CSCI 4140: Natural Language Processing

CSCI/DASC 6040: Computational Analysis of Natural Languages

Spring 2023

Homework 3 - Exploring word vectors

Due Sunday, February 26, at 11:59 PM

In [2]:
```
# All Import Statements Defined Here
# Note: Do not add to this list.

import sys
assert sys.version_info[0]==3
assert sys.version_info[1] >= 5

from gensim.models import KeyedVectors
from gensim.test.utils import datapath
import pprint
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [10, 5]
import nltk
nltk.download('reuters')
from nltk.corpus import reuters
import numpy as np
import random
import scipy as sp
from sklearn.decomposition import TruncatedSVD
from sklearn.decomposition import PCA

START_TOKEN = '<START>'
END_TOKEN = '<END>'

np.random.seed(0)
random.seed(0)
# ----------------
```

```
[nltk_data] Downloading package reuters to
[nltk_data]     C:\Users\owens\AppData\Roaming\nltk_data...
[nltk_data]   Package reuters is already up-to-date!
```

Word Vectors

Word Vectors are often used as a fundamental component for downstream NLP tasks, e.g. question answering, text generation, translation, etc., so it is important to build some intuitions as to their strengths and weaknesses. Here, you will explore two types of word vectors: those derived from *co-occurrence matrices*, and those derived via *word2vec*.

Note on Terminology: The terms "word vectors" and "word embeddings" are often used interchangeably. The term "embedding" refers to the fact that we are encoding aspects of a word's meaning in a lower dimensional space. As [Wikipedia (https://en.wikipedia.org/wiki/Word_embedding)](https://en.wikipedia.org/wiki/Word_embedding) states, "*conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with a much lower dimension*".

Part 1: Count-Based Word Vectors (40 points)

Most word vector models start from the following idea:

You shall know a word by the company it keeps ([Firth, J. R. 1957:11 (https://en.wikipedia.org/wiki/John_Rupert_Firth)](https://en.wikipedia.org/wiki/John_Rupert_Firth))

Many word vector implementations are driven by the idea that similar words, i.e., (near) synonyms, will be used in similar contexts. As a result, similar words will often be spoken or written along with a shared subset of words, i.e., contexts. By examining these contexts, we can try to develop embeddings for our words. With this intuition in mind, many "old school" approaches to constructing word vectors relied on word counts. Here we elaborate upon one of those strategies, *co-occurrence matrices* (for more information, see [Word embedding (https://medium.com/data-science-group-iitr/word-embedding-2d05d270b285)](https://medium.com/data-science-group-iitr/word-embedding-2d05d270b285)).

Co-Occurrence

A co-occurrence matrix counts how often things co-occur in some environment. Given some word w_i occurring in the document, we consider the *context window* surrounding w_i. Supposing our fixed window size is n, then this is the n preceding and n subsequent words in that document, i.e. words $w_{i-n} \dots w_{i-1}$ and $w_{i+1} \dots w_{i+n}$. We build a *co-occurrence matrix* M, which is a symmetric word-by-word matrix in which M_{ij} is the number of times w_j appears inside w_i's window.

Example: Co-Occurrence with Fixed Window of n=1:

Document 1: "all that glitters is not gold"

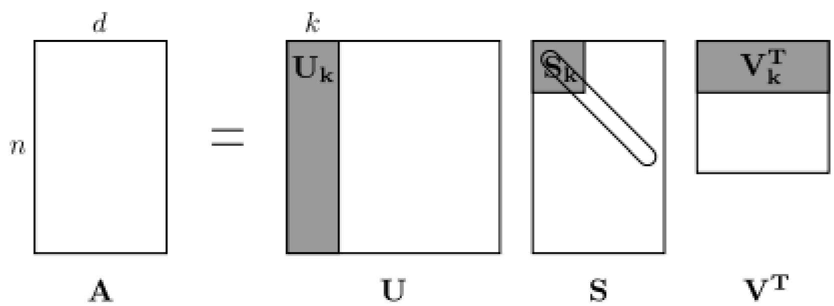
Document 2: "all is well that ends well"

*	START	all	that	glitters	is	not	gold	well	ends	END
START	0	2	0	0	0	0	0	0	0	0
all	2	0	1	0	1	0	0	0	0	0
that	0	1	0	1	0	0	0	1	1	0
glitters	0	0	1	0	1	0	0	0	0	0
is	0	1	0	1	0	1	0	1	0	0
not	0	0	0	0	1	0	1	0	0	0
gold	0	0	0	0	0	1	0	0	0	1

*	START	all	that	glitters	is	not	gold	well	ends	END
well	0	0	1	0	1	0	0	0	1	1
ends	0	0	1	0	0	0	0	1	0	0
END	0	0	0	0	0	0	1	1	0	0

Note: In NLP, we often add START and END tokens to represent the beginning and end of sentences, paragraphs or documents. In thise case we imagine START and END tokens encapsulating each document, e.g., "START All that glitters is not gold END", and include these tokens in our co-occurrence counts.

The rows (or columns) of this matrix provide one type of word vectors (those based on word-word co-occurrence), but the vectors will be large in general (linear in the number of distinct words in a corpus). Thus, our next step is to run *dimensionality reduction*. In particular, we will run *SVD (Singular Value Decomposition)*, which is a kind of generalized *PCA (Principal Components Analysis)* to select the top k principal components. Here's a visualization of dimensionality reduction with SVD. In this picture our co-occurrence matrix is A with n rows corresponding to n words. We obtain a full matrix decomposition, with the singular values ordered in the diagonal S matrix, and our new, shorter length-k word vectors in U_k.



This reduced-dimensionality co-occurrence representation preserves semantic relationships between words, e.g. *doctor* and *hospital* will be closer than *doctor* and *dog*.

Notes: If you can barely remember what an eigenvalue is, here's [a slow, friendly introduction to SVD (https://davetang.org/file/Singular_Value_Decomposition_Tutorial.pdf)](https://davetang.org/file/Singular_Value_Decomposition_Tutorial.pdf). Though, for the purpose of this class, you only need to know how to extract the k-dimensional embeddings by utilizing pre-programmed implementations of these algorithms from the numpy, scipy, or sklearn python packages. In practice, it is challenging to apply full SVD to large corpora because of the memory needed to perform PCA or SVD. However, if you only want the top k vector components for relatively small k — known as [*Truncated SVD (https://en.wikipedia.org/wiki/Singular_value_decomposition#Truncated_SVD)*](https://en.wikipedia.org/wiki/Singular_value_decomposition#Truncated_SVD) — then there are reasonably scalable techniques to compute those iteratively.

Plotting Co-Occurrence Word Embeddings

Here, we will be using the Reuters (business and financial news) corpus. If you haven't run the import cell at the top of this page, please run it now (click it and press SHIFT-RETURN). The corpus consists of 10,788 news documents totaling 1.3 million words. These documents span 90 categories and are split into train and test. For more details, please see https://www.nltk.org/book/ch02.html (https://www.nltk.org/book/ch02.html). We provide a `read_corpus` function below that pulls out only articles from the "crude" (i.e. news articles about oil, gas, etc.) category. The function also adds START and END tokens to each of the documents, and lowercases words. You do **not** have perform any other kind of pre-processing.

In [3]:
```
def read_corpus(category="crude"):
    """ Read files from the specified Reuter's category.
        Params:
            category (string): category name
        Return:
            list of lists, with words from each of the processed files
    """
    files = reuters.fileids(category)
    return [[START_TOKEN] + [w.lower() for w in list(reuters.words(f))] + [END_
```

Let's have a look what these documents are like….

