CS224N Assignment 1: Exploring Word Vectors (25 Points)

Welcome to CS224n!

Before you start, make sure you read the README.txt in the same directory as this notebook and enter your SUID below.

Please Enter Your SUID Here: 06349270

In [1]:

```
# All Import Statements Defined Here
# Note: Do not add to this list.
# All the dependencies you need, can be installed by running .
# ----------------

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 /home/zheng/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*.

Assignment Notes: Please make sure to save the notebook as you go along. Submission Instructions are located at the bottom of the notebook.

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) 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 (10 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))

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 here (http://web.stanford.edu/class/cs124/lec/vectorsemantics.video.pdf) or here (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:

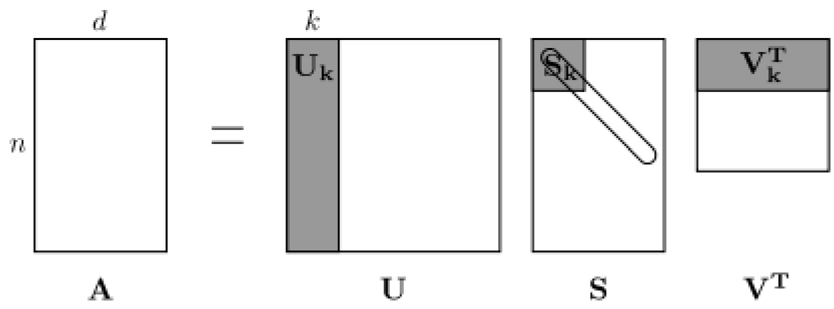
Document 1: "all that glitters is not gold"

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
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). If you want to learn more thoroughly about PCA or SVD, feel free to check out lectures [7 (https://web.stanford.edu/class/cs168/l/l7.pdf)](https://web.stanford.edu/class/cs168/l/l7.pdf), [8 (http://theory.stanford.edu/~tim/s15/l/l8.pdf)](http://theory.stanford.edu/~tim/s15/l/l8.pdf), and [9 (https://web.stanford.edu/class/cs168/l/l9.pdf)](https://web.stanford.edu/class/cs168/l/l9.pdf) of CS168. These course notes provide a great high-level treatment of these general purpose algorithms. Though, for the purpose of this class, you only need to know how to extract the k-dimensional embeddings by utilizing pre-

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)](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 [2]:
```
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_TOKEN] for f in files]
```

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

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

```
[['<START>', 'japan', 'to', 'revise', 'long', '-', 'term', 'energy',
'demand', 'downwards', 'the',
  'ministry', 'of', 'international', 'trade', 'and', 'industry', '(',
'miti', ')', 'will', 'revise',
  'its', 'long', '-', 'term', 'energy', 'supply', '/', 'demand', 'out
look', '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', '.', 'th
e', 'decision', 'follows',
  'the', 'emergence', 'of', 'structural', 'changes', 'in', 'japanes
e', 'industry', 'following',
  'the', 'rise', 'in', 'the', 'value', 'of', 'the', 'yen', 'and',
'a', 'decline', 'in', 'domestic',
  'electric', 'power', 'demand', '.', 'miti', 'is', 'planning', 'to',
'work', 'out', 'a', 'revised',
  'energy', 'supply', '/', 'demand', 'outlook', 'through', 'deliberat
ions', 'of', 'committee',
  'meetings', 'of', 'the', 'agency', 'of', 'natural', 'resources', 'a
nd', 'energy', ',', 'the',
  'officials', 'said', '.', 'they', 'said', 'miti', 'will', 'also',
'review', 'the', 'breakdown',
  'of', 'energy', 'supply', 'sources', ',', 'including', 'oil', ',',
'nuclear', ',', 'coal', 'and',
  'natural', 'gas', '.', 'nuclear', 'energy', 'provided', 'the', 'bul
k', 'of', 'japan', "'", 's',
  'electric', 'power', 'in', 'the', 'fiscal', 'year', 'ended', 'marc
h', '31', ',', 'supplying',
  'an', 'estimated', '27', 'pct', 'on', 'a', 'kilowatt', '/', 'hour',
'basis', ',', 'followed',
  'by', 'oil', '(', '23', 'pct', ')', 'and', 'liquefied', 'natural',
'gas', '(', '21', 'pct', '),',
  'they', 'noted', '.', '<END>'],
 ['<START>', 'energy', '/', 'u', '.', 's', '.', 'petrochemical', 'ind
ustry', 'cheap', 'oil',
  'feedstocks', ',', 'the', 'weakened', 'u', '.', 's', '.', 'dollar',
'and', 'a', 'plant',
  'utilization', 'rate', 'approaching', '90', 'pct', 'will', 'prope
l', 'the', 'streamlined', 'u',
  '.', 's', '.', 'petrochemical', 'industry', 'to', 'record', 'profit
s', 'this', 'year', ',',
  'with', 'growth', 'expected', 'through', 'at', 'least', '1990',
',', 'major', 'company',
  'executives', 'predicted', '.', 'this', 'bullish', 'outlook', 'fo
r', 'chemical', 'manufacturing',
  'and', 'an', 'industrywide', 'move', 'to', 'shed', 'unrelated', 'bu
sinesses', 'has', 'prompted',
  'gaf', 'corp', '&', 'lt', ';', 'gaf', '>,', 'privately', '-', 'hel
d', 'cain', 'chemical', 'inc',
  ',', 'and', 'other', 'firms', 'to', 'aggressively', 'seek', 'acquis
itions', 'of', 'petrochemical',
  'plants', '.', 'oil', 'companies', 'such', 'as', 'ashland', 'oil',
'inc', '&', 'lt', ';', 'ash',
```

```
  '>,', 'the', 'kentucky', '-', 'based', 'oil', 'refiner', 'and', 'ma
rketer', ',', 'are', 'also',
  'shopping', 'for', 'money', '-', 'making', 'petrochemical', 'busine
sses', 'to', 'buy', '.', '"',
  'i', 'see', 'us', 'poised', 'at', 'the', 'threshold', 'of', 'a', 'g
olden', 'period', ',"', 'said',
  'paul', 'oreffice', ',', 'chairman', 'of', 'giant', 'dow', 'chemica
l', 'co', '&', 'lt', ';',
  'dow', '>,', 'adding', ',', '"', 'there', "'", 's', 'no', 'major',
'plant', 'capacity', 'being',
  'added', 'around', 'the', 'world', 'now', '.', 'the', 'whole', 'gam
e', 'is', 'bringing', 'out',
  'new', 'products', 'and', 'improving', 'the', 'old', 'ones', '."',
'analysts', 'say', 'the',
  'chemical', 'industry', "'", 's', 'biggest', 'customers', ',', 'aut
omobile', 'manufacturers',
  'and', 'home', 'builders', 'that', 'use', 'a', 'lot', 'of', 'paint
s', 'and', 'plastics', ',',
  'are', 'expected', 'to', 'buy', 'quantities', 'this', 'year', '.',
'u', '.', 's', '.',
  'petrochemical', 'plants', 'are', 'currently', 'operating', 'at',
'about', '90', 'pct',
  'capacity', ',', 'reflecting', 'tighter', 'supply', 'that', 'coul
d', 'hike', 'product', 'prices',
  'by', '30', 'to', '40', 'pct', 'this', 'year', ',', 'said', 'john',
'dosher', ',', 'managing',
  'director', 'of', 'pace', 'consultants', 'inc', 'of', 'houston',
'.', 'demand', 'for', 'some',
  'products', 'such', 'as', 'styrene', 'could', 'push', 'profit', 'ma
rgins', 'up', 'by', 'as',
  'much', 'as', '300', 'pct', ',', 'he', 'said', '.', 'oreffice',
',', 'speaking', 'at', 'a',
  'meeting', 'of', 'chemical', 'engineers', 'in', 'houston', ',', 'sa
id', 'dow', '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', 'barr
el', 'and', 'chemical',
  'exports', 'were', 'adversely', 'affected', 'by', 'the', 'strong',
'u', '.', 's', '.', 'dollar',
  ',', 'dow', 'had', 'profits', 'of', '58', 'mln', 'dlrs', '.', '"',
'i', 'believe', 'the',
  'entire', 'chemical', 'industry', 'is', 'headed', 'for', 'a', 'reco
rd', 'year', 'or', 'close',
  'to', 'it', ',"', 'oreffice', 'said', '.', 'gaf', 'chairman', 'samu
el', 'heyman', 'estimated',
  'that', 'the', 'u', '.', 's', '.', 'chemical', 'industry', 'would',
'report', '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', 'combination', 'of', 'luck',
  'and', 'planning', ',', 'said', 'pace', "'", 's', 'john', 'dosher',
'.', 'dosher', 'said', 'last',
  'year', "'", 's', 'fall', 'in', 'oil', 'prices', 'made', 'feedstock
s', 'dramatically', 'cheaper',
  'and', 'at', 'the', 'same', 'time', 'the', 'american', 'dollar', 'w
as', 'weakening', 'against',
  'foreign', 'currencies', '.', 'that', 'helped', 'boost', 'u', '.',
's', '.', 'chemical',
  'exports', '.', 'also', 'helping', 'to', 'bring', 'supply', 'and',
'demand', 'into', 'balance',
  'has', 'been', 'the', 'gradual', 'market', 'absorption', 'of', 'th
e', 'extra', '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', 'exten
sive', 'corporate',
  'restructuring', 'program', 'to', 'mothball', 'inefficient', 'plant
s', ',', 'trim', 'the',
  'payroll', 'and', 'eliminate', 'unrelated', 'businesses', '.', 'th
e', 'restructuring', 'touched',
  'off', 'a', 'flurry', 'of', 'friendly', 'and', 'hostile', 'takeove
r', 'attempts', '.', 'gaf', ',',
  'which', 'made', 'an', 'unsuccessful', 'attempt', 'in', '1985', 't
o', 'acquire', 'union',
  'carbide', 'corp', '&', 'lt', ';', 'uk', '>,', 'recently', 'offere
d', 'three', 'billion', 'dlrs',
  'for', 'borg', 'warner', 'corp', '&', 'lt', ';', 'bor', '>,', 'a',
'chicago', 'manufacturer',
  'of', 'plastics', 'and', 'chemicals', '.', 'another', 'industry',
'powerhouse', ',', 'w', '.',
  'r', '.', 'grace', '&', 'lt', ';', 'gra', '>', 'has', 'divested',
'its', 'retailing', ',',
  'restaurant', 'and', 'fertilizer', 'businesses', 'to', 'raise', 'ca
sh', 'for', 'chemical',
  'acquisitions', '.', 'but', 'some', 'experts', 'worry', 'that', 'th
e', 'chemical', 'industry',
  'may', 'be', 'headed', 'for', 'trouble', 'if', 'companies', 'contin
ue', 'turning', 'their',
  'back', 'on', 'the', 'manufacturing', 'of', 'staple', 'petrochemica
l', 'commodities', ',', 'such',
  'as', 'ethylene', ',', 'in', 'favor', 'of', 'more', 'profitable',
'specialty', 'chemicals',
  'that', 'are', 'custom', '-', 'designed', 'for', 'a', 'small', 'gro
up', 'of', 'buyers', '.', '"',
  'companies', 'like', 'dupont', '&', 'lt', ';', 'dd', '>', 'and', 'm
onsanto', 'co', '&', 'lt', ';',
  'mtc', '>', 'spent', 'the', 'past', 'two', 'or', 'three', 'years',
'trying', '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', 'profi
table', 'chemicals', 'in',
```

```
  'the', 'niche', 'market', '."', 'some', 'top', 'chemical', 'executi
ves', 'share', 'the',
  'concern', '.', '"', 'the', 'challenge', 'for', 'our', 'industry',
'is', 'to', 'keep', 'from',
  'getting', 'carried', 'away', 'and', 'repeating', 'past', 'mistake
s', ',"', 'gaf', "'", 's',
  'heyman', 'cautioned', '.', '"', 'the', 'shift', 'from', 'commodit
y', 'chemicals', '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', 'i
t', 'can', 'generate', '700',
  'mln', 'dlrs', 'in', 'annual', 'sales', 'by', 'bucking', 'the', 'in
dustry', 'trend', '.',
  'chairman', 'gordon', 'cain', ',', 'who', 'previously', 'led', 'a',
'leveraged', 'buyout', 'of',
  'dupont', "'", 's', 'conoco', 'inc', "'", 's', 'chemical', 'busines
s', ',', 'has', 'spent', '1',
  '.', '1', 'billion', 'dlrs', 'since', 'january', 'to', 'buy', 'seve
n', 'petrochemical', 'plants',
  'along', 'the', 'texas', 'gulf', 'coast', '.', 'the', 'plants', 'pr
oduce', 'only', 'basic',
  'commodity', 'petrochemicals', 'that', 'are', 'the', 'building', 'b
locks', 'of', 'specialty',
  'products', '.', '"', 'this', 'kind', 'of', 'commodity', 'chemica
l', 'business', 'will', 'never',
  'be', 'a', 'glamorous', ',', 'high', '-', 'margin', 'business',
',"', 'cain', 'said', ',',
  'adding', 'that', 'demand', 'is', 'expected', 'to', 'grow', 'by',
'about', 'three', 'pct',
  'annually', '.', 'garo', 'armen', ',', 'an', 'analyst', 'with', 'de
an', 'witter', 'reynolds', ',',
  'said', 'chemical', 'makers', 'have', 'also', 'benefitted', 'by',
'increasing', 'demand', 'for',
  'plastics', 'as', 'prices', 'become', 'more', 'competitive', 'wit
h', 'aluminum', ',', 'wood',
  'and', 'steel', 'products', '.', 'armen', 'estimated', 'the', 'uptu
rn', 'in', 'the', 'chemical',
  'business', 'could', 'last', 'as', 'long', 'as', 'four', 'or', 'fiv
e', 'years', ',', 'provided',
  'the', 'u', '.', 's', '.', 'economy', 'continues', 'its', 'modest',
'rate', 'of', 'growth', '.',
  '<END>'],
 ['<START>', 'turkey', 'calls', 'for', 'dialogue', 'to', 'solve', 'di
spute', 'turkey', 'said',
  'today', 'its', 'disputes', 'with', 'greece', ',', 'including', 'ri
ghts', 'on', 'the',
  'continental', 'shelf', 'in', 'the', 'aegean', 'sea', ',', 'shoul
d', 'be', 'solved', 'through',
  'negotiations', '.', 'a', 'foreign', 'ministry', 'statement', 'sai
d', 'the', 'latest', 'crisis',
  'between', 'the', 'two', 'nato', 'members', 'stemmed', 'from', 'th
e', 'continental', 'shelf',
  'dispute', 'and', 'an', 'agreement', 'on', 'this', 'issue', 'woul
d', 'effect', 'the', 'security',
```

```
  ',', 'economy', 'and', 'other', 'rights', 'of', 'both', 'countrie
s', '.', '"', 'as', 'the',
  'issue', 'is', 'basicly', 'political', ',', 'a', 'solution', 'can',
'only', 'be', 'found', 'by',
  'bilateral', 'negotiations', ',"', 'the', 'statement', 'said', '.',
'greece', 'has', 'repeatedly',
  'said', 'the', 'issue', 'was', 'legal', 'and', 'could', 'be', 'solv
ed', 'at', 'the',
  'international', 'court', 'of', 'justice', '.', 'the', 'two', 'coun
tries', 'approached', 'armed',
  'confrontation', 'last', 'month', 'after', 'greece', 'announced',
'it', 'planned', 'oil',
  'exploration', 'work', 'in', 'the', 'aegean', 'and', 'turkey', 'sai
d', 'it', 'would', 'also',
  'search', 'for', 'oil', '.', 'a', 'face', '-', 'off', 'was', 'avert
ed', 'when', 'turkey',
  'confined', 'its', 'research', 'to', 'territorrial', 'waters', '.',
'"', 'the', 'latest',
  'crises', 'created', 'an', 'historic', 'opportunity', 'to', 'solv
e', 'the', 'disputes', 'between',
  'the', 'two', 'countries', ',"', 'the', 'foreign', 'ministry', 'sta
tement', 'said', '.', 'turkey',
  "'", 's', 'ambassador', 'in', 'athens', ',', 'nazmi', 'akiman',
',', 'was', 'due', 'to', 'meet',
  'prime', 'minister', 'andreas', 'papandreou', 'today', 'for', 'th
e', 'greek', 'reply', 'to', 'a',
  'message', 'sent', 'last', 'week', 'by', 'turkish', 'prime', 'minis
ter', 'turgut', 'ozal', '.',
  'the', 'contents', 'of', 'the', 'message', 'were', 'not', 'disclose
d', '.', '<END>']]
```

Question 1.1: Implement `distinct_words` [code] (2 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 [4]:
```
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 ac
ross the corpus, sorted (using python 'sorted' function)
            num_corpus_words (integer): number of distinct words acro
ss the corpus
    """
    corpus_words = []
    num_corpus_words = -1

    # ------------------
    # Write your implementation here.
    corpus_words = sorted(list(set(j for i in corpus for j in i)))
    num_corpus_words = len(corpus_words)

    # ------------------
    return corpus_words, num_corpus_words
```

In [5]:
```
# ---------------------
# 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's well that ends well END".split(" ")]
test_corpus_words, num_corpus_words = distinct_words(test_corpus)

# Correct answers
ans_test_corpus_words = sorted(list(set(["START", "All", "ends", "that", "gold", "All's", "glitters", "isn't", "well", "END"])))
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 distinct words. Correct: {}. Yours: {}".format(ans_num_corpus_words, num_corpus_words)

# Test correct words
assert (test_corpus_words == ans_test_corpus_words), "Incorrect corpus_words.\nCorrect: {}\nYours:   {}".format(str(ans_test_corpus_words), str(test_corpus_words))

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

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

Question 1.2: Implement `compute_co_occurrence_matrix` [code] (3 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 NumPy tutorial in the second half of this cs231n [Python NumPy tutorial (http://cs231n.github.io/python-numpy-tutorial/)](http://cs231n.github.io/python-numpy-tutorial/).

In [6]:
```
def compute_co_occurrence_matrix(corpus, window_size=4):
    """ Compute co-occurrence matrix for the given corpus and window_
size (default of 4).

        Note: Each word in a document should be at the center of a wi
ndow. Words near edges will have a smaller
              number of co-occurring words.

              For example, if we take the document "START All that gl
itters is not gold END" with window size of 4,
              "All" will co-occur with "START", "that", "glitters",
 "is", and "not".

        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 number of corpus words)):
                Co-occurence matrix of word counts.
                The ordering of the words in the rows/columns should
 be the same as the ordering of the words given by the distinct_words
 function.
            word2Ind (dict): dictionary that maps word to index (i.e.
 row/column number) for matrix M.
    """
    words, num_words = distinct_words(corpus)
    M = None

    word2Ind = {}

    # ------------------
    # Write your implementation here.
    M = np.zeros((num_words, num_words))

    # generate word2Ind for M indexing
    for i, word in enumerate(words):
        word2Ind[word] = i

    for corp in corpus:
        for s,sub_string in enumerate(corp):
            sub_string_word_to_check = corp[0 if s-window_size < 0 el
se s-window_size :s]
            sub_string_word_to_check.extend(corp[s+1:s+window_size+1
])
            i = word2Ind[sub_string]
            for string in sub_string_word_to_check:
                M[i][word2Ind[string]] = M[i][word2Ind[string]] + 1.0

    # ------------------

    return M, word2Ind
```

In [7]:

```
# ---------------------
# 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's well that ends well END".split(" ")]
M_test, word2Ind_test = compute_co_occurrence_matrix(test_corpus, win
dow_size=1)

# 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
, 'glitters': 5, 'gold': 6, "isn't": 7, 'that': 8, 'well': 9}

# Test correct word2Ind
assert (word2Ind_ans == word2Ind_test), "Your word2Ind is incorrect:
\nCorrect: {}\nYours: {}".format(word2Ind_ans, word2Ind_test)

# Test correct M shape
assert (M_test.shape == M_test_ans.shape), "M matrix has incorrect sh
ape.\nCorrect: {}\nYours: {}".format(M_test.shape, M_test_ans.shape)

# 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 ({}, {})=(
{}, {}) in matrix M. Yours has {} but should have {}.".format(idx1, i
dx2, w1, w2, student, correct))

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

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

Question 1.3: Implement `reduce_to_k_dim` [code] (1 point)

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 [8]:
```python
def reduce_to_k_dim(M, k=2):
    """ Reduce a co-occurence count matrix of dimensionality (num_corpus_words, num_corpus_words)
        to a matrix of dimensionality (num_corpus_words, k) using the following SVD function from Scikit-Learn:
            - http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

        Params:
            M (numpy matrix of shape (number of corpus words, number of number of corpus words)): co-occurence matrix of word counts
            k (int): embedding size of each word after dimension reduction
        Return:
            M_reduced (numpy matrix of shape (number of corpus words, k)): matrix of k-dimensioal word embeddings.
                    In terms of the SVD from math class, this actually returns U * S
    """
    n_iters = 10     # Use this parameter in your call to `TruncatedSVD`
    M_reduced = None
    print("Running Truncated SVD over %i words..." % (M.shape[0]))

    # ------------------
    # Write your implementation here.
    svd = TruncatedSVD(n_components=k, n_iter=n_iters)
    M_reduced = svd.fit_transform(M)
    # ------------------

    print("Done.")
    return M_reduced
```

In [9]:
```
# ---------------------
# 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's well that ends well END".split(" ")]
M_test, word2Ind_test = compute_co_occurrence_matrix(test_corpus, window_size=1)
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 {}".format(M_test_reduced.shape[0], 10)
assert (M_test_reduced.shape[1] == 2), "M_reduced has {} columns; should have {}".format(M_test_reduced.shape[1], 2)

