

The entire process can be done running the task7.ipynb. There are code cells in it and we have runned it already and get the files we need.

1. we now have generated 500 fake papers including features like authors, titles and texts, and have converted them into a csv file.

	author	title	text
0	Abagael Abagail, Abbe Abbey, Abbi Abbie, Abby ...	Pulser vs. Vulcan	A). Blood samples were collected by centrifuga...
1	Ada Adah, Adaline Adara, Addie Addis, Adel Ade...	The caspase-1-active caspase-2 is a self-incom...	I was asked to give my brief opinion on the su...
2	Adelind Adeline, Adella Adelle, Adena Adey, Ad...	cmc_kills_bp(9) -cKlaxg2 -Dpk_Kills__p+_c_k...	.n\nProbimin and psilocybin: a cross-species ...
3	Adorne Adrea, Adria Adriaens, Adrian Adriana, ...	I like the 3.5-in. R9 270 a lot, but I'm not a...	SUBJECTS\n\nThe following were selected as a s...
4	Aeriel Aeriela, Aeriell Ag, Agace Agata, Agath...	In their analysis, they found that vegetarians...	A new report suggests that severe infections c...

Now we need to combine some features such as affiliation, citation and highest degree of the author.

2. From task3, we have generated a sample csv file including those features and we can complete the new csv file using those rows and elements in it. Then we need to combine those columns to form a dataframe.

Author	Title	Text	Citation	DOI	Year
Dedie Dedra, Dee Dee Dee, Deeann Deeanne, Deed...	"I am ready to go and play in the Premier Leag...	While the American Psychiatric Association (AP...	PLOS ONE, July 2013 Volume 8 Issue 7 e69...	10.1128/IAI.06056-11	2013
Domenic Domenico, Dominic Dominick, Dominique ...	The Koryo Robot will be able to detect and fie...	SPRINGFIELD, Mass. (AP) — A medical examiner o...	Oncogene (2015) 34, 826Ä837	10.1186/bcr3322	2009
Rica Ricca, Rici Ricki, Rickie Ricky, Riki Rik...	As such, the newly formed Captain.	Methylated\n\nmetallopropylthiothreitol.\n\...	Evidence-Based Complementary, and Alternative ...	10.1016/j.jaut.2006.05.001	2013

And then write it into a new output csv file called feature_add.csv.

3. To get the tsv file considering we already have the csv file, we only need to convert it to another file type.