CTAB: Corpus of Tunisian Arabizi

Produced by *Data Engineering and Semantics Research Unit, University of Sfax, Tunisia*

Creators: Amina Amara, Houcemeddine Turki, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha and Kaouthar Ellouze

Abstract:

This dataset has been created between 2017 and 2021 to provide a textual resource that can be used to study the behaviors of Tunisian people in writing Tunisian Arabic (ISO 693-3: aeb) in Latin Script. This corpus is constituted from messages written using Tunisian Arabic Chat Alphabet or Arabizi and is developed to solve the matter of the lack of NLP databases about the use of the Latin Script for transcribing Tunisian Arabic. The messages are automatically pulled using web scraping of Facebook public pages and are kept as they are without any annotation, spelling adjustments or morphological and syntactic labeling. Then, messages that are written in Latin Script but not in Tunisian Arabic are manually eliminated. Finally, every collection of messages that are retrieved from the same Facebook page in the same period is included in the same text file where every message is featured as one line.

The corpus includes the following samples:

Sample	Chars	Words	Messages	Source	Timespan
CTAB-SAMPLE0001	21693	3658	313	tv.labes	01 Sep. 2016 - 31 Mar. 2017
CTAB-SAMPLE0002	5636	946	130	AttessiaTV	31 Mar. 2016 - 02 Apr. 2017
CTAB-SAMPLE0003	2906	501	100	Tn.1.Blid	01 Jan. 2014 - 01 Apr. 2017
CTAB-SAMPLE0004	22465	3788	227	tv.labes	01 May 2017 - 16 May 2017
CTAB-SAMPLE0005	22661	3857	322	EttounsiyaTV.AndiMankolek	29 Apr. 2017 - 19 May 2017
CTAB-SAMPLE0006	20893	3287	411	AttessiaTV	16 May 2017 - 18 May 2017
CTAB-SAMPLE0007	4890	828	91	AttessiaTV	21 Dec. 2017 - 21 Dec. 2017
CTAB-SAMPLE0008	57403	9458	1335	AttessiaTV	13 May 2021 - 16 May 2021
Overall	158547	26323	2929		