# 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] (1 point)

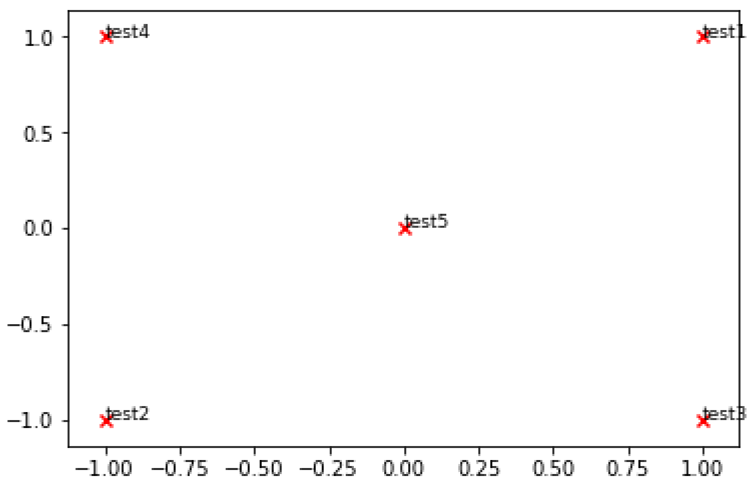
Here you will write a function to plot a set of 2D vectors in 2D space. For graphs, we will use Matplotlib (`plt`).

For this example, you may find it useful to adapt [this code (https://www.pythonmembers.club/2018/05/08/matplotlib-scatter-plot-annotate-set-text-at-label-each-point/)](https://www.pythonmembers.club/2018/05/08/matplotlib-scatter-plot-annotate-set-text-at-label-each-point/). 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 [10]:
```
def plot_embeddings(M_reduced, word2Ind, words):
    """ Plot in a scatterplot the embeddings of the words specified i
n the list "words".
        NOTE: do not plot all the words listed in M_reduced / word2In
d.
        Include a label next to each point.

        Params:
            M_reduced (numpy matrix of shape (number of corpus words,
 k)): matrix of k-dimensioal word embeddings
            word2Ind (dict): dictionary that maps word to indices for
 matrix M
            words (list of strings): words whose embeddings we want t
o visualize
    """

    # ------------------
    # Write your implementation here.
    for word in words:
        coords = M_reduced[word2Ind[word]]
        x = coords[0]
        y = coords[1]
        plt.scatter(x, y, marker='x', color='red')
        plt.text(x, y, word, fontsize=9)
    plt.show()

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

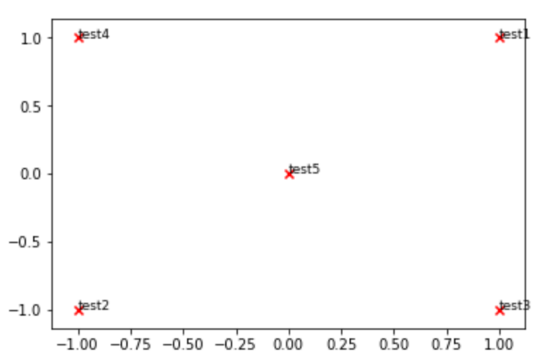
In [11]:
```
# ---------------------
# Run this sanity check
# Note that this not an exhaustive check for correctness.
# The plot produced should look like the "test solution plot" depicte
d below.
# ---------------------

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': 4}
words = ['test1', 'test2', 'test3', 'test4', 'test5']
plot_embeddings(M_reduced_plot_test, word2Ind_plot_test, words)

print ("-" * 80)
```

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




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

Test Plot Solution



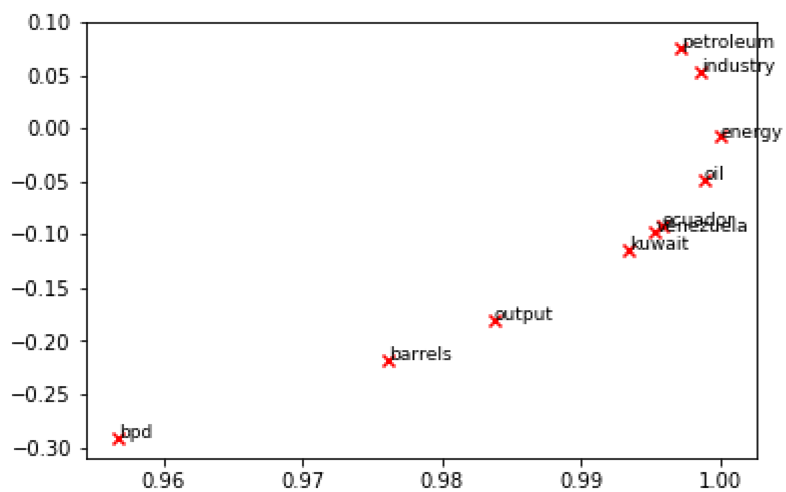
Question 1.5: Co-Occurrence Plot Analysis [written] (3 points)

Now we will put together all the parts you have written! We will compute the co-occurrence matrix with fixed window of 5, over the Reuters "crude" corpus. Then we will use TruncatedSVD to compute 2-dimensional embeddings of each word. TruncatedSVD returns U*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 [12]:
```
# ------------------------------
# Run This Cell to Produce Your Plot
# ------------------------------
reuters_corpus = read_corpus()
M_co_occurrence, word2Ind_co_occurrence = compute_co_occurrence_matri
x(reuters_corpus)
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] # b
roadcasting

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

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



Answer:

kuwait,venezuela,ecuador which are all countries (names) have clustered together. Those countries also close to oil, since those country produces a lot of oil. energy and oil are closed since oil can generate genergy so it is related to energy.Petroleum and industry are closed since petroleum is kind of industry,so industry is generalizion of petroleum.

oil and petroleum should clustered, since they have similar meaning. bpd and output should have clustered as they could have samilar meaning of the productitity outputs of petroleum industry

Part 2: Prediction-Based Word Vectors (15 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).

First make sure that you have downloaded the word2vec embeddings from [https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit (https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit)](https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit)

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

In [13]:
```
# Fill this variable with the path to your downloaded and unzipped em
beddings (`GoogleNews-vectors-negative300.bin` file).
#
# For Windows users place the `GoogleNews-vectors-negative300.bin` fi
le in your conda environment's installation of gensim:
# `envs/{conda_env_name}/lib/site-packages/gensim/test/test_data`
#
# For Mac/Linux users, you can place the `GoogleNews-vectors-negative
300.bin` file anywhere on your machine.
#

embeddings_fp = "/home/zheng/jupyter_notebook/a1/GoogleNews-vectors-n
egative300.bin"
```

In [14]:
```
def load_word2vec(embeddings_fp=embeddings_fp):
    """ Load Word2Vec Vectors
        Param:
            embeddings_fp (string) - path to .bin file of pretrained
 word vectors
        Return:
            wv_from_bin: All 3 million embeddings, each lengh 300
                This is the KeyedVectors format: https://radimrehure
k.com/gensim/models/deprecated/keyedvectors.html
    """
    embed_size = 300
    print("Loading 3 million word vectors from file...")
    wv_from_bin = KeyedVectors.load_word2vec_format(datapath(embeddin
gs_fp), binary=True)
    vocab = list(wv_from_bin.vocab.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()
```

```
Loading 3 million word vectors from file...
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
2. Run reduce_to_k_dim (your Truncated SVD function) to reduce the vectors from 300-dimensional to 2-dimensional.

In [16]:
```
def get_matrix_of_vectors(wv_from_bin):
    """ Put the word2vec vectors into a matrix M.
        Param:
            wv_from_bin: KeyedVectors object; the 3 million word2vec
 vectors loaded from file
        Return:
            M: numpy matrix shape (num words, 300) containing the vec
tors
            word2Ind: dictionary mapping each word to its row number
 in M
    """
    words = list(wv_from_bin.vocab.keys())
    print("Putting %i words into word2Ind and matrix M..." % len(word
s))
    word2Ind = {}
    M = []
    curInd = 0
    for w in words:
        try:
            M.append(wv_from_bin.word_vec(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-Dimensinal 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)
```

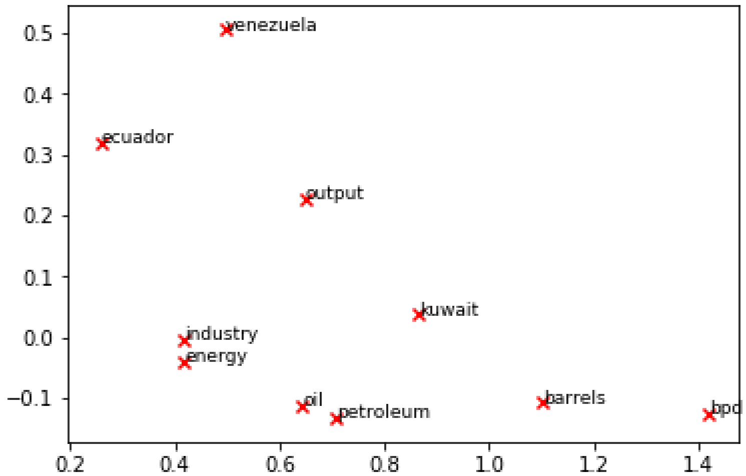
```
Putting 3000000 words into word2Ind and matrix M...
Done.
Running Truncated SVD over 3000000 words...
Done.
```

Question 2.1: Word2Vec Plot Analysis [written] (4 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?

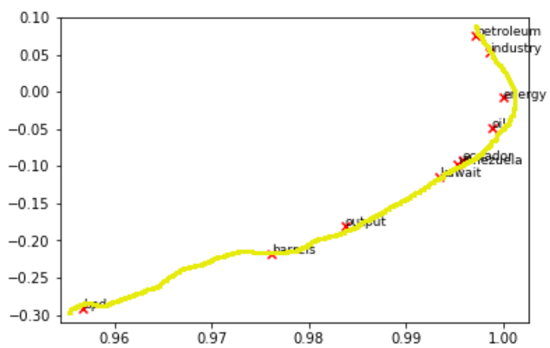
In [18]:
```
words = ['barrels', 'bpd', 'ecuador', 'energy', 'industry', 'kuwait',
 'oil', 'output', 'petroleum', 'venezuela']
plot_embeddings(M_reduced, word2Ind, words)
```

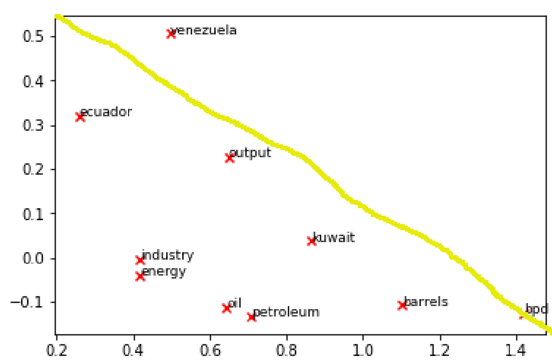


Answer:

oil and petroleum which have similar meaning are clustered. industry and energy are clustered. #### ecuador, kuwait, venezuela whitch are all petroleum production country name should have clustered,output and bpd have samilar meaning for petroleum production amount should have clustered

The difference this plot from the one generated eariler is: First, the scale(X, Y) is different, which shows the two embedding spaces are different. second, the word distribution is different between those 2 plot, in the plot generate eariler we could see all the words distrion is along/closed to a curved line, for the second one we just generated is spread in the left down side of the plot(See following 2figures). Third different is the evaluation result is different. in the plot generated earilier shows the oil contry are clustered together and close to oil. in the new one, we didn`t see this. Forth, for same word in those 2 word embedding plot, the word vector weight/value (coordinates) for same word is different. For example, for oil in first plot, it is (1.00, -0.05), the oil in second plot has (0.62, -0.11)

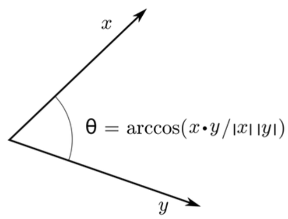




Cosine Similarity

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) 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 (2 points) [code + written]

Find a polysemous (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.keyedvectors.FastTextKeye**

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

wv_from_bin.most_similar("left")

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

Out[19]:
```
[('leaving', 0.6707000732421875),
 ('leave', 0.525093138217926),
 ('leaves', 0.5228644609451294),
 ('returned', 0.5059226751327515),
 ('right', 0.49213993549346924),
 ('departed', 0.49109700322151184),
 ('limped', 0.48599502444267273),
 ('went', 0.4719873070716858),
 ('remaining', 0.4650370478630066),
 ('empty', 0.4546155333518982)]
```

Answer:

The word I choose is "left", which has "leave" and "remaining".

Most of the meanings of the polysemous have are unrelated, it is possible that the meaning cluster for those meanings in the embedding spaces are not close to each other. and the number of similar words for each cluster could have more than the polysemous words and more closer to each other in word embedding spaces.which also means the one or some polysemous in the corpus traing the wv are more often used or "contexted" than others. and also only pick top 10, so it even reduces the chance to show the simility for polysemous in the list.

Question 2.3: Synonyms & Antonyms (2 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.keyedvectors.FastTextKeye**](https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvectors.FastTextKeye) for further assistance.

In [20]:
```
# ------------------
# Write your synonym & antonym exploration code here.

