

Embedding

August 18, 2019

```
[17]: import numpy as np
import pickle
import json
import tensorflow as tf
```

```
[18]: def get_embedding(data, model_type, vector_type, embed_size):
    embed_filename = output_folder_name + data + "_" + model_type + "_" +
    →vector_type + "_" + embed_size + ".pkl"
    vocab_filename = output_folder_name + data + "_" + model_type + "_" +
    →vector_type + "_" + embed_size + "_dict.json"
    reverse_vocab_filename = output_folder_name + data + "_" + model_type + "_" +
    →vector_type + "_" + embed_size + "_reversedict.json"

    embedding = np.load(embed_filename, allow_pickle=True)

    with open(vocab_filename, "r") as f:
        vocab = json.load(f)
    with open(reverse_vocab_filename, "r") as f:
        reverse_vocab = json.load(f)

    data_dict = {
        "data": data,
        "embed": embedding,
        "vocab": vocab,
        "reverse_vocab": reverse_vocab
    }

    return data_dict
```

```
[19]: from sklearn.manifold import TSNE
import matplotlib.pyplot as plt

def plot_with_labels(low_dim_embs, labels, filename='tsne.png'):
    assert low_dim_embs.shape[0] >= len(labels), 'More labels than embeddings'
    plt.figure(figsize=(14, 14)) # in inches
    for i, label in enumerate(labels):
        x, y = low_dim_embs[i, :]
```

```

plt.scatter(x, y)
plt.annotate(label,
             xy=(x, y),
             xytext=(5, 2),
             textcoords='offset points',
             ha='right',
             va='bottom')

plt.show()

def plot_embedding(embedding, m, n, reverse_dictionary):
    tsne = TSNE(perplexity=30, n_components=2, init='pca', n_iter=5000)
    low_dim_embs = tsne.fit_transform(embedding[m:n, :])
    labels = [reverse_dictionary[i] for i in range(m,n)]
    plot_with_labels(low_dim_embs, labels)
    return low_dim_embs

def top_similar(low_dim_embs, mapping, reverse_mapping, valid_words):
    similarity = tf.matmul(
        low_dim_embs, low_dim_embs, transpose_b=True)

    sim = tf.Session().run(similarity)
    for i in range(len(valid_words)):
        if(valid_words[i] in mapping):
            index = mapping[valid_words[i]]
            valid_word = reverse_mapping[index]
            top_k = 20 # number of nearest neighbors
            nearest = (-sim[index, :][:500]).argsort()[1:top_k + 1]
            log_str = 'Nearest to %s:' % valid_word
            for k in range(top_k):
                close_word = reverse_mapping[nearest[k]]
                log_str = '%s %s,' % (log_str, close_word)
            print(log_str)

```

```

[20]: def similar_clusters(data, words, embed_range=[0,500]):
    embedding = data["embed"]
    data_dict = data["vocab"]
    data_reverse_dict = data["reverse_vocab"]
    low_dim_embs = plot_embedding(embedding, embed_range[0], embed_range[1],
    ↪data_reverse_dict)
    top_similar(low_dim_embs, data_dict, data_reverse_dict, words)

```

```

[21]: data = "formspring"
model_type = "blstm_attention"
vector_type = "random"
embed_size = "50"
output_folder_name = "/home/sujeendra/Desktop/results/"

words = ['gay', 'slave', 'evidence', 'hate', 'fat']

```



```
[23]: def get_similarityMatrix(low_dim_embs):
    similarity = tf.matmul(
        low_dim_embs, low_dim_embs, transpose_b=True)

    sim = tf.Session().run(similarity)
    return sim

def similarWords(word, sim, mapping, reverse_mapping, top_k):
    index = mapping[word]
    log_str = ''
    try:
        valid_word = reverse_mapping[index]
        nearest = (-sim[index, :][:1000]).argsort()[1:top_k + 1]

        for k in range(top_k):
            close_word = reverse_mapping[nearest[k]]
            log_str = '%s%s, ' % (log_str, close_word)
    except:
        print ("Word beyond 10k")
    return log_str

def get_tsneembedding(embedding, m, n):
    tsne = TSNE(perplexity=30, n_components=2, init='pca', n_iter=5000)
    low_dim_embs = tsne.fit_transform(embedding[m:n, :])
    return low_dim_embs

def comparison(dict_1, dict_2, dict_3, words, top_k):
    for word in words:
        print("Word: " + str(word))
        print(dict_1["data"] + ": " + similarWords(word,
            ↪get_similarityMatrix(dict_1["embed"]), dict_1["vocab"],
            ↪dict_1["reverse_vocab"], top_k))
        print(dict_2["data"] + ": " + similarWords(word,
            ↪get_similarityMatrix(dict_2["embed"]), dict_2["vocab"],
            ↪dict_2["reverse_vocab"], top_k))
        print(dict_3["data"] + ": " + similarWords(word,
            ↪get_similarityMatrix(dict_3["embed"]), dict_3["vocab"],
            ↪dict_3["reverse_vocab"], top_k))
        print("-----")
```

```
[24]: words = ['gay', 'slave', 'evidence', 'fat', 'religion', 'ass']
```

```
[26]: data_dict
```

```
[26]: {'data': 'formspring',
      'embed': array([[ 0.01143891, -0.02900364,  0.04330887, ...,  0.02242133,
                        -0.04362371, -0.04626227],
                      [ 0.02064788,  0.0383586 , -0.04162367, ..., -0.06183602,
```

```

        0.00959281, -0.00566038],
        [-0.00244074, -0.03423123, 0.00414955, ..., 0.00354305,
        -0.03292997, 0.03247664],
        ...,
        [ 0.01077047, -0.07009816, 0.0494134 , ..., 0.09908214,
        -0.08192186, 0.10202785],
        [-0.00997376, 0.02495752, -0.03498416, ..., -0.04293251,
        0.08547857, -0.09842607],
        [-0.04248628, 0.09873257, -0.0873236 , ..., -0.09732826,
        0.08273684, -0.05167283]], dtype=float32),
'vocab': {'<UNK>': 0,
'you': 1,
'i': 2,
'to': 3,
'a': 4,
'the': 5,
'and': 6,
'do': 7,
'is': 8,
'it': 9,
'not': 10,
'your': 11,
'have': 12,
'what': 13,
'that': 14,
'of': 15,
'in': 16,
'are': 17,
'my': 18,
'me': 19,
'lol': 20,
'like': 21,
'u': 22,
'would': 23,
'if': 24,
'so': 25,
'on': 26,
'be': 27,
'but': 28,
'for': 29,
'or': 30,
'was': 31,
'with': 32,
'im': 33,
'how': 34,
'no': 35,
'think': 36,

```

'know': 37,
'r': 38,
'd': 39,
'who': 40,
'just': 41,
'love': 42,
'get': 43,
'haha': 44,
'one': 45,
'why': 46,
'all': 47,
'they': 48,
'about': 49,
'can': 50,
'when': 51,
'its': 52,
'at': 53,
'up': 54,
'dont': 55,
'ever': 56,
'really': 57,
'out': 58,
'this': 59,
'3': 60,
'yes': 61,
'p': 62,
'people': 63,
'go': 64,
'did': 65,
'as': 66,
'had': 67,
'we': 68,
'she': 69,
'them': 70,
'will': 71,
'there': 72,
'now': 73,
'then': 74,
'well': 75,
'her': 76,
'want': 77,
'time': 78,
'say': 79,
'lmao': 80,
'good': 81,
'too': 82,
'ur': 83,

'life': 84,
'from': 85,
'were': 86,
'because': 87,
"i'm": 88,
'am': 89,
'been': 90,
'someone': 91,
'an': 92,
'he': 93,
'more': 94,
'o': 95,
'make': 96,
'some': 97,
'much': 98,
'going': 99,
'bitch': 100,
'could': 101,
'person': 102,
'something': 103,
'name': 104,
'does': 105,
'any': 106,
'right': 107,
'thats': 108,
'off': 109,
'never': 110,
'thing': 111,
'back': 112,
'got': 113,
'fake': 114,
'see': 115,
'tell': 116,
'only': 117,
'girl': 118,
'most': 119,
'last': 120,
'shit': 121,
'yeah': 122,
'should': 123,
'fuck': 124,
'day': 125,
'oh': 126,
'being': 127,
'where': 128,
'even': 129,
'way': 130,

'has': 131,
'still': 132,
'by': 133,
'questions': 134,
'things': 135,
'pretty': 136,
'many': 137,
'mean': 138,
'talk': 139,
'look': 140,
'favorite': 141,
'thanks': 142,
'feel': 143,
'stop': 144,
'ass': 145,
'sure': 146,
'need': 147,
'over': 148,
'here': 149,
'always': 150,
'ask': 151,
'first': 152,
'than': 153,
'hey': 154,
'hahaha': 155,
'other': 156,
'hate': 157,
'anything': 158,
'actually': 159,
'cause': 160,
'friends': 161,
'question': 162,
'take': 163,
'2': 164,
'best': 165,
'said': 166,
'which': 167,
'him': 168,
'give': 169,
'sex': 170,
'fucking': 171,
'nice': 172,
'their': 173,
'better': 174,
'okay': 175,
'nope': 176,
'n': 177,

'though': 178,
'ok': 179,
'come': 180,
'ha': 181,
'down': 182,
'friend': 183,
'thank': 184,
'idk': 185,
'sorry': 186,
'ill': 187,
'maybe': 188,
'spam': 189,
'lmfao': 190,
'let': 191,
'school': 192,
'xd': 193,
'answer': 194,
'find': 195,
'formspring': 196,
'ugly': 197,
'everyone': 198,
'little': 199,
'very': 200,
'yourself': 201,
'wanna': 202,
'whats': 203,
'anyone': 204,
'face': 205,
'work': 206,
'again': 207,
'big': 208,
'told': 209,
'night': 210,
'probably': 211,
'believe': 212,
'funny': 213,
'girls': 214,
'kind': 215,
'live': 216,
'same': 217,
'talking': 218,
'long': 219,
'before': 220,
'baby': 221,
'guess': 222,
'bad': 223,
'old': 224,

'into': 225,
'cool': 226,
'song': 227,
'made': 228,
'both': 229,
'ca': 230,
'nothing': 231,
'please': 232,
'today': 233,
'guys': 234,
'put': 235,
'own': 236,
'around': 237,
'call': 238,
'rather': 239,
'stupid': 240,
'guy': 241,
'lot': 242,
'saying': 243,
'done': 244,
'eat': 245,
's': 246,
'went': 247,
'man': 248,
'try': 249,
'cant': 250,
'didnt': 251,
'leave': 252,
'watch': 253,
'world': 254,
'yo': 255,
'cute': 256,
'new': 257,
'gonna': 258,
'doing': 259,
'thought': 260,
'wat': 261,
'year': 262,
'happy': 263,
'bed': 264,
'care': 265,
'hot': 266,
'real': 267,
'wow': 268,
'hard': 269,
'wtf': 270,
'family': 271,

'miss': 272,
'everything': 273,
'stuff': 274,
'car': 275,
'else': 276,
'his': 277,
'hoe': 278,
'home': 279,
'myself': 280,
'1': 281,
'hmm': 282,
'makes': 283,
'wish': 284,
'money': 285,
'com': 286,
'gay': 287,
'show': 288,
'ya': 289,
'damn': 290,
'our': 291,
'aww': 292,
'seen': 293,
'hmmm': 294,
'keep': 295,
'movie': 296,
'getting': 297,
'http': 298,
'change': 299,
'drink': 300,
'hair': 301,
'hope': 302,
'ppl': 303,
'years': 304,
'use': 305,
'those': 306,
'dick': 307,
'next': 308,
'while': 309,
'place': 310,
'away': 311,
'omg': 312,
'trying': 313,
'two': 314,
'house': 315,
'hell': 316,
'mind': 317,
'fun': 318,

'hehe': 319,
'k': 320,
'sometimes': 321,
'times': 322,
'y': 323,
'date': 324,
'having': 325,
'suck': 326,
'yew': 327,
'yu': 328,
'after': 329,
'id': 330,
'5': 331,
'sleep': 332,
'used': 333,
'www': 334,
'asked': 335,
'asking': 336,
'l': 337,
'already': 338,
'without': 339,
'cuz': 340,
'hi': 341,
'another': 342,
'youu': 343,
'every': 344,
'yeh': 345,
'end': 346,
'huh': 347,
'wanted': 348,
'food': 349,
'parents': 350,
'such': 351,
'dad': 352,
'sexy': 353,
'called': 354,
'hahah': 355,
'these': 356,
'4': 357,
'enough': 358,
'wear': 359,
'remember': 360,
'alone': 361,
'great': 362,
'33': 363,
'choose': 364,
'music': 365,

'usually': 366,
'once': 367,
'depends': 368,
'dumb': 369,
'sister': 370,
'wait': 371,
'god': 372,
'ummm': 373,
'least': 374,
'phone': 375,
'play': 376,
'different': 377,
'kids': 378,
'often': 379,
'umm': 380,
'wrong': 381,
'course': 382,
'dnt': 383,
'saw': 384,
'word': 385,
'b': 386,
'relationship': 387,
'through': 388,
'10': 389,
'help': 390,
'pussy': 391,
'read': 392,
'age': 393,
'laugh': 394,
'yuh': 395,
'body': 396,
'eyes': 397,
'sweet': 398,
'bit': 399,
'hahahaha': 400,
'head': 401,
'hurt': 402,
'days': 403,
'ones': 404,
'alot': 405,
'mad': 406,
'us': 407,
'yet': 408,
'c': 409,
'seriously': 410,
'soo': 411,
'weird': 412,

'either': 413,
'opinion': 414,
'also': 415,
'die': 416,
'nd': 417,
'stfu': 418,
'beautiful': 419,
'dude': 420,
'v': 421,
'bout': 422,
'lady': 423,
'means': 424,
'send': 425,
'start': 426,
'etc': 427,
'happened': 428,
'morning': 429,
'shut': 430,
'yea': 431,
' ': 432,
'awww': 433,
'explain': 434,
'favourite': 435,
'mom': 436,
'moment': 437,
'tabi': 438,
'true': 439,
'havent': 440,
'job': 441,
'kill': 442,
'kiss': 443,
'listen': 444,
'm': 445,
'anyway': 446,
'fat': 447,
'reason': 448,
'tomorrow': 449,
'amazing': 450,
'came': 451,
'dream': 452,
'kinda': 453,
'lil': 454,
'mine': 455,
'myspace': 456,
'ahh': 457,
'aint': 458,
'e': 459,

'facebook': 460,
'far': 461,
'wit': 462,
'yours': 463,
'knew': 464,
'together': 465,
'twitter': 466,
'may': 467,
'women': 468,
'x': 469,
'blue': 470,
'fight': 471,
'white': 472,
'worst': 473,
'ive': 474,
'loved': 475,
'prefer': 476,
't': 477,
'week': 478,
'happen': 479,
'hello': 480,
'stay': 481,
'half': 482,
'heard': 483,
'married': 484,
'must': 485,
'shes': 486,
'since': 487,
'beer': 488,
'buy': 489,
'dead': 490,
'kid': 491,
'making': 492,
'might': 493,
'room': 494,
'top': 495,
'able': 496,
'anymore': 497,
'few': 498,
'hit': 499,
'looking': 500,
'random': 501,
'wont': 502,
'whole': 503,
'anonymous': 504,
'says': 505,
'yep': 506,

'7': 507,
'child': 508,
'heart': 509,
'sense': 510,
'annoying': 511,
'color': 512,
'doesnt': 513,
'page': 514,
'fs': 515,
'ily': 516,
'started': 517,
'answers': 518,
'beach': 519,
'da': 520,
'dat': 521,
'grow': 522,
'jealous': 523,
'mi': 524,
'nah': 525,
'sad': 526,
'um': 527,
'whore': 528,
'yup': 529,
'0': 530,
'hear': 531,
'until': 532,
'chocolate': 533,
'dog': 534,
'each': 535,
'h': 536,
'knows': 537,
'laura': 538,
'matter': 539,
'tonight': 540,
'6': 541,
'country': 542,
'grade': 543,
'special': 544,
'super': 545,
'thinking': 546,
'water': 547,
'fine': 548,
'free': 549,
'hah': 550,
'honestly': 551,
'idea': 552,
'obviously': 553,

'tv': 554,
'type': 555,
'yuu': 556,
'awesome': 557,
'chicken': 558,
'coming': 559,
'describe': 560,
'enjoy': 561,
'found': 562,
'justin': 563,
'lame': 564,
'perfect': 565,
'side': 566,
'straight': 567,
'ago': 568,
'couple': 569,
'gaga': 570,
'gone': 571,
'hes': 572,
'important': 573,
'inbox': 574,
'jk': 575,
'meet': 576,
'part': 577,
'picture': 578,
'sooo': 579,
'spend': 580,
'333': 581,
'bf': 582,
'black': 583,
'crazy': 584,
'deal': 585,
'dirty': 586,
'honest': 587,
'looks': 588,
'pictures': 589,
'pink': 590,
'telling': 591,
'turn': 592,
'ahaha': 593,
'drunk': 594,
'learn': 595,
'men': 596,
'past': 597,
'whatever': 598,
'close': 599,
'definitely': 600,

'gets': 601,
'gf': 602,
'later': 603,
'short': 604,
'story': 605,
'walk': 606,
'wants': 607,
'words': 608,
'8': 609,
'cold': 610,
'park': 611,
'ready': 612,
'soon': 613,
'using': 614,
'w': 615,
'youtube': 616,
'agree': 617,
'btw': 618,
'fan': 619,
'fav': 620,
'fuk': 621,
'ice': 622,
'lost': 623,
'okies': 624,
'outside': 625,
'post': 626,
'slut': 627,
'speak': 628,
'weekend': 629,
'between': 630,
'book': 631,
'daughter': 632,
'drive': 633,
'fall': 634,
'mood': 635,
'pick': 636,
'quite': 637,
'sounds': 638,
'tht': 639,
'u2665': 640,
'act': 641,
'boyfriend': 642,
'cream': 643,
'left': 644,
'pathetic': 645,
'dd': 646,
'dinner': 647,

'goes': 648,
'ima': 649,
'inside': 650,
'kno': 651,
'lie': 652,
'others': 653,
'party': 654,
'q': 655,
'small': 656,
'took': 657,
'wouldnt': 658,
'babe': 659,
'boy': 660,
'clothes': 661,
'dark': 662,
'gorgeous': 663,
'high': 664,
'movies': 665,
'nigga': 666,
'point': 667,
'seem': 668,
'self': 669,
'smile': 670,
'stand': 671,
'wud': 672,
'check': 673,
'famous': 674,
'feeling': 675,
'gurl': 676,
'happens': 677,
'lets': 678,
'serious': 679,
'taken': 680,
'totally': 681,
'unless': 682,
'virgin': 683,
'bye': 684,
'comes': 685,
'game': 686,
'gave': 687,
'kelly': 688,
'met': 689,
'mum': 690,
'number': 691,
'rock': 692,
'sick': 693,
'suppose': 694,

'tired': 695,
'under': 696,
'watching': 697,
'16': 698,
'emily': 699,
'follow': 700,
'message': 701,
'move': 702,
'rest': 703,
'sort': 704,
'three': 705,
'wana': 706,
'woman': 707,
'children': 708,
'hmmmm': 709,
'living': 710,
'problem': 711,
'second': 712,
'write': 713,
'add': 714,
'bring': 715,
'chance': 716,
'red': 717,
'skins': 718,
'smell': 719,
'3333': 720,
'calling': 721,
'currently': 722,
'given': 723,
'kisses': 724,
'none': 725,
'older': 726,
'perhaps': 727,
'run': 728,
'thee': 729,
'video': 730,
'wife': 731,
'worry': 732,
'answering': 733,
'birthday': 734,
'butt': 735,
'likes': 736,
'summer': 737,
'texas': 738,
'break': 739,
'bxtch': 740,
'dance': 741,

'gotta': 742,
'list': 743,
'mouth': 744,
'muj': 745,
'online': 746,
'open': 747,
'text': 748,
'till': 749,
'wake': 750,
'f': 751,
'fuckin': 752,
'interesting': 753,
'mr': 754,
'vv': 755,
'vvv': 756,
'waiting': 757,
'win': 758,
'y0': 759,
'alright': 760,
'bet': 761,
'bored': 762,
'crap': 763,
'death': 764,
'hun': 765,
'pay': 766,
'rude': 767,
'smoke': 768,
'wo': 769,
'worse': 770,
'wrote': 771,
'yur': 772,
'along': 773,
'choice': 774,
'colour': 775,
'ex': 776,
'feet': 777,
'following': 778,
'hours': 779,
'needs': 780,
'ohh': 781,
'sound': 782,
'alive': 783,
'chinese': 784,
'cook': 785,
'late': 786,
'leaving': 787,
'ma': 788,

'member': 789,
'naahh': 790,
'soooo': 791,
'trust': 792,
'visit': 793,
'whos': 794,
'boo': 795,
'dating': 796,
'fly': 797,
'hang': 798,
'hop': 799,
'imagine': 800,
'named': 801,
'retarded': 802,
'save': 803,
'spamming': 804,
'stuck': 805,
'thankyou': 806,
'tried': 807,
'won': 808,
'young': 809,
'20': 810,
'8th': 811,
'anyways': 812,
'aw': 813,
'cheese': 814,
'cutie': 815,
'eating': 816,
'especially': 817,
'green': 818,
'isnt': 819,
'jail': 820,
'lmbo': 821,
'lots': 822,
'personality': 823,
'single': 824,
'tabitha': 825,
'tho': 826,
'ways': 827,
'working': 828,
'yahoo': 829,
'beat': 830,
'certain': 831,
'class': 832,
'cut': 833,
'deep': 834,
'fast': 835,

'female': 836,
'film': 837,
'possible': 838,
'taking': 839,
'trouble': 840,
'yeahh': 841,
'apart': 842,
'background': 843,
'behind': 844,
'bitchh': 845,
'comment': 846,
'felt': 847,
'hands': 848,
'ish': 849,
'liked': 850,
'personally': 851,
'silly': 852,
'star': 853,
'sweetie': 854,
'victor': 855,
'waht': 856,
'13': 857,
'15': 858,
'afraid': 859,
'almost': 860,
'bieber': 861,
'biggest': 862,
'boobs': 863,
'bother': 864,
'computer': 865,
'cry': 866,
'fool': 867,
'full': 868,
'girlfriend': 869,
'light': 870,
'months': 871,
'seeing': 872,
'shay': 873,
'welcome': 874,
'yess': 875,
'account': 876,
'aha': 877,
'asleep': 878,
'attention': 879,
'bitches': 880,
'brother': 881,
'chomp': 882,

'door': 883,
'during': 884,
'front': 885,
'hand': 886,
'instead': 887,
'noe': 888,
'opposite': 889,
'partner': 890,
'piece': 891,
'scared': 892,
'shall': 893,
'step': 894,
'theres': 895,
'touch': 896,
'uh': 897,
'animal': 898,
'austin': 899,
'band': 900,
'easily': 901,
'hold': 902,
'human': 903,
'lovee': 904,
'missed': 905,
'pic': 906,
'places': 907,
'quit': 908,
'sexual': 909,
'shower': 910,
'situation': 911,
'somewhere': 912,
'state': 913,
'thoughts': 914,
'ummmm': 915,
'writing': 916,
'yesterday': 917,
'9': 918,
'born': 919,
'chill': 920,
'control': 921,
'creepy': 922,
'cunt': 923,
'dreams': 924,
'drinking': 925,
'dum': 926,
'fair': 927,
'fave': 928,
'feelings': 929,

'kool': 930,
'middle': 931,
'million': 932,
'month': 933,
'ohhh': 934,
'prolly': 935,
'road': 936,
'songs': 937,
'spell': 938,
'stick': 939,
'tew': 940,
'understand': 941,
'watched': 942,
'yay': 943,
'30': 944,
'ate': 945,
'become': 946,
'boys': 947,
'completely': 948,
'delete': 949,
'difference': 950,
'dork': 951,
'dress': 952,
'easy': 953,
'forever': 954,
'french': 955,
'fucked': 956,
'glad': 957,
'hahahah': 958,
'hour': 959,
'juss': 960,
'la': 961,
'language': 962,
'meant': 963,
'mother': 964,
'naw': 965,
'ohm': 966,
'share': 967,
'sit': 968,
'starting': 969,
'thnx': 970,
'wasnt': 971,
'yall': 972,
'younger': 973,
'33333': 974,
'celebrity': 975,
'clothing': 976,

'club': 977,
'count': 978,
'crush': 979,
'cum': 980,
'driving': 981,
'fakes': 982,
'fell': 983,
'forget': 984,
'jeans': 985,
'kick': 986,
'looked': 987,
'moo': 988,
'set': 989,
'street': 990,
'weeks': 991,
'american': 992,
'cat': 993,
'changed': 994,
'college': 995,
'eva': 996,
'everyday': 997,
'ew': 998,
'fact': 999,
...},
'reverse_vocab': ['<UNK>',
'you',
'i',
'to',
'a',
'the',
'and',
'do',
'is',
'it',
'not',
'your',
'have',
'what',
'that',
'of',
'in',
'are',
'my',
'me',
'lol',
'like',
'u',

'would',
'if',
'so',
'on',
'be',
'but',
'for',
'or',
'was',
'with',
'im',
'how',
'no',
'think',
'know',
'r',
'd',
'who',
'just',
'love',
'get',
'haha',
'one',
'why',
'all',
'they',
'about',
'can',
'when',
'its',
'at',
'up',
'dont',
'ever',
'really',
'out',
'this',
'3',
'yes',
'p',
'people',
'go',
'did',
'as',
'had',
'we',
'she',

'them',
'will',
'there',
'now',
'then',
'well',
'her',
'want',
'time',
'say',
'lmao',
'good',
'too',
'ur',
'life',
'from',
'were',
'because',
"i'm",
'am',
'been',
'someone',
'an',
'he',
'more',
'o',
'make',
'some',
'much',
'going',
'bitch',
'could',
'person',
'something',
'name',
'does',
'any',
'right',
'thats',
'off',
'never',
'thing',
'back',
'got',
'fake',
'see',
'tell',

'only',
'girl',
'most',
'last',
'shit',
'yeah',
'should',
'fuck',
'day',
'oh',
'being',
'where',
'even',
'way',
'has',
'still',
'by',
'questions',
'things',
'pretty',
'many',
'mean',
'talk',
'look',
'favorite',
'thanks',
'feel',
'stop',
'ass',
'sure',
'need',
'over',
'here',
'always',
'ask',
'first',
'than',
'hey',
'hahaha',
'other',
'hate',
'anything',
'actually',
'cause',
'friends',
'question',
'take',

'2',
'best',
'said',
'which',
'him',
'give',
'sex',
'fucking',
'nice',
'their',
'better',
'okay',
'nope',
'n',
'though',
'ok',
'come',
'ha',
'down',
'friend',
'thank',
'idk',
'sorry',
'ill',
'maybe',
'spam',
'lmfao',
'let',
'school',
'xd',
'answer',
'find',
'formspring',
'ugly',
'everyone',
'little',
'very',
'yourself',
'wanna',
'whats',
'anyone',
'face',
'work',
'again',
'big',
'told',
'night',

'probably',
'believe',
'funny',
'girls',
'kind',
'live',
'same',
'talking',
'long',
'before',
'baby',
'guess',
'bad',
'old',
'into',
'cool',
'song',
'made',
'both',
'ca',
'nothing',
'please',
'today',
'guys',
'put',
'own',
'around',
'call',
'rather',
'stupid',
'guy',
'lot',
'saying',
'done',
'eat',
's',
'went',
'man',
'try',
'cant',
'didnt',
'leave',
'watch',
'world',
'yo',
'cute',
'new',

'gonna',
'doing',
'thought',
'wat',
'year',
'happy',
'bed',
'care',
'hot',
'real',
'wow',
'hard',
'wtf',
'family',
'miss',
'everything',
'stuff',
'car',
'else',
'his',
'hoe',
'home',
'myself',
'1',
'hmm',
'makes',
'wish',
'money',
'com',
'gay',
'show',
'ya',
'damn',
'our',
'aww',
'seen',
'hmmm',
'keep',
'movie',
'getting',
'http',
'change',
'drink',
'hair',
'hope',
'ppl',
'years',

'use',
'those',
'dick',
'next',
'while',
'place',
'away',
'omg',
'trying',
'two',
'house',
'hell',
'mind',
'fun',
'hehe',
'k',
'sometimes',
'times',
'y',
'date',
'having',
'suck',
'yew',
'yu',
'after',
'id',
'5',
'sleep',
'used',
'www',
'asked',
'asking',
'l',
'already',
'without',
'cuz',
'hi',
'another',
'youu',
'every',
'yeh',
'end',
'huh',
'wanted',
'food',
'parents',
'such',

'dad',
'sexy',
'called',
'hahah',
'these',
'4',
'enough',
'wear',
'remember',
'alone',
'great',
'33',
'choose',
'music',
'usually',
'once',
'depends',
'dumb',
'sister',
'wait',
'god',
'ummm',
'least',
'phone',
'play',
'different',
'kids',
'often',
'umm',
'wrong',
'course',
'dnt',
'saw',
'word',
'b',
'relationship',
'through',
'10',
'help',
'pussy',
'read',
'age',
'laugh',
'yuh',
'body',
'eyes',
'sweet',

'bit',
'hahahaha',
'head',
'hurt',
'days',
'ones',
'alot',
'mad',
'us',
'yet',
'c',
'seriously',
'soo',
'weird',
'either',
'opinion',
'also',
'die',
'nd',
'stfu',
'beautiful',
'dude',
'v',
'bout',
'lady',
'means',
'send',
'start',
'etc',
'happened',
'morning',
'shut',
'yea',
""',
'awww',
'explain',
'favourite',
'mom',
'moment',
'tabi',
'true',
'havent',
'job',
'kill',
'kiss',
'listen',
'm',

'anyway',
'fat',
'reason',
'tomorrow',
'amazing',
'came',
'dream',
'kinda',
'lil',
'mine',
'myspace',
'ahh',
'aint',
'e',
'facebook',
'far',
'wit',
'yours',
'knew',
'together',
'twitter',
'may',
'women',
'x',
'blue',
'fight',
'white',
'worst',
'ive',
'loved',
'prefer',
't',
'week',
'happen',
'hello',
'stay',
'half',
'heard',
'married',
'must',
'shes',
'since',
'beer',
'buy',
'dead',
'kid',
'making',

'might',
'room',
'top',
'able',
'anymore',
'few',
'hit',
'looking',
'random',
'wont',
'whole',
'anonymous',
'says',
'yep',
'7',
'child',
'heart',
'sense',
'annoying',
'color',
'doesnt',
'page',
'fs',
'ily',
'started',
'answers',
'beach',
'da',
'dat',
'grow',
'jealous',
'mi',
'nah',
'sad',
'um',
'whore',
'yup',
'0',
'hear',
'until',
'chocolate',
'dog',
'each',
'h',
'knows',
'laura',
'matter',

'tonight',
'6',
'country',
'grade',
'special',
'super',
'thinking',
'water',
'fine',
'free',
'hah',
'honestly',
'idea',
'obviously',
'tv',
'type',
'yuu',
'awesome',
'chicken',
'coming',
'describe',
'enjoy',
'found',
'justin',
'lame',
'perfect',
'side',
'straight',
'ago',
'couple',
'gaga',
'gone',
'hes',
'important',
'inbox',
'jk',
'meet',
'part',
'picture',
'sooo',
'spend',
'333',
'bf',
'black',
'crazy',
'deal',
'dirty',

'honest',
'looks',
'pictures',
'pink',
'telling',
'turn',
'ahaha',
'drunk',
'learn',
'men',
'past',
'whatever',
'close',
'definitely',
'gets',
'gf',
'later',
'short',
'story',
'walk',
'wants',
'words',
'8',
'cold',
'park',
'ready',
'soon',
'using',
'w',
'youtube',
'agree',
'btw',
'fan',
'fav',
'fuk',
'ice',
'lost',
'okies',
'outside',
'post',
'slut',
'speak',
'weekend',
'between',
'book',
'daughter',
'drive',

'fall',
'mood',
'pick',
'quite',
'sounds',
'tht',
'u2665',
'act',
'boyfriend',
'cream',
'left',
'pathetic',
'dd',
'dinner',
'goes',
'ima',
'inside',
'kno',
'lie',
'others',
'party',
'q',
'small',
'took',
'wouldnt',
'babe',
'boy',
'clothes',
'dark',
'gorgeous',
'high',
'movies',
'nigga',
'point',
'seem',
'self',
'smile',
'stand',
'wud',
'check',
'famous',
'feeling',
'gurl',
'happens',
'lets',
'serious',
'taken',

'totally',
'unless',
'virgin',
'bye',
'comes',
'game',
'gave',
'kelly',
'met',
'mum',
'number',
'rock',
'sick',
'suppose',
'tired',
'under',
'watching',
'16',
'emily',
'follow',
'message',
'move',
'rest',
'sort',
'three',
'wana',
'woman',
'children',
'hmmmm',
'living',
'problem',
'second',
'write',
'add',
'bring',
'chance',
'red',
'skins',
'smell',
'3333',
'calling',
'currently',
'given',
'kisses',
'none',
'older',
'perhaps',

'run',
'thee',
'video',
'wife',
'worry',
'answering',
'birthday',
'butt',
'likes',
'summer',
'texas',
'break',
'bxtch',
'dance',
'gotta',
'list',
'mouth',
'muj',
'online',
'open',
'text',
'till',
'wake',
'f',
'fuckin',
'interesting',
'mr',
'vv',
'vvv',
'waiting',
'win',
'y0',
'alright',
'bet',
'bored',
'crap',
'death',
'hun',
'pay',
'rude',
'smoke',
'wo',
'worse',
'wrote',
'yur',
'along',
'choice',

'colour',
'ex',
'feet',
'following',
'hours',
'needs',
'ohh',
'sound',
'alive',
'chinese',
'cook',
'late',
'leaving',
'ma',
'member',
'nahh',
'soooo',
'trust',
'visit',
'whos',
'boo',
'dating',
'fly',
'hang',
'hop',
'imagine',
'named',
'retarded',
'save',
'spamming',
'stuck',
'thankyou',
'tried',
'won',
'young',
'20',
'8th',
'anyways',
'aw',
'cheese',
'cutie',
'eating',
'especially',
'green',
'isnt',
'jail',
'lmbo',

'lots',
'personality',
'single',
'tabitha',
'tho',
'ways',
'working',
'yahoo',
'beat',
'certain',
'class',
'cut',
'deep',
'fast',
'female',
'film',
'possible',
'taking',
'trouble',
'yeahh',
'apart',
'background',
'behind',
'bitchh',
'comment',
'felt',
'hands',
'ish',
'liked',
'personally',
'silly',
'star',
'sweetie',
'victor',
'waht',
'13',
'15',
'afraid',
'almost',
'bieber',
'biggest',
'boobs',
'bother',
'computer',
'cry',
'fool',
'full',

'girlfriend',
'light',
'months',
'seeing',
'shay',
'welcome',
'yess',
'account',
'aha',
'asleep',
'attention',
'bitches',
'brother',
'chomp',
'door',
'during',
'front',
'hand',
'instead',
'noe',
'opposite',
'partner',
'piece',
'scared',
'shall',
'step',
'theres',
'touch',
'uh',
'animal',
'austin',
'band',
'easily',
'hold',
'human',
'lovee',
'missed',
'pic',
'places',
'quit',
'sexual',
'shower',
'situation',
'somewhere',
'state',
'thoughts',
'ummmm',

'writing',
'yesterday',
'9',
'born',
'chill',
'control',
'creepy',
'cunt',
'dreams',
'drinking',
'dum',
'fair',
'fave',
'feelings',
'kool',
'middle',
'million',
'month',
'ohhh',
'prolly',
'road',
'songs',
'spell',
'stick',
'tew',
'understand',
'watched',
'yay',
'30',
'ate',
'become',
'boys',
'completely',
'delete',
'difference',
'dork',
'dress',
'easy',
'forever',
'french',
'fucked',
'glad',
'hahahah',
'hour',
'juss',
'la',
'language',

```
'meant',  
'mother',  
'naw',  
'ohm',  
'share',  
'sit',  
'starting',  
'thnx',  
'wasnt',  
'yall',  
'younger',  
'33333',  
'celebrity',  
'clothing',  
'club',  
'count',  
'crush',  
'cum',  
'driving',  
'fakes',  
'fell',  
'forget',  
'jeans',  
'kick',  
'looked',  
'moo',  
'set',  
'street',  
'weeks',  
'american',  
'cat',  
'changed',  
'college',  
'eva',  
'everyday',  
'ew',  
'fact',  
...]]}
```

[]: