" + 0.042*"مخالفة" + 0.037*"طفل"'),
     (6,
      '0.229*"سيارة" + 0.061*"شخص" + 0.044*"أحد" + 0.038*"طفل" + 0.035*"وجد" + 0.035*"أطفال" + 0.030*"خلف" + 0.029*"وحد" + 0.028*"هرب" + 0.026*"حك"')]
```

Second Model:
NMF (Non-negative Matrix Factorization)
First apply TF_IDF factorize, fits it to NMF, then print the result.



```
[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']
```

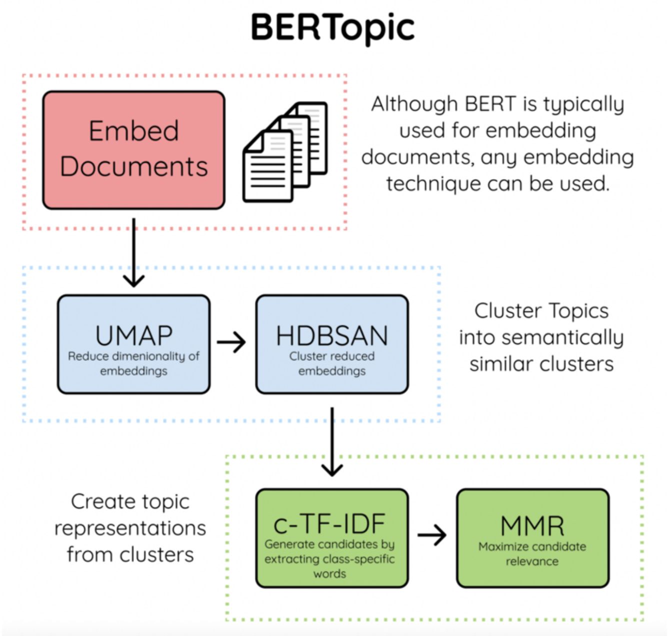
Third Model (The chosen model):
BERT_for_Arabic_Topic_Modeling_ACLing2021

First, we have to know how the BERTopic modeling technique works? it uses transformers (BERT embeddings) and class-based TF-IDF (Term Frequency — Inverse Document Frequency) to create dense clusters. It also allows you to easily interpret and visualize the topics generated.
BerTopic algorithm contains 3 stages:

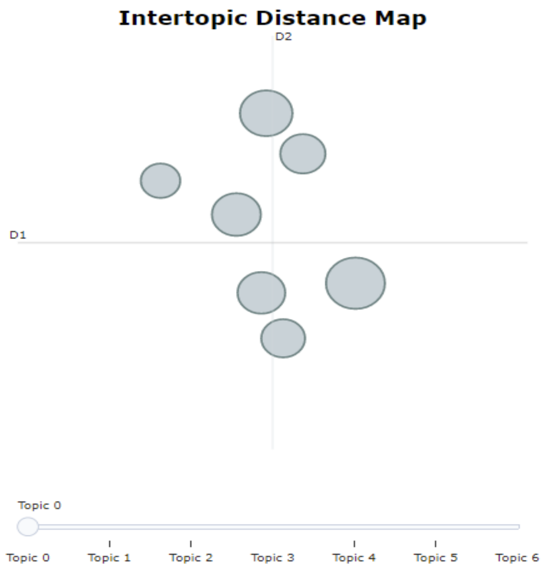
1. **Embed the textual data (documents)**
   In this step, the algorithm extracts document embeddings with BERT, or it can use any other embedding technique.

2. **Cluster Documents**
   It uses UMAP to reduce the dimensionality of embeddings and the HDBSCAN technique to cluster reduced embeddings and create clusters of semantically similar documents.
3. **Create a topic representation**
   The last step is to extract and reduce topics with class-based TF-IDF

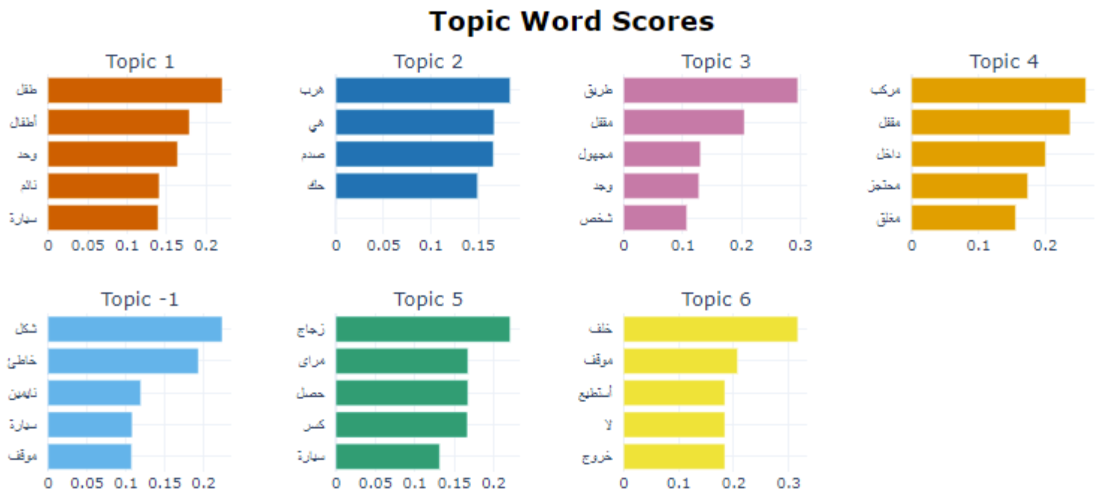
The figure below represents how the model will be processed:


BERTopic
Embed Documents
Although BERT is typically used for embedding documents, any embedding technique can be used.
UMAP
Reduce dimenionality of embeddings
HDBSAN
Cluster reduced embeddings
Cluster Topics into semantically similar clusters
Create topic representations from clusters
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

Overthrow, In our project the BERTopic Pre-Trained Arabic Language Model is embedded, and compare its results against LDA (Latent Dirichlet Allocation) and NMF (Non-negative Matrix Factorization) techniques.
We used Normalized Pointwise Mutual Information (NPMI) measure to evaluate the results of topic modeling techniques. The overall results generated by BERTopic showed better results compared to NMF and LDA.
The big plus for this model that it supports the different Arabic dialects especially the Gulf countries dialects, for example the word "لقيت" is conver to "وجد" and so on on many local words.



This figure above represents the intertopic distance map, in this project the group agreed on choosing seven as a fixed number for the number of clusters (topics).



- The figure above represents the BarH plot, the plot shows the frequency of word in each topic,the plot shows seven diffrent topics, every subplot shows the most words occured within the topic.
- The (-1 topic) represent topic has no relation with the other topics (outlier), this topic occured due to that data collected from users by google form, which some of them did not filled the form correctly.