w1 = "large"
w2 = "gaint"
w3 = "small"
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 large, gaint have cosine distance: 0.8168608269597246
Antonyms large, small have cosine distance: 0.26688501049911195
```

Answer

The word I found is large, gaint, small.

Here`s the reason why this conter-intuitive result may have happened:
if we run wv_from_bin.most_similar("large") like question 2.2, we will get:

[('sizeable', 0.7341436147689819),
('small', 0.733115017414093),
('sizable', 0.7325241565704346),
('Large', 0.6654301285743713),
('huge', 0.6589166522026062),
('larger', 0.6517370939254761),
('massive', 0.6127927899360657),
('smaller', 0.6025481224060059),
('substantial', 0.5873902440071106),
('gigantic', 0.5858762860298157)] ,

if we run wv_from_bin.most_similar("small") we will get

[('large', 0.733115017414093),
('tiny', 0.718792736530304),
('medium_sized', 0.6426598429679871),
('Small', 0.6270469427108765),
('smaller', 0.619428813457489),
('minuscule', 0.5676835775375366),
('larger', 0.5446805953979492),
('mid_sized', 0.5328439474105835),
(midsized', 0.519831120967865),
('sizable', 0.5175076723098755)],

if we run wv_from_bin.most_similar("gaint"), we will get

[('giant', 0.5869466066360474),
('BOMBAY_XFN_ASIA', 0.4832102656364441),
('jt_venture', 0.4771811068058014),
('Videocon_Industries_Ltd.', 0.4737812578678131)
('TPV_Technology_Ltd', 0.46327781677246094),
('outsourcer_Wipro', 0.45941978693008423),
('RELI.BO', 0.45818576216697693),
('Ltd._Indiaā', 0.4561852216720581),
('Jindal_Saw_Ltd', 0.4555988311767578),
('retailer_Gome', 0.45485517382621765)]

For result above , we could see for large and small, the top 10 words are almost same and most focusing on the size property description. for gaint, which is also size property desction word, expecting showing similar result, but top 10 is showing most of them are company names. which means in the word embedding space, word vector for word "gaint" are closer to the company name clusters. and word vector for "large" and "small" are closer to size descriptions clusters. which also means in the corpus for traing w2v , the gaint are most often being used for comanies or it self represents big companines rather than being used for size.

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.FastTextKeye](https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvectors.FastTextKeye)**

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

```
[('queen', 0.7118192315101624),
 ('monarch', 0.6189675331115723),
 ('princess', 0.5902431011199951),
 ('crown_prince', 0.5499460697174072),
 ('prince', 0.5377321243286133),
 ('kings', 0.5236844420433044),
 ('Queen_Consort', 0.5235946178436279),
 ('queens', 0.5181134343147278),
 ('sultan', 0.5098592638969421),
 ('monarchy', 0.5087411999702454)]
```

Question 2.4: Finding Analogies [code + written] (2 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 [22]:
```
# ------------------
# Write your analogy exploration code here.

pprint.pprint(wv_from_bin.most_similar(positive=['man', 'actress'], n
egative=['woman']))
print()
pprint.pprint(wv_from_bin.most_similar(positive=['man', 'girl'], nega
tive=['woman']))
print()
pprint.pprint(wv_from_bin.most_similar(positive=['large', 'low'], neg
ative=['small']))


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

```
[('actor', 0.8202018141746521),
 ('Actor', 0.6390689611434937),
 ('funnyman', 0.6047691106796265),
 ('thespian', 0.5847228765487671),
 ('thesp', 0.5750559568405151),
 ('heart_throb', 0.5708590745925903),
 ('starlet', 0.5672707557678223),
 ('singer', 0.5657444000244141),
 ('Actress', 0.5615350008010864),
 ('ADRIEN_BRODY', 0.5556015372276306)]

[('boy', 0.8748372793197632),
 ('teenager', 0.6932151913642883),
 ('teenage_girl', 0.6277071237564087),
 ('lad', 0.624101996421814),
 ('kid', 0.6167631149291992),
 ('schoolboy', 0.600092351436615),
 ('teen_ager', 0.5852149724960327),
 ('boys', 0.5807873606681824),
 ('youngster', 0.5773991346359253),
 ('guy', 0.5459756255149841)]

[('high', 0.6629930734634399),
 ('Low', 0.5137864947319031),
 ('lower', 0.5012373924255371),
 ('higher', 0.49018242955207825),
 ('lowest', 0.48971161246299744),
 ('abysmally_low', 0.4538779556751251),
 ('highest', 0.4514465928077698),
 ('levels', 0.4280760586261749),
 ('rising', 0.4265591502189636),
 ('ahigh', 0.4251363277435303)]
```

Answer:

woman:actress::man:actor

woman:girl::man:boy

large:low::small:high

Question 2.5: Incorrect Analogy [code + written] (1 point)

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 [23]:
```
# ------------------
# Write your incorrect analogy exploration code here.

pprint.pprint(wv_from_bin.most_similar(positive=['woman','doctor'], n
egative=['man']))

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

```
[('gynecologist', 0.7093890905380249),
 ('nurse', 0.6477286219596863),
 ('doctors', 0.6471461653709412),
 ('physician', 0.64389967918396),
 ('pediatrician', 0.6249487400054932),
 ('nurse_practitioner', 0.6218313574790955),
 ('obstetrician', 0.6072014570236206),
 ('ob_gyn', 0.5986713171005249),
 ('midwife', 0.5927063226699829),
 ('dermatologist', 0.5739566683769226)]
```

Answer:

Intended: man:doctor::woman:physician, actual: man:doctor::woman:gynecologist, gynecologist should be more generalized.

Question 2.6: Guided Analysis of Bias in Word Vectors [written] (1 point)

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 [24]:
```
# Run this cell
# Here `positive` indicates the list of words to be similar to and `n
egative` indicates the list of words to be
# most dissimilar from.
2
print()
pprint.pprint(wv_from_bin.most_similar(positive=['man', 'boss'], nega
tive=['woman']))
```

```
[('supremo', 0.6097398400306702),
 ('MOTHERWELL_boss', 0.5489562153816223),
 ('CARETAKER_boss', 0.5375303626060486),
 ('Bully_Wee_boss', 0.533397376537323),
 ('YEOVIL_Town_boss', 0.5321705341339111),
 ('head_honcho', 0.5281979441642761),
 ('manager_Stan_Ternent', 0.5259714722633362),
 ('Viv_Busby', 0.5256163477897644),
 ('striker_Gabby_Agbonlahor', 0.5250812768936157),
 ('BARNSLEY_boss', 0.5238943696022034)]
```

Answer:

The prediction of positive=['woman', 'boss'], negative=['man']) is less appropriate than positive=['man', 'boss'], negative=['woman']， “receptionist”, "Coronation_Stree_actree", "coworker" are inappropriate even disrespectful like cornation_stree_actress, bosses is complex of boss, however man and woman are all singular

Question 2.7: Independent Analysis of Bias in Word Vectors [code + written] (2 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 [25]:
```
# ------------------
# Write your bias exploration code here.

pprint.pprint(wv_from_bin.most_similar(positive=['woman','doctor'], n
egative=['man']))
print()
pprint.pprint(wv_from_bin.most_similar(positive=['man','doctor'], neg
ative=['woman']))

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

```
[('gynecologist', 0.7093890905380249),
 ('nurse', 0.6477286219596863),
 ('doctors', 0.6471461653709412),
 ('physician', 0.64389967918396),
 ('pediatrician', 0.6249487400054932),
 ('nurse_practitioner', 0.6218313574790955),
 ('obstetrician', 0.6072014570236206),
 ('ob_gyn', 0.5986713171005249),
 ('midwife', 0.5927063226699829),
 ('dermatologist', 0.5739566683769226)]

[('physician', 0.646366536617279),
 ('doctors', 0.5858404636383057),
 ('surgeon', 0.572394073009491),
 ('dentist', 0.552364706993103),
 ('cardiologist', 0.5413815975189209),
 ('neurologist', 0.5271126627922058),
 ('neurosurgeon', 0.5249835848808289),
 ('urologist', 0.5247740149497986),
 ('Doctor', 0.5240625143051147),
 ('internist', 0.5183224081993103)]
```

Answer:

The prediction for positive=['man','doctor'], negative=['woman'] is much more accure than positive=['woman','doctor'], negative=['man'], we expect man:doctor:woman:physician for both two above. Woman and man should have equal opportunity of being doctor/physician rather than just gynecologist at the most, which is not equal for woman and inappropriate, it`s same problem as shown in question 2.6

Question 2.8: Thinking About Bias [written] (1 point)

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

Answer:

The cause of the biases may come from the corpus for training the word vector, some words has less context content existance consistent with the common human sense(correct answer)，so it is similar like when a machine leaning model training with a small number of dataset may have bias. For example, in question Qw.2.7, man:doctor::woman:gynecologist, however it should be physician as man:doctor::woman::physician, it may because woman have most context with gynecologist in the corpus due to most of the OB doctor is woman. so the wv for gynecologist is closer than physician

Submission Instructions

1. Click the Save button at the top of the Jupyter Notebook.
2. Please make a Gradescope account using your @stanford.edu email address (this is very important to help us enter your grade at the end of the quarter), and ensure your SUID is entered at the top of this notebook too.
3. Select Cell -> All Output -> Clear. This will clear all the outputs from all cells (but will keep the content of all cells).
4. Select Cell -> Run All. This will run all the cells in order, and will take several minutes.
5. Once you've rerun everything, select File -> Download as -> PDF via LaTeX
6. Look at the PDF file and make sure all your solutions are there, displayed correctly. The PDF is the only thing your graders will see!
7. Submit your PDF on Gradescope.

In [ ]: