Using Pre-trained Word Embeddings

In this notebook we will show some operations on pre-trained word embeddings to gain an intuition about them.

We will be using the pre-trained GloVe embeddings that can be found in the [official website](#). In particular, we will use the file `glove.6B.300d.txt` contained in this [zip file](#).

We will first load the GloVe embeddings using [Gensim](#). Specifically, we will use `KeyedVectors`'s `load_word2vec_format()` classmethod, which supports the original word2vec file format. However, there is a difference in the file formats used by GloVe and word2vec, which is a header used by word2vec to indicate the number of embeddings and dimensions stored in the file. The file that stores the GloVe embeddings doesn't have this header, so we will have to address that when loading the embeddings.

Loading the embeddings may take a little bit, so hang in there!

```python
from gensim.models import KeyedVectors

fname = "/kaggle/input/glove-fasttext-embedding-for-medium-articles/glove.6B.300d.txt"
glove = KeyedVectors.load_word2vec_format(fname, no_header=True)
glove.vectors.shape
```

Out[1]: `(400000, 300)`

Word similarity

One attribute of word embeddings that makes them useful is the ability to compare them using cosine similarity to find how similar they are. `KeyedVectors` objects provide a method called `most_similar()` that we can use to find the closest words to a particular word of interest. By default, `most_similar()` returns the 10 most similar words, but this can be changed using the `topn` parameter.

Below we test this function using a few different words.

```python
# common noun
glove.most_similar("cactus")
```

Out[2]:
```
[('cacti', 0.663456380367279),
 ('saguaro', 0.6195855140686035),
 ('pear', 0.5233485698699951),
 ('cactuses', 0.5178281664848328),
 ('prickly', 0.515631914138794),
 ('mesquite', 0.48448556661605835),
 ('opuntia', 0.4540084898471832),
 ('shrubs', 0.45362064242362976),
 ('peyote', 0.45344963669776917),
 ('succulents', 0.4512787461280823)]
```

```
In [3]:   # common noun
          glove.most_similar("cake")
```

```
Out[3]:   [('cakes', 0.7506030201911926),
           ('chocolate', 0.6965583562850952),
           ('dessert', 0.6440261006355286),
           ('pie', 0.608742892742157),
           ('cookies', 0.6082394123077393),
           ('frosting', 0.601721465587616),
           ('bread', 0.5954801440238953),
           ('cookie', 0.593381941318512),
           ('recipe', 0.5827102661132812),
           ('baked', 0.5819962620735168)]
```

```
In [4]:   # adjective
          glove.most_similar("angry")
```

```
Out[4]:   [('enraged', 0.7087873816490173),
           ('furious', 0.7078357934951782),
           ('irate', 0.6938743591308594),
           ('outraged', 0.6705204248428345),
           ('frustrated', 0.6515549421310425),
           ('angered', 0.635320246219635),
           ('provoked', 0.5827428102493286),
           ('annoyed', 0.581898033618927),
           ('incensed', 0.5751833319664001),
           ('indignant', 0.5704443454742432)]
```

```
In [5]:   # adverb
          glove.most_similar("quickly")
```

```
Out[5]:   [('soon', 0.766185998916626),
           ('rapidly', 0.7216640114784241),
           ('swiftly', 0.7197349667549133),
           ('eventually', 0.7043026685714722),
           ('finally', 0.6900882124900818),
           ('immediately', 0.6842609643936157),
           ('then', 0.6697486042976379),
           ('slowly', 0.6645645499229431),
           ('gradually', 0.6401675939559937),
           ('when', 0.6347666382789612)]
```

```
In [6]:   # preposition
          glove.most_similar("between")
```

```
Out[6]:   [('sides', 0.5867610573768616),
           ('both', 0.5843431949615479),
           ('two', 0.5652360916137695),
           ('differences', 0.514071524143219),
           ('which', 0.5120179057121277),
           ('conflict', 0.5115456581115723),
           ('relationship', 0.5022751092910767),
           ('and', 0.498425155878067),
           ('in', 0.4970666766166687),
           ('relations', 0.4970114529132843)]
```

```
In [7]:   # determiner
          glove.most_similar("the")
```

```
Out[7]:  [('of', 0.7057957649230957),
          ('which', 0.6992015838623047),
          ('this', 0.6747026443481445),
          ('part', 0.6727458238601685),
          ('same', 0.6592389345169067),
          ('its', 0.6446539759635925),
          ('first', 0.6398990750312805),
          ('in', 0.6361348032951355),
          ('one', 0.6245334148406982),
          ('that', 0.6176422834396362)]
```

Word analogies

Another characteristic of word embeddings is their ability to solve analogy problems. The same `most_similar()` method can be used for this task, by passing two lists of words: a `positive` list with the words that should be added and a `negative` list with the words that should be subtracted. Using these arguments, the famous example $\vec{king} - \vec{man} + \vec{woman} \approx \vec{queen}$ can be executed as follows:

```
In [8]:  # king - man + woman
         glove.most_similar(positive=["king", "woman"], negative=["man"])
```

```
Out[8]:  [('queen', 0.6713277101516724),
          ('princess', 0.5432624816894531),
          ('throne', 0.5386103987693787),
          ('monarch', 0.5347574949264526),
          ('daughter', 0.49802514910697937),
          ('mother', 0.49564430117607117),
          ('elizabeth', 0.4832652509212494),
          ('kingdom', 0.47747090458869934),
          ('prince', 0.4668239951133728),
          ('wife', 0.46473270654678345)]
```

Here are a few other interesting analogies:

```
In [9]:  # car - drive + fly
         glove.most_similar(positive=["car", "fly"], negative=["drive"])
```

```
Out[9]:  [('airplane', 0.5897148251533508),
          ('flying', 0.5675230026245117),
          ('plane', 0.5317023992538452),
          ('flies', 0.5172374248504639),
          ('flown', 0.514790415763855),
          ('airplanes', 0.5091356635093689),
          ('flew', 0.5011662244796753),
          ('planes', 0.4970923364162445),
          ('aircraft', 0.4957723915576935),
          ('helicopter', 0.45859551429748535)]
```

```
In [10]:  # berlin - germany + australia
          glove.most_similar(positive=["berlin", "australia"], negative=["germany"])
```