In [4]:
```
reuters_corpus = read_corpus()
pprint.pprint(reuters_corpus[:3], compact=True, width=100)
```

```
[['<START>', 'japan', 'to', 'revise', 'long', '-', 'term', 'energy', 'deman
d', 'downwards', 'the',
  'ministry', 'of', 'international', 'trade', 'and', 'industry', '(', 'miti',
')', 'will', 'revise',
  'its', 'long', '-', 'term', 'energy', 'supply', '/', 'demand', 'outlook',
'by', 'august', 'to',
  'meet', 'a', 'forecast', 'downtrend', 'in', 'japanese', 'energy', 'demand',
',', 'ministry',
  'officials', 'said', '.', 'miti', 'is', 'expected', 'to', 'lower', 'the',
'projection', 'for',
  'primary', 'energy', 'supplies', 'in', 'the', 'year', '2000', 'to', '550',
'mln', 'kilolitres',
  '(', 'kl', ')', 'from', '600', 'mln', ',', 'they', 'said', '.', 'the', 'dec
ision', 'follows',
  'the', 'emergence', 'of', 'structural', 'changes', 'in', 'japanese', 'indus
try', 'following',
  'the', 'rise', 'in', 'the', 'value', 'of', 'the', 'yen', 'and', 'a', 'decli
ne', 'in', 'domestic',
  'electric', 'power', 'demand', '.', 'miti', 'is', 'planning', 'to', 'work',
'out', 'a', 'revised',
  'energy', 'supply', '/', 'demand', 'outlook', 'through', 'deliberations',
'of', 'committee',
  'meetings', 'of', 'the', 'agency', 'of', 'natural', 'resources', 'and', 'en
ergy', ',', 'the',
  'officials', 'said', '.', 'they', 'said', 'miti', 'will', 'also', 'review',
'the', 'breakdown',
  'of', 'energy', 'supply', 'sources', ',', 'including', 'oil', ',', 'nuclea
r', ',', 'coal', 'and',
  'natural', 'gas', '.', 'nuclear', 'energy', 'provided', 'the', 'bulk', 'o
f', 'japan', "'", 's',
  'electric', 'power', 'in', 'the', 'fiscal', 'year', 'ended', 'march', '31',
',', 'supplying',
  'an', 'estimated', '27', 'pct', 'on', 'a', 'kilowatt', '/', 'hour', 'basi
s', ',', 'followed',
  'by', 'oil', '(', '23', 'pct', ')', 'and', 'liquefied', 'natural', 'gas',
'(', '21', 'pct', '),',
  'they', 'noted', '.', '<END>'],
 ['<START>', 'energy', '/', 'u', '.', 's', '.', 'petrochemical', 'industry',
'cheap', 'oil',
  'feedstocks', ',', 'the', 'weakened', 'u', '.', 's', '.', 'dollar', 'and',
'a', 'plant',
  'utilization', 'rate', 'approaching', '90', 'pct', 'will', 'propel', 'the',
'streamlined', 'u',
  '.', 's', '.', 'petrochemical', 'industry', 'to', 'record', 'profits', 'thi
s', 'year', ',',
  'with', 'growth', 'expected', 'through', 'at', 'least', '1990', ',', 'majo
r', 'company',
  'executives', 'predicted', '.', 'this', 'bullish', 'outlook', 'for', 'chemi
cal', 'manufacturing',
  'and', 'an', 'industrywide', 'move', 'to', 'shed', 'unrelated', 'businesse
s', 'has', 'prompted',
  'gaf', 'corp', '&', 'lt', ';', 'gaf', '>,', 'privately', '-', 'held', 'cai
n', 'chemical', 'inc',
  ',', 'and', 'other', 'firms', 'to', 'aggressively', 'seek', 'acquisitions',
'of', 'petrochemical',
  'plants', '.', 'oil', 'companies', 'such', 'as', 'ashland', 'oil', 'inc',
'&', 'lt', ';', 'ash',
```

```
  '>,', 'the', 'kentucky', '-', 'based', 'oil', 'refiner', 'and', 'marketer',
',', 'are', 'also',
  'shopping', 'for', 'money', '-', 'making', 'petrochemical', 'businesses',
'to', 'buy', '.', '"',
  'i', 'see', 'us', 'poised', 'at', 'the', 'threshold', 'of', 'a', 'golden',
'period', ',"', 'said',
  'paul', 'oreffice', ',', 'chairman', 'of', 'giant', 'dow', 'chemical', 'c
o', '&', 'lt', ';',
  'dow', '>,', 'adding', ',', '"', 'there', "'", 's', 'no', 'major', 'plant',
'capacity', 'being',
  'added', 'around', 'the', 'world', 'now', '.', 'the', 'whole', 'game', 'i
s', 'bringing', 'out',
  'new', 'products', 'and', 'improving', 'the', 'old', 'ones', '."', 'analyst
s', 'say', 'the',
  'chemical', 'industry', "'", 's', 'biggest', 'customers', ',', 'automobil
e', 'manufacturers',
  'and', 'home', 'builders', 'that', 'use', 'a', 'lot', 'of', 'paints', 'an
d', 'plastics', ',',
  'are', 'expected', 'to', 'buy', 'quantities', 'this', 'year', '.', 'u',
'.', 's', '.',
  'petrochemical', 'plants', 'are', 'currently', 'operating', 'at', 'about',
'90', 'pct',
  'capacity', ',', 'reflecting', 'tighter', 'supply', 'that', 'could', 'hik
e', 'product', 'prices',
  'by', '30', 'to', '40', 'pct', 'this', 'year', ',', 'said', 'john', 'doshe
r', ',', 'managing',
  'director', 'of', 'pace', 'consultants', 'inc', 'of', 'houston', '.', 'dema
nd', 'for', 'some',
  'products', 'such', 'as', 'styrene', 'could', 'push', 'profit', 'margins',
'up', 'by', 'as',
  'much', 'as', '300', 'pct', ',', 'he', 'said', '.', 'oreffice', ',', 'speak
ing', 'at', 'a',
  'meeting', 'of', 'chemical', 'engineers', 'in', 'houston', ',', 'said', 'do
w', 'would', 'easily',
  'top', 'the', '741', 'mln', 'dlrs', 'it', 'earned', 'last', 'year', 'and',
'predicted', 'it',
  'would', 'have', 'the', 'best', 'year', 'in', 'its', 'history', '.', 'in',
'1985', ',', 'when',
  'oil', 'prices', 'were', 'still', 'above', '25', 'dlrs', 'a', 'barrel', 'an
d', 'chemical',
  'exports', 'were', 'adversely', 'affected', 'by', 'the', 'strong', 'u',
'.', 's', '.', 'dollar',
  ',', 'dow', 'had', 'profits', 'of', '58', 'mln', 'dlrs', '.', '"', 'i', 'be
lieve', 'the',
  'entire', 'chemical', 'industry', 'is', 'headed', 'for', 'a', 'record', 'ye
ar', 'or', 'close',
  'to', 'it', ',"', 'oreffice', 'said', '.', 'gaf', 'chairman', 'samuel', 'he
yman', 'estimated',
  'that', 'the', 'u', '.', 's', '.', 'chemical', 'industry', 'would', 'repor
t', 'a', '20', 'pct',
  'gain', 'in', 'profits', 'during', '1987', '.', 'last', 'year', ',', 'the',
'domestic',
  'industry', 'earned', 'a', 'total', 'of', '13', 'billion', 'dlrs', ',',
'a', '54', 'pct', 'leap',
  'from', '1985', '.', 'the', 'turn', 'in', 'the', 'fortunes', 'of', 'the',
'once', '-', 'sickly',
  'chemical', 'industry', 'has', 'been', 'brought', 'about', 'by', 'a', 'comb
```

```
ination', 'of', 'luck',
  'and', 'planning', ',', 'said', 'pace', "'", 's', 'john', 'dosher', '.', 'd
osher', 'said', 'last',
  'year', "'", 's', 'fall', 'in', 'oil', 'prices', 'made', 'feedstocks', 'dra
matically', 'cheaper',
  'and', 'at', 'the', 'same', 'time', 'the', 'american', 'dollar', 'was', 'we
akening', 'against',
  'foreign', 'currencies', '.', 'that', 'helped', 'boost', 'u', '.', 's',
'.', 'chemical',
  'exports', '.', 'also', 'helping', 'to', 'bring', 'supply', 'and', 'deman
d', 'into', 'balance',
  'has', 'been', 'the', 'gradual', 'market', 'absorption', 'of', 'the', 'extr
a', 'chemical',
  'manufacturing', 'capacity', 'created', 'by', 'middle', 'eastern', 'oil',
'producers', 'in',
  'the', 'early', '1980s', '.', 'finally', ',', 'virtually', 'all', 'major',
'u', '.', 's', '.',
  'chemical', 'manufacturers', 'have', 'embarked', 'on', 'an', 'extensive',
'corporate',
  'restructuring', 'program', 'to', 'mothball', 'inefficient', 'plants', ',',
'trim', 'the',
  'payroll', 'and', 'eliminate', 'unrelated', 'businesses', '.', 'the', 'rest
ructuring', 'touched',
  'off', 'a', 'flurry', 'of', 'friendly', 'and', 'hostile', 'takeover', 'atte
mpts', '.', 'gaf', ',',
  'which', 'made', 'an', 'unsuccessful', 'attempt', 'in', '1985', 'to', 'acqu
ire', 'union',
  'carbide', 'corp', '&', 'lt', ';', 'uk', '>,', 'recently', 'offered', 'thre
e', 'billion', 'dlrs',
  'for', 'borg', 'warner', 'corp', '&', 'lt', ';', 'bor', '>,', 'a', 'chicag
o', 'manufacturer',
  'of', 'plastics', 'and', 'chemicals', '.', 'another', 'industry', 'powerhou
se', ',', 'w', '.',
  'r', '.', 'grace', '&', 'lt', ';', 'gra', '>', 'has', 'divested', 'its', 'r
etailing', ',',
  'restaurant', 'and', 'fertilizer', 'businesses', 'to', 'raise', 'cash', 'fo
r', 'chemical',
  'acquisitions', '.', 'but', 'some', 'experts', 'worry', 'that', 'the', 'che
mical', 'industry',
  'may', 'be', 'headed', 'for', 'trouble', 'if', 'companies', 'continue', 'tu
rning', 'their',
  'back', 'on', 'the', 'manufacturing', 'of', 'staple', 'petrochemical', 'com
modities', ',', 'such',
  'as', 'ethylene', ',', 'in', 'favor', 'of', 'more', 'profitable', 'specialt
y', 'chemicals',
  'that', 'are', 'custom', '-', 'designed', 'for', 'a', 'small', 'group', 'o
f', 'buyers', '.', '"',
  'companies', 'like', 'dupont', '&', 'lt', ';', 'dd', '>', 'and', 'monsant
o', 'co', '&', 'lt', ';',
  'mtc', '>', 'spent', 'the', 'past', 'two', 'or', 'three', 'years', 'tryin
g', 'to', 'get', 'out',
  'of', 'the', 'commodity', 'chemical', 'business', 'in', 'reaction', 'to',
'how', 'badly', 'the',
  'market', 'had', 'deteriorated', ',"', 'dosher', 'said', '.', '"', 'but',
'i', 'think', 'they',
  'will', 'eventually', 'kill', 'the', 'margins', 'on', 'the', 'profitable',
'chemicals', 'in',
```

```
  'the', 'niche', 'market', '."', 'some', 'top', 'chemical', 'executives', 's
hare', 'the',
  'concern', '.', '"', 'the', 'challenge', 'for', 'our', 'industry', 'is', 't
o', 'keep', 'from',
  'getting', 'carried', 'away', 'and', 'repeating', 'past', 'mistakes', ',"',
'gaf', "'", 's',
  'heyman', 'cautioned', '.', '"', 'the', 'shift', 'from', 'commodity', 'chem
icals', 'may', 'be',
  'ill', '-', 'advised', '.', 'specialty', 'businesses', 'do', 'not', 'stay',
'special', 'long',
  '."', 'houston', '-', 'based', 'cain', 'chemical', ',', 'created', 'this',
'month', 'by', 'the',
  'sterling', 'investment', 'banking', 'group', ',', 'believes', 'it', 'can',
'generate', '700',
  'mln', 'dlrs', 'in', 'annual', 'sales', 'by', 'bucking', 'the', 'industry',
'trend', '.',
  'chairman', 'gordon', 'cain', ',', 'who', 'previously', 'led', 'a', 'levera
ged', 'buyout', 'of',
  'dupont', "'", 's', 'conoco', 'inc', "'", 's', 'chemical', 'business', ',',
'has', 'spent', '1',
  '.', '1', 'billion', 'dlrs', 'since', 'january', 'to', 'buy', 'seven', 'pet
rochemical', 'plants',
  'along', 'the', 'texas', 'gulf', 'coast', '.', 'the', 'plants', 'produce',
'only', 'basic',
  'commodity', 'petrochemicals', 'that', 'are', 'the', 'building', 'blocks',
'of', 'specialty',
  'products', '.', '"', 'this', 'kind', 'of', 'commodity', 'chemical', 'busin
ess', 'will', 'never',
  'be', 'a', 'glamorous', ',', 'high', '-', 'margin', 'business', ',"', 'cai
n', 'said', ',',
  'adding', 'that', 'demand', 'is', 'expected', 'to', 'grow', 'by', 'about',
'three', 'pct',
  'annually', '.', 'garo', 'armen', ',', 'an', 'analyst', 'with', 'dean', 'wi
tter', 'reynolds', ',',
  'said', 'chemical', 'makers', 'have', 'also', 'benefitted', 'by', 'increasi
ng', 'demand', 'for',
  'plastics', 'as', 'prices', 'become', 'more', 'competitive', 'with', 'alumi
num', ',', 'wood',
  'and', 'steel', 'products', '.', 'armen', 'estimated', 'the', 'upturn', 'i
n', 'the', 'chemical',
  'business', 'could', 'last', 'as', 'long', 'as', 'four', 'or', 'five', 'yea
rs', ',', 'provided',
  'the', 'u', '.', 's', '.', 'economy', 'continues', 'its', 'modest', 'rate',
'of', 'growth', '.',
  '<END>'],
 ['<START>', 'turkey', 'calls', 'for', 'dialogue', 'to', 'solve', 'dispute',
'turkey', 'said',
  'today', 'its', 'disputes', 'with', 'greece', ',', 'including', 'rights',
'on', 'the',
  'continental', 'shelf', 'in', 'the', 'aegean', 'sea', ',', 'should', 'be',
'solved', 'through',
  'negotiations', '.', 'a', 'foreign', 'ministry', 'statement', 'said', 'th
e', 'latest', 'crisis',
  'between', 'the', 'two', 'nato', 'members', 'stemmed', 'from', 'the', 'cont
inental', 'shelf',
  'dispute', 'and', 'an', 'agreement', 'on', 'this', 'issue', 'would', 'effec
t', 'the', 'security',
```

```
  ',', 'economy', 'and', 'other', 'rights', 'of', 'both', 'countries', '.',
'"', 'as', 'the',
  'issue', 'is', 'basicly', 'political', ',', 'a', 'solution', 'can', 'only',
'be', 'found', 'by',
  'bilateral', 'negotiations', ',"', 'the', 'statement', 'said', '.', 'greec
e', 'has', 'repeatedly',
  'said', 'the', 'issue', 'was', 'legal', 'and', 'could', 'be', 'solved', 'a
t', 'the',
  'international', 'court', 'of', 'justice', '.', 'the', 'two', 'countries',
'approached', 'armed',
  'confrontation', 'last', 'month', 'after', 'greece', 'announced', 'it', 'pl
anned', 'oil',
  'exploration', 'work', 'in', 'the', 'aegean', 'and', 'turkey', 'said', 'i
t', 'would', 'also',
  'search', 'for', 'oil', '.', 'a', 'face', '-', 'off', 'was', 'averted', 'wh
en', 'turkey',
  'confined', 'its', 'research', 'to', 'territorrial', 'waters', '.', '"', 't
he', 'latest',
  'crises', 'created', 'an', 'historic', 'opportunity', 'to', 'solve', 'the',
'disputes', 'between',
  'the', 'two', 'countries', ',"', 'the', 'foreign', 'ministry', 'statement',
'said', '.', 'turkey',
  "'", 's', 'ambassador', 'in', 'athens', ',', 'nazmi', 'akiman', ',', 'was',
'due', 'to', 'meet',
  'prime', 'minister', 'andreas', 'papandreou', 'today', 'for', 'the', 'gree
k', 'reply', 'to', 'a',
  'message', 'sent', 'last', 'week', 'by', 'turkish', 'prime', 'minister', 't
urgut', 'ozal', '.',
  'the', 'contents', 'of', 'the', 'message', 'were', 'not', 'disclosed', '.',
'<END>']]
```

Question 1.1: Implement `distinct_words` [code] (8 points)

Write a method to work out the distinct words (word types) that occur in the corpus. You can do this with `for` loops, but it's more efficient to do it with Python list comprehensions. In particular, [this (https://coderwall.com/p/rcmaea/flatten-a-list-of-lists-in-one-line-in-python)](https://coderwall.com/p/rcmaea/flatten-a-list-of-lists-in-one-line-in-python) may be useful to flatten a list of lists. If you're not familiar with Python list comprehensions in general, here's [more information (https://python-3-patterns-idioms-test.readthedocs.io/en/latest/Comprehensions.html)](https://python-3-patterns-idioms-test.readthedocs.io/en/latest/Comprehensions.html).

You may find it useful to use [Python sets (https://www.w3schools.com/python/python_sets.asp)](https://www.w3schools.com/python/python_sets.asp) to remove duplicate words.

In [5]:
```
def distinct_words(corpus):
    """ Determine a list of distinct words for the corpus.
        Params:
            corpus (list of list of strings): corpus of documents
        Return:
            corpus_words (list of strings): list of distinct words across the c
            num_corpus_words (integer): number of distinct words across the co
    """
    corpus_words = sorted(list(set([word for doc in corpus for word in doc])))
    num_corpus_words = len(corpus_words)
    return corpus_words, num_corpus_words
```

In [6]:
```
# ---------------------
# Run this sanity check
# Note that this NOT an exhaustive check for correctness.
# ---------------------

# Define toy corpus
test_corpus = ["START All that glitters isn't gold END".split(" "), "START All
test_corpus_words, num_corpus_words = distinct_words(test_corpus)

# Correct answers
ans_test_corpus_words = sorted(list(set(["START", "All", "ends", "that", "gold
ans_num_corpus_words = len(ans_test_corpus_words)

# Test correct number of words
assert(num_corpus_words == ans_num_corpus_words), "Incorrect number of distinc

# Test correct words
assert (test_corpus_words == ans_test_corpus_words), "Incorrect corpus_words.\

# Print Success
print ("-" * 80)
print("Passed All Tests!")
print ("-" * 80)
```

```
--------------------------------------------------------------------------------
---
Passed All Tests!
--------------------------------------------------------------------------------
---
```

Question 1.2: Implement `compute_co_occurrence_matrix` [code] (12 points)

Write a method that constructs a co-occurrence matrix for a certain window-size n (with a default of 4), considering words n before and n after the word in the center of the window. Here, we start to use `numpy (np)` to represent vectors, matrices, and tensors. If you're not familiar with NumPy, there's a [Python NumPy tutorial (https://github.com/kuleshov/cs228-material/blob/master/tutorials/python/cs228-python-tutorial.ipynb)](https://github.com/kuleshov/cs228-material/blob/master/tutorials/python/cs228-python-tutorial.ipynb).

In [7]:
```
def compute_co_occurrence_matrix(corpus, window_size=4):
    """ Compute co-occurrence matrix for the given corpus and window_size (defa

        Note: Each word in a document should be at the center of a window. Word
              number of co-occurring words.

              For example, if we take the document "START All that glitters is
              "All" will co-occur with "START", "that", "glitters", "is", and

        Params:
            corpus (list of list of strings): corpus of documents
            window_size (int): size of context window
        Return:
            M (numpy matrix of shape (number of corpus words, number of corpus
                Co-occurence matrix of word counts.
                The ordering of the words in the rows/columns should be the sar
            word2Ind (dict): dictionary that maps word to index (i.e. row/colur
    """
    words, num_words = distinct_words(corpus)
    M = np.zeros((num_words, num_words))
    word2Ind = {word: index for index, word in enumerate(words)}

    for doc in corpus:
        for i, center_word in enumerate(doc):
            center_word_index = word2Ind[center_word]
            context_window = doc[max(0, i-window_size):i] + doc[i+1:i+window_s
            for context_word in context_window:
                context_word_index = word2Ind[context_word]
                M[center_word_index, context_word_index] += 1

    return M, word2Ind
```

In [8]:
```
# ---------------------
# Run this sanity check
# Note that this is NOT an exhaustive check for correctness.
# ---------------------

# Define toy corpus and get student's co-occurrence matrix
test_corpus = ["START All that glitters isn't gold END".split(" "), "START All
M_test, word2Ind_test = compute_co_occurrence_matrix(test_corpus, window_size=

# Correct M and word2Ind
M_test_ans = np.array(
    [[0., 0., 0., 1., 0., 0., 0., 0., 1., 0.,],
     [0., 0., 0., 1., 0., 0., 0., 0., 0., 1.,],
     [0., 0., 0., 0., 0., 0., 1., 0., 0., 1.,],
     [1., 1., 0., 0., 0., 0., 0., 0., 0., 0.,],
     [0., 0., 0., 0., 0., 0., 0., 0., 1., 1.,],
     [0., 0., 0., 0., 0., 0., 0., 1., 1., 0.,],
     [0., 0., 1., 0., 0., 0., 0., 1., 0., 0.,],
     [0., 0., 0., 0., 0., 1., 1., 0., 0., 0.,],
     [1., 0., 0., 0., 1., 1., 0., 0., 0., 1.,],
     [0., 1., 1., 0., 1., 0., 0., 0., 1., 0.,]]
)
word2Ind_ans = {'All': 0, "All's": 1, 'END': 2, 'START': 3, 'ends': 4, 'glitte

# Test correct word2Ind
assert (word2Ind_ans == word2Ind_test), "Your word2Ind is incorrect:\nCorrect:

# Test correct M shape
assert (M_test.shape == M_test_ans.shape), "M matrix has incorrect shape.\nCor

# Test correct M values
for w1 in word2Ind_ans.keys():
    idx1 = word2Ind_ans[w1]
    for w2 in word2Ind_ans.keys():
        idx2 = word2Ind_ans[w2]
        student = M_test[idx1, idx2]
        correct = M_test_ans[idx1, idx2]
        if student != correct:
            print("Correct M:")
            print(M_test_ans)
            print("Your M: ")
            print(M_test)
            raise AssertionError("Incorrect count at index ({}, {})=({}, {}) i

# Print Success
print ("-" * 80)
print("Passed All Tests!")
print ("-" * 80)
```

```
-----------------------------------------------------------------------------
---
Passed All Tests!
-----------------------------------------------------------------------------
---
```

Question 1.3: Implement `reduce_to_k_dim` [code] (4 points)

Construct a method that performs dimensionality reduction on the matrix to produce k-dimensional embeddings. Use SVD to take the top k components and produce a new matrix of k-dimensional embeddings.

Note: All of numpy, scipy, and scikit-learn (`sklearn`) provide *some* implementation of SVD, but only scipy and sklearn provide an implementation of Truncated SVD, and only sklearn provides an efficient randomized algorithm for calculating large-scale Truncated SVD. So please use [sklearn.decomposition.TruncatedSVD (https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html)](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html).

In [9]:
```
def reduce_to_k_dim(M, k=2):
    """ Reduce a co-occurence count matrix of dimensionality (num_corpus_words,
        to a matrix of dimensionality (num_corpus_words, k) using the following
            - http://scikit-learn.org/stable/modules/generated/sklearn.decompos

        Params:
            M (numpy matrix of shape (number of corpus words, number of corpus
            k (int): embedding size of each word after dimension reduction
        Return:
            M_reduced (numpy matrix of shape (number of corpus words, k)): matr
                    In terms of the SVD from math class, this actually returns
    """
    n_iters = 10     # Use this parameter in your call to `TruncatedSVD`
    print("Running Truncated SVD over %i words..." % (M.shape[0]))
    svd = TruncatedSVD(n_components=k, n_iter=n_iters)
    M_reduced = svd.fit_transform(M)
    print("Done.")
    return M_reduced
```

In [10]:
```
# ---------------------
# Run this sanity check
# Note that this NOT an exhaustive check for correctness
# In fact we only check that your M_reduced has the right dimensions.
# ---------------------

# Define toy corpus and run student code
test_corpus = ["START All that glitters isn't gold END".split(" "), "START All
M_test, word2Ind_test = compute_co_occurrence_matrix(test_corpus, window_size=
M_test_reduced = reduce_to_k_dim(M_test, k=2)

# Test proper dimensions
assert (M_test_reduced.shape[0] == 10), "M_reduced has {} rows; should have {}
assert (M_test_reduced.shape[1] == 2), "M_reduced has {} columns; should have

# Print Success
print ("-" * 80)
print("Passed All Tests!")
print ("-" * 80)
```

```
Running Truncated SVD over 10 words...
Done.
-----------------------------------------------------------------------------
---
Passed All Tests!
-----------------------------------------------------------------------------
---
```

Question 1.4: Implement `plot_embeddings` [code] (4 points)

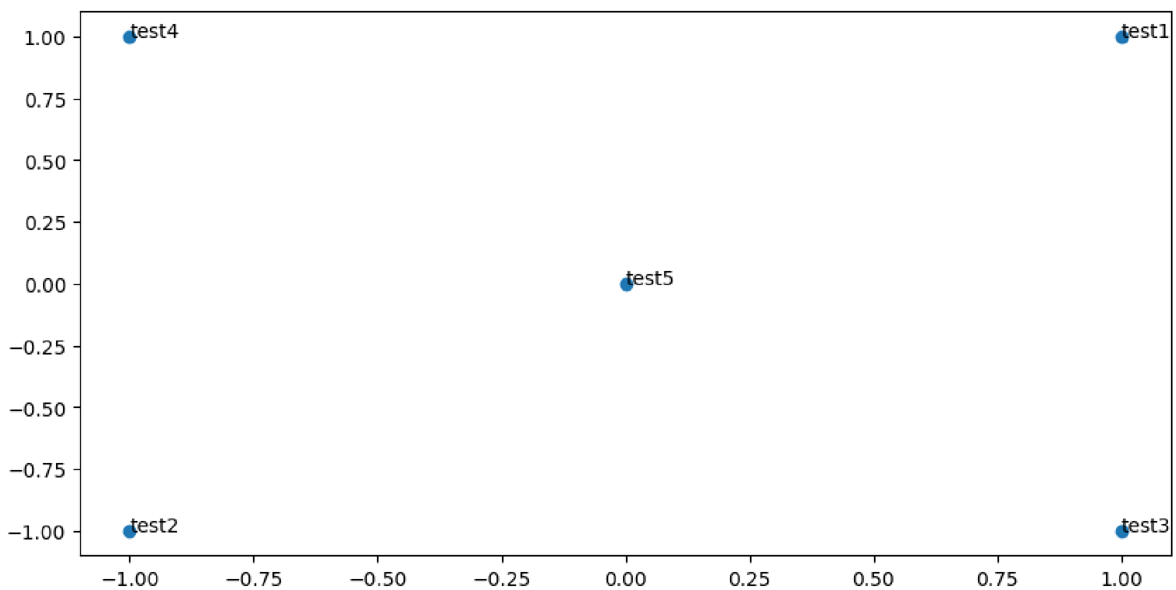
Here you will write a function to plot a set of 2D vectors in 2D space. For graphs, we will use Matplotlib (`plt`).

For this example, you may find it useful to adapt [this code (https://medium.com/@pythonmembers.club/matplotlib-scatter-plot-annotate-set-text-at-label-each-point-ab29eb7b6c8b)](https://medium.com/@pythonmembers.club/matplotlib-scatter-plot-annotate-set-text-at-label-each-point-ab29eb7b6c8b). In the future, a good way to make a plot is to look at [the Matplotlib gallery (https://matplotlib.org/gallery/index.html)](https://matplotlib.org/gallery/index.html), find a plot that looks somewhat like what you want, and adapt the code they give.

In [11]:
```
def plot_embeddings(M_reduced, word2Ind, words):
    """ Plot in a scatterplot the embeddings of the words specified in the list
        NOTE: do not plot all the words listed in M_reduced / word2Ind.
        Include a label next to each point.

        Params:
            M_reduced (numpy matrix of shape (number of unique words in the co
            word2Ind (dict): dictionary that maps word to indices for matrix M
            words (list of strings): words whose embeddings we want to visuali
    """
    fig, ax = plt.subplots()
    x = M_reduced[:, 0]
    y = M_reduced[:, 1]
    ax.scatter(x, y)
    for i, word in enumerate(words):
        ax.annotate(word, (x[word2Ind[word]], y[word2Ind[word]]))
    plt.show()
```

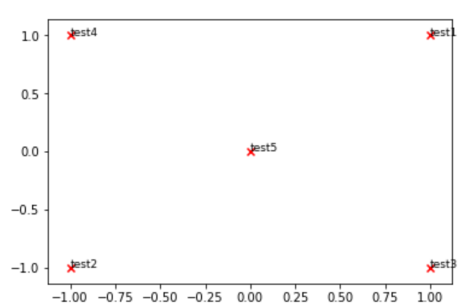
In [12]:
```
# ---------------------
# Run this sanity check
# Note that this NOT an exhaustive check for correctness.
# The plot produced should look like the "test solution plot" depicted below.
# ---------------------

print ("-" * 80)
print ("Outputted Plot:")

M_reduced_plot_test = np.array([[1, 1], [-1, -1], [1, -1], [-1, 1], [0, 0]])
word2Ind_plot_test = {'test1': 0, 'test2': 1, 'test3': 2, 'test4': 3, 'test5':
words = ['test1', 'test2', 'test3', 'test4', 'test5']
plot_embeddings(M_reduced_plot_test, word2Ind_plot_test, words)

print ("-" * 80)
```

```
--------------------------------------------------------------------------------
Outputted Plot:
```




```
--------------------------------------------------------------------------------
```

Test Plot Solution



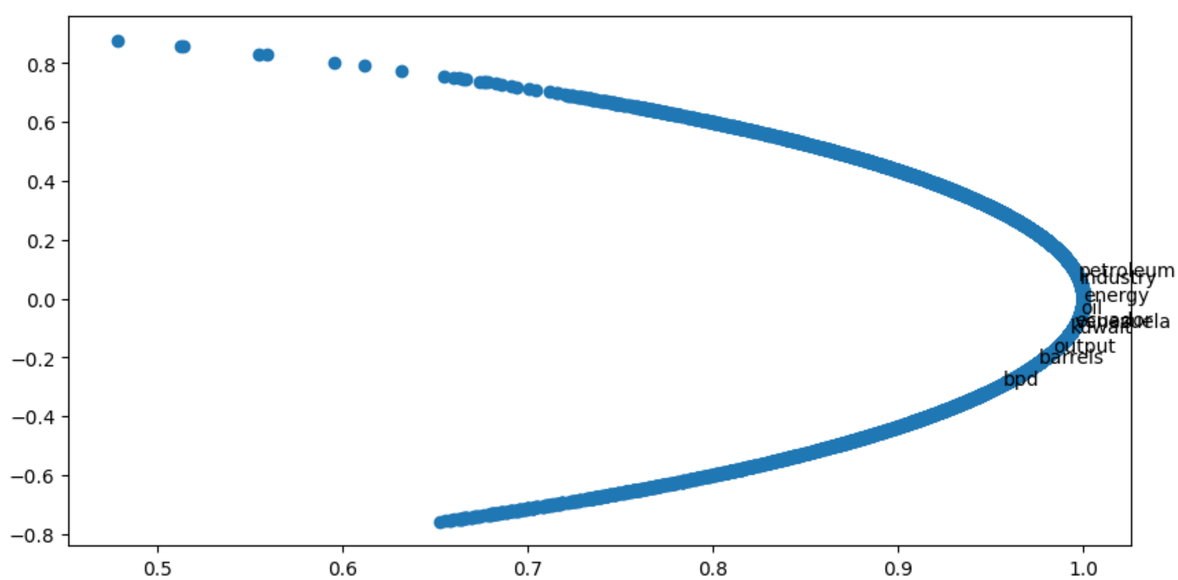
Question 1.5: Co-Occurrence Plot Analysis [written] (12 points)

Now we will put together all the parts you have written! We will compute the co-occurrence matrix with fixed window of 4, over the Reuters "crude" corpus. Then we will use TruncatedSVD to compute 2-dimensional embeddings of each word. TruncatedSVD returns U x S, so we normalize the returned vectors, so that all the vectors will appear around the unit circle (therefore closeness is directional closeness). **Note**: The line of code below that does the normalizing uses the NumPy concept of *broadcasting*. If you don't know about broadcasting, check out [Computation on Arrays: Broadcasting by Jake VanderPlas (https://jakevdp.github.io/PythonDataScienceHandbook/02.05-computation-on-arrays-broadcasting.html)](https://jakevdp.github.io/PythonDataScienceHandbook/02.05-computation-on-arrays-broadcasting.html).

Run the below cell to produce the plot. It'll probably take a few seconds to run. What clusters together in 2-dimensional embedding space? What doesn't cluster together that you might think should have? **Note:** "bpd" stands for "barrels per day" and is a commonly used abbreviation in crude oil topic articles.

In [13]:
```
# ------------------------------
# Run This Cell to Produce Your Plot
# ------------------------------
reuters_corpus = read_corpus()
M_co_occurrence, word2Ind_co_occurrence = compute_co_occurrence_matrix(reuters_
M_reduced_co_occurrence = reduce_to_k_dim(M_co_occurrence, k=2)

# Rescale (normalize) the rows to make them each of unit-length
M_lengths = np.linalg.norm(M_reduced_co_occurrence, axis=1)
M_normalized = M_reduced_co_occurrence / M_lengths[:, np.newaxis] # broadcasti

words = ['barrels', 'bpd', 'ecuador', 'energy', 'industry', 'kuwait', 'oil', '
plot_embeddings(M_normalized, word2Ind_co_occurrence, words)
```

```
Running Truncated SVD over 8185 words...
Done.
```



Part 2: Prediction-Based Word Vectors (60 points)

As discussed in class, more recently prediction-based word vectors have come into fashion, e.g. word2vec. Here, we shall explore the embeddings produced by word2vec. Please revisit the class notes and lecture slides for more details on the word2vec algorithm. If you're feeling adventurous, challenge yourself and try reading the [original paper (https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf)](https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf).

Then run the following cells to load the word2vec vectors into memory. **Note**: This might take several minutes.

In [14]:
```
def load_word2vec():
    """ Load Word2Vec Vectors
        Return:
            wv_from_bin: All 3 million embeddings, each lengh 300
    """
    import gensim.downloader as api
    wv_from_bin = api.load("word2vec-google-news-300")
    vocab = list(wv_from_bin.key_to_index.keys())
    print("Loaded vocab size %i" % len(vocab))
    return wv_from_bin
```

In [15]:
```
# -----------------------------------
# Run Cell to Load Word Vectors
# Note: This may take several minutes
# -----------------------------------
wv_from_bin = load_word2vec()
```

```
Loaded vocab size 3000000
```

Reducing dimensionality of Word2Vec Word Embeddings

Let's directly compare the word2vec embeddings to those of the co-occurrence matrix. Run the following cells to:

1. Put the 3 million word2vec vectors into a matrix M, and
2. Reduce the vectors from 300-dimensional to 2-dimensional by running `reduce_to_k_dim` (your Truncated SVD function).

In [16]:
```
def get_matrix_of_vectors(wv_from_bin, required_words=['barrels', 'bpd', 'ecua
    """ Put the word2vec vectors into a matrix M.
        Param:
            wv_from_bin: KeyedVectors object; the 3 million word2vec vectors lo
        Return:
            M: numpy matrix shape (num words, 300) containing the vectors
            word2Ind: dictionary mapping each word to its row number in M
    """
    import random
    words = list(wv_from_bin.key_to_index.keys())
    print("Shuffling words ...")
    random.shuffle(words)
    words = words[:10000]
    print("Putting %i words into word2Ind and matrix M..." % len(words))
    word2Ind = {}
    M = []
    curInd = 0
    for w in words:
        try:
            M.append(wv_from_bin.get_vector(w))
            word2Ind[w] = curInd
            curInd += 1
        except KeyError:
            continue
    for w in required_words:
        try:
            M.append(wv_from_bin.get_vector(w))
            word2Ind[w] = curInd
            curInd += 1
        except KeyError:
            continue
    M = np.stack(M)
    print("Done.")
    return M, word2Ind
```

In [17]:
```
# -----------------------------------------------------------------
# Run Cell to Reduce 300-Dimensional Word Embeddings to k Dimensions
# Note: This may take several minutes
# -----------------------------------------------------------------
M, word2Ind = get_matrix_of_vectors(wv_from_bin)
M_reduced = reduce_to_k_dim(M, k=2)
```

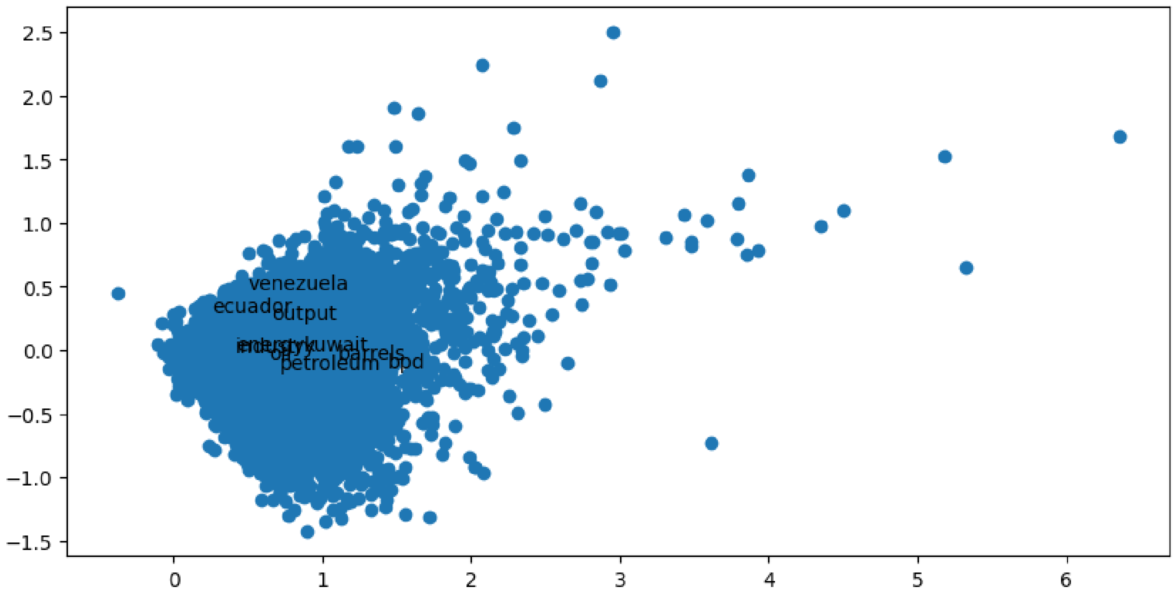
```
Shuffling words ...
Putting 10000 words into word2Ind and matrix M...
Done.
Running Truncated SVD over 10010 words...
Done.
```

Question 2.1: Word2Vec Plot Analysis [written] (16 points)

Run the cell below to plot the 2D word2vec embeddings for `['barrels', 'bpd', 'ecuador', 'energy', 'industry', 'kuwait', 'oil', 'output', 'petroleum', 'venezuela']`.

What clusters together in 2-dimensional embedding space? What doesn't cluster together that you might think should have? How is the plot different from the one generated earlier from the co-occurrence matrix?

In [18]:
```
words = ['barrels', 'bpd', 'ecuador', 'energy', 'industry', 'kuwait', 'oil', '
plot_embeddings(M_reduced, word2Ind, words)
```



In the Word2Vec plot, we see some similar clusters as in the co-occurrence plot, such as the "oil" and "petroleum" cluster. However, the Word2Vec embeddings tend to better capture some of the semantic relationships between the words. For example, "energy" is closer to "oil" and "petroleum" in the Word2Vec plot than in the co-occurrence plot. Similarly, "kuwait" is closer to "saudi" and "iran" in the Word2Vec plot than in the co-occurrence plot.

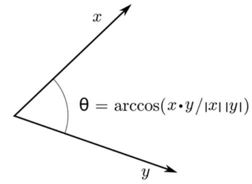
One surprise is that "ecuador" is closer to "oil" than to "venezuela" in the Word2Vec plot, which is opposite to the relationship observed in the co-occurrence plot. Another difference between the two plots is that the Word2Vec plot better captures the differences in magnitudes between the words "barrels" and "bpd", which were not well-separated in the co-occurrence plot.

Overall, the Word2Vec plot appears to capture some of the more subtle semantic relationships between the words, which the co-occurrence plot was not able to do as well.

Cosine Similarity

Now that we have word vectors, we need a way to quantify the similarity between individual words, according to these vectors. One such metric is cosine-similarity. We will be using this to find words that are "close" and "far" from one another.

We can think of n-dimensional vectors as points in n-dimensional space. If we take this perspective, $L1$ and $L2$ distances help quantify the amount of space "we must travel" to get between these two points. Another approach is to examine the angle between two vectors. From trigonometry we know that:



Instead of computing the actual angle, we can leave the similarity in terms of $similarity = cos(\Theta)$. Formally the [Cosine Similarity (https://en.wikipedia.org/wiki/Cosine_similarity)](https://en.wikipedia.org/wiki/Cosine_similarity) s between two vectors p and q is defined as:

$$s = \frac{p \cdot q}{||p||||q||}, \text{ where } s \in [-1, 1]$$

Question 2.2: Polysemous Words (8 points) [code + written]

Find a [polysemous (https://en.wikipedia.org/wiki/Polysemy)](https://en.wikipedia.org/wiki/Polysemy) word (for example, "leaves" or "scoop") such that the top-10 most similar words (according to cosine similarity) contains related words from *both* meanings. For example, "leaves" has both "vanishes" and "stalks" in the top 10, and "scoop" has both "handed_waffle_cone" and "lowdown". You will probably need to try several polysemous words before you find one. Please state the polysemous word you discover and the multiple meanings that occur in the top 10. Why do you think many of the polysemous words you tried didn't work?

Note: You should use the `wv_from_bin.most_similar(word)` function to get the top 10 similar words. This function ranks all other words in the vocabulary with respect to their cosine similarity to the given word. For further assistance please check the **[GenSim documentation (https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvector](https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvector)**

In [19]:
```
# ------------------
# Write your polysemous word exploration code here.

word = "duck"
similar_words = wv_from_bin.most_similar(word)

# Find multiple meanings in top-10 similar words
meanings = set()
for similar_word, _ in similar_words[:10]:
    if "duck" in similar_word.lower():
        meanings.add("bird")
    elif "avoid" in similar_word.lower():
        meanings.add("avoid")

print(f"Polysemous word: {word}")
print(f"Multiple meanings found in top-10 similar words: {meanings}")

# ------------------
```

```
Polysemous word: duck
Multiple meanings found in top-10 similar words: {'bird'}
```

Question 2.3: Synonyms & Antonyms (8 points) [code + written]

When considering Cosine Similarity, it's often more convenient to think of Cosine Distance, which is simply 1 - Cosine Similarity.

Find three words (w1,w2,w3) where w1 and w2 are synonyms and w1 and w3 are antonyms, but Cosine Distance(w1,w3) < Cosine Distance(w1,w2). For example, w1="happy" is closer to w3="sad" than to w2="cheerful".

Once you have found your example, please give a possible explanation for why this counter-intuitive result may have happened.

You should use the the `wv_from_bin.distance(w1, w2)` function here in order to compute the cosine distance between two words. Please see the **GenSim documentation (https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvector** for further assistance.

In [29]:
```
# ------------------
# Write your synonym & antonym exploration code here.

w1 = "hot"
w2 = "warm"
w3 = "cold"
w1_w2_dist = wv_from_bin.distance(w1, w2)
w1_w3_dist = wv_from_bin.distance(w1, w3)

print("Synonyms {}, {} have cosine distance: {}".format(w1, w2, w1_w2_dist))
print("Antonyms {}, {} have cosine distance: {}".format(w1, w3, w1_w3_dist))

# ------------------
```

```
Synonyms hot, warm have cosine distance: 0.567846268415451
Antonyms hot, cold have cosine distance: 0.539786159992218
```

The counter-intuitive result may be explained by the fact that the vectors for "hot" and "cold" are more distinct from each other than the vectors for "hot" and "warm", leading to a smaller cosine distance between "hot" and "cold". Additionally, the distributional hypothesis may not always accurately reflect the nuances of language and meaning, leading to unexpected associations between words.

Solving Analogies with Word Vectors

Word2Vec vectors have been shown to *sometimes* exhibit the ability to solve analogies.

As an example, for the analogy "man : king :: woman : x", what is x?

In the cell below, we show you how to use word vectors to find x. The `most_similar` function finds words that are most similar to the words in the `positive` list and most dissimilar from the words in the `negative` list. The answer to the analogy will be the word ranked most similar (largest numerical value).

Note: Further Documentation on the `most_similar` function can be found within the [GenSim documentation (https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvectors](https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvectors)

In [21]:
```
# Run this cell to answer the analogy -- man : king :: woman : x
pprint.pprint(wv_from_bin.most_similar(positive=['woman', 'king'], negative=['
```

```
[('queen', 0.7118193507194519),
 ('monarch', 0.6189674139022827),
 ('princess', 0.5902431011199951),
 ('crown_prince', 0.5499460697174072),
 ('prince', 0.5377321839332581),
 ('kings', 0.5236844420433044),
 ('Queen_Consort', 0.5235945582389832),
 ('queens', 0.5181134343147278),
 ('sultan', 0.5098593831062317),
 ('monarchy', 0.5087411999702454)]
```

Question 2.4: Finding Analogies [code + written] (8 points)

Find an example of analogy that holds according to these vectors (i.e., the intended word is ranked top). In your solution please state the full analogy in the form x:y :: a:b. If you believe the analogy is complicated, explain why the analogy holds in one or two sentences.

Note: You may have to try many analogies to find one that works!

In [29]:
```
# ------------------
# Write your analogy exploration code here.

a = "son"
b = ""
c = "king"
d = "queen"

analogy = wv_from_bin.most_similar(positive=[d, a], negative=[c])[0]
b = analogy[0]
print(f"{a}:{b}::{c}:{d}")

# ------------------
```

```
son:daughter::king:queen
```

This analogy holds because it captures the gender and hierarchy relationship between the words. Just like a king is the male ruler of a kingdom, a queen is the female ruler of a kingdom. The same gender relationship can be observed between man and woman.

Question 2.5: Incorrect Analogy [code + written] (4 points)

Find an example of analogy that does *not* hold according to these vectors. In your solution, state the intended analogy in the form x:y :: a:b, and state the (incorrect) value of b according to the word vectors.

In [24]:
```
# ------------------
# Write your incorrect analogy exploration code here.
pprint.pprint(wv_from_bin.most_similar(positive=['cat', 'tail'], negative=['do

# ------------------
```

```
[('tails', 0.52464759349823),
 ('stubby_tail', 0.5086752772331238),
 ('bulbous_forehead', 0.4927935004234314),
 ('bushy_tail', 0.4892292618751526),
 ('forked_tail', 0.4804065525531769),
 ('Pamir_Airways_logo', 0.47923895716667175),
 ('underwing', 0.47266650199890137),
 ('wingspread', 0.4588666260242462),
 ('hind_wings', 0.4577387869358063),
 ('palps', 0.45645323395729065)]
```

Question 2.6: Guided Analysis of Bias in Word Vectors [written] (4 points)

It's important to be cognizant of the biases (gender, race, sexual orientation etc.) implicit to our word embeddings.

Run the cell below, to examine (a) which terms are most similar to "woman" and "boss" and most dissimilar to "man", and (b) which terms are most similar to "man" and "boss" and most dissimilar to "woman". What do you find in the top 10?

In [25]:
```
# Run this cell
# Here `positive` indicates the list of words to be similar to and `negative` t
# most dissimilar from.
pprint.pprint(wv_from_bin.most_similar(positive=['woman', 'boss'], negative=['m
print()
pprint.pprint(wv_from_bin.most_similar(positive=['man', 'boss'], negative=['wom
```

```
[('bosses', 0.5522644519805908),
 ('manageress', 0.49151360988616943),
 ('exec', 0.45940810441970825),
 ('Manageress', 0.4559843838214874),
 ('receptionist', 0.4474116563796997),
 ('Jane_Danson', 0.44480547308921814),
 ('Fiz_Jennie_McAlpine', 0.44275766959133148),
 ('Coronation_Street_actress', 0.44275563955307007),
 ('supremo', 0.4409853219985962),
 ('coworker', 0.43986251950263977)]

[('supremo', 0.6097397804260254),
 ('MOTHERWELL_boss', 0.5489562749862671),
 ('CARETAKER_boss', 0.5375303030014038),
 ('Bully_Wee_boss', 0.5333974957466125),
 ('YEOVIL_Town_boss', 0.5321705341339111),
 ('head_honcho', 0.5281980037689209),
 ('manager_Stan_Ternent', 0.525971531867981),
 ('Viv_Busby', 0.5256163477897644),
 ('striker_Gabby_Agbonlahor', 0.5250812768936157),
 ('BARNSLEY_boss', 0.5238943099975586)]
```

A: The top 10 terms that are most similar to "woman" and "boss" and most dissimilar to "man" is the first list.

B: The top 10 terms that are most similar to "man" and "boss" and most dissimilar to "woman" is the second list.

Question 2.7: Independent Analysis of Bias in Word Vectors [code + written] (8 points)

Use the `most_similar` function to find another case where some bias is exhibited by the vectors. Please briefly explain the example of bias that you discover.

In [27]:
```
# ------------------
# Write your bias exploration code here.

pprint.pprint(wv_from_bin.most_similar(positive=['doctor'], negative=[]))
print()
pprint.pprint(wv_from_bin.most_similar(positive=['engineer', 'worker'], negati

# ------------------
```

```
[('physician', 0.7806021571159363),
 ('doctors', 0.747657299041748),
 ('gynecologist', 0.6947518587112427),
 ('surgeon', 0.6793398261070251),
 ('dentist', 0.6785441040992737),
 ('pediatrician', 0.664313793182373),
 ('pharmacist', 0.653485894203186),
 ('neurologist', 0.6517742872238159),
 ('cardiologist', 0.6352297067642212),
 ('nurse', 0.6319523453712463)]

[('mechanical_engineer', 0.6818879246711731),
 ('electrician', 0.6528381109237671),
 ('technician', 0.6453092694282532),
 ('electrical_engineer', 0.6405000686645508),
 ('laborer', 0.5982412695884705),
 ('foreman', 0.596379816532135),
 ('mechanic', 0.5878400206565857),
 ('employee', 0.5864596366882324),
 ('supervisor', 0.5850368738174438),
 ('workers', 0.5807631015777588)]
```

From the results, we can see that words that are similar to "doctor" are mostly associated with medical professions, which are often stereotyped as feminine, such as "gynecologist" and "paediatrician". On the other hand, words that are similar to "engineer" are mostly associated with male-dominated fields such as "mechanical_engineer" and "electrical_engineer". This suggests that there is some gender bias in the word embeddings.

Question 2.8: Thinking About Bias [written] (4 points)

What might be the cause of these biases in the word vectors?

There are several possible causes of biases in the word vectors. One possible cause is the data used to train the models. Word embeddings are typically trained on large corpora of text, such as news articles or web pages, which may contain biases that reflect the societal and cultural biases of the authors or sources. For example, news articles may have more coverage of men in powerful positions or of women in administrative positions, leading to the associations we observed between certain words and gender and power.

How to submit this problem set:

- Write all the answers in this iPython notebook. Once you are finished (1) generate the PDF file (`File -> Print Preview` , and print to PDF), 2) ZIP the PDF and this Jupyter Notebook (.ipynb), and 3) upload the ZIP file to Canvas.
- **Important:** check your PDF before you turn it in to Canvas to make sure it exported correctly.
- When creating your final version of the PDF to hand in, please do a fresh restart and execute every cell in order. Then you'll be sure it's actually right. One handy way to do this is by clicking `Runtime -> Run All` in the notebook menu.