XCS224N Assignment 1: Exploring Word Vectors (24 Points)

Before you start, make sure you read "XCS224N HW1 - Handout".

In [0]:

```
# All Import Statements Defined Here
# Note: Do not add to this list.
# All the dependencies you need can be installed by running this cell.
# Throughout this notebook you can run a cell by hitting CTRL+RETURN or the Play
button/icon at left
# ----------------

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 /root/nltk_data...
```

Your Name: Shravan Shetty

Your SCPD XID Number: X459416

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

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) 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:

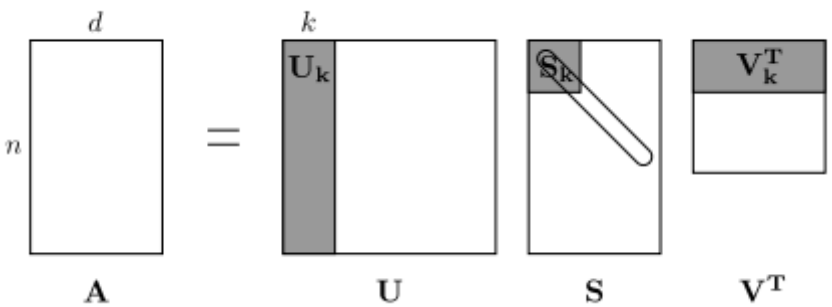
Document 1: "all that glitters is not gold"

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 these 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). 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), 8 (http://theory.stanford.edu/~tim/s15/l/l8.pdf), and 9 (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-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)* — 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 CTRL-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 [0]:

```
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_T
OKEN] for f in files]
```

Let's have a look what these documents are like…. (run the cell below)

In [0]:

```
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', 'ou
tlook', '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', 't
o', '550', 'mln', 'kilolitres',
  '(', 'kl', ')', 'from', '600', 'mln', ',', 'they', 'said', '.', 't
he', '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', 't
o', 'work', 'out', 'a', 'revised',
  'energy', 'supply', '/', 'demand', 'outlook', 'through', 'delibera
tions', 'of', 'committee',
  'meetings', 'of', 'the', 'agency', 'of', 'natural', 'resources',
'and', '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', 'bu
lk', 'of', 'japan', "'", 's',
  'electric', 'power', 'in', 'the', 'fiscal', 'year', 'ended', 'marc
h', '31', ',', 'supplying',
  'an', 'estimated', '27', 'pct', 'on', 'a', 'kilowatt', '/', 'hou
r', 'basis', ',', 'followed',
  'by', 'oil', '(', '23', 'pct', ')', 'and', 'liquefied', 'natural',
'gas', '(', '21', 'pct', '),',
  'they', 'noted', '.', '<END>'],
 ['<START>', 'energy', '/', 'u', '.', 's', '.', 'petrochemical', 'in
dustry', 'cheap', 'oil',
  'feedstocks', ',', 'the', 'weakened', 'u', '.', 's', '.', 'dolla
r', 'and', 'a', 'plant',
  'utilization', 'rate', 'approaching', '90', 'pct', 'will', 'prope
l', 'the', 'streamlined', 'u',
  '.', 's', '.', 'petrochemical', 'industry', 'to', 'record', 'profi
ts', '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', 'b
usinesses', 'has', 'prompted',
  'gaf', 'corp', '&', 'lt', ';', 'gaf', '>,', 'privately', '-', 'hel
d', 'cain', 'chemical', 'inc',
  ',', 'and', 'other', 'firms', 'to', 'aggressively', 'seek', 'acqui
sitions', 'of', 'petrochemical',
  'plants', '.', 'oil', 'companies', 'such', 'as', 'ashland', 'oil',
'inc', '&', 'lt', ';', 'ash',
  '>,', 'the', 'kentucky', '-', 'based', 'oil', 'refiner', 'and', 'm
arketer', ',', 'are', 'also',
  'shopping', 'for', 'money', '-', 'making', 'petrochemical', 'busin
esses', 'to', 'buy', '.', '"',
```

```
  'i', 'see', 'us', 'poised', 'at', 'the', 'threshold', 'of', 'a',
'golden', 'period', ',"', 'said',
  'paul', 'oreffice', ',', 'chairman', 'of', 'giant', 'dow', 'chemic
al', 'co', '&', 'lt', ';',
  'dow', '>,', 'adding', ',', '"', 'there', "'", 's', 'no', 'major',
'plant', 'capacity', 'being',
  'added', 'around', 'the', 'world', 'now', '.', 'the', 'whole', 'ga
me', 'is', 'bringing', 'out',
  'new', 'products', 'and', 'improving', 'the', 'old', 'ones', '."',
'analysts', 'say', 'the',
  'chemical', 'industry', "'", 's', 'biggest', 'customers', ',', 'au
tomobile', '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', 'joh
n', 'dosher', ',', 'managing',
  'director', 'of', 'pace', 'consultants', 'inc', 'of', 'houston',
'.', 'demand', 'for', 'some',
  'products', 'such', 'as', 'styrene', 'could', 'push', 'profit', 'm
argins', 'up', 'by', 'as',
  'much', 'as', '300', 'pct', ',', 'he', 'said', '.', 'oreffice',
',', 'speaking', 'at', 'a',
  'meeting', 'of', 'chemical', 'engineers', 'in', 'houston', ',', 's
aid', 'dow', 'would', 'easily',
  'top', 'the', '741', 'mln', 'dlrs', 'it', 'earned', 'last', 'yea
r', 'and', 'predicted', 'it',
  'would', 'have', 'the', 'best', 'year', 'in', 'its', 'history',
'.', 'in', '1985', ',', 'when',
  'oil', 'prices', 'were', 'still', 'above', '25', 'dlrs', 'a', 'bar
rel', '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', 'rec
ord', 'year', 'or', 'close',
  'to', 'it', ',"', 'oreffice', 'said', '.', 'gaf', 'chairman', 'sam
uel', 'heyman', 'estimated',
  'that', 'the', 'u', '.', 's', '.', 'chemical', 'industry', 'woul
d', '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', 'doshe
r', '.', 'dosher', 'said', 'last',
  'year', "'", 's', 'fall', 'in', 'oil', 'prices', 'made', 'feedstoc
ks', 'dramatically', 'cheaper',
  'and', 'at', 'the', 'same', 'time', 'the', 'american', 'dollar',
'was', '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', 'exte
nsive', 'corporate',
  'restructuring', 'program', 'to', 'mothball', 'inefficient', 'plan
ts', ',', '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', 'c
ash', 'for', 'chemical',
  'acquisitions', '.