Out[10]:
```
[('sydney', 0.6780862212181091),
 ('melbourne', 0.6499180793762207),
 ('australian', 0.594883143901825),
 ('perth', 0.5828553438186646),
 ('canberra', 0.5610732436180115),
 ('brisbane', 0.5523110628128052),
 ('zealand', 0.5240115523338318),
 ('queensland', 0.5193883180618286),
 ('adelaide', 0.5027671456336975),
 ('london', 0.4644604027271271)]
```

In [11]:
```
# england - London + baghdad
glove.most_similar(positive=["england", "baghdad"], negative=["london"])
```

Out[11]:
```
[('iraq', 0.5320571660995483),
 ('fallujah', 0.4834090769290924),
 ('iraqi', 0.47287362813949585),
 ('mosul', 0.464663565158844),
 ('iraqis', 0.43555372953414917),
 ('najaf', 0.4352763295173645),
 ('baqouba', 0.42063194513320923),
 ('basra', 0.41905173659324646),
 ('samarra', 0.4125366508960724),
 ('saddam', 0.40791556239128113)]
```

In [12]:
```
# japan - yen + peso
glove.most_similar(positive=["japan", "peso"], negative=["yen"])
```

Out[12]:
```
[('mexico', 0.5726832151412964),
 ('philippines', 0.5445368885993958),
 ('peru', 0.48382261395454407),
 ('venezuela', 0.4816672205924988),
 ('brazil', 0.4664309620857239),
 ('argentina', 0.45490506291389465),
 ('philippine', 0.4417841136455536),
 ('chile', 0.43960973620414734),
 ('colombia', 0.4386259913444519),
 ('thailand', 0.43396785855293274)]
```

In [13]:
```
# best - good + tall
glove.most_similar(positive=["best", "tall"], negative=["good"])
```

Out[13]:
```
[('tallest', 0.5077419281005859),
 ('taller', 0.47616496682167053),
 ('height', 0.46000057458877563),
 ('metres', 0.4584786891937256),
 ('cm', 0.45212721824645996),
 ('meters', 0.44067248702049255),
 ('towering', 0.42784252762794495),
 ('centimeters', 0.4234543442726135),
 ('inches', 0.41745859384536743),
 ('erect', 0.4087314009666443)]
```

Looking under the hood

Now that we are more familiar with the `most_similar()` method, it is time to implement its functionality ourselves. But first, we need to take a look at the different parts of the

`KeyedVectors` object that we will need. Obviously, we will need the vectors themselves. They are stored in the `vectors` attribute.

```
In [14]:   glove.vectors.shape
```

```
Out[14]:   (400000, 300)
```

As we can see above, `vectors` is a 2-dimensional matrix with 400,000 rows and 300 columns. Each row corresponds to a 300-dimensional word embedding. These embeddings are not normalized, but normalized embeddings can be obtained using the `get_normed_vectors()` method.

```
In [15]:   normed_vectors = glove.get_normed_vectors()
           normed_vectors.shape
```

```
Out[15]:   (400000, 300)
```

Now we need to map the words in the vocabulary to rows in the `vectors` matrix, and vice versa. The `KeyedVectors` object has the attributes `index_to_key` and `key_to_index` which are a list of words and a dictionary of words to indices, respectively.

```
In [16]:   glove.index_to_key
```

```
Out[16]:  ['the',
           ',',
           '.',
           'of',
           'to',
           'and',
           'in',
           'a',
           '"',
           "'s",
           'for',
           '-',
           'that',
           'on',
           'is',
           'was',
           'said',
           'with',
           'he',
           'as',
           'it',
           'by',
           'at',
           '(',
           ')',
           'from',
           'his',
           "''",
           '``',
           'an',
           'be',
           'has',
           'are',
           'have',
           'but',
           'were',
           'not',
           'this',
           'who',
           'they',
           'had',
           'i',
           'which',
           'will',
           'their',
           ':',
           'or',
           'its',
           'one',
           'after',
           'new',
           'been',
           'also',
           'we',
           'would',
           'two',
           'more',
           "'",
           'first',
           'about',
```

```
'up',
'when',
'year',
'there',
'all',
'--',
'out',
'she',
'other',
'people',
"n't",
'her',
'percent',
'than',
'over',
'into',
'last',
'some',
'government',
'time',
'$',
'you',
'years',
'if',
'no',
'world',
'can',
'three',
'do',
';',
'president',
'only',
'state',
'million',
'could',
'us',
'most',
'_',
'against',
'u.s.',
'so',
'them',
'what',
'him',
'united',
'during',
'before',
'may',
'since',
'many',
'while',
'where',
'states',
'because',
'now',
'city',
'made',
'like',
'between',
'did',
```

```
            'just',
            'national',
            'day',
            'country',
            'under',
            'such',
            'second',
            'then',
            'company',
            'group',
            'any',
            'through',
            'china',
            'four',
            'being',
            'down',
            'war',
            'back',
            'off',
            'south',
            'american',
            'minister',
            'police',
            'well',
            'including',
            'team',
            'international',
            'week',
            'officials',
            'still',
            'both',
            'even',
            'high',
            'part',
            'told',
            'those',
            'end',
            'former',
            'these',
            'make',
            'billion',
            'work',
            'our',
            'home',
            'school',
            'party',
            'house',
            'old',
            'later',
            'get',
            'another',
            'tuesday',
            'news',
            'long',
            'five',
            'called',
            '1',
            'wednesday',
            'military',
            'way',
```

```
          'used',
          'much',
          'next',
          'monday',
          'thursday',
          'friday',
          'game',
          'here',
          '?',
          'should',
          'take',
          'very',
          'my',
          'north',
          'security',
          'season',
          'york',
          'how',
          'public',
          'early',
          'according',
          'several',
          'court',
          'say',
          'around',
          'foreign',
          '10',
          'until',
          'set',
          'political',
          'says',
          'market',
          'however',
          'family',
          'life',
          'same',
          'general',
          '-',
          'left',
          'good',
          'top',
          'university',
          'going',
          'number',
          'major',
          'known',
          'points',
          'won',
          'six',
          'month',
          'dollars',
          'bank',
          '2',
          'iraq',
          'use',
          'members',
          'each',
          'area',
          'found',
          'official',
```

```
                'sunday',
                'place',
                'go',
                'based',
                'among',
                'third',
                'times',
                'took',
                'right',
                'days',
                'local',
                'economic',
                'countries',
                'see',
                'best',
                'report',
                'killed',
                'held',
                'business',
                'west',
                'does',
                'own',
                '%',
                'came',
                'law',
                'months',
                'women',
                "'re",
                'power',
                'think',
                'service',
                'children',
                'bush',
                'show',
                '/',
                'help',
                'chief',
                'saturday',
                'system',
                'john',
                'support',
                'series',
                'play',
                'office',
                'following',
                'me',
                'meeting',
                'expected',
                'late',
                'washington',
                'games',
                'european',
                'league',
                'reported',
                'final',
                'added',
                'without',
                'british',
                'white',
                'history',
```

```
'man',
'men',
'became',
'want',
'march',
'case',
'few',
'run',
'money',
'began',
'open',
'name',
'trade',
'center',
'3',
'israel',
'oil',
'too',
'al',
'film',
'win',
'led',
'east',
'central',
'20',
'air',
'come',
'chinese',
'town',
'leader',
'army',
'line',
'never',
'little',
'played',
'prime',
'death',
'companies',
'least',
'put',
'forces',
'past',
'de',
'half',
'june',
'saying',
'know',
'federal',
'french',
'peace',
'earlier',
'capital',
'force',
'great',
'union',
'near',
'released',
'small',
'department',
'every',
```

```
                    'health',
                    'japan',
                    'head',
                    'ago',
                    'night',
                    'big',
                    'cup',
                    'election',
                    'region',
                    'director',
                    'talks',
                    'program',
                    'far',
                    'today',
                    'statement',
                    'july',
                    'although',
                    'district',
                    'again',
                    'born',
                    'development',
                    'leaders',
                    'council',
                    'close',
                    'record',
                    'along',
                    'county',
                    'france',
                    'went',
                    'point',
                    'must',
                    'spokesman',
                    'your',
                    'member',
                    'plan',
                    'financial',
                    'april',
                    'recent',
                    'campaign',
                    'become',
                    'troops',
                    'whether',
                    'lost',
                    'music',
                    '15',
                    'got',
                    'israeli',
                    '30',
                    'need',
                    '4',
                    'lead',
                    'already',
                    'russia',
                    'though',
                    'might',
                    'free',
                    'hit',
                    'rights',
                    '11',
                    'information',
```

```
'away',
'12',
'5',
'others',
'control',
'within',
'large',
'economy',
'press',
'agency',
'water',
'died',
'career',
'making',
'...',
'deal',
'attack',
'side',
'seven',
'better',
'less',
'september',
'once',
'clinton',
'main',
'due',
'committee',
'building',
'conference',
'club',
'january',
'decision',
'stock',
'america',
'given',
'give',
'often',
'announced',
'television',
'industry',
'order',
'young',
"'ve",
'palestinian',
'age',
'start',
'administration',
'russian',
'prices',
'round',
'december',
'nations',
"'m",
'human',
'india',
'defense',
'asked',
'total',
'october',
'players',
```

```
'bill',
'important',
'southern',
'move',
'fire',
'population',
'rose',
'november',
'include',
'further',
'nuclear',
'street',
'taken',
'media',
'different',
'issue',
'received',
'secretary',
'return',
'college',
'working',
'community',
'eight',
'groups',
'despite',
'level',
'largest',
'whose',
'attacks',
'germany',
'august',
'change',
'church',
'nation',
'german',
'station',
'london',
'weeks',
'having',
'18',
'research',
'black',
'services',
'story',
'6',
'europe',
'sales',
'policy',
'visit',
'northern',
'lot',
'across',
'per',
'current',
'board',
'football',
'ministry',
'workers',
'vote',
'book',
```

```
'fell',
'seen',
'role',
'students',
'shares',
'iran',
'process',
'agreement',
'quarter',
'full',
'match',
'started',
'growth',
'yet',
'moved',
'possible',
'western',
'special',
'100',
'plans',
'interest',
'behind',
'strong',
'england',
'named',
'food',
'period',
'real',
'authorities',
'car',
'term',
'rate',
'race',
'nearly',
'korea',
'enough',
'site',
'opposition',
'keep',
'25',
'call',
'future',
'taking',
'island',
'2008',
'2006',
'road',
'outside',
'really',
'century',
'democratic',
'almost',
'single',
'share',
'leading',
'trying',
'find',
'album',
'senior',
'minutes',
```

```
'together',
'congress',
'index',
'australia',
'results',
'hard',
'hours',
'land',
'action',
'higher',
'field',
'cut',
'coach',
'elections',
'san',
'issues',
'executive',
'february',
'production',
'areas',
'river',
'face',
'using',
'japanese',
'province',
'park',
'price',
'commission',
'california',
'father',
'son',
'education',
'7',
'village',
'energy',
'shot',
'short',
'africa',
'key',
'red',
'association',
'average',
'pay',
'exchange',
'eu',
'something',
'gave',
'likely',
'player',
'george',
'2007',
'victory',
'8',
'low',
'things',
'2010',
'pakistan',
'14',
'post',
'social',
```

```
'continue',
'ever',
'look',
'chairman',
'job',
'2000',
'soldiers',
'able',
'parliament',
'front',
'himself',
'problems',
'private',
'lower',
'list',
'built',
'13',
'efforts',
'dollar',
'miles',
'included',
'radio',
'live',
'form',
'david',
'african',
'increase',
'reports',
'sent',
'fourth',
'always',
'king',
'50',
'tax',
'taiwan',
'britain',
'16',
'playing',
'title',
'middle',
'meet',
'global',
'wife',
'2009',
'position',
'located',
'clear',
'ahead',
'2004',
'2005',
'iraqi',
'english',
'result',
'release',
'violence',
'goal',
'project',
'closed',
'border',
'body',
```

```
                    'soon',
                    'crisis',
                    'division',
                    '&amp;',
                    'served',
                    'tour',
                    'hospital',
                    'kong',
                    'test',
                    'hong',
                    'u.n.',
                    'inc.',
                    'technology',
                    'believe',
                    'organization',
                    'published',
                    'weapons',
                    'agreed',
                    'why',
                    'nine',
                    'summer',
                    'wanted',
                    'republican',
                    'act',
                    'recently',
                    'texas',
                    'course',
                    'problem',
                    'senate',
                    'medical',
                    'un',
                    'done',
                    'reached',
                    'star',
                    'continued',
                    'investors',
                    'living',
                    'care',
                    'signed',
                    '17',
                    'art',
                    'provide',
                    'worked',
                    'presidential',
                    'gold',
                    'obama',
                    'morning',
                    'dead',
                    'opened',
                    "'ll",
                    'event',
                    'previous',
                    'cost',
                    'instead',
                    'canada',
                    'band',
                    'teams',
                    'daily',
                    '2001',
                    'available',
```

```
'drug',
'coming',
'2003',
'investment',
'’s',
'michael',
'civil',
'woman',
'training',
'appeared',
'9',
'involved',
'indian',
'similar',
'situation',
'24',
'los',
'running',
'fighting',
'mark',
'40',
'trial',
'hold',
'australian',
'thought',
'!',
'study',
'fall',
'mother',
'met',
'relations',
'anti',
'2002',
'song',
'popular',
'base',
'tv',
'ground',
'markets',
'ii',
'newspaper',
'staff',
'saw',
'hand',
'hope',
'operations',
'pressure',
'americans',
'eastern',
'st.',
'legal',
'asia',
'budget',
'returned',
'considered',
'love',
'wrote',
'stop',
'fight',
'currently',
```

```
'charges',
'try',
'aid',
'ended',
'management',
'brought',
'cases',
'decided',
'failed',
'network',
'works',
'gas',
'turned',
'fact',
'vice',
'ca',
'mexico',
'trading',
'especially',
'reporters',
'afghanistan',
'common',
'looking',
'space',
'rates',
'manager',
'loss',
'2011',
'justice',
'thousands',
'james',
'rather',
'fund',
'thing',
'republic',
'opening',
'accused',
'winning',
'scored',
'championship',
'example',
'getting',
'biggest',
'performance',
'sports',
'1998',
'let',
'allowed',
'schools',
'means',
'turn',
'leave',
'no.',
'robert',
'personal',
'stocks',
'showed',
'light',
'arrested',
'person',
```

```
'either',
'offer',
'majority',
'battle',
'19',
'class',
'evidence',
'makes',
'society',
'products',
'regional',
'needed',
'stage',
'am',
'doing',
'families',
'construction',
'various',
'1996',
'sold',
'independent',
'kind',
'airport',
'paul',
'judge',
'internet',
'movement',
'room',
'followed',
'original',
'angeles',
'italy',
'`',
'data',
'comes',
'parties',
'nothing',
'sea',
'bring',
'2012',
'annual',
'officer',
'beijing',
'present',
'remain',
'nato',
'1999',
'22',
'remains',
'allow',
'florida',
'computer',
'21',
'contract',
'coast',
'created',
'demand',
'operation',
'events',
'islamic',
```

```
                'beat',
                'analysts',
                'interview',
                'helped',
                'child',
                'probably',
                'spent',
                'asian',
                'effort',
                'cooperation',
                'shows',
                'calls',
                'investigation',
                'lives',
                'video',
                'yen',
                'runs',
                'tried',
                'bad',
                'described',
                '1994',
                'toward',
                'written',
                'throughout',
                'established',
                'mission',
                'associated',
                'buy',
                'growing',
                'green',
                'forward',
                'competition',
                'poor',
                'latest',
                'banks',
                'question',
                '1997',
                'prison',
                'feel',
                'attention',
                ...]
```

In [17]: `glove.key_to_index`

```
Out[17]:  {'the': 0,
           ',': 1,
           '.': 2,
           'of': 3,
           'to': 4,
           'and': 5,
           'in': 6,
           'a': 7,
           '"': 8,
           "'s": 9,
           'for': 10,
           '-': 11,
           'that': 12,
           'on': 13,
           'is': 14,
           'was': 15,
           'said': 16,
           'with': 17,
           'he': 18,
           'as': 19,
           'it': 20,
           'by': 21,
           'at': 22,
           '(': 23,
           ')': 24,
           'from': 25,
           'his': 26,
           "''": 27,
           '``': 28,
           'an': 29,
           'be': 30,
           'has': 31,
           'are': 32,
           'have': 33,
           'but': 34,
           'were': 35,
           'not': 36,
           'this': 37,
           'who': 38,
           'they': 39,
           'had': 40,
           'i': 41,
           'which': 42,
           'will': 43,
           'their': 44,
           ':': 45,
           'or': 46,
           'its': 47,
           'one': 48,
           'after': 49,
           'new': 50,
           'been': 51,
           'also': 52,
           'we': 53,
           'would': 54,
           'two': 55,
           'more': 56,
           "'": 57,
           'first': 58,
           'about': 59,
```

```
            'up': 60,
            'when': 61,
            'year': 62,
            'there': 63,
            'all': 64,
            '--': 65,
            'out': 66,
            'she': 67,
            'other': 68,
            'people': 69,
            "n't": 70,
            'her': 71,
            'percent': 72,
            'than': 73,
            'over': 74,
            'into': 75,
            'last': 76,
            'some': 77,
            'government': 78,
            'time': 79,
            '$': 80,
            'you': 81,
            'years': 82,
            'if': 83,
            'no': 84,
            'world': 85,
            'can': 86,
            'three': 87,
            'do': 88,
            ';': 89,
            'president': 90,
            'only': 91,
            'state': 92,
            'million': 93,
            'could': 94,
            'us': 95,
            'most': 96,
            '_': 97,
            'against': 98,
            'u.s.': 99,
            'so': 100,
            'them': 101,
            'what': 102,
            'him': 103,
            'united': 104,
            'during': 105,
            'before': 106,
            'may': 107,
            'since': 108,
            'many': 109,
            'while': 110,
            'where': 111,
            'states': 112,
            'because': 113,
            'now': 114,
            'city': 115,
            'made': 116,
            'like': 117,
            'between': 118,
            'did': 119,
```

```
'just': 120,
'national': 121,
'day': 122,
'country': 123,
'under': 124,
'such': 125,
'second': 126,
'then': 127,
'company': 128,
'group': 129,
'any': 130,
'through': 131,
'china': 132,
'four': 133,
'being': 134,
'down': 135,
'war': 136,
'back': 137,
'off': 138,
'south': 139,
'american': 140,
'minister': 141,
'police': 142,
'well': 143,
'including': 144,
'team': 145,
'international': 146,
'week': 147,
'officials': 148,
'still': 149,
'both': 150,
'even': 151,
'high': 152,
'part': 153,
'told': 154,
'those': 155,
'end': 156,
'former': 157,
'these': 158,
'make': 159,
'billion': 160,
'work': 161,
'our': 162,
'home': 163,
'school': 164,
'party': 165,
'house': 166,
'old': 167,
'later': 168,
'get': 169,
'another': 170,
'tuesday': 171,
'news': 172,
'long': 173,
'five': 174,
'called': 175,
'1': 176,
'wednesday': 177,
'military': 178,
'way': 179,
```

```
'used': 180,
'much': 181,
'next': 182,
'monday': 183,
'thursday': 184,
'friday': 185,
'game': 186,
'here': 187,
'?': 188,
'should': 189,
'take': 190,
'very': 191,
'my': 192,
'north': 193,
'security': 194,
'season': 195,
'york': 196,
'how': 197,
'public': 198,
'early': 199,
'according': 200,
'several': 201,
'court': 202,
'say': 203,
'around': 204,
'foreign': 205,
'10': 206,
'until': 207,
'set': 208,
'political': 209,
'says': 210,
'market': 211,
'however': 212,
'family': 213,
'life': 214,
'same': 215,
'general': 216,
'-': 217,
'left': 218,
'good': 219,
'top': 220,
'university': 221,
'going': 222,
'number': 223,
'major': 224,
'known': 225,
'points': 226,
'won': 227,
'six': 228,
'month': 229,
'dollars': 230,
'bank': 231,
'2': 232,
'iraq': 233,
'use': 234,
'members': 235,
'each': 236,
'area': 237,
'found': 238,
'official': 239,
```

```
'sunday': 240,
'place': 241,
'go': 242,
'based': 243,
'among': 244,
'third': 245,
'times': 246,
'took': 247,
'right': 248,
'days': 249,
'local': 250,
'economic': 251,
'countries': 252,
'see': 253,
'best': 254,
'report': 255,
'killed': 256,
'held': 257,
'business': 258,
'west': 259,
'does': 260,
'own': 261,
'%': 262,
'came': 263,
'law': 264,
'months': 265,
'women': 266,
"'re": 267,
'power': 268,
'think': 269,
'service': 270,
'children': 271,
'bush': 272,
'show': 273,
'/': 274,
'help': 275,
'chief': 276,
'saturday': 277,
'system': 278,
'john': 279,
'support': 280,
'series': 281,
'play': 282,
'office': 283,
'following': 284,
'me': 285,
'meeting': 286,
'expected': 287,
'late': 288,
'washington': 289,
'games': 290,
'european': 291,
'league': 292,
'reported': 293,
'final': 294,
'added': 295,
'without': 296,
'british': 297,
'white': 298,
'history': 299,
```

```
'man': 300,
'men': 301,
'became': 302,
'want': 303,
'march': 304,
'case': 305,
'few': 306,
'run': 307,
'money': 308,
'began': 309,
'open': 310,
'name': 311,
'trade': 312,
'center': 313,
'3': 314,
'israel': 315,
'oil': 316,
'too': 317,
'al': 318,
'film': 319,
'win': 320,
'led': 321,
'east': 322,
'central': 323,
'20': 324,
'air': 325,
'come': 326,
'chinese': 327,
'town': 328,
'leader': 329,
'army': 330,
'line': 331,
'never': 332,
'little': 333,
'played': 334,
'prime': 335,
'death': 336,
'companies': 337,
'least': 338,
'put': 339,
'forces': 340,
'past': 341,
'de': 342,
'half': 343,
'june': 344,
'saying': 345,
'know': 346,
'federal': 347,
'french': 348,
'peace': 349,
'earlier': 350,
'capital': 351,
'force': 352,
'great': 353,
'union': 354,
'near': 355,
'released': 356,
'small': 357,
'department': 358,
'every': 359,
```

```
'health': 360,
'japan': 361,
'head': 362,
'ago': 363,
'night': 364,
'big': 365,
'cup': 366,
'election': 367,
'region': 368,
'director': 369,
'talks': 370,
'program': 371,
'far': 372,
'today': 373,
'statement': 374,
'july': 375,
'although': 376,
'district': 377,
'again': 378,
'born': 379,
'development': 380,
'leaders': 381,
'council': 382,
'close': 383,
'record': 384,
'along': 385,
'county': 386,
'france': 387,
'went': 388,
'point': 389,
'must': 390,
'spokesman': 391,
'your': 392,
'member': 393,
'plan': 394,
'financial': 395,
'april': 396,
'recent': 397,
'campaign': 398,
'become': 399,
'troops': 400,
'whether': 401,
'lost': 402,
'music': 403,
'15': 404,
'got': 405,
'israeli': 406,
'30': 407,
'need': 408,
'4': 409,
'lead': 410,
'already': 411,
'russia': 412,
'though': 413,
'might': 414,
'free': 415,
'hit': 416,
'rights': 417,
'11': 418,
'information': 419,
```

```
'away': 420,
'12': 421,
'5': 422,
'others': 423,
'control': 424,
'within': 425,
'large': 426,
'economy': 427,
'press': 428,
'agency': 429,
'water': 430,
'died': 431,
'career': 432,
'making': 433,
'...': 434,
'deal': 435,
'attack': 436,
'side': 437,
'seven': 438,
'better': 439,
'less': 440,
'september': 441,
'once': 442,
'clinton': 443,
'main': 444,
'due': 445,
'committee': 446,
'building': 447,
'conference': 448,
'club': 449,
'january': 450,
'decision': 451,
'stock': 452,
'america': 453,
'given': 454,
'give': 455,
'often': 456,
'announced': 457,
'television': 458,
'industry': 459,
'order': 460,
'young': 461,
"'ve": 462,
'palestinian': 463,
'age': 464,
'start': 465,
'administration': 466,
'russian': 467,
'prices': 468,
'round': 469,
'december': 470,
'nations': 471,
"'m": 472,
'human': 473,
'india': 474,
'defense': 475,
'asked': 476,
'total': 477,
'october': 478,
'players': 479,
```

```
'bill': 480,
'important': 481,
'southern': 482,
'move': 483,
'fire': 484,
'population': 485,
'rose': 486,
'november': 487,
'include': 488,
'further': 489,
'nuclear': 490,
'street': 491,
'taken': 492,
'media': 493,
'different': 494,
'issue': 495,
'received': 496,
'secretary': 497,
'return': 498,
'college': 499,
'working': 500,
'community': 501,
'eight': 502,
'groups': 503,
'despite': 504,
'level': 505,
'largest': 506,
'whose': 507,
'attacks': 508,
'germany': 509,
'august': 510,
'change': 511,
'church': 512,
'nation': 513,
'german': 514,
'station': 515,
'london': 516,
'weeks': 517,
'having': 518,
'18': 519,
'research': 520,
'black': 521,
'services': 522,
'story': 523,
'6': 524,
'europe': 525,
'sales': 526,
'policy': 527,
'visit': 528,
'northern': 529,
'lot': 530,
'across': 531,
'per': 532,
'current': 533,
'board': 534,
'football': 535,
'ministry': 536,
'workers': 537,
'vote': 538,
'book': 539,
```

```
            'fell': 540,
            'seen': 541,
            'role': 542,
            'students': 543,
            'shares': 544,
            'iran': 545,
            'process': 546,
            'agreement': 547,
            'quarter': 548,
            'full': 549,
            'match': 550,
            'started': 551,
            'growth': 552,
            'yet': 553,
            'moved': 554,
            'possible': 555,
            'western': 556,
            'special': 557,
            '100': 558,
            'plans': 559,
            'interest': 560,
            'behind': 561,
            'strong': 562,
            'england': 563,
            'named': 564,
            'food': 565,
            'period': 566,
            'real': 567,
            'authorities': 568,
            'car': 569,
            'term': 570,
            'rate': 571,
            'race': 572,
            'nearly': 573,
            'korea': 574,
            'enough': 575,
            'site': 576,
            'opposition': 577,
            'keep': 578,
            '25': 579,
            'call': 580,
            'future': 581,
            'taking': 582,
            'island': 583,
            '2008': 584,
            '2006': 585,
            'road': 586,
            'outside': 587,
            'really': 588,
            'century': 589,
            'democratic': 590,
            'almost': 591,
            'single': 592,
            'share': 593,
            'leading': 594,
            'trying': 595,
            'find': 596,
            'album': 597,
            'senior': 598,
            'minutes': 599,
```

```
'together': 600,
'congress': 601,
'index': 602,
'australia': 603,
'results': 604,
'hard': 605,
'hours': 606,
'land': 607,
'action': 608,
'higher': 609,
'field': 610,
'cut': 611,
'coach': 612,
'elections': 613,
'san': 614,
'issues': 615,
'executive': 616,
'february': 617,
'production': 618,
'areas': 619,
'river': 620,
'face': 621,
'using': 622,
'japanese': 623,
'province': 624,
'park': 625,
'price': 626,
'commission': 627,
'california': 628,
'father': 629,
'son': 630,
'education': 631,
'7': 632,
'village': 633,
'energy': 634,
'shot': 635,
'short': 636,
'africa': 637,
'key': 638,
'red': 639,
'association': 640,
'average': 641,
'pay': 642,
'exchange': 643,
'eu': 644,
'something': 645,
'gave': 646,
'likely': 647,
'player': 648,
'george': 649,
'2007': 650,
'victory': 651,
'8': 652,
'low': 653,
'things': 654,
'2010': 655,
'pakistan': 656,
'14': 657,
'post': 658,
'social': 659,
```

```
'continue': 660,
'ever': 661,
'look': 662,
'chairman': 663,
'job': 664,
'2000': 665,
'soldiers': 666,
'able': 667,
'parliament': 668,
'front': 669,
'himself': 670,
'problems': 671,
'private': 672,
'lower': 673,
'list': 674,
'built': 675,
'13': 676,
'efforts': 677,
'dollar': 678,
'miles': 679,
'included': 680,
'radio': 681,
'live': 682,
'form': 683,
'david': 684,
'african': 685,
'increase': 686,
'reports': 687,
'sent': 688,
'fourth': 689,
'always': 690,
'king': 691,
'50': 692,
'tax': 693,
'taiwan': 694,
'britain': 695,
'16': 696,
'playing': 697,
'title': 698,
'middle': 699,
'meet': 700,
'global': 701,
'wife': 702,
'2009': 703,
'position': 704,
'located': 705,
'clear': 706,
'ahead': 707,
'2004': 708,
'2005': 709,
'iraqi': 710,
'english': 711,
'result': 712,
'release': 713,
'violence': 714,
'goal': 715,
'project': 716,
'closed': 717,
'border': 718,
'body': 719,
```

```
'soon': 720,
'crisis': 721,
'division': 722,
'&amp;': 723,
'served': 724,
'tour': 725,
'hospital': 726,
'kong': 727,
'test': 728,
'hong': 729,
'u.n.': 730,
'inc.': 731,
'technology': 732,
'believe': 733,
'organization': 734,
'published': 735,
'weapons': 736,
'agreed': 737,
'why': 738,
'nine': 739,
'summer': 740,
'wanted': 741,
'republican': 742,
'act': 743,
'recently': 744,
'texas': 745,
'course': 746,
'problem': 747,
'senate': 748,
'medical': 749,
'un': 750,
'done': 751,
'reached': 752,
'star': 753,
'continued': 754,
'investors': 755,
'living': 756,
'care': 757,
'signed': 758,
'17': 759,
'art': 760,
'provide': 761,
'worked': 762,
'presidential': 763,
'gold': 764,
'obama': 765,
'morning': 766,
'dead': 767,
'opened': 768,
"'ll": 769,
'event': 770,
'previous': 771,
'cost': 772,
'instead': 773,
'canada': 774,
'band': 775,
'teams': 776,
'daily': 777,
'2001': 778,
'available': 779,
```

```
                    'drug': 780,
                    'coming': 781,
                    '2003': 782,
                    'investment': 783,
                    ''s': 784,
                    'michael': 785,
                    'civil': 786,
                    'woman': 787,
                    'training': 788,
                    'appeared': 789,
                    '9': 790,
                    'involved': 791,
                    'indian': 792,
                    'similar': 793,
                    'situation': 794,
                    '24': 795,
                    'los': 796,
                    'running': 797,
                    'fighting': 798,
                    'mark': 799,
                    '40': 800,
                    'trial': 801,
                    'hold': 802,
                    'australian': 803,
                    'thought': 804,
                    '!': 805,
                    'study': 806,
                    'fall': 807,
                    'mother': 808,
                    'met': 809,
                    'relations': 810,
                    'anti': 811,
                    '2002': 812,
                    'song': 813,
                    'popular': 814,
                    'base': 815,
                    'tv': 816,
                    'ground': 817,
                    'markets': 818,
                    'ii': 819,
                    'newspaper': 820,
                    'staff': 821,
                    'saw': 822,
                    'hand': 823,
                    'hope': 824,
                    'operations': 825,
                    'pressure': 826,
                    'americans': 827,
                    'eastern': 828,
                    'st.': 829,
                    'legal': 830,
                    'asia': 831,
                    'budget': 832,
                    'returned': 833,
                    'considered': 834,
                    'love': 835,
                    'wrote': 836,
                    'stop': 837,
                    'fight': 838,
                    'currently': 839,
```

```
'charges': 840,
'try': 841,
'aid': 842,
'ended': 843,
'management': 844,
'brought': 845,
'cases': 846,
'decided': 847,
'failed': 848,
'network': 849,
'works': 850,
'gas': 851,
'turned': 852,
'fact': 853,
'vice': 854,
'ca': 855,
'mexico': 856,
'trading': 857,
'especially': 858,
'reporters': 859,
'afghanistan': 860,
'common': 861,
'looking': 862,
'space': 863,
'rates': 864,
'manager': 865,
'loss': 866,
'2011': 867,
'justice': 868,
'thousands': 869,
'james': 870,
'rather': 871,
'fund': 872,
'thing': 873,
'republic': 874,
'opening': 875,
'accused': 876,
'winning': 877,
'scored': 878,
'championship': 879,
'example': 880,
'getting': 881,
'biggest': 882,
'performance': 883,
'sports': 884,
'1998': 885,
'let': 886,
'allowed': 887,
'schools': 888,
'means': 889,
'turn': 890,
'leave': 891,
'no.': 892,
'robert': 893,
'personal': 894,
'stocks': 895,
'showed': 896,
'light': 897,
'arrested': 898,
'person': 899,
```

```
'either': 900,
'offer': 901,
'majority': 902,
'battle': 903,
'19': 904,
'class': 905,
'evidence': 906,
'makes': 907,
'society': 908,
'products': 909,
'regional': 910,
'needed': 911,
'stage': 912,
'am': 913,
'doing': 914,
'families': 915,
'construction': 916,
'various': 917,
'1996': 918,
'sold': 919,
'independent': 920,
'kind': 921,
'airport': 922,
'paul': 923,
'judge': 924,
'internet': 925,
'movement': 926,
'room': 927,
'followed': 928,
'original': 929,
'angeles': 930,
'italy': 931,
'`': 932,
'data': 933,
'comes': 934,
'parties': 935,
'nothing': 936,
'sea': 937,
'bring': 938,
'2012': 939,
'annual': 940,
'officer': 941,
'beijing': 942,
'present': 943,
'remain': 944,
'nato': 945,
'1999': 946,
'22': 947,
'remains': 948,
'allow': 949,
'florida': 950,
'computer': 951,
'21': 952,
'contract': 953,
'coast': 954,
'created': 955,
'demand': 956,
'operation': 957,
'events': 958,
'islamic': 959,
```

```
  'beat': 960,
  'analysts': 961,
  'interview': 962,
  'helped': 963,
  'child': 964,
  'probably': 965,
  'spent': 966,
  'asian': 967,
  'effort': 968,
  'cooperation': 969,
  'shows': 970,
  'calls': 971,
  'investigation': 972,
  'lives': 973,
  'video': 974,
  'yen': 975,
  'runs': 976,
  'tried': 977,
  'bad': 978,
  'described': 979,
  '1994': 980,
  'toward': 981,
  'written': 982,
  'throughout': 983,
  'established': 984,
  'mission': 985,
  'associated': 986,
  'buy': 987,
  'growing': 988,
  'green': 989,
  'forward': 990,
  'competition': 991,
  'poor': 992,
  'latest': 993,
  'banks': 994,
  'question': 995,
  '1997': 996,
  'prison': 997,
  'feel': 998,
  'attention': 999,
  ...}
```

Word similarity from scratch

Now we have everything we need to implement a `most_similar_words()` function that takes a word, the vector matrix, the `index_to_key` list, and the `key_to_index` dictionary. This function will return the 10 most similar words to the provided word, along with their similarity scores.

In [19]:
```python
import numpy as np

def most_similar_words(word, vectors, index_to_key, key_to_index, topn=10):
    # Retrieve word_id corresponding to the given word
    word_id = key_to_index.get(word)
    if word_id is None:
        raise ValueError(f"Word '{word}' not found in the vocabulary.")
```

```python
        # Retrieve embedding for the given word
        word_vector = vectors[word_id]

        # Calculate similarities to all words in our vocabulary (hint: use @)
        similarities = vectors @ word_vector.T   # Matrix multiplication to get cosine simi

        # Get word_ids in ascending order with respect to similarity score
        similar_word_ids = np.argsort(similarities)[::-1]   # Sort in descending order

        # Get boolean array with element corresponding to word_id set to false
        is_word_id = np.arange(len(index_to_key)) != word_id   # All ids except for the tar

        # Obtain new array of indices that doesn't contain word_id
        filtered_indices = similar_word_ids[is_word_id][0:topn]   # Filter and take topn

        # Get topn word_ids
        top_words_ids = filtered_indices[:topn]

        # Retrieve topn words with their corresponding similarity score
        top_words = [(index_to_key[idx], similarities[idx]) for idx in top_words_ids]

        # Return results
        return top_words
```

Now let's try the same example that we used above: the most similar words to "cactus".

```python
In [20]:   vectors = glove.get_normed_vectors()
           index_to_key = glove.index_to_key
           key_to_index = glove.key_to_index
           most_similar_words("cactus", vectors, index_to_key, key_to_index)
```

```python
Out[20]:   [('cactus', 1.0000001),
            ('cacti', 0.66345644),
            ('saguaro', 0.6195854),
            ('pear', 0.5233487),
            ('cactuses', 0.5178282),
            ('prickly', 0.51563185),
            ('mesquite', 0.4844855),
            ('opuntia', 0.45400843),
            ('shrubs', 0.45362067),
            ('peyote', 0.4534496)]
```

Analogies from scratch

The `most_similar_words()` function behaves as expected. Now let's implement a function to perform the analogy task. We will give it the very creative name `analogy`. This function will get two lists of words (one for positive words and one for negative words), just like the `most_similar()` method we discussed above.

```python
In [22]:   from numpy.linalg import norm
           import numpy as np

           def analogy(positive, negative, vectors, index_to_key, key_to_index, topn=10):
               # Find ids for positive and negative words
               pos_ids = [key_to_index[word] for word in positive]
               neg_ids = [key_to_index[word] for word in negative]
```

```python
        # Get embeddings for positive and negative words
        pos_emb = np.sum(vectors[pos_ids], axis=0)  # Sum embeddings for positive words
        neg_emb = np.sum(vectors[neg_ids], axis=0)  # Sum embeddings for negative words

        # Get embedding for analogy
        emb = pos_emb - neg_emb  # Create the analogy vector

        # Normalize embedding
        emb = emb / norm(emb)  # Normalize the vector

        # Calculate similarities to all words in our vocabulary
        similarities = vectors @ emb.T  # Matrix multiplication to get cosine similarities

        # Get word_ids in ascending order with respect to similarity score
        ids_ascending = np.argsort(similarities)

        # Reverse word_ids
        ids_descending = ids_ascending[::-1]  # Sort in descending order

        # Get boolean array with elements corresponding to any of the given_word_ids set t
        given_word_ids = pos_ids + neg_ids
        given_words_mask = np.isin(np.arange(len(index_to_key)), given_word_ids, invert=Tr

        # Obtain new array of indices that doesn't contain any of the given_word_ids
        filtered_ids = ids_descending[given_words_mask]

        # Get topn word_ids
        top_ids = filtered_ids[:topn]

        # Retrieve topn words with their corresponding similarity scores
        top_words = [(index_to_key[idx], similarities[idx]) for idx in top_ids]

        # Return results
        return top_words
```

Let's try this function with the $\vec{king} - \vec{man} + \vec{woman} \approx \vec{queen}$ example we discussed above.

```
In [23]:  positive = ["king", "woman"]
          negative = ["man"]
          vectors = glove.get_normed_vectors()
          index_to_key = glove.index_to_key
          key_to_index = glove.key_to_index
          analogy(positive, negative, vectors, index_to_key, key_to_index)
```

```
Out[23]:  [('king', 0.75726104),
           ('queen', 0.67132777),
           ('princess', 0.5432625),
           ('throne', 0.5386105),
           ('monarch', 0.53475755),
           ('daughter', 0.49802518),
           ('mother', 0.49564433),
           ('elizabeth', 0.48326525),
           ('kingdom', 0.47747087),
           ('prince', 0.466824)]
```

```
In [ ]:
```