Author : Amir

Written at : 2/10/19

Preprocess / Enrich data.

I have written three spark job scripts to enrich the tables : nlu, nbr, sdk.

You can locate the scripts at <https://github.com/amirmmuze/work/tree/master/enrich_data>

Currently their input is a tsv file and they write into parquet format which could be altered easily.

The purpose of this enrichment is to serve as a base for analysis and modeling.

The next steps could be:

1. Create a report that visualize different aspects of the data.
2. Write an aggregation script that joins all three tables into one table that will have all of the info regarding a conversation.
3. Research and deploy different modelling and architecture to improve conversations.

Fields currently added by the scripts:

Nbr:

|  |  |  |
| --- | --- | --- |
| Name | Type | Origin \ meaning |
| ﻿date | datetime | ﻿date and time from unix |
| ﻿ack\_text | str | ﻿bot text response. fetched from 'json\_response' |
| response\_code | str | ﻿bot response code fetched from 'json\_response'. |
| possible\_values | ﻿struct type | ﻿quick replies |
| ts\_plus\_response | Int | Ts + response time |
| time\_diff | int | ﻿int difference between this event compared to previous one in seconds |
| ﻿is\_start\_session | Int (binary) | ﻿after 6 hours from previous engagement a new session starts. |
| ﻿conv | Int | ﻿incremental conv id per sender\_id by is\_session\_starts |

Sdk:

|  |  |  |
| --- | --- | --- |
| Name | Type | Origin \ meaning |
| date | datetime | ﻿date and time from the unix timestamp. |
| browser | str | ﻿browser fetched form user agent. |
| ﻿os | str | os fetched form user agent. |
| device | str | device fetched form user agent. |
| Time\_diff | Int | ﻿difference between this event compared to previous one in seconds. |
| is\_start\_session | Int (binary) | ﻿after 6 hours from previous engagement a new session starts. |
| conv | int | ﻿incremental conv id per my\_id by is\_session\_starts |

Nlu:

|  |  |  |
| --- | --- | --- |
| Name | Type | Origin \ meaning |
| date | datetime | ﻿date and time from the unix timestamp. |
| Year | Int | Year from date |
| Month | Int | Month from date |
| Day\_of\_week | Int | Dow from date |
| Year\_motth | Str | Year + “\_” + month |
| ﻿age\_group | Str | fetched from json\_response. |
| ﻿gender | str | fetched from json\_response. |
| Latest | Boolean | fetched from json\_response. |
| Intents\_list | List of str | fetched from json\_response. |
| Subvertical | Str | fetched from json\_response. |
| Positive\_aspects | List of str | fetched from json\_response. |
| ﻿positive\_brands | List of str | fetched from json\_response. |
| ﻿positive\_product\_type | Str | fetched from json\_response. |
| ﻿positive\_models | str | ﻿fetched from json\_response. |
| ﻿negative\_aspects | List of str | ﻿fetched from json\_response. |
| negative\_brands | ﻿Struct type | ﻿containing few fields , fetched from json\_response. |
| negative\_product\_type | ﻿list of str | ﻿fetched from json\_response. |
| ﻿negative\_models | list of str | ﻿fetched from json\_response. |
| ﻿occasion\_tags | list of str | ﻿fetched from json\_response. |
| occasion\_tags\_len | int | Len of occasions |
| ﻿product\_search\_result\_len | Int | #products shown |
| Is\_sku | Bolean | Is sku search |
| Is\_sku0 | List of str | The skus |
| My\_id | Str | Gi if exist, si otherwise |
| Time\_diff | Int | Difference in time from lag event |
| Is\_start\_session | Int binary | ﻿after 6 hours from previous engagement a new session starts. |
| Conv | Int | ﻿incremental conv id per my\_id by is\_session\_starts |