TP2 - Sentiment Analysis

June 11, 2019

1 Sentiment Analysis In Textual Movie Reviews

Maxime Tchibozo

```
In [150]: import string
```

```
In [151]: import os
          os.chdir('C:\\Users\\Max Tchibozo\\Desktop\\SD-TSIA214\\TP2\\data\\data')
```

```
In [152]: # Authors: Alexandre Gramfort
          #          Chloe Clavel
          # License: BSD Style.
          # TP Cours ML Telecom ParisTech MDI343

          import os.path as op
          import numpy as np

          from sklearn.base import BaseEstimator, ClassifierMixin

          ###############################################################################
          # Load data
          print("Loading dataset")

          from glob import glob
          filenames_neg = sorted(glob(op.join('..', 'data', 'imdb1', 'neg', '*.txt')))
          filenames_pos = sorted(glob(op.join('..', 'data', 'imdb1', 'pos', '*.txt')))
          texts_neg = [open(f).read() for f in filenames_neg]
          texts_pos = [open(f).read() for f in filenames_pos]
          texts = texts_neg + texts_pos
          y = np.ones(len(texts), dtype=np.int)
          y[:len(texts_neg)] = 0.

          print("%d documents" % len(texts))
```

```
Loading dataset
2000 documents
```

```
In [153]: y[999],y[1000]#This is the moment when we go from the negative reviews to the positi

Out[153]: (0, 1)
```

2 Question 1

```
In [154]: import string
          ################################################################################
          # Start part to fill in

          def count_words(texts):
              """Vectorize text : return count of each word in the text snippets

              Parameters
              ----------
              texts : list of str
                  The texts

              Returns
              -------
              vocabulary : dict
                  A dictionary that points to an index in counts for each word.
              counts : ndarray, shape (n_samples, n_features)
                  The counts of each word in each text.
                  n_samples == number of documents.
                  n_features == number of words in vocabulary.
              """
              punctuation = list(string.punctuation)+['\n'] #We also want to remove the newlin

              total_word_list = [] #will contain the words of all texts without separation
              text_word_list = [] #will separate the words of each text in a different list it

              for text in texts:
                  for punct in punctuation:
                      text = text.replace(punct,' ')
                  word_list = text.split(' ')
                  word_list = [x for x in word_list if x != ''] #We remove the empty strings :

                  total_word_list += word_list # We build vocabulary thanks to the list of all
                  text_word_list.append(word_list)

              words = list(set(total_word_list))
              vocabulary = {}

              for i in range(len(words)):
                  vocabulary[words[i]] = i
```

```
    counts = np.zeros((2000,len(words))) #there are 2000 documents

    for i in range(len(text_word_list)):
        for j in range(len(text_word_list[i])):
            index = vocabulary[text_word_list[i][j]] #This is the index of the word
            counts[i][index] += 1

    return vocabulary, counts


count_words(texts)
```

```
Out[154]: ({'graded': 0,
            'sayles': 1,
            'stubby': 2,
            'wholesome': 3,
            'rachmaninov': 4,
            'aftermath': 5,
            'rubs': 6,
            'mnemonic': 7,
            'directional': 8,
            'rich': 9,
            'alarmed': 10,
            'inversion': 11,
            '1865': 12,
            'shards': 13,
            'gosnell': 14,
            'orwellian': 15,
            'synchs': 16,
            'buffoonish': 17,
            'hailed': 18,
            'motions': 19,
            'mtcts1': 20,
            'greenhouse': 21,
            'hole\x14': 22,
            'osmond': 23,
            'phenomenas': 24,
            'planets': 25,
            'lingered': 26,
            'standards': 27,
            'treetops': 28,
            'steadiocam': 29,
            'excavating': 30,
            'weaponesque': 31,
            'imaginary': 32,
            'truant': 33,
            'wiper': 34,
            'zoe': 35,
            'barenboim': 36,
            'cromwell': 37,
            'residential': 38,
            'machinist': 39,
            'impregnating': 40,
            'precedes': 41,
            'saigon': 42,
            'oaf': 43,
            'ferguson': 44,
            'roberts': 45,
            'flicker': 46,
            'treacherous': 47,
            'motley': 48,
            'starphoenix': 49,
            'sails': 50,
            'dejection': 51,
            'subsidies': 52,
            'svenwara': 53,
            'ferrell': 54,
            'squashed': 55,
            'luggage': 56,
            'burke': 57,
            'sittings': 58,
            'sued': 59,
            'touching': 60,
            'extremel': 61,
            'shape': 62,
            'recourse': 63,
            'projectioner': 64,
            'partnerships': 65,
            'dreamworld': 66,
            'norad': 67,
            'slashfest': 68,
            'expressionists': 69,
            'sexpot': 70,
            'aboard': 71,
            'indulging': 72,
            'folk': 73,
            'steele': 74,
            'abbe': 75,
            'unselfconsciously': 76,
            'hellhole': 77,
            'judge': 78,
            'meditative': 79,
            'whorehouses': 80,
            'overtly': 81,
            'debauchery': 82,
            'romulus': 83,
            'arbuthnot': 84,
            'dislikable': 85,
            'movement': 86,
            'freedoms': 87,
            'highpoint': 88,
            'sludge': 89,
            'hips': 90,
            'many': 91,
            'sling': 92,
            'wraps': 93,
            'collector': 94,
            'revolucionario': 95,
            'adventure': 96,
            'lechery': 97,
            'counterfeit': 98,
            'biziou': 99,
            'milton': 100,
            'puritanical': 101,
            'washer': 102,
            'assert': 103,
            'swamps': 104,
            'underwritten': 105,
            'repent': 106,
            'ly': 107,
            'blackout': 108,
            'truer': 109,
            'playfulness': 110,
            'transcending': 111,
            'intercom': 112,
            'parks': 113,
            'belting': 114,
            'messy': 115,
            'hurt': 116,
            'sleuthing': 117,
            'mamma': 118,
            'dignities': 119,
            'danna': 120,
            'flatmate': 121,
            'exotica': 122,
            'tinges': 123,
            'subconscious': 124,
            'piet': 125,
            'banished': 126,
            '180': 127,
            'define': 128,
            'discouraging': 129,
            'sequences': 130,
            'gods': 131,
            'affirmation': 132,
            'retardation': 133,
            'mesquida': 134,
            'desert': 135,
            'goldenberg': 136,
            'szwarc': 137,
            'heigh': 138,
            'dominated': 139,
            'infinitum': 140,
            'confusing': 141,
            'geography': 142,
            'nfl': 143,
            'cowering': 144,
            'artifact': 145,
            'substantially': 146,
            'sv2': 147,
            'graduates': 148,
            'wasn': 149,
            'katt': 150,
            'conservationist': 151,
            'grandest': 152,
            'tableware': 153,
            'cultish': 154,
            'bongos': 155,
            'negotiable': 156,
            'steinberg': 157,
            'disobeys': 158,
            'invulnerability': 159,
            'countryside': 160,
            'faison': 161,
            'fluoro': 162,
            'entangled': 163,
            'unopposed': 164,
            'messinger': 165,
            'incestuous': 166,
            'nominal': 167,
            'bumblingly': 168,
            'aciton': 169,
            'irrepressible': 170,
            'decide': 171,
            'autumn': 172,
            'resource': 173,
            'eulogy': 174,
            'champagne': 175,
            'tangent': 176,
            'jagger': 177,
            'songwriting': 178,
            'footnotes': 179,
            'cyberkillers': 180,
            'winnie': 181,
            'ventricle': 182,
            'mulholland': 183,
            'malloy': 184,
            'linearity': 185,
            'xander': 186,
            'timex': 187,
            'sexiest': 188,
            'dublin': 189,
            'bounteous': 190,
            'obligated': 191,
            'electronically': 192,
            'lemme': 193,
            'crony': 194,
            'veritably': 195,
            '152': 196,
            'multidimensionality': 197,
            'apparel': 198,
            'criticisms': 199,
            'shorter': 200,
            'cries': 201,
            'dictionary': 202,
            'landscapes': 203,
            'bass': 204,
            'canran': 205,
            'mispronounces': 206,
            'tempest': 207,
            'admission': 208,
            'overplayed': 209,
            'unbridled': 210,
            'etymology': 211,
            'parked': 212,
            'abs': 213,
            'layered': 214,
            'boneheaded': 215,
            'juke': 216,
            'rob': 217,
            'honey': 218,
            'await': 219,
            'ramses': 220,
            'ww2': 221,
            'typically': 222,
            'observers': 223,
            'unfunny': 224,
            'poking': 225,
            'angered': 226,
            'lamanna': 227,
            'klingon': 228,
            'dive': 229,
            'tooth': 230,
            'donate': 231,
            'starbuck': 232,
            'grays': 233,
            'biographies': 234,
            'democratic': 235,
            'revolting': 236,
            'train': 237,
            'extravaganza': 238,
            'astoundingly': 239,
            'thawed': 240,
            '14': 241,
            'friedkin': 242,
            'saul': 243,
            'posess': 244,
            'spacey': 245,
            'obscurity': 246,
            'effect': 247,
            'droves': 248,
            'hoist': 249,
            'sorvino': 250,
            'effects': 251,
            'ice': 252,
            'jungle2jungle': 253,
            'commodities': 254,
            'montages': 255,
            'feeble': 256,
            'limestone': 257,
            'buehler': 258,
            'preposterousness': 259,
            'guitars': 260,
            'melee': 261,
            'androgony': 262,
            'borders': 263,
            'goatee': 264,
            'hamster': 265,
            'notion': 266,
            'kowtowing': 267,
            'policed': 268,
            'teeter': 269,
            'tailor': 270,
            'floods': 271,
            'obtained': 272,
            'plumber': 273,
            'torment': 274,
            'stonily': 275,
            'drowned': 276,
            'gears': 277,
            'stormare': 278,
            'decks': 279,
            'arrive': 280,
            'loutish': 281,
            'glave': 282,
            'petitioned': 283,
            'glorify': 284,
            'salability': 285,
            'southerners': 286,
            'resevoir': 287,
            'estella': 288,
            'mark': 289,
            'hughley': 290,
            'ripe': 291,
            'coenesque': 292,
            'vampira': 293,
            'nordoff': 294,
            'madeliene': 295,
            'businesswomen': 296,
            'morquio': 297,
            'ing': 298,
            'adaptions': 299,
            'crooked': 300,
            'crams': 301,
            'discontented': 302,
            'desparate': 303,
            'various': 304,
            'obstetrician': 305,
            'dianne': 306,
            'constant': 307,
            'responsibility': 308,
            'maestro': 309,
            'leavins': 310,
            'improvise': 311,
            'cameraman': 312,
            'electroshock': 313,
            'stead': 314,
            'valjean': 315,
            'coolest': 316,
            'ignore': 317,
            'names': 318,
            'sleaze': 319,
            'pronounces': 320,
            'picturing': 321,
            'o': 322,
            'comaprison': 323,
            'conked': 324,
            'labute': 325,
            'caberat': 326,
            'distracted': 327,
            'mandatory': 328,
            'matrix': 329,
            'chainsmokes': 330,
            'sandefur': 331,
            'masaya': 332,
            'refund': 333,
            'thermopolis': 334,
            'shrewd': 335,
            'timelines': 336,
            'postmodernism': 337,
            'nominees': 338,
            'assasination': 339,
            'faced': 340,
            'contents': 341,
            'vexatiousness': 342,
            'mustered': 343,
            'berle': 344,
            'deemphasize': 345,
            'rosy': 346,
            'augustin': 347,
            'both': 348,
            'treason': 349,
            'certainly': 350,
            'mufasa': 351,
            'beasts': 352,
            'whispery': 353,
            'subdued': 354,
            'sims': 355,
            'projected': 356,
            'transportation': 357,
            'straight': 358,
            'counts': 359,
            'waylon': 360,
            'outrun': 361,
            'pratfalling': 362,
            'colored': 363,
            'sexless': 364,
            'vig': 365,
            'cannot': 366,
            'trampoline': 367,
            'robs': 368,
            'proclaimed': 369,
            'dub': 370,
            'mistaken': 371,
            'horseplay': 372,
            'either': 373,
            'dispose': 374,
            'knockout': 375,
            'outlands': 376,
            'detritus': 377,
            'bloodier': 378,
            'jackee': 379,
            'hypocritical': 380,
            'limps': 381,
            'drumroll': 382,
            'bogg': 383,
            'squarely': 384,
            'facing': 385,
            'lovesick': 386,
            'appended': 387,
            'boone': 388,
            'stepsister': 389,
            'drape': 390,
            'intermingle': 391,
            'mediating': 392,
            'zoologist': 393,
            'pina': 394,
            'motor': 395,
            'observance': 396,
            'wabbit': 397,
            'allegedly': 398,
            'as': 399,
            'actual': 400,
            'briers': 401,
            'trickster': 402,
            'slightest': 403,
            'delegates': 404,
            'eviction': 405,
            'pitted': 406,
            'schwalbach': 407,
            'winces': 408,
            'anecdotes': 409,
            'measly': 410,
            'brennan': 411,
            'lorry': 412,
            'mouthed': 413,
            'block': 414,
            'stored': 415,
            'deconstruction': 416,
            'gramercy': 417,
            'sheer': 418,
            'angers': 419,
            'stimulates': 420,
            'gulliver': 421,
            'fatigues': 422,
            'chloe': 423,
            'crystal': 424,
            'stadium': 425,
            'restored': 426,
            'closets': 427,
            'duelling': 428,
            'pego': 429,
            'reindeer': 430,
            'hanif': 431,
            'allison': 432,
            'plunder': 433,
            '8a': 434,
            'chiding': 435,
            'survivor': 436,
            'rider': 437,
            'phantoms': 438,
            'longbaugh': 439,
            'fleshes': 440,
            'donovan': 441,
            'patterned': 442,
            'puffy': 443,
            'midwestern': 444,
            'plow': 445,
            'fogs': 446,
            'mesmerizes': 447,
            'testing': 448,
            'tick': 449,
            'lifting': 450,
            'computech': 451,
            'unrecognizable': 452,
            'ramblings': 453,
            'macho': 454,
            'chairs': 455,
            'wittliff': 456,
            'council': 457,
            'crushing': 458,
            'rapier': 459,
            'fearful': 460,
            'bud': 461,
            'staged': 462,
            'wildly': 463,
            'credentials': 464,
            'receipt': 465,
            'frankiln': 466,
            'snorting': 467,
            'scoopfuls': 468,
            'dramatism': 469,
            'balthazar': 470,
            'override': 471,
            'thrusting': 472,
            'reinforced': 473,
            'proposal': 474,
            'phillipe': 475,
            'casket': 476,
            'intoxicated': 477,
            'masterson': 478,
            'freelance': 479,
            'vicent': 480,
            'truths': 481,
            'caddy': 482,
            'snorri': 483,
            'ethnic': 484,
            'liveliness': 485,
            'demme': 486,
            'convulsing': 487,
            'harumph': 488,
            'misstep': 489,
            'unveil': 490,
            'theme': 491,
            'voiceovers': 492,
            'inually': 493,
            'poolboy': 494,
            'realization': 495,
            'paralells': 496,
            'maneuvers': 497,
            'russkies': 498,
            'references': 499,
            'chooses': 500,
            'blowing': 501,
            'renshaw': 502,
            'incessant': 503,
            'spiraling': 504,
            'knifed': 505,
            'politically': 506,
            'bossy': 507,
            'tinkering': 508,
            'severance': 509,
            'disapproves': 510,
            'hawks': 511,
            'maturity': 512,
            'portray': 513,
            'ie': 514,
            'maternity': 515,
            'mantlepiece': 516,
            'identital': 517,
            'bouts': 518,
            'summon': 519,
            'vonnegut': 520,
            'fragmentary': 521,
            'liking': 522,
            'ears': 523,
            'preaches': 524,
            'misinterprets': 525,
            'jackson': 526,
            'patronizing': 527,
            'slicker': 528,
            'compatible': 529,
            'cuarsn': 530,
            'bravery': 531,
            'uniqe': 532,
            'motorcylce': 533,
            'politicians': 534,
            'terpiece': 535,
            'inhabiting': 536,
            'home': 537,
            'responsibilities': 538,
            'boats': 539,
            'stabilize': 540,
            'advancements': 541,
            'invested': 542,
            'ww': 543,
            'outtakes': 544,
            'cds': 545,
            'potts': 546,
            'depraved': 547,
            'stargher': 548,
            'stirrings': 549,
            'serpico': 550,
            'soavi': 551,
            'traffiking': 552,
            'peaks': 553,
            'crunchem': 554,
            'aggravate': 555,
            'mandy': 556,
            'mcjob': 557,
            'esuqe': 558,
            'tek': 559,
            'uncovering': 560,
            'affliction': 561,
            'awk': 562,
            'paulina': 563,
            'klieg': 564,
            'race': 565,
            'authorial': 566,
            'swoops': 567,
            'heenan': 568,
            'slurping': 569,
            'archetype': 570,
            'dissappointed': 571,
            'handle': 572,
            'crust': 573,
            'moi': 574,
            'dwayne': 575,
            'spice': 576,
            'prejudge': 577,
            'curious': 578,
            'animosity': 579,
            'uganda': 580,
            'selects': 581,
            'sloth': 582,
            'morose': 583,
            'weiss': 584,
            'blofeld': 585,
            'rigorous': 586,
            'slowing': 587,
            'pseudo': 588,
            'lashing': 589,
            'acquired': 590,
            'goodtimes': 591,
            'testings': 592,
            'needling': 593,
            'relished': 594,
            'moores': 595,
            'surname': 596,
            'obtain': 597,
            'impossibility': 598,
            'tycoon': 599,
            'stragely': 600,
            'wo': 601,
            'caleb': 602,
            'racism': 603,
            'protovision': 604,
            'newton': 605,
            'wallenberg': 606,
            'reeking': 607,
            'screen': 608,
            'hours': 609,
            'pitfalls': 610,
            'rapists': 611,
            'parlays': 612,
            'ritchie': 613,
            'arctic': 614,
            'beatles': 615,
            'venting': 616,
            'gos': 617,
            'maidservant': 618,
            'chewing': 619,
            'cousin': 620,
            'punk': 621,
            'ridicously': 622,
            'tires': 623,
            'munching': 624,
            'gellar': 625,
            'frightfulness': 626,
            'lite': 627,
            'toenails': 628,
            'bosom': 629,
            'matchmaking': 630,
            'badger': 631,
            'grasshoppers': 632,
            'slathering': 633,
            'flops': 634,
            'servo': 635,
            'shadows': 636,
            'proval': 637,
            'ryuichi': 638,
            'janssen': 639,
            'conqueror': 640,
            'stefanson': 641,
            'rally': 642,
            'moonlit': 643,
            'preteen': 644,
            'cream': 645,
            'watermelons': 646,
            'easily': 647,
            'classes': 648,
            'mildred': 649,
            'fragmented': 650,
            'rename': 651,
            'misogyny': 652,
            'wardens': 653,
            'century': 654,
            'winnebago': 655,
            'barnacles': 656,
            'biceps': 657,
            'wheat': 658,
            'wouldn': 659,
            'mise': 660,
            'hauff': 661,
            'seldon': 662,
            'zone': 663,
            'undifferentiated': 664,
            'militant': 665,
            'cruising': 666,
            'fashionable': 667,
            'hughes': 668,
            'panoramic': 669,
            'wizened': 670,
            'products': 671,
            'hates': 672,
            'ebony': 673,
            'forbids': 674,
            'sicken': 675,
            'sprightly': 676,
            'parfitt': 677,
            'disappear': 678,
            'courageous': 679,
            'downsides': 680,
            'snorts': 681,
            'clooney': 682,
            'protgaonist': 683,
            'termination': 684,
            'countenance': 685,
            'shaved': 686,
            'universe': 687,
            'stepahne': 688,
            'brett': 689,
            'maxx': 690,
            'whisked': 691,
            'bochner': 692,
            '1st': 693,
            'proclivity': 694,
            'match': 695,
            'spungen': 696,
            'mostel': 697,
            'easygoing': 698,
            'defensive': 699,
            'elmaloglou': 700,
            'cherot': 701,
            'eerily': 702,
            'guesses': 703,
            'accomplishes': 704,
            'tighten': 705,
            'excesses': 706,
            'uplifting': 707,
            'bump': 708,
            'heralded': 709,
            'morissey': 710,
            'tested': 711,
            'degredation': 712,
            'marienbad': 713,
            'bibile': 714,
            'duper': 715,
            'dantes': 716,
            'improper': 717,
            'regurgitate': 718,
            'fiercer': 719,
            'cheesefest': 720,
            'hokum': 721,
            'frequency': 722,
            'dialectic': 723,
            'decalogue': 724,
            'canoes': 725,
            'bunny': 726,
            'remastered': 727,
            'waterboy': 728,
            '555': 729,
            'raw': 730,
            'abel': 731,
            'jeez': 732,
            'orgasmically': 733,
            'isn\x12t': 734,
            'bashes': 735,
            'tying': 736,
            'hooch': 737,
            'loose': 738,
            'misspelled': 739,
            'minimum': 740,
            'adeptness': 741,
            'channeling': 742,
            'spelling': 743,
            'problem': 744,
            'posturing': 745,
            'acquit': 746,
            'cessation': 747,
            'unearthed': 748,
            'courtyards': 749,
            'benoit': 750,
            'moland': 751,
            'cohn': 752,
            'barcalow': 753,
            'disneyfied': 754,
            'squirt': 755,
            'slick': 756,
            'spend': 757,
            'congenial': 758,
            'mckee': 759,
            'spiritless': 760,
            'drafted': 761,
            'abusers': 762,
            'ogle': 763,
            'presidential': 764,
            'organizer': 765,
            'salvage': 766,
            'taught': 767,
            'lobotomise': 768,
            'auteurs': 769,
            'horrendous': 770,
            'chintzy': 771,
            'cohesive': 772,
            'hooky': 773,
            'clandestine': 774,
            'appelation': 775,
            'seann': 776,
            'vanquished': 777,
            'whereupon': 778,
            'spock': 779,
            'hypotheitically': 780,
            'gags': 781,
            'fluidly': 782,
            'deschanel': 783,
            'frankfurters': 784,
            'sadie': 785,
            'grandfather': 786,
            'commandeer': 787,
            'mikel': 788,
            'instrument': 789,
            '1930': 790,
            'casted': 791,
            'damian': 792,
            'squaddie': 793,
            'preconceptions': 794,
            'dramatic': 795,
            'chromium': 796,
            'stalwart': 797,
            'villagers': 798,
            'bogeyman': 799,
            'movingly': 800,
            'rate': 801,
            'derail': 802,
            'widows': 803,
            'gutierrez': 804,
            'boni': 805,
            'ribbing': 806,
            'attic': 807,
            'satiate': 808,
            'delinquent': 809,
            'directives': 810,
            'slumping': 811,
            'grumpier': 812,
            'ironing': 813,
            'tabbed': 814,
            'jermaine': 815,
            'contestants': 816,
            'penney': 817,
            'systematically': 818,
            'miriam': 819,
            'guenveur': 820,
            'mentoring': 821,
            'crumpled': 822,
            'sensual': 823,
            'nunez': 824,
            'abominable': 825,
            'gauntlet': 826,
            'unformal': 827,
            'fasano': 828,
            'paid': 829,
            'erroneous': 830,
            'instalment': 831,
            'cabbie': 832,
            'asswhole': 833,
            'tate': 834,
            'within': 835,
            'stubbornness': 836,
            'delpy': 837,
            'abort': 838,
            'edt': 839,
            'intensified': 840,
            'schoolchildren': 841,
            'galvanizing': 842,
            'pricked': 843,
            'sk': 844,
            'quicktime': 845,
            'songwriter': 846,
            'hallmark': 847,
            'crud': 848,
            'independence': 849,
            'beaver': 850,
            'flickers': 851,
            'kang': 852,
            'electing': 853,
            'correspondence': 854,
            'schoolteacher': 855,
            'necessarily': 856,
            'glossing': 857,
            'uncharismatic': 858,
            'sides': 859,
            'planet': 860,
            'stogie': 861,
            'applauded': 862,
            'cabin': 863,
            'strikeout': 864,
            'hankie': 865,
            'arresting': 866,
            'inter': 867,
            'funk': 868,
            'claw': 869,
            'corrupted': 870,
            'caricatures': 871,
            'sugary': 872,
            'petter': 873,
            'inoffensive': 874,
            'stalked': 875,
            'trial': 876,
            'indies': 877,
            'uninterest': 878,
            'mgm': 879,
            'butt': 880,
            'wiretaps': 881,
            'charted': 882,
            'neglecting': 883,
            'discussions': 884,
            'basking': 885,
            'dwindle': 886,
            'flexes': 887,
            'baby': 888,
            'lifted': 889,
            'amercian': 890,
            'moaning': 891,
            'arangements': 892,
            'beaumont': 893,
            'outstanding': 894,
            'lacklustre': 895,
            'drawer': 896,
            'lets': 897,
            'corvino': 898,
            'saves': 899,
            'amidst': 900,
            'sfx': 901,
            'prodigy': 902,
            'pitchfork': 903,
            'infiltrated': 904,
            'lineman': 905,
            'roam': 906,
            'achieve': 907,
            'unchecked': 908,
            'squished': 909,
            'rangoon': 910,
            'ende': 911,
            'heffron': 912,
            'brave': 913,
            'consonants': 914,
            'commandos': 915,
            'fractured': 916,
            'suitcase': 917,
            'counterbalanced': 918,
            'resonant': 919,
            'finklestein': 920,
            'kivilo': 921,
            'millionare': 922,
            'teaser': 923,
            'rank': 924,
            'deconstruct': 925,
            'copycats': 926,
            'naifeh': 927,
            'iran': 928,
            'teenagers': 929,
            'unacceptable': 930,
            'lundgren': 931,
            'regularity': 932,
            'vitality': 933,
            'yielded': 934,
            'bujold': 935,
            'droppingly': 936,
            'indulgently': 937,
            'halves': 938,
            'unforgiveably': 939,
            'mariner': 940,
            'dispensing': 941,
            'waving': 942,
            'birch': 943,
            'tease': 944,
            'slaughtering': 945,
            'cutbacks': 946,
            'ladened': 947,
            'greek': 948,
            'inches': 949,
            'bran': 950,
            'laptop': 951,
            'touches': 952,
            'trite': 953,
            'dullness': 954,
            'chaste': 955,
            'ketchup': 956,
            'raeeyain': 957,
            'aug': 958,
            'sluttish': 959,
            'reiner': 960,
            'serialised': 961,
            'petaluma': 962,
            'circulation': 963,
            'ab': 964,
            'courtoom': 965,
            'durable': 966,
            'bostonians': 967,
            'ninjaman': 968,
            'weirdoes': 969,
            'cuteness': 970,
            'harra': 971,
            'interpretation': 972,
            'toneless': 973,
            'jungle': 974,
            'bombay': 975,
            'smirk': 976,
            'excess': 977,
            'beesley': 978,
            'knob': 979,
            'anjelica': 980,
            'downsized': 981,
            'eyebrows': 982,
            'gamorreans': 983,
            'lively': 984,
            'acupuncture': 985,
            'rooftop': 986,
            'graduate': 987,
            'tomlin': 988,
            'infectious': 989,
            'resignedly': 990,
            'proclivities': 991,
            'fished': 992,
            'juror': 993,
            'amalgam': 994,
            'humourous': 995,
            'trough': 996,
            'sublte': 997,
            'wooed': 998,
            'eighty': 999,
            ...},
           array([[0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  ...,
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.]]))
```

```
In [155]: vocabulary , counts = count_words(texts)
```

There are 39443 different words in the vocabulary of the IMDB Movie Review database.

It is important to realise that some of these words do not hold much semantic meaning because of the pre-processing we have done:

Composite words (i.e "Jean-Claude") and more generally words with any type of punctuation (i.e: O-M-G) are split into several individal sub-words ("Jean","Claude" and "O","M","G").

3 Question 2

The class is attributed to reviews is recognized through the first explicit and identifiable rating.

These ratings are specified through star and numerical values i.e : "8/10", "four out of five", and "OUT OF ****: ***"

There is one noteworthy specificality, which is that in the case where the identified rating is a 5 star rating with decimal points (i.e 2.5 stars, 3.5 stars), the associated rating will be the truncated value of the rating (resp. 2 stars, 3 stars).

The authors explain that this is not problematic, given that the class output of ± 1 will be erroneous only when the rating was 2.5. And even then, it could be said that average reviews are negative reviews.

4 Question 3

```
In [157]: class NB(BaseEstimator, ClassifierMixin):
              def __init__(self):
                  pass

              def fit(self, X, y): #This corresponds to TrainMultinomialNB
                  #X = counts
                  #y = vecteur de {0,1}  label de la classe
                  N = X.shape[0]
                  self.prior = np.zeros(2)
                  counts_neg = X[:N//2][:]
                  counts_pos = X[N//2:][:]
```

```
        counts = [counts_neg,counts_pos]
        self.condprob = np.zeros((X.shape[1],2))
        for c in range(2): #for each class (0 or 1)
            Nc = len(y)/2 #half of the docs are positive, the other half are negativ
            self.prior[c] = Nc/N
            current_counts = counts[c] #This is the counts matrix of the given class
            c_total_counts = 0

            T = np.zeros((2, X.shape[1]))
            sums = np.zeros(2)
            for t in range(current_counts.shape[1]):
                T[c][t] = np.sum(list(current_counts[:, t]))
                sums[c] += T[c][t]

            for t in range(current_counts.shape[1]):
                self.condprob[t][c] = (T[c][t] + 1) / (sums[c] + current_counts.shap
        return self

    def predict(self, X):

        n = X.shape[0]
        predictions = np.zeros(n, dtype=int)

        for i in range(X.shape[0]):
            W = np.argwhere(X[i] != 0).flatten() #Where X points are != 0

            score = np.zeros(2)
            for c in range(2):
                score[c] = np.log(self.prior[c])
                for t in W:
                    score[c] += np.log(self.condprob[t][c])

            predictions[i] = np.argmax(score)

        return predictions

    def score(self, X, y):
        return np.mean(self.predict(X) == y)

# Count words in text
vocabulary, X = count_words(texts)

# Try to fit, predict and score
nb = NB()
nb.fit(X[::2], y[::2])
print('The score on the complete X dataset is : '+str(nb.score(X[1::2], y[1::2])))
```

```
The score on the complete X dataset is : 0.82
```

5 Cross-Validation 5-Folds

```
In [158]: from sklearn.model_selection import cross_val_score

          print('The 5-fold cross-validation score is : '+str(cross_val_score(NB(), X, y, cv=5
```

```
The 5-fold cross-validation score is : 0.8255000000000001
```

6 Stop-Words

```
In [159]: with open ('english.stop','r') as f:
              lines = f.readlines()
          stop_words = [x[:-1] for x in lines]
          stop_words
```

```
Out[159]: ['a',
           "a's",
           'able',
           'about',
           'above',
           'according',
           'accordingly',
           'across',
           'actually',
           'after',
           'afterwards',
           'again',
           'against',
           "ain't",
           'all',
           'allow',
           'allows',
           'almost',
           'alone',
           'along',
           'already',
           'also',
           'although',
           'always',
           'am',
           'among',
           'amongst',
           'an',
           'and',
'another',
'any',
'anybody',
'anyhow',
'anyone',
'anything',
'anyway',
'anyways',
'anywhere',
'apart',
'appear',
'appreciate',
'appropriate',
'are',
"aren't",
'around',
'as',
'aside',
'ask',
'asking',
'associated',
'at',
'available',
'away',
'awfully',
'b',
'be',
'became',
'because',
'become',
'becomes',
'becoming',
'been',
'before',
'beforehand',
'behind',
'being',
'believe',
'below',
'beside',
'besides',
'best',
'better',
'between',
'beyond',
'both',
'brief',
'but',
'by',
'c',
"c'mon",
"c's",
'came',
'can',
"can't",
'cannot',
'cant',
'cause',
'causes',
'certain',
'certainly',
'changes',
'clearly',
'co',
'com',
'come',
'comes',
'concerning',
'consequently',
'consider',
'considering',
'contain',
'containing',
'contains',
'corresponding',
'could',
"couldn't",
'course',
'currently',
'd',
'definitely',
'described',
'despite',
'did',
"didn't",
'different',
'do',
'does',
"doesn't",
'doing',
"don't",
'done',
'down',
'downwards',
'during',
'e',
'each',
'edu',
'eg',
'eight',
'either',
'else',
'elsewhere',
'enough',
'entirely',
'especially',
'et',
'etc',
'even',
'ever',
'every',
'everybody',
'everyone',
'everything',
'everywhere',
'ex',
'exactly',
'example',
'except',
'f',
'far',
'few',
'fifth',
'first',
'five',
'followed',
'following',
'follows',
'for',
'former',
'formerly',
'forth',
'four',
'from',
'further',
'furthermore',
'g',
'get',
'gets',
'getting',
'given',
'gives',
'go',
'goes',
'going',
'gone',
'got',
'gotten',
'greetings',
'h',
'had',
"hadn't",
'happens',
'hardly',
'has',
"hasn't",
'have',
"haven't",
'having',
'he',
"he's",
'hello',
'help',
'hence',
'her',
'here',
"here's",
'hereafter',
'hereby',
'herein',
'hereupon',
'hers',
'herself',
'hi',
'him',
'himself',
'his',
'hither',
'hopefully',
'how',
'howbeit',
'however',
'i',
"i'd",
"i'll",
"i'm",
"i've",
'ie',
'if',
'ignored',
'immediate',
'in',
'inasmuch',
'inc',
'indeed',
'indicate',
'indicated',
'indicates',
'inner',
'insofar',
'instead',
'into',
'inward',
'is',
"isn't",
'it',
"it'd",
"it'll",
"it's",
'its',
'itself',
'j',
'just',
'k',
'keep',
'keeps',
'kept',
'know',
'knows',
'known',
'l',
'last',
'lately',
'later',
'latter',
'latterly',
'least',
'less',
'lest',
'let',
"let's",
'like',
'liked',
'likely',
'little',
'look',
'looking',
'looks',
'ltd',
'm',
'mainly',
'many',
'may',
'maybe',
'me',
'mean',
'meanwhile',
'merely',
'might',
'more',
'moreover',
'most',
'mostly',
'much',
'must',
'my',
'myself',
'n',
'name',
'namely',
'nd',
'near',
'nearly',
'necessary',
'need',
'needs',
'neither',
'never',
'nevertheless',
'new',
'next',
'nine',
'no',
'nobody',
'non',
'none',
'noone',
'nor',
'normally',
'not',
'nothing',
'novel',
'now',
'nowhere',
'o',
'obviously',
'of',
'off',
'often',
'oh',
'ok',
'okay',
'old',
'on',
'once',
'one',
'ones',
'only',
'onto',
'or',
'other',
'others',
'otherwise',
'ought',
'our',
'ours',
'ourselves',
'out',
'outside',
'over',
'overall',
'own',
'p',
'particular',
'particularly',
'per',
'perhaps',
'placed',
'please',
'plus',
'possible',
'presumably',
'probably',
'provides',
'q',
'que',
'quite',
'qv',
'r',
'rather',
'rd',
're',
'really',
'reasonably',
'regarding',
'regardless',
'regards',
'relatively',
'respectively',
'right',
's',
'said',
'same',
'saw',
'say',
'saying',
'says',
'second',
'secondly',
'see',
'seeing',
'seem',
'seemed',
'seeming',
'seems',
'seen',
'self',
'selves',
'sensible',
'sent',
'serious',
'seriously',
'seven',
'several',
'shall',
'she',
'should',
"shouldn't",
'since',
'six',
'so',
'some',
'somebody',
'somehow',
'someone',
'something',
'sometime',
'sometimes',
'somewhat',
'somewhere',
'soon',
'sorry',
'specified',
'specify',
'specifying',
'still',
'sub',
'such',
'sup',
'sure',
't',
"t's",
'take',
'taken',
'tell',
'tends',
'th',
'than',
'thank',
'thanks',
'thanx',
'that',
"that's",
'thats',
'the',
'their',
'theirs',
'them',
'themselves',
'then',
'thence',
'there',
"there's",
'thereafter',
'thereby',
'therefore',
'therein',
'theres',
'thereupon',
'these',
'they',
"they'd",
"they'll",
"they're",
"they've",
'think',
'third',
'this',
'thorough',
'thoroughly',
'those',
'though',
'three',
'through',
'throughout',
'thru',
'thus',
'to',
'together',
'too',
'took',
'toward',
'towards',
'tried',
'tries',
'truly',
'try',
'trying',
'twice',
'two',
'u',
'un',
'under',
'unfortunately',
'unless',
'unlikely',
'until',
'unto',
'up',
'upon',
'us',
'use',
'used',
'useful',
'uses',
'using',
'usually',
'uucp',
'v',
'value',
'various',
'very',
'via',
'viz',
'vs',
'w',
'want',
'wants',
'was',
"wasn't",
    'way',
    'we',
    "we'd",
    "we'll",
    "we're",
    "we've",
    'welcome',
    'well',
    'went',
    'were',
    "weren't",
    'what',
    "what's",
    'whatever',
    'when',
    'whence',
    'whenever',
    'where',
    "where's",
    'whereafter',
    'whereas',
    'whereby',
    'wherein',
    'whereupon',
    'wherever',
    'whether',
    'which',
    'while',
    'whither',
    'who',
    "who's",
    'whoever',
    'whole',
    'whom',
    'whose',
    'why',
    'will',
    'willing',
    'wish',
    'with',
    'within',
    'without',
    "won't",
    'wonder',
    'would',
    'would',
    "wouldn't",
    'x',
    'y',
    'yes',
    'yet',
    'you',
    "you'd",
    "you'll",
    "you're",
    "you've",
    'your',
    'yours',
    'yourself',
    'yourselves',
    'z',
    'zero']
```

```
In [160]: def count_words(texts):
              """Vectorize text : return count of each word in the text snippets

              Parameters
              ----------
              texts : list of str
                  The texts

              Returns
              -------
              vocabulary : dict
                  A dictionary that points to an index in counts for each word.
              counts : ndarray, shape (n_samples, n_features)
                  The counts of each word in each text.
                  n_samples == number of documents.
                  n_features == number of words in vocabulary.
              """

              punctuation = list(string.punctuation)+['\n']+stop_words #We now remove the stop

              total_word_list = [] #will contain the words of all texts without separation
              text_word_list = [] #will separate the words of each text in a different list it

              for text in texts:
                  for punct in punctuation:
                      text = text.replace(punct,' ')

                  word_list = text.split(' ')
                  word_list = [x for x in word_list if x != ''] #We remove the empty strings :

                  total_word_list += word_list # We build vocabulary thanks to the list of all
                  text_word_list.append(word_list)
```

```
    words = list(set(total_word_list))
    vocabulary = {}

    for i in range(len(words)):
        vocabulary[words[i]] = i

    counts = np.zeros((2000,len(words))) #there are 2000 documents

    for i in range(len(text_word_list)):
        for j in range(len(text_word_list[i])):
            index = vocabulary[text_word_list[i][j]] #This is the index of the word
            counts[i][index] += 1

    return vocabulary, counts


count_words(texts)
```

```
Out[160]: ({'1935': 0,
            '1912': 1,
            '300': 2,
            '50000': 3,
            '1865': 4,
            '1980': 5,
            '44': 6,
            '209': 7,
            '122': 8,
            '1956': 9,
            '65': 10,
            '640': 11,
            '118': 12,
            '1986': 13,
            '\x13': 14,
            '1952': 15,
            '1400': 16,
            '230': 17,
            '20': 18,
            '\x12': 19,
            '1938': 20,
            '700': 21,
            '1942': 22,
            '75': 23,
            '1925': 24,
            '\x05\x05': 25,
            '67': 26,
            '1972': 27,
            '1871': 28,
'111': 29,
'983': 30,
'2259': 31,
'1960': 32,
'125': 33,
'2654': 34,
'30': 35,
'1932': 36,
'54': 37,
'747': 38,
'460': 39,
'1982': 40,
'1961': 41,
'357': 42,
'1862': 43,
'1800': 44,
'8216': 45,
'41': 46,
'19': 47,
'1984': 48,
'87': 49,
'175': 50,
'81': 51,
'49': 52,
'1957': 53,
'140': 54,
'1975': 55,
'1987': 56,
'3654': 57,
'2050': 58,
'1830': 59,
'1991': 60,
'1792': 61,
'007': 62,
'999': 63,
'63': 64,
'254': 65,
'5000': 66,
'1995': 67,
'1985': 68,
'1923': 69,
'105': 70,
'98': 71,
'1922': 72,
'143': 73,
'133': 74,
'126': 75,
'1965': 76,
'1138': 77,
'104': 78,
'2018': 79,
'79': 80,
'1583': 81,
'56': 82,
'802': 83,
'180': 84,
'4960': 85,
'135': 86,
'1999': 87,
'61': 88,
'115': 89,
'\x14': 90,
'138': 91,
'1989': 92,
'1983': 93,
'90210': 94,
'26': 95,
'69': 96,
'1967': 97,
'2001': 98,
'555': 99,
'161': 100,
'1298': 101,
'2000': 102,
'107': 103,
'426': 104,
'206': 105,
'750': 106,
'167': 107,
'1500': 108,
'1700': 109,
'62': 110,
'1992': 111,
'1962': 112,
'1928': 113,
'23': 114,
'1959': 115,
'777': 116,
'05425': 117,
'1839': 118,
'86': 119,
'911': 120,
'2036': 121,
'55': 122,
'2058': 123,
'59': 124,
'779': 125,
'102': 126,
'152': 127,
'1709': 128,
'00': 129,
'216': 130,
'1998': 131,
'1930': 132,
'1946': 133,
'70': 134,
'2007': 135,
'1791': 136,
'5': 137,
'1305': 138,
'1933': 139,
'83': 140,
'80': 141,
'1994': 142,
'1963': 143,
'710': 144,
'24': 145,
'666': 146,
'90': 147,
'14': 148,
'3411': 149,
'137': 150,
'3000': 151,
'8': 152,
'2040': 153,
'1847': 154,
'1692': 155,
'18': 156,
'3465': 157,
'1908': 158,
'170': 159,
'7': 160,
'1939': 161,
'900': 162,
'165': 163,
'2020': 164,
'1794': 165,
'5671': 166,
'100': 167,
'1869': 168,
'280': 169,
'57': 170,
'51': 171,
'25': 172,
'53': 173,
'2056': 174,
'85': 175,
'127': 176,
'1898': 177,
'600': 178,
'6': 179,
'29': 180,
'9': 181,
'04': 182,
'1979': 183,
'16': 184,
'1977': 185,
'1914': 186,
'2099': 187,
'132': 188,
'1966': 189,
'1988': 190,
'234': 191,
'113': 192,
'128': 193,
'189': 194,
'1990': 195,
'22': 196,
'114': 197,
'1916': 198,
'449': 199,
'89': 200,
'91': 201,
'1': 202,
'1978': 203,
'121': 204,
'11': 205,
'1934': 206,
'1969': 207,
'106': 208,
'10000': 209,
'250': 210,
'108': 211,
'7000': 212,
'2400': 213,
'172': 214,
'400': 215,
'1947': 216,
'109': 217,
'92': 218,
'144': 219,
'1951': 220,
'360': 221,
'1943': 222,
'139': 223,
'153': 224,
'157': 225,
'2024': 226,
'1773': 227,
'1997': 228,
'76': 229,
'32': 230,
'1974': 231,
'28': 232,
'58': 233,
'94': 234,
'1968': 235,
'13': 236,
'96': 237,
'1996': 238,
'1920': 239,
'200': 240,
'1971': 241,
'1993': 242,
'800': 243,
'112': 244,
'17': 245,
'1948': 246,
'21': 247,
'\x16': 248,
'130': 249,
'60': 250,
'48': 251,
'1976': 252,
'1964': 253,
'35': 254,
'1970': 255,
'39': 256,
'6000': 257,
'1955': 258,
'131': 259,
'2015': 260,
'03': 261,
'712': 262,
'2013': 263,
'1981': 264,
'2002': 265,
'939': 266,
'1919': 267,
'78': 268,
'2': 269,
'50': 270,
'1885': 271,
'12': 272,
'123': 273,
'37': 274,
'1793': 275,
'52': 276,
'150': 277,
'2176': 278,
'42': 279,
'151': 280,
'129': 281,
'1958': 282,
'34': 283,
'99': 284,
'40': 285,
'88': 286,
'43': 287,
'77': 288,
'1954': 289,
'66': 290,
'84': 291,
'701': 292,
'10': 293,
'2010': 294,
'1944': 295,
'4': 296,
'101': 297,
'2023': 298,
'45': 299,
'97': 300,
'1888': 301,
'2017': 302,
'1799': 303,
'1554': 304,
'73': 305,
'500': 306,
'1945': 307,
'0009': 308,
'2470': 309,
'1272': 310,
'117': 311,
'36': 312,
'\x05': 313,
'000': 314,
'1913': 315,
'33': 316,
 '0': 317,
 '38': 318,
 '05': 319,
 '1937': 320,
 '1926': 321,
 '1941': 322,
 '93': 323,
 '64': 324,
 '3': 325,
 '1950': 326,
 '103': 327,
 '47': 328,
 '1521': 329,
 '160': 330,
 '1896': 331,
 '95': 332,
 '1000': 333,
 '110': 334,
 '8034': 335,
 '2293': 336,
 '289': 337,
 '1973': 338,
 '27': 339,
 '1590': 340,
 '1953': 341,
 '68': 342,
 '1940': 343,
 '1812': 344,
 '1899': 345,
 '607': 346,
 '571': 347,
 '2029': 348,
 '15': 349,
 '1949': 350,
 '1900': 351,
 '82': 352,
 '1600': 353,
 '310': 354,
 '2065': 355,
 '155': 356,
 '31': 357,
 '1903': 358},
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.]]))
```

```
In [161]: from sklearn.model_selection import cross_val_score
          vocabulary, X = count_words(texts)
          nb = NB()

          print('The 5-fold cross-validation score WITHOUT the stop-words is : '+str(cross_val_
```

```
The 5-fold cross-validation score WITHOUT the stop-words is : 0.5405
```

Removing the stop-words worsens the performance.

7 Scikit-Learn Use

```
In [162]: from sklearn.naive_bayes import MultinomialNB
          from sklearn.feature_extraction.text import CountVectorizer
          from sklearn.pipeline import Pipeline
```

```
In [163]: vectorizer =  CountVectorizer()
          X = vectorizer.fit_transform(texts)
          nb = MultinomialNB()
          print('The 5-fold cross-validation score of the sklearn function WITHOUT the stop-wo
```

```
The 5-fold cross-validation score of the sklearn function WITHOUT the stop-words is : 0.8145
```

The default scikitlearn libraries yield a much better score than our hand-made estimator.

8 LinearSVC

```
In [164]: from sklearn.svm import LinearSVC
          clf = LinearSVC()
          clf.fit(X,y)
          print('The 5-fold cross-validation score of the Linear SVC WITHOUT the stop-words is
```

```
The 5-fold cross-validation score of the Linear SVC WITHOUT the stop-words is : 0.832500000000
```

LinearSVC yields a similar score to the Sklearn Bayesian approach. However, this score is slightly higher than our hand-made Bayesian estimator.

9 NLTK Stemming

```
In [165]: from nltk import SnowballStemmer
```

```
In [166]: stemmer = SnowballStemmer(language="english")
          stemmer.stem('congratulations')
```

```
Out[166]: 'congratul'
```

```
In [167]: def count_words(texts):
              """Vectorize text : return count of each word in the text snippets

              Parameters
              ----------
              texts : list of str
                  The texts

              Returns
              -------
              vocabulary : dict
                  A dictionary that points to an index in counts for each word.
              counts : ndarray, shape (n_samples, n_features)
                  The counts of each word in each text.
                  n_samples == number of documents.
                  n_features == number of words in vocabulary.
              """
              punctuation = list(string.punctuation)+['\n'] #We also want to remove the newlin

              total_word_list = [] #will contain the words of all texts without separation
              text_word_list = [] #will separate the words of each text in a different list it

              for text in texts:
                  for punct in punctuation:
                      text = text.replace(punct,' ')
                  word_list = text.split(' ')
                  word_list = [stemmer.stem(x) for x in word_list if x != ''] #We remove the e

                  total_word_list += word_list # We build vocabulary thanks to the list of all
                  text_word_list.append(word_list)

              words = list(set(total_word_list))
              vocabulary = {}

              for i in range(len(words)):
                  vocabulary[words[i]] = i

              counts = np.zeros((2000,len(words))) #there are 2000 documents

              for i in range(len(text_word_list)):
                  for j in range(len(text_word_list[i])):
                      index = vocabulary[text_word_list[i][j]] #This is the index of the word
                      counts[i][index] += 1

              return vocabulary, counts
```

```
count_words(texts)
```

```
Out[167]: ({'shadi': 0,
            'rachmaninov': 1,
            'reput': 2,
            'aftermath': 3,
            'unwieldi': 4,
            'rich': 5,
            '1865': 6,
            'orwellian': 7,
            'depriv': 8,
            'buffoonish': 9,
            'andi': 10,
            'mtcts1': 11,
            'hole\x14': 12,
            'osmond': 13,
            'rhapsodi': 14,
            'titshot': 15,
            'allegori': 16,
            'glu': 17,
            'remast': 18,
            'instantan': 19,
            'steadiocam': 20,
            'slobber': 21,
            'truant': 22,
            'wiper': 23,
            'zoe': 24,
            'barenboim': 25,
            'machinist': 26,
            'introduc': 27,
            'saigon': 28,
            'oaf': 29,
            'ferguson': 30,
            'flicker': 31,
            'roberts': 32,
            'inanim': 33,
            'motley': 34,
            'starphoenix': 35,
            'svenwara': 36,
            'mullal': 37,
            'extremel': 38,
            'shape': 39,
            'dreamworld': 40,
            'norad': 41,
            'slashfest': 42,
            'fullyload': 43,
            'sexpot': 44,
            'aboard': 45,
            'folk': 46,
            'dodg': 47,
            'lighthead': 48,
            'bartlebi': 49,
            'potboil': 50,
            'romulus': 51,
            'arbuthnot': 52,
            'movement': 53,
            'highpoint': 54,
            'attribut': 55,
            'litani': 56,
            'sling': 57,
            'luftwaff': 58,
            'unworthi': 59,
            'oasi': 60,
            'compil': 61,
            'collector': 62,
            'subtlti': 63,
            'revolucionario': 64,
            'counterfeit': 65,
            'biziou': 66,
            'milton': 67,
            'bookstor': 68,
            'washer': 69,
            'assert': 70,
            'underwritten': 71,
            'arsine': 72,
            'repent': 73,
            'ly': 74,
            'blackout': 75,
            'dummi': 76,
            'truer': 77,
            'redund': 78,
            'intercom': 79,
            'hurt': 80,
            'mamma': 81,
            'autobiograph': 82,
            'stuyves': 83,
            'danna': 84,
            'exotica': 85,
            'intrins': 86,
            'piet': 87,
            '180': 88,
            'dullsvill': 89,
            'mesquida': 90,
            'desert': 91,
            'aquamarin': 92,
            'goldenberg': 93,
            'szwarc': 94,
            'heigh': 95,
            'infinitum': 96,
            'nfl': 97,
            'artifact': 98,
            'sonatin': 99,
            'disadvantag': 100,
            'fiefdom': 101,
            'sv2': 102,
            'warbl': 103,
            'wasn': 104,
            'katt': 105,
            'conservationist': 106,
            'grandest': 107,
            'underw': 108,
            'unfulfil': 109,
            'abomin': 110,
            'cultish': 111,
            'steinberg': 112,
            'kerdelhu': 113,
            'wors': 114,
            'faison': 115,
            'escap': 116,
            'shabbi': 117,
            'fluoro': 118,
            'aciton': 119,
            'hatti': 120,
            'snowi': 121,
            'statesid': 122,
            'reciev': 123,
            'autumn': 124,
            'recless': 125,
            'tangent': 126,
            'jagger': 127,
            'boathous': 128,
            'fasten': 129,
            'mulholland': 130,
            'malloy': 131,
            'narcolepsi': 132,
            'timex': 133,
            'xander': 134,
            'sexiest': 135,
            'dublin': 136,
            'bounteous': 137,
            'firework': 138,
            '152': 139,
            'etymolog': 140,
            'apparel': 141,
            'shorter': 142,
            'bass': 143,
            'canran': 144,
            'tempest': 145,
            'ukrain': 146,
            'visibl': 147,
            'chanc': 148,
            'vinc': 149,
            'juke': 150,
            'incit': 151,
            'phenomen': 152,
            'rob': 153,
            'honey': 154,
            'await': 155,
            'ww2': 156,
            'alchohol': 157,
            'lamanna': 158,
            'klingon': 159,
            'dive': 160,
            'tooth': 161,
            'starbuck': 162,
            'darnel': 163,
            'deceiv': 164,
            'train': 165,
            'extravaganza': 166,
            'applianc': 167,
            '14': 168,
            'grassi': 169,
            'friedkin': 170,
            'motherfuck': 171,
            'saul': 172,
            'posess': 173,
            'deedl': 174,
            'comedienn': 175,
            'spacey': 176,
            'effect': 177,
            'hoist': 178,
            'sorvino': 179,
            'ice': 180,
            'specifi': 181,
            'buehler': 182,
            'hamster': 183,
            'notion': 184,
            'teeter': 185,
            'tailor': 186,
            'nurs': 187,
            'nitrit': 188,
            'plumber': 189,
            'travelogu': 190,
            'torment': 191,
            'remad': 192,
            'cuddi': 193,
            'loutish': 194,
            'glave': 195,
            'pubesc': 196,
            'councilmemb': 197,
            'resevoir': 198,
            'estella': 199,
            'missil': 200,
            'appris': 201,
            'mark': 202,
            'hughley': 203,
            'ripe': 204,
            'merpeopl': 205,
            'vampira': 206,
            'hellhol': 207,
            'nordoff': 208,
            'businesswomen': 209,
            'morquio': 210,
            'ing': 211,
            'mifun': 212,
            'fillmor': 213,
            'prestig': 214,
            'various': 215,
            'duveyri': 216,
            'trampl': 217,
            'obstetrician': 218,
            'constant': 219,
            'veterinari': 220,
            'maestro': 221,
            'cameraman': 222,
            'aphrodiasiat': 223,
            'electroshock': 224,
            'stead': 225,
            'valjean': 226,
            'coolest': 227,
            'acupunctur': 228,
            'o': 229,
            'comaprison': 230,
            'caberat': 231,
            'distracted': 232,
            'matrix': 233,
            'sandefur': 234,
            'masaya': 235,
            'refund': 236,
            'tendanc': 237,
            'shrewd': 238,
            'towner': 239,
            'advertis': 240,
            'augustin': 241,
            'unholi': 242,
            'both': 243,
            'treason': 244,
            'mufasa': 245,
            'thanksgiv': 246,
            'waylon': 247,
            'straight': 248,
            'outrun': 249,
            'sexless': 250,
            'vig': 251,
            'cannot': 252,
            'immediat': 253,
            'understat': 254,
            'dub': 255,
            'mistaken': 256,
            'horseplay': 257,
            'deflow': 258,
            'either': 259,
            'happi': 260,
            'knockout': 261,
            'detritus': 262,
            'bloodier': 263,
            'favorit': 264,
            'compatriot': 265,
            'bogg': 266,
            'lovesick': 267,
            'diabol': 268,
            'drape': 269,
            'zoologist': 270,
            'pina': 271,
            'motor': 272,
            'sporti': 273,
            'wabbit': 274,
            'as': 275,
            'actual': 276,
            'trickster': 277,
            'slightest': 278,
            'damnat': 279,
            'booti': 280,
            'schwalbach': 281,
            'pector': 282,
            'primarili': 283,
            'brennan': 284,
            'prinz': 285,
            'caddi': 286,
            'block': 287,
            'sheer': 288,
            'inquisit': 289,
            'chloe': 290,
            'stoke': 291,
            'crystal': 292,
            'stadium': 293,
            'pego': 294,
            'reindeer': 295,
            'hanif': 296,
            'allison': 297,
            'plunder': 298,
            '8a': 299,
            'sabotag': 300,
            'survivor': 301,
            'centenni': 302,
            'unimport': 303,
            'rider': 304,
            'longbaugh': 305,
            'donovan': 306,
            'midwestern': 307,
            'plow': 308,
            'stipul': 309,
            'grubbi': 310,
            'tick': 311,
            'computech': 312,
            'identif': 313,
            'macho': 314,
            'lustr': 315,
            'heali': 316,
            'wittliff': 317,
            'council': 318,
            'constrict': 319,
            'rapier': 320,
            'charlott': 321,
            'bud': 322,
            'receipt': 323,
            'frankiln': 324,
            'clientel': 325,
            'pornographi': 326,
            'unfilm': 327,
            'woolli': 328,
            'balthazar': 329,
            'casket': 330,
            'masterson': 331,
            'vicent': 332,
            'oliv': 333,
            'geopolit': 334,
            'cancerogen': 335,
            'firesid': 336,
            'francoi': 337,
            'snorri': 338,
            'ethnic': 339,
            'rosmari': 340,
            'harumph': 341,
            'misstep': 342,
            'unveil': 343,
            'theme': 344,
            'expatri': 345,
            'poolboy': 346,
            'shagwel': 347,
            'vampyr': 348,
            'renshaw': 349,
            'filmmka': 350,
            'hech': 351,
            'broncobust': 352,
            'styleless': 353,
            'portray': 354,
            'ie': 355,
            'bruskott': 356,
            'cessat': 357,
            'summon': 358,
            'vonnegut': 359,
            'weddel': 360,
            'jackson': 361,
            'pressur': 362,
            'slicker': 363,
            'cuarsn': 364,
            'substant': 365,
            'pastri': 366,
            'gravedigg': 367,
            'home': 368,
            'ww': 369,
            'graci': 370,
            'cds': 371,
            'collat': 372,
            'stargher': 373,
            'serpico': 374,
            'soavi': 375,
            'percol': 376,
            'crunchem': 377,
            'stylewis': 378,
            'mcjob': 379,
            'veini': 380,
            'tek': 381,
            'unrel': 382,
            'awk': 383,
            'fulli': 384,
            'paulina': 385,
            'obsequi': 386,
            'gaiti': 387,
            'klieg': 388,
            'race': 389,
            'totat': 390,
            'heenan': 391,
            'monogami': 392,
            'crust': 393,
            'moi': 394,
            'spice': 395,
            'curious': 396,
            'comrad': 397,
            'uncar': 398,
            'uganda': 399,
            'sloth': 400,
            'weiss': 401,
            'blofeld': 402,
            'hassid': 403,
            'pseudo': 404,
            'placard': 405,
            'courtesi': 406,
            'unassoci': 407,
            'agoni': 408,
            'obtain': 409,
            'chimpanze': 410,
            'tycoon': 411,
            'dissuas': 412,
            'wo': 413,
            'caleb': 414,
            'racism': 415,
            'derid': 416,
            'newton': 417,
            'wallenberg': 418,
            'watt': 419,
            'screen': 420,
            'premonit': 421,
            'glorif': 422,
            'lanc': 423,
            'arctic': 424,
            'schandl': 425,
            'gos': 426,
            'potenc': 427,
            'cousin': 428,
            'punk': 429,
            'gellar': 430,
            'lite': 431,
            'bosom': 432,
            'badger': 433,
            'servo': 434,
            'proval': 435,
            'ryuichi': 436,
            'inextric': 437,
            'janssen': 438,
            'conqueror': 439,
            'shipment': 440,
            'stefanson': 441,
            'headach': 442,
            'moonlit': 443,
            'preteen': 444,
            'diver': 445,
            'cream': 446,
            'mozambiqu': 447,
            'scalvag': 448,
            'winnebago': 449,
            'woodwind': 450,
            'wheat': 451,
            'wouldn': 452,
            'mise': 453,
            'cloy': 454,
            'hauff': 455,
            'gawd': 456,
            'seldon': 457,
            'zone': 458,
            'disgrac': 459,
            'vehement': 460,
            'shopkeep': 461,
            'viabil': 462,
            'sourc': 463,
            'sicken': 464,
            'parfitt': 465,
            'disappear': 466,
            'tempestu': 467,
            'articul': 468,
            'clooney': 469,
            'protgaonist': 470,
            'necessarili': 471,
            'brett': 472,
            'approv': 473,
            'maxx': 474,
            'bochner': 475,
            'abysm': 476,
            '1st': 477,
            'match': 478,
            'inan': 479,
            'knowl': 480,
            'spungen': 481,
            'nearbi': 482,
            'mostel': 483,
            'elmaloglou': 484,
            'cherot': 485,
            'tighten': 486,
            'disloy': 487,
            'bump': 488,
            'morissey': 489,
            'realis': 490,
            'marienbad': 491,
            'duper': 492,
            'fiercer': 493,
            'cheesefest': 494,
            'hokum': 495,
            'heartwarm': 496,
            'waterboy': 497,
            'adulteri': 498,
            'cheeki': 499,
            'outpour': 500,
            '555': 501,
            'raw': 502,
            'abel': 503,
            'jeez': 504,
            'isn\x12t': 505,
            'hooch': 506,
            'streem': 507,
            'organis': 508,
            'minimum': 509,
            'problem': 510,
            'acquit': 511,
            'benoit': 512,
            'moland': 513,
            'cohn': 514,
            'barcalow': 515,
            'panick': 516,
            'squirt': 517,
            'slick': 518,
            'spend': 519,
            'mckee': 520,
            'spiritless': 521,
            'miscalcul': 522,
            'narrat': 523,
            'taught': 524,
            'ratcliff': 525,
            'overpopul': 526,
            'stoppabl': 527,
            'seann': 528,
            'whereupon': 529,
            'spock': 530,
            'gwynn': 531,
            'workabl': 532,
            'josian': 533,
            'deschanel': 534,
            'refineri': 535,
            'mikel': 536,
            'huggabl': 537,
            'instrument': 538,
            '1930': 539,
            'damian': 540,
            'clune': 541,
            'unload': 542,
            'preposter': 543,
            'chromium': 544,
            'bustl': 545,
            'stalwart': 546,
            'bogeyman': 547,
            'winteri': 548,
            'baroqu': 549,
            'headdress': 550,
            'rate': 551,
            'derail': 552,
            'rebuff': 553,
            'gutierrez': 554,
            'boni': 555,
            'salabl': 556,
            'attic': 557,
            'footag': 558,
            'grumpier': 559,
            'penney': 560,
            'figurin': 561,
            'miriam': 562,
            'guenveur': 563,
            'ambianc': 564,
            'sensual': 565,
            'irrefut': 566,
            'nunez': 567,
            'animatron': 568,
            'manti': 569,
            'gauntlet': 570,
            'unseason': 571,
            'fasano': 572,
            'paid': 573,
            'tate': 574,
            'within': 575,
            'societi': 576,
            'peg': 577,
            'regrett': 578,
            'abort': 579,
            'edt': 580,
            'schoolchildren': 581,
            'sk': 582,
            'dystop': 583,
            'unimpress': 584,
            'longstand': 585,
            'hallmark': 586,
            'explant': 587,
            'crud': 588,
            'pentamet': 589,
            'extraordinarili': 590,
            'beaver': 591,
            'obliqu': 592,
            'kang': 593,
            'biographi': 594,
            'planet': 595,
            'woefulli': 596,
            'nonsens': 597,
            'cabin': 598,
            'strikeout': 599,
            'repar': 600,
            'inter': 601,
            'funk': 602,
            'unrestrain': 603,
            'claw': 604,
            'schoolmat': 605,
            'petter': 606,
            'fratern': 607,
            'fisburn': 608,
            'trial': 609,
            'uninterest': 610,
            'margiull': 611,
            'mgm': 612,
            'butt': 613,
            'overdub': 614,
            'innov': 615,
            'wispi': 616,
            'presenc': 617,
            'psitiv': 618,
            'privileg': 619,
            'amercian': 620,
            'contagi': 621,
            'beaumont': 622,
            'cryogen': 623,
            'chasm': 624,
            'drawer': 625,
            'corvino': 626,
            'amidst': 627,
            'slingblad': 628,
            'sfx': 629,
            'pitchfork': 630,
            'lavern': 631,
            'humphri': 632,
            'lineman': 633,
            'roam': 634,
            'rangoon': 635,
            'heffron': 636,
            'brave': 637,
            'finklestein': 638,
            'overstat': 639,
            'kivilo': 640,
            'teaser': 641,
            'rank': 642,
            'deconstruct': 643,
            'backfir': 644,
            'naifeh': 645,
            'iran': 646,
            'shorelin': 647,
            'lundgren': 648,
            'dornack': 649,
            'bujold': 650,
            'puni': 651,
            'conneri': 652,
            'birch': 653,
            'addi': 654,
            'greek': 655,
            'bran': 656,
            'laptop': 657,
            'digg': 658,
            'trite': 659,
            'traver': 660,
            'ketchup': 661,
            'raeeyain': 662,
            'aug': 663,
            'sluttish': 664,
            'brundag': 665,
            'reiner': 666,
            'petaluma': 667,
            'ab': 668,
            'courtoom': 669,
            'counterbal': 670,
            'ninjaman': 671,
            'harra': 672,
            'toneless': 673,
            'gwythant': 674,
            'bombay': 675,
            'smirk': 676,
            'excess': 677,
            'beesley': 678,
            'knob': 679,
            'anjelica': 680,
            'featur': 681,
            'goodi': 682,
            'gossam': 683,
            'monarchi': 684,
            'uniti': 685,
            'triniti': 686,
            'rooftop': 687,
            'emblem': 688,
            'tomlin': 689,
            'juror': 690,
            'heebi': 691,
            'profan': 692,
            'amalgam': 693,
            'subtl': 694,
            'trough': 695,
            'yesterday': 696,
            'courteney': 697,
            'scribbl': 698,
            'erupt': 699,
            'conniv': 700,
            'misguid': 701,
            'splish': 702,
            'concordia': 703,
            'fakeri': 704,
            'pullam': 705,
            'forti': 706,
            'circumst': 707,
            'uncharacterist': 708,
            'hum': 709,
            'outdoor': 710,
            'boil': 711,
            'dialoug': 712,
            'gloopett': 713,
            'stupend': 714,
            'enraptur': 715,
            'subsidi': 716,
            'valor': 717,
            'groupi': 718,
            'niro': 719,
            'conscienc': 720,
            'ineffectu': 721,
            'skateboard': 722,
            'mayhew': 723,
            'lawford': 724,
            'brimley': 725,
            'antionio': 726,
            'gunsling': 727,
            'javert': 728,
            'trey': 729,
            'atmostpher': 730,
            'livingston': 731,
            'overlight': 732,
            'amour': 733,
            'parallax': 734,
            'naomi': 735,
            'scalpel': 736,
            'turk': 737,
            'cocoon': 738,
            'jester': 739,
            'sh': 740,
            'lorenzo': 741,
            'foam': 742,
            'gutsi': 743,
            'aoki': 744,
            'utah': 745,
            'alejandro': 746,
            'ambigu': 747,
            'ashman': 748,
            'sherlockian': 749,
            'characterless': 750,
            'huddl': 751,
            'simmon': 752,
            'gyrat': 753,
            'inport': 754,
            'crass': 755,
            'gleba': 756,
            'fledgl': 757,
            'woburn': 758,
            'overwritten': 759,
            'borderlin': 760,
            'lemmi': 761,
            'skater': 762,
            'abber': 763,
            'factor': 764,
            'seasid': 765,
            'rosalba': 766,
            'pentecost': 767,
            'bundl': 768,
            'bravado': 769,
            'tay': 770,
            'guild': 771,
            'homosexu': 772,
            'proxima': 773,
            'session': 774,
            'inher': 775,
            'petey': 776,
            'estefan': 777,
            'bumbl': 778,
            'saddl': 779,
            'marrow': 780,
            'cellar': 781,
            'amateur': 782,
            'note': 783,
            'stalingrad': 784,
            'sheikh': 785,
            'booz': 786,
            'skanki': 787,
            'szubanski': 788,
            'postmodern': 789,
            'salvati': 790,
            'yo': 791,
            'flutter': 792,
            'trauma': 793,
            'timothi': 794,
            'stucker': 795,
            'rink': 796,
            'laureat': 797,
            'downer': 798,
            '44': 799,
            'boston': 800,
            'reprehens': 801,
            'marlon': 802,
            'exagger': 803,
            'amerocentr': 804,
            'strode': 805,
            'enchant': 806,
            'ur': 807,
            'breck': 808,
            'forb': 809,
            'grievanc': 810,
            'phylida': 811,
            'pastier': 812,
            'slide': 813,
            'vacuum': 814,
            'ringwood': 815,
            'vacant': 816,
            'plotlin': 817,
            'digimon': 818,
            'hardcov': 819,
            'breezi': 820,
            'griffith': 821,
            'elvira': 822,
            'batman': 823,
            'eventu': 824,
            'galosh': 825,
            'scalper': 826,
            'exhilar': 827,
            'unfinish': 828,
            'pluss': 829,
            'leprechaun': 830,
            'cargo': 831,
            'portho': 832,
            'lrighti': 833,
            'sicili': 834,
            'torpedo': 835,
            'clanci': 836,
            'askew': 837,
            'dewi': 838,
            'solondz': 839,
            'unsatisfi': 840,
            'sensationalizt': 841,
            'cash': 842,
            'edgefield': 843,
            'sputter': 844,
            'taxi': 845,
            'aloof': 846,
            'meant': 847,
            'unarm': 848,
            'concious': 849,
            'lugosi': 850,
            'ari': 851,
            'orvill': 852,
            'morrow': 853,
            'harmon': 854,
            'whenc': 855,
            'nativ': 856,
            'motherf': 857,
            'ledg': 858,
            'rimini': 859,
            'cuisin': 860,
            'directori': 861,
            'pepe': 862,
            'leder': 863,
            'ungain': 864,
            'horrifi': 865,
            'suit': 866,
            'magnuson': 867,
            'macneil': 868,
            'growl': 869,
            'overtur': 870,
            'cheryl': 871,
            'eileen': 872,
            'tomoto': 873,
            'terminolog': 874,
            'discothequ': 875,
            'zipper': 876,
            'viterelli': 877,
            'pas': 878,
            'packer': 879,
            'contrapt': 880,
            'asap': 881,
            'perch': 882,
            'frog': 883,
            'titus': 884,
            'logan': 885,
            'misti': 886,
            'cyborg': 887,
            'terrorist': 888,
            'mastabatori': 889,
            'snitch': 890,
            'def': 891,
            'plaudit': 892,
            'gangland': 893,
            'bungler': 894,
            'adolesc': 895,
            'irrevoc': 896,
            'urbano': 897,
            'gallifrey': 898,
            'surplus': 899,
            'shock': 900,
            'thieveri': 901,
            'asthmat': 902,
            'montreal': 903,
            'medfield': 904,
            'scroung': 905,
            'ride': 906,
            'zack': 907,
            'harelik': 908,
            'bicep': 909,
            'satirist': 910,
            'ivori': 911,
            'khanijian': 912,
            'sleazebal': 913,
            'grumpiest': 914,
            'crosbi': 915,
            'intercours': 916,
            'casserol': 917,
            'enthusiasm': 918,
            'parillaud': 919,
            'proposit': 920,
            'aveng': 921,
            '34th': 922,
            'unforc': 923,
            'alani': 924,
            'crucifi': 925,
            'overdirect': 926,
            'heist': 927,
            'spotless': 928,
            'dolenz': 929,
            'parodyish': 930,
            'whodunit': 931,
            'dinsdal': 932,
            'famish': 933,
            'biz': 934,
            'stodgi': 935,
            'proxi': 936,
            'peckinpah': 937,
            '3411': 938,
            'smack': 939,
            'persuad': 940,
            'pipe': 941,
            'subsum': 942,
            'pritchett': 943,
            'walkway': 944,
            'totoro': 945,
            'unattain': 946,
            'hi': 947,
            'chenoa': 948,
            'diner': 949,
            'entrepeneur': 950,
            'snap': 951,
            'finest': 952,
            'yogi': 953,
            'hamburg': 954,
            'fenster': 955,
            'workshop': 956,
            'volcano': 957,
            'foregin': 958,
            'emo': 959,
            'phobia': 960,
            'fate': 961,
            'phobic': 962,
            'rama': 963,
            'lyonn': 964,
            'hangov': 965,
            'coe': 966,
            'tie': 967,
            'loudmouth': 968,
            'katanga': 969,
            'dani': 970,
            'regail': 971,
            'astrid': 972,
            'ali': 973,
            'deform': 974,
            'angelina': 975,
            'cronenbergto': 976,
            'feder': 977,
            'mapplethorp': 978,
            'bugg': 979,
            'croni': 980,
            'squirrel': 981,
            'guidelin': 982,
            '132': 983,
            'nuclear': 984,
            'natti': 985,
            'hypnotist': 986,
            'do': 987,
            'dinosaur': 988,
            'commend': 989,
            'bystand': 990,
            'milo': 991,
            'vini': 992,
            'firefight': 993,
            'pine': 994,
            'melang': 995,
            'bellami': 996,
            'advers': 997,
            'theoret': 998,
            'tawdri': 999,
            ...},
           array([[0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  ...,
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.]]))
```

```
In [168]: vocabulary, X = count_words(texts)

          # Try to fit, predict and score
          nb = NB()
          print('The 5-fold cross-validation score of our hand-made estimator after stemming is
```

```
The 5-fold cross-validation score of our hand-made estimator after stemming is : 0.82100000000
```

Conclusion :

We observe that both the stemmed and non-stemming lead to comparable results : ~0.82 on a 5-fold cross-validation. This is extremely interesting, as it means that we analyze a much smaller text dataset.

The stemmed text dataset contains only partial words, and all words which have the same root will become identical. Our set of words is much smaller as it only contains those very roots.

Stemming amounts to compressing the information, meaning we can process large amounts of text in a shorter amount of time, with comparable results.

10 Part of Speech

```
In [169]: import nltk
          nltk.download('punkt')
          nltk.download('averaged_perceptron_tagger')
          nltk.download('universal_tagset')
```

```
[nltk_data] Downloading package punkt to C:\Users\Max
[nltk_data]     Tchibozo\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\Max Tchibozo\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package universal_tagset to C:\Users\Max
[nltk_data]     Tchibozo\AppData\Roaming\nltk_data...
[nltk_data]   Package universal_tagset is already up-to-date!
```

```
Out[169]: True
```

```
In [170]: from nltk import pos_tag, word_tokenize
```

```
In [171]: pos_tag(word_tokenize("John's big idea isn't all that bad very strong. To eating, to
Out[171]: [('John', 'NOUN'),
           ("'s", 'PRT'),
           ('big', 'ADJ'),
           ('idea', 'NOUN'),
           ('is', 'VERB'),
           ("n't", 'ADV'),
           ('all', 'DET'),
           ('that', 'ADP'),
           ('bad', 'ADJ'),
           ('very', 'ADV'),
           ('strong', 'ADJ'),
           ('.', '.'),
           ('To', 'PRT'),
           ('eating', 'VERB'),
           (',', '.'),
           ('to', 'PRT'),
           ('be', 'VERB'),
           ('or', 'CONJ'),
           ('not', 'ADV'),
           ('to', 'PRT'),
           ('be', 'VERB')]
```

```
In [172]: accepted_words = ['NOUN','VERB','ADV','ADJ']
```

```
In [173]: def count_words(texts):
              """Vectorize text : return count of each word in the text snippets

              Parameters
              ----------
              texts : list of str
                  The texts

              Returns
              -------
              vocabulary : dict
                  A dictionary that points to an index in counts for each word.
              counts : ndarray, shape (n_samples, n_features)
                  The counts of each word in each text.
                  n_samples == number of documents.
                  n_features == number of words in vocabulary.
              """
              total_word_list = [] #will contain the words of all texts without separation
              text_word_list = [] #will separate the words of each text in a different list it

              for text in texts:
                  sentence = pos_tag(word_tokenize(text),tagset='universal')
                  word_list = [x[0] for x in sentence if x[1] in accepted_words] #We retain on
                  total_word_list += word_list # We build vocabulary thanks to the list of all
```

```
            text_word_list.append(word_list)

        words = list(set(total_word_list))
        vocabulary = {}

        for i in range(len(words)):
            vocabulary[words[i]] = i

        counts = np.zeros((2000,len(words))) #there are 2000 documents

        for i in range(len(text_word_list)):
            for j in range(len(text_word_list[i])):
                index = vocabulary[text_word_list[i][j]] #This is the index of the word
                counts[i][index] += 1

        return vocabulary, counts


    count_words(texts)
```

```
Out[173]: ({"o'barr": 0,
            'graded': 1,
            'anti-depressant': 2,
            'sayles': 3,
            'stubby': 4,
            'tag-line': 5,
            'wholesome': 6,
            'rachmaninov': 7,
            'aftermath': 8,
            'rubs': 9,
            'mnemonic': 10,
            'directional': 11,
            'jewelry-sporting': 12,
            'rich': 13,
            'alarmed': 14,
            'inversion': 15,
            'straight-out': 16,
            'shards': 17,
            'gosnell': 18,
            'sub-inspired': 19,
            'orwellian': 20,
            'buffoonish': 21,
            'jack/rose': 22,
            'hailed': 23,
            'motions': 24,
            'mtcts1': 25,
            'darkness_': 26,
            'greenhouse': 27,
            'hole\x14': 28,
            'osmond': 29,
            'phenomenas': 30,
            'planets': 31,
            'lingered': 32,
            'standards': 33,
            'chinese-american': 34,
            'treetops': 35,
            'out-of-body': 36,
            'steadiocam': 37,
            'excavating': 38,
            'weaponesque': 39,
            'imaginary': 40,
            'truant': 41,
            'wiper': 42,
            'zoe': 43,
            'barenboim': 44,
            'cromwell': 45,
            'residential': 46,
            'non-fans': 47,
            'machinist': 48,
            'impregnating': 49,
            'precedes': 50,
            'saigon': 51,
            'oaf': 52,
            'ferguson': 53,
            'roberts': 54,
            'flicker': 55,
            'treacherous': 56,
            'motley': 57,
            'starphoenix': 58,
            'sails': 59,
            'dejection': 60,
            'subsidies': 61,
            'straight-shooting': 62,
            'svenwara': 63,
            'ferrell': 64,
            'squashed': 65,
            'luggage': 66,
            'nation-wide': 67,
            'burke': 68,
            'sittings': 69,
            "'thing": 70,
            'sued': 71,
            'touching': 72,
            'extremel': 73,
            'shape': 74,
            'recourse': 75,
            'projectioner': 76,
            'partnerships': 77,
            'dreamworld': 78,
            'norad': 79,
            'slashfest': 80,
            'expressionists': 81,
            'sexpot': 82,
            'one-character-decides-not-to-return': 83,
            'anti-woman': 84,
            'aboard': 85,
            'benigni+s': 86,
            'indulging': 87,
            'folk': 88,
            'innocuous-enough-on-the-surface': 89,
            'steele': 90,
            'abbe': 91,
            'unselfconsciously': 92,
            'tape-to-film': 93,
            'hellhole': 94,
            'judge': 95,
            'meditative': 96,
            'whorehouses': 97,
            'century-fox': 98,
            'overtly': 99,
            'self-discoveries': 100,
            'debauchery': 101,
            'romulus': 102,
            'arbuthnot': 103,
            'dislikable': 104,
            'movement': 105,
            'freedoms': 106,
            'highpoint': 107,
            'sludge': 108,
            'hips': 109,
            'lambs_': 110,
            'child-hating': 111,
            'many': 112,
            'bio-lab': 113,
            'sling': 114,
            'heart-strings': 115,
            'pre-teens': 116,
            'wraps': 117,
            'collector': 118,
            'revolucionario': 119,
            'adventure': 120,
            'lechery': 121,
            'counterfeit': 122,
            'biziou': 123,
            'milton': 124,
            'puritanical': 125,
            'over-zealous': 126,
            'assert': 127,
            'swamps': 128,
            'underwritten': 129,
            'repent': 130,
            'ly': 131,
            'blackout': 132,
            'truer': 133,
            'hardy-har-har': 134,
            'playfulness': 135,
            'transcending': 136,
            'intercom': 137,
            'parks': 138,
            'belting': 139,
            'messy': 140,
            'hurt': 141,
            'sleuthing': 142,
            'mamma': 143,
            'dignities': 144,
            'danna': 145,
            'big-guy': 146,
            'flatmate': 147,
            'exotica': 148,
            'tinges': 149,
            'subconscious': 150,
            'piet': 151,
            'banished': 152,
            'define': 153,
            'discouraging': 154,
            'sequences': 155,
            'gods': 156,
            'affirmation': 157,
            'retardation': 158,
            'beast-people': 159,
            'mesquida': 160,
            'desert': 161,
            'goldenberg': 162,
            'szwarc': 163,
            'heigh': 164,
            'dominated': 165,
            'infinitum': 166,
            'confusing': 167,
            'geography': 168,
            'nfl': 169,
            'cowering': 170,
            'artifact': 171,
            'substantially': 172,
            'sv2': 173,
            'graduates': 174,
            'katt': 175,
            'conservationist': 176,
            'grandest': 177,
            'anti-fascist': 178,
            'one_': 179,
            'co-owner': 180,
            'tableware': 181,
            '_little': 182,
            'cultish': 183,
            'bongos': 184,
            'negotiable': 185,
            'steinberg': 186,
            'disobeys': 187,
            'self-propelled': 188,
            '30-minute': 189,
            'invulnerability': 190,
            'countryside': 191,
            'vah-vah-voom': 192,
            'full-of-tension': 193,
            'six-million': 194,
            'faison': 195,
            'fluoro': 196,
            'entangled': 197,
            'unopposed': 198,
            'messinger': 199,
            'super-senses': 200,
            'psychological-romance-thriller': 201,
            'incestuous': 202,
            'nominal': 203,
            'bumblingly': 204,
            'aciton': 205,
            'irrepressible': 206,
            'decide': 207,
            'autumn': 208,
            'resource': 209,
            'eulogy': 210,
            'champagne': 211,
            'tangent': 212,
            'jagger': 213,
            'songwriting': 214,
            'footnotes': 215,
            'cyberkillers': 216,
            'winnie': 217,
            'ventricle': 218,
            'mulholland': 219,
            'malloy': 220,
            'linearity': 221,
            'co-ordinator': 222,
            'xander': 223,
            'timex': 224,
            'sexiest': 225,
            'dublin': 226,
            'bounteous': 227,
            'obligated': 228,
            'electronically': 229,
            'crony': 230,
            'veritably': 231,
            'hand-to-hand': 232,
            'boy-next-door': 233,
            'multidimensionality': 234,
            'apparel': 235,
            'fight-scenes': 236,
            'criticisms': 237,
            'shorter': 238,
            'cries': 239,
            'dictionary': 240,
            'landscapes': 241,
            'bass': 242,
            'less-than-original': 243,
            'tempest': 244,
            'admission': 245,
            'overplayed': 246,
            'unbridled': 247,
            'etymology': 248,
            'parked': 249,
            'abs': 250,
            'ant-eaters': 251,
            'layered': 252,
            'boneheaded': 253,
            'juke': 254,
            'rob': 255,
            'pseudo-intellectual': 256,
            'honey': 257,
            'await': 258,
            'ww2': 259,
            'typically': 260,
            'observers': 261,
            'ramses': 262,
            'semi-brainless': 263,
            'unfunny': 264,
            'poking': 265,
            'angered': 266,
            'lamanna': 267,
            'klingon': 268,
            'dive': 269,
            'tooth': 270,
            'donate': 271,
            'starbuck': 272,
            'humor-free': 273,
            'grays': 274,
            'biographies': 275,
            'democratic': 276,
            'go-ahead': 277,
            'revolting': 278,
            'low-powered': 279,
            'train': 280,
            'extravaganza': 281,
            'astoundingly': 282,
            'thawed': 283,
            'self-possessed': 284,
            'friedkin': 285,
            'spin-off': 286,
            'personality-impaired': 287,
            'saul': 288,
            'posess': 289,
            'spacey': 290,
            'obscurity': 291,
            'effect': 292,
            'droves': 293,
            'hoist': 294,
            'bare-knuckle': 295,
            'sorvino': 296,
            'effects': 297,
            'ice': 298,
            'ghost-like': 299,
            'jungle2jungle': 300,
            'commodities': 301,
            'high-tech': 302,
            'montages': 303,
            'feeble': 304,
            'limestone': 305,
            'buehler': 306,
            'preposterousness': 307,
            'guitars': 308,
            'memories-': 309,
            'melee': 310,
            'androgony': 311,
            'twenty-one': 312,
            'borders': 313,
            'goatee': 314,
            'hamster': 315,
            'notion': 316,
            'kowtowing': 317,
            'policed': 318,
            '18-plus': 319,
            'minute-and': 320,
            'teeter': 321,
            'tailor': 322,
            'floods': 323,
            'obtained': 324,
            'plumber': 325,
            'torment': 326,
            'class-conscious': 327,
            'stonily': 328,
            'drowned': 329,
            'gears': 330,
            'stormare': 331,
            'decks': 332,
            'arrive': 333,
            'loutish': 334,
            'glave': 335,
            'petitioned': 336,
            'glorify': 337,
            'salability': 338,
            'southerners': 339,
            'resevoir': 340,
            'estella': 341,
            'mark': 342,
            'hughley': 343,
            'ripe': 344,
            'go-around': 345,
            'and/or': 346,
            'media-saturated': 347,
            'coenesque': 348,
            '_hustler_': 349,
            'vampira': 350,
            'madeliene': 351,
            'businesswomen': 352,
            'morquio': 353,
            'ing': 354,
            'adaptions': 355,
            'crooked': 356,
            'crams': 357,
            'discontented': 358,
            'desparate': 359,
            'various': 360,
            'obstetrician': 361,
            'dianne': 362,
            'constant': 363,
            'responsibility': 364,
            'maestro': 365,
            'camera-work': 366,
            'leavins': 367,
            'improvise': 368,
            'cameraman': 369,
            'electroshock': 370,
            'stead': 371,
            'valjean': 372,
            'coolest': 373,
            'ignore': 374,
            'names': 375,
            'sleaze': 376,
            'pronounces': 377,
            'picturing': 378,
            'o': 379,
            'comaprison': 380,
            'part-comedy': 381,
            'conked': 382,
            'labute': 383,
            'caberat': 384,
            'kidnap/ransom': 385,
            'ex-military': 386,
            'distracted': 387,
            'mandatory': 388,
            'matrix': 389,
            'chainsmokes': 390,
            'sandefur': 391,
            'refund': 392,
            'movie/road': 393,
            'thermopolis': 394,
            'shrewd': 395,
            'timelines': 396,
            'postmodernism': 397,
            'scene-and': 398,
            'mis-': 399,
            '_roxbury_': 400,
            'hard-fought': 401,
            'self-vanity': 402,
            'vexatiousness': 403,
            'assasination': 404,
            'faced': 405,
            'contents': 406,
            'mustered': 407,
            'nominees': 408,
            'berle': 409,
            'deemphasize': 410,
            'rosy': 411,
            'new-kid-on-the-block': 412,
            'treason': 413,
            'certainly': 414,
            'mufasa': 415,
            'male/female': 416,
            'toupee-sporting': 417,
            'beasts': 418,
            'subdued': 419,
            'sims': 420,
            'shifty-eyed': 421,
            'projected': 422,
            'transportation': 423,
            'straight': 424,
            'counts': 425,
            'waylon': 426,
            'outrun': 427,
            'pratfalling': 428,
            'colored': 429,
            'sexless': 430,
            "doens't": 431,
            'vig': 432,
            'trampoline': 433,
            'robs': 434,
            'thing__about': 435,
            'proclaimed': 436,
            'dub': 437,
            'mistaken': 438,
            'dewy-eyed': 439,
            'three-time': 440,
            'horseplay': 441,
            'either': 442,
            "'normal": 443,
            'strong-worded': 444,
            'dispose': 445,
            'knockout': 446,
            'outlands': 447,
            'closed-in': 448,
            'detritus': 449,
            'bloodier': 450,
            'jackee': 451,
            'hypocritical': 452,
            'face-whippings': 453,
            'limps': 454,
            'drumroll': 455,
            'bogg': 456,
            'squarely': 457,
            'tell-all/show-all': 458,
            'facing': 459,
            'lovesick': 460,
            'appended': 461,
            'boone': 462,
            'stepsister': 463,
            'intermingle': 464,
            'mediating': 465,
            'zoologist': 466,
            'pina': 467,
            'self-assuredness': 468,
            'motor': 469,
            'bread-and-butter': 470,
            'observance': 471,
            'wabbit': 472,
            'allegedly': 473,
            'as': 474,
            'actual': 475,
            'briers': 476,
            'trickster': 477,
            'slightest': 478,
            'delegates': 479,
            'eviction': 480,
            'pitted': 481,
            'schwalbach': 482,
            'winces': 483,
            'best-looking': 484,
            'anecdotes': 485,
            'measly': 486,
            'brennan': 487,
            'lorry': 488,
            'seventy-eight': 489,
            'mouthed': 490,
            'block': 491,
            'stored': 492,
            'deconstruction': 493,
            'gramercy': 494,
            'sheer': 495,
            'angers': 496,
            'stimulates': 497,
            'gulliver': 498,
            'fatigues': 499,
            'chloe': 500,
            'crystal': 501,
            'singer/alcoholic': 502,
            'stadium': 503,
            'restored': 504,
            'closets': 505,
            'duelling': 506,
            'pego': 507,
            'reindeer': 508,
            'hanif': 509,
            'allison': 510,
            'plunder': 511,
            'survivor': 512,
            'now-tired': 513,
            'rider': 514,
            'phantoms': 515,
            'longbaugh': 516,
            'fleshes': 517,
            'donovan': 518,
            'patterned': 519,
            'puffy': 520,
            'midwestern': 521,
            'plow': 522,
            'fogs': 523,
            'mesmerizes': 524,
            'testing': 525,
            'tick': 526,
            'lifting': 527,
            'computech': 528,
            'unrecognizable': 529,
            'ramblings': 530,
            'macho': 531,
            'chairs': 532,
            'wittliff': 533,
            'jean-claude': 534,
            'council': 535,
            'crushing': 536,
            'rapier': 537,
            'fearful': 538,
            'bud': 539,
            'staged': 540,
            'worm-eaten': 541,
            'wildly': 542,
            'credentials': 543,
            'receipt': 544,
            'frankiln': 545,
            'snorting': 546,
            "'blabbed": 547,
            'scoopfuls': 548,
            'dramatism': 549,
            'balthazar': 550,
            'override': 551,
            'thrusting': 552,
            'reinforced': 553,
            'manson-type': 554,
            'mean-spiritedness': 555,
            'gladiator-like': 556,
            'proposal': 557,
            'phillipe': 558,
            'casket': 559,
            'intoxicated': 560,
            'masterson': 561,
            'freelance': 562,
            'vicent': 563,
            'truths': 564,
            'triple-bladed': 565,
            'caddy': 566,
            'snorri': 567,
            'ethnic': 568,
            'liveliness': 569,
            'demme': 570,
            'convulsing': 571,
            'harumph': 572,
            'misstep': 573,
            'unveil': 574,
            'theme': 575,
            'voiceovers': 576,
            'inually': 577,
            'poolboy': 578,
            'flat-top': 579,
            'anti-government': 580,
            'realization': 581,
            'dreamy-eyed': 582,
            'paralells': 583,
            'maneuvers': 584,
            'russkies': 585,
            'references': 586,
            'chooses': 587,
            'blowing': 588,
            'renshaw': 589,
            'incessant': 590,
            'net-surfing': 591,
            'spiraling': 592,
            'knifed': 593,
            'politically': 594,
            'bossy': 595,
            'tinkering': 596,
            'frank-n-furter': 597,
            'severance': 598,
            'disapproves': 599,
            'hawks': 600,
            'maturity': 601,
            'portray': 602,
            'drug-addicted': 603,
            'ie': 604,
            'maternity': 605,
            'mantlepiece': 606,
            'identital': 607,
            'bouts': 608,
            'summon': 609,
            'vonnegut': 610,
            'fragmentary': 611,
            '/10': 612,
            '8-ball': 613,
            'liking': 614,
            'ears': 615,
            'preaches': 616,
            'misinterprets': 617,
            'jackson': 618,
            'patronizing': 619,
            'slicker': 620,
            'compatible': 621,
            'cuarsn': 622,
            'bravery': 623,
            'uniqe': 624,
            'motorcylce': 625,
            'politicians': 626,
            'ex-minister': 627,
            'bulls-eye': 628,
            'bio-rhythms': 629,
            'inhabiting': 630,
            'home': 631,
            'responsibilities': 632,
            'boats': 633,
            'stabilize': 634,
            'advancements': 635,
            'invested': 636,
            'ww': 637,
            'outtakes': 638,
            'iron-fisted': 639,
            'cds': 640,
            'potts': 641,
            'depraved': 642,
            'stargher': 643,
            'stirrings': 644,
            'serpico': 645,
            'soavi': 646,
            'traffiking': 647,
            'peaks': 648,
            'crunchem': 649,
            'aggravate': 650,
            'mandy': 651,
            'mcjob': 652,
            'uncovering': 653,
            'affliction': 654,
            'awk': 655,
            'paulina': 656,
            'klieg': 657,
            'race': 658,
            'authorial': 659,
            'judge-jury-and': 660,
            'bullet-shaped': 661,
            'swoops': 662,
            'heenan': 663,
            'slurping': 664,
            'archetype': 665,
            'dissappointed': 666,
            'handle': 667,
            'crust': 668,
            'moi': 669,
            'dwayne': 670,
            'spice': 671,
            'prejudge': 672,
            'curious': 673,
            'retro-garbo': 674,
            'animosity': 675,
            'uganda': 676,
            'shit-terpiece': 677,
            'selects': 678,
            'morose': 679,
            'weiss': 680,
            'blofeld': 681,
            'rigorous': 682,
            'slowing': 683,
            'pseudo': 684,
            'lashing': 685,
            'acquired': 686,
            'goodtimes': 687,
            'not-so-cheap': 688,
            'testings': 689,
            'needling': 690,
            'relished': 691,
            'moores': 692,
            'surname': 693,
            'obtain': 694,
            'impossibility': 695,
            'tycoon': 696,
            'stragely': 697,
            'wo': 698,
            'caleb': 699,
            'racism': 700,
            'protovision': 701,
            'newton': 702,
            'wallenberg': 703,
            'reeking': 704,
            'screen': 705,
            'hours': 706,
            'pitfalls': 707,
            'rapists': 708,
            'parlays': 709,
            'ritchie': 710,
            'arctic': 711,
            'beatles': 712,
            'venting': 713,
            'gos': 714,
            'maidservant': 715,
            'chewing': 716,
            'cousin': 717,
            'punk': 718,
            'ridicously': 719,
            'tires': 720,
            'gellar': 721,
            'munching': 722,
            'frightfulness': 723,
            'lite': 724,
            'toenails': 725,
            'bosom': 726,
            'matchmaking': 727,
            'shock-value': 728,
            'badger': 729,
            'grasshoppers': 730,
            'slathering': 731,
            'flops': 732,
            'servo': 733,
            'shadows': 734,
            'proval': 735,
            'ryuichi': 736,
            "'em": 737,
            'janssen': 738,
            'conqueror': 739,
            'stefanson': 740,
            'rally': 741,
            'moonlit': 742,
            'preteen': 743,
            'cream': 744,
            'watermelons': 745,
            'easily': 746,
            'classes': 747,
            'mildred': 748,
            'fragmented': 749,
            'rename': 750,
            'misogyny': 751,
            'wardens': 752,
            'sex-starved': 753,
            'highly-populated': 754,
            'century': 755,
            'winnebago': 756,
            'barnacles': 757,
            'biceps': 758,
            'wheat': 759,
            'thick-bladed': 760,
            'hauff': 761,
            'seldon': 762,
            'zone': 763,
            'undifferentiated': 764,
            'militant': 765,
            'cruising': 766,
            'fashionable': 767,
            'hughes': 768,
            'panoramic': 769,
            'off-camera': 770,
            'wizened': 771,
            'products': 772,
            'hates': 773,
            'ebony': 774,
            'forbids': 775,
            'candy-coated': 776,
            'ever-bemused': 777,
            'non-spoiler': 778,
            'sicken': 779,
            'sprightly': 780,
            'parfitt': 781,
            'disappear': 782,
            'borrr-ring': 783,
            'real-life': 784,
            'courageous': 785,
            'downsides': 786,
            'snorts': 787,
            'pin-up': 788,
            'clooney': 789,
            'pick-ups': 790,
            'protgaonist': 791,
            'termination': 792,
            'countenance': 793,
            'shaved': 794,
            'universe': 795,
            'stepahne': 796,
            'brett': 797,
            'maxx': 798,
            'whisked': 799,
            'bochner': 800,
            'proclivity': 801,
            'drug-addict': 802,
            'match': 803,
            'non-drinking': 804,
            't-birds': 805,
            'spungen': 806,
            'mostel': 807,
            'authors-i': 808,
            'well-supported': 809,
            'defensive': 810,
            'easygoing': 811,
            'elmaloglou': 812,
            'cherot': 813,
            'eerily': 814,
            'guesses': 815,
            'accomplishes': 816,
            'tighten': 817,
            'excesses': 818,
            'potty-mouthed': 819,
            'uplifting': 820,
            'non-stereotyped': 821,
            'bump': 822,
            'heralded': 823,
            'morissey': 824,
            'tested': 825,
            'degredation': 826,
            'bibile': 827,
            'improper': 828,
            'tree-surfs': 829,
            'regurgitate': 830,
            'children_': 831,
            'cheesefest': 832,
            'hokum': 833,
            'fiercer': 834,
            'frequency': 835,
            'dialectic': 836,
            'decalogue': 837,
            'canoes': 838,
            'casting-against-type': 839,
            'high-wire': 840,
            'bunny': 841,
            'producer/lover': 842,
            'remastered': 843,
            'waterboy': 844,
            'raw': 845,
            'abel': 846,
            'jeez': 847,
            'orgasmically': 848,
            'isn\x12t': 849,
            'leverage-using': 850,
            'tying': 851,
            'hooch': 852,
            'heavily-armed': 853,
            'loose': 854,
            'misspelled': 855,
            'tear-jerking': 856,
            'minimum': 857,
            'adeptness': 858,
            'channeling': 859,
            'spelling': 860,
            'problem': 861,
            'posturing': 862,
            'acquit': 863,
            'cessation': 864,
            'unearthed': 865,
            'courtyards': 866,
            'benoit': 867,
            'moland': 868,
            'cohn': 869,
            'barcalow': 870,
            'disneyfied': 871,
            'squirt': 872,
            'slick': 873,
            'spend': 874,
            'jetsons-like': 875,
            'fund-raiser': 876,
            'spiritless': 877,
            'drafted': 878,
            'abusers': 879,
            'mckee': 880,
            'ogle': 881,
            'organizer': 882,
            'presidential': 883,
            'salvage': 884,
            'taught': 885,
            'lobotomise': 886,
            'auteurs': 887,
            'horrendous': 888,
            'chintzy': 889,
            'cohesive': 890,
            'hooky': 891,
            'clandestine': 892,
            'appelation': 893,
            'seann': 894,
            'props-strategically-positioned-between-naked-actors-and-camera': 895,
            'sandler-annoying': 896,
            'vanquished': 897,
            'non-nude': 898,
            'spock': 899,
            "'help": 900,
            'mental/physical': 901,
            'whereupon': 902,
            'francisco-based': 903,
            'hypotheitically': 904,
            'gags': 905,
            'fluidly': 906,
            'deschanel': 907,
            'nasty-tempered': 908,
            'frankfurters': 909,
            'sadie': 910,
            'grandfather': 911,
            'commandeer': 912,
            'mikel': 913,
            'instrument': 914,
            'angel-related': 915,
            'casted': 916,
            'damian': 917,
            'minor-league': 918,
            'squaddie': 919,
            'preconceptions': 920,
            'dramatic': 921,
            'chromium': 922,
            'stalwart': 923,
            'villagers': 924,
            'bogeyman': 925,
            'movingly': 926,
            'rate': 927,
            'derail': 928,
            'widows': 929,
            "'dude": 930,
            'animal-rights': 931,
            'gutierrez': 932,
            'ribbing': 933,
            'attic': 934,
            'player/musician/composer': 935,
            'satiate': 936,
            'delinquent': 937,
            'directives': 938,
            '_boom_': 939,
            'slumping': 940,
            'grumpier': 941,
            'ironing': 942,
            'tabbed': 943,
            'jermaine': 944,
            'contestants': 945,
            'penney': 946,
            'systematically': 947,
            'miriam': 948,
            'guenveur': 949,
            'mentoring': 950,
            'crumpled': 951,
            'sensual': 952,
            'nunez': 953,
            'abominable': 954,
            'head-strong': 955,
            'poison-': 956,
            'long-running': 957,
            'd-day': 958,
            'hook-laden': 959,
            'gauntlet': 960,
            'faux-mysticism': 961,
            'fasano': 962,
            'unformal': 963,
            'mid-1500s': 964,
            'paid': 965,
            'erroneous': 966,
            'instalment': 967,
            'cabbie': 968,
            'life-affirming': 969,
            'asswhole': 970,
            'tate': 971,
            'five-hundred-pound': 972,
            'double-zero': 973,
            'stubbornness': 974,
            'delpy': 975,
            'near-eden-like': 976,
            'abort': 977,
            'cd-rom': 978,
            'edt': 979,
            'schoolchildren': 980,
            'galvanizing': 981,
            'pricked': 982,
            'sk': 983,
            'protect-earth-from-destruction': 984,
            'quicktime': 985,
            'songwriter': 986,
            'hallmark': 987,
            'crud': 988,
            'independence': 989,
            'noodle-headed': 990,
            'flickers': 991,
            'kang': 992,
            'electing': 993,
            'correspondence': 994,
            'schoolteacher': 995,
            'necessarily': 996,
            'glossing': 997,
            'movie-you': 998,
            'uncharismatic': 999,
            ...},
           array([[0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  ...,
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.],
                  [0., 0., 0., ..., 0., 0., 0.]]))
```

```
In [174]: vocabulary, X = count_words(texts)

          # Try to fit, predict and score
          nb = NB()
          print('The 5-fold cross-validation score of our hand-made estimator after stemming i
```

```
The 5-fold cross-validation score of our hand-made estimator after stemming is : 0.8355
```

Using the ntlk library and keeping only nouns, verbs, adverbs and adjectives yields the best results!

One might still prefer the stemming approach for computational and compression reasons.