# Compare which custom tokenizer works better on more examples

Samples to run in python console or testing:

text1

El médico argentino Eduardo Sosa en el hat-trick de su twitter @yesyes y #happy en https://www.lanacion.com.ar/e y su esposa Edith gat@gmail.com fueron a dejar todo en la Argentina para mudarse con su familia a Bielorrusia, la ex república de la Unión Soviética que había sufrido las mayores consecuencias de la explosión de la central de Chernobyl, ubicada a pocos kilómetros de la frontera con Ucrania.

text2

El médico argentino Eduardo Sosa en hat-trick de su twitter @yesyes y #happy

Original Result:

['El', 'médico', 'argentino', 'Eduardo', 'Sosa', 'en', 'el', 'hat-trick', 'de', 'su', 'twitter', '@yesyes', 'y', '#happy', 'en', '\n ', 'https://www.lanacion.com.ar/e', 'y', 'su', 'esposa', 'Edith', 'gat@gmail.com', 'fueron', 'a', 'dejar', 'todo', 'en', 'la', 'Argentina', 'para', 'mudarse', 'con', 'su', '\n ', 'familia', 'a', 'Bielorrusia', ',', 'la', 'ex', 'república', 'de', 'la', 'Unión', 'Soviética', 'que', 'había', 'sufrido', 'las', 'mayores', 'consecuencias', 'de', 'la', '\n ', 'explosión', 'de', 'la', 'central', 'de', 'Chernobyl', ',', 'ubicada', 'a', 'pocos', 'kilómetros', 'de', 'la', 'frontera', 'con', 'Ucrania', '.']

['El', 'médico', 'argentino', 'Eduardo', 'Sosa', 'en', 'hat-trick', 'de', 'su', 'twitter', '@yesyes', 'y', '#happy']

Modified Result:

['El', 'médico', 'argentino', 'Eduardo', 'Sosa', 'en', 'el', 'hat-trick', 'de', 'su', 'twitter', '@yesyes', 'y', '#happy', 'en', '\n ', 'https://www.lanacion.com.ar/e', 'y', 'su', 'esposa', 'Edith', 'gat@gmail.com', 'fueron', 'a', 'dejar', 'todo', 'en', 'la', 'Argentina', 'para', 'mudarse', 'con', 'su', '\n ', 'familia', 'a', 'Bielorrusia', ',', 'la', 'ex', 'república', 'de', 'la', 'Unión', 'Soviética', 'que', 'había', 'sufrido', 'las', 'mayores', 'consecuencias', 'de', 'la', '\n ', 'explosión', 'de', 'la', 'central', 'de', 'Chernobyl', ',', 'ubicada', 'a', 'pocos', 'kilómetros', 'de', 'la', 'frontera', 'con', 'Ucrania', '.']

['El', 'médico', 'argentino', 'Eduardo', 'Sosa', 'en', 'hat-trick', 'de', 'su', 'twitter', '@yesyes', 'y', '#happy']

Way to test: Copy and paste text content into function custom\_tokenizer\_to\_df. Run all the code and check things in csv file generated in the same dictionary of the code.