CS121/INF141: Bidyuk
Winter 2016

Team Members:
Cassie Jeansonne 18923914
Kevin Ko 56956077
Samuel Lin 52478518
Sophia Chan 33196560

Milestone #1: Implement Indexer and Build an Index

1. Number of Documents: 44,536 documents
2. Number of [unique] words: 381,786 unique words
3. Sample index: key-value pairs:
    - docid2termlist.txt:
        - 1: [10518, 19618, 2451, 104, 151795, 6536, 185, 10543, 10518, 4278, 5465, 4657, 767, 6175, 298961, 853, 3993, 1013, 298208, 298323, 880, 2785, 5727, 194408, 298187, 297970, 275521, 109393, 32941, 298566, 298567, 56, 298050, 298568, 225977, 13738, 298051, 298775, 298193, 298319]
        - 2: [841, 898, 14548, 20443, 12915, 20444, 3984, 134, 1826, 56, 4852, 51, 268, 17385, 566, 48, 10347, 2775, 591, 119, 480, 134, 1112, 119, 228, 566, 73, 20445, 2889, 18757, 20446, 7380, 20447, 20448, 20449]
        - 3: [96743, 96743, 98600, 4506, 9081, 87216, 87215, 658, 158, 86654, 9202, 3510, 83593, 83594, 9, 10342, 85020, 8086, 86801, 4842, 86326, 98595, 5067, 87210, 87211, 98596, 299, 93248, 1406, 5785, 1667, 93249, 493, 83435, 460, 8714, 979, 460, 83434, 7652, 2889, 115, 493, 83435, 37, 38, 19, 39, 53, 446, 42]
        - 4: [15399, 3868, 38, 303647, 119783, 19618, 303647, 119783, 80, 88376, 8698, 104, 14150, 419, 185, 303653, 303647, 119783, 15399, 3868, 38, 303647, 119783, 4278, 5059, 4407, 451, 871, 301660, 853, 3993, 1013, 297994, 303659, 880, 2785, 5727, 194408, 297969, 298005, 202693, 109393, 2126, 298063, 298056, 298035, 225977, 158712, 298051, 300086, 303660, 194547, 10620]
    - term2termid.txt:
        - : 13072
        - frowning: 123490
        - undermining: 354052
        - orgitrrescuedataguardindexingindex: 361638
        - 34pm: 188097
        - 3640: 262950
        - pamey: 235599
        - abrupt: 253882
        - nationalkonomie: 94617
        - hardtoanswer: 377160
        - dkyqjmwrohe0hyu19h39uqeozuqppyg4xq7: 58365
        - rtenyear: 147071
    - termid2docidTFIDF.txt:
        - 1: [18384, 34037, 29058, 2712, 23208, 15001, 23546, 3135, 5754, 6638, 11425, 2927, 2421, 4319, 21319, 15494, 7196, 1604, 10263, 21697, 575, 40541, 79, 12920, 42132, 31028, 42742, 28093, 34240, 43891, 16219, 27327, 24814, 12653, 28160, 33661, 11480,

24020, 31640, 12888, 30083, 33604, 223, 3578, 44531, 16667, 41956, 37189, 41707, 9926, 29790, 7131, 15275, 7155, 36733, 42292, 12886, 27465, 27702, 27701, 34974, 26947, 32199, 6402, 39591, 3432, 355, 3624, 27772, 24513, 33843, 27330, 37582, 24969, 27680, 29786, 34755, 34010, 14940, 2828, 38203, 31142, 3068, 15702, 43781, 8771, 21378, 34662, 37137, 33420, 12838, 14721, 949, 4858, 33660, 40131, 42227, 15816, 26370, 354, 20675, 16399, 27577, 42802, 10610, 20908, 20013, 38904, 17932, 25399, 38982, 31195, 18974, 3340, 34528, 21822, 15186, 16050, 41798, 12525, 43554, 9761, 4911, 4360, 7233, 24917, 3039, 280, 10416, 25352, 6956, 22095, 26706, 40645, 18817, 198, 4907, 18315, 1990, 8569, 34726, …]

termid2doclist.txt:

1: [18384, 34037, 29058, 2712, 23208, 15001, 23546, 3135, 5754, 6638, 11425, 2927, 2421, 4319, 21319, 15494, 7196, 1604, 10263, 21697, 575, 40541, 79, 12920, 42132, 31028, 42742, 28093, 34240, 43891, 16219, 27327, 24814, 12653, 28160, 33661, 11480, 24020, 31640, 12888, 30083, 33604, 223, 3578, 44531, 16667, 41956, 37189, 41707, 9926, 29790, 7131, 15275, 7155, 36733, 42292, 12886, 27465, 27702, 27701, 34974, 26947, 32199, 6402, 39591, 3432, 355, 3624, 27772, 24513, 33843, 27330, 37582, 24969, 27680, 29786, 34755, 34010, 14940, 2828, 38203, 31142, 3068, 15702, 43781, 8771, 21378, 34662, 37137, 33420, 12838, 14721, 949, 4858, …]

termid2term.txt:

1: contact
2: us
3: alliance
4: for
5: california
6: computing
7: education
8: students
9: and
10: schools

4. Total size (in KB) of your index on disk:

docid2termlist.txt : 71,181 KB
term2termid.txt : 7,875 KB
termid2docidTFIDF.txt : 133,358 KB
termid2doclist.txt : 38,096 KB
termid2term.txt : 7,875 KB

Total size (in KB) : 258,385 KB

5. Time taken to create your index:

```
"C:\Program Files\Java\jdk1.8.0_66\bin\java" ...
Runtime: 3419611 ms

Process finished with exit code 0
```

56.9935 minutes.