CSE3063 OOSD Project #2 – Python Project – Fall 2017

Simple Text Mining Application

Input: a set of documents
Your program will read a set of documents (Word or pdf or text files) from a folder

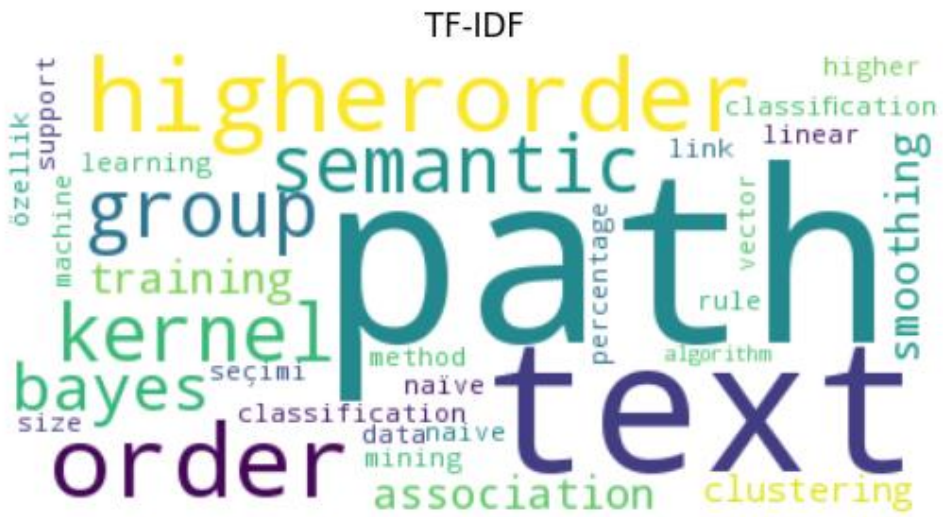
Output: several documents (text files)
Your program will output following files:

1- tf_list.csv: Most frequent 50 words in the input set of documents, sorted descending by their term frequency (tf) coupled with their tf values (comma seperated file, example: document;7)
2- tf_wordCloud.pdf: Word cloud of the these words
3- tfidf_list.csv: Most frequent 50 words in the input set of documents, sorted descending by their term frequency*inverse document frequency (tf-idf) coupled with their tf-idf values (comma seperated file, example: document;2.8)
4- tfidf_wordCloud.pdf: Word cloud of the these words

Description:

- Use your existing GitHub repositories but open a new folder called "project2" and put your python code into this directory.
- You can use any Python libraries to ease your job
- Transforming all text to lowercase may help in your calculations
- You need to filter very common words which are called stopwords before your calculations. You can use stopword list file for this purpose. However, please extend your stopword list with the common words in scientific publications such as abstract, introduction, conclusions, related work, author, university, etc, https://en.wikipedia.org/wiki/Stop_words
- Info for tf-idf calculations: https://en.wikipedia.org/wiki/Tf–idf
- You need to download the publications of a particular faculty member in our department or Computer Engineering departments of İTÜ and BOUN. You will use the set of publication documents of this faculty member as your input. You can use faculty members home page (e.g. http://mimoza.marmara.edu.tr/~murat.ganiz/) or Google scholar profile (https://scholar.google.com/citations?user=Ho93Fu0AAAAJ&hl=en). You must download as many publications of the faculty as possible. Use Google or Google scholar to search for publication names and to access pdf's. You can download and prepare your input files manually if you like. If you automatize this process you will get +15 Extra Credit.
- No analysis or design documents are required fort his Project
- Your projects will be evaluated based on the correctness and the quality of the outputs as well as the object oriented design. Please note that we can use a different input (for example a folder of publication pdf's) to evaluate your project

- An example word cloud can be found below



List of faculty – group assignments:

#	Group	Assigned Faculty
1	https://github.com/mgokceer/cse3063f17p1_gokceer_ebrukizilkiren	http://mimoza.marmara.edu.tr/~haluk/publications.html
2	https://github.com/mehmetcanyuney/cse3063f17p1_ercerkaya_mcyuney	https://scholar.google.com/citations?user=Ho93Fu0AAAAJ&hl=en
3	https://github.com/sumeyragulsoyy/cse3063f17p1_SUMEYRAGULSOY_ERMANHAVUC	https://scholar.google.com.au/citations?hl=en&user=5m6IrpgAAAAJ
4	https://github.com/ezgicinan/cse3063f17p1_ecinan_syildiz	http://mimoza.marmara.edu.tr/~falkaya/publications.htm
5	https://github.com/alpakseyma/cse3063f17p1-alpak_seyma-	http://mimoza.marmara.edu.tr/~borahan.tumer/publications.html
6	https://github.com/ali-mercan/cse3063f17p1_aimercan_ocelik	http://mimoza.marmara.edu.tr/~betul.demiroz/
7	mertcan	
8	https://github.com/hilalbalci/cse3063f17p1-sgumustas_aulgen_hbalci	http://mimoza.marmara.edu.tr/~fatma.ergin/
9	https://github.com/mertKelkit/cse3063f17p1_mkelkit_relioz_fozkan	http://mimoza.marmara.edu.tr/~omer.korcak/
10	https://github.com/ulgacemre/Monopoly_Project	
11	https://github.com/cerenbattal/cse3063f17p1_cbattal_snsari_ayilmaz	http://www.ahozer.com/publications.html
12	https://github.com/OkanEke/cse3063f17p1_oeke_madogan_aozturk	http://mimoza.marmara.edu.tr/~mujdat.soyturk/#publications
13	https://github.com/YasinEmreOZBARUT/Monopoly_Project	
14	https://github.com/Gulsahyilmaaz/cse3063f17p1_gyilmaz_hsahin_amustafa-	https://scholar.google.com/citations?user=Ho93Fu0AAAAJ&hl=en
15	https://github.com/tayfun-yurdaer/cse3063f17p1_tyurdaer_hyalcin	https://scholar.google.com.au/citations?hl=en&user=5m6IrpgAAAAJ
16	https://github.com/ozgegunay/cse3063f17p1_ogunay_mhaskan	http://mimoza.marmara.edu.tr/~falkaya/publications.htm
17	https://github.com/berkdagli/cse3063f17p1_huseyincanErbayraktar_berkdagli	http://mimoza.marmara.edu.tr/~borahan.tumer/publications.html
18	https://github.com/farukfurkan/cse3063f17p1_atay_bavincan_ffsisman	http://mimoza.marmara.edu.tr/~betul.demiroz/
19	bilgehan	
20	https://github.com/efeertugrul/cse3063f17p1_eecoskun_tsarp_moizmitlioglu	http://mimoza.marmara.edu.tr/~fatma.ergin/
21	Erman Kundakçioğlu ve Halid Seyfullah Sert	
22	https://github.com/muhammeddkilicc/cse3063f17p1_mkilic_obayraktar_masakalli	http://mimoza.marmara.edu.tr/~omer.korcak/
23	https://github.com/tahabilal/cse3063f17p1_tbozbey_aacebeci_sgulbetekin	http://www.ahozer.com/publications.html
24	https://github.com/omerfarukmercan/cse3063f17p1_ofmercan_isporhan	http://mimoza.marmara.edu.tr/~mujdat.soyturk/#publications
25	https://github.com/ezeybekoglu/cse3063f17p1_byasar_ezeybekoglu	http://mimoza.marmara.edu.tr/~haluk/publications.html
26	https://github.com/hilalekinci/cse3063f17p1_efmemis_hekinci_nkoroglu	https://scholar.google.com/citations?user=Ho93Fu0AAAAJ&hl=en
27	https://github.com/efeertugrul/cse3063f17p1_eecoskun_tsarp_moizmitlioglu	https://scholar.google.com.au/citations?hl=en&user=5m6IrpgAAAAJ
28	https://github.com/mehmethangur/cse3063f17p1_yTor_mGur_nIlgin	http://mimoza.marmara.edu.tr/~falkaya/publications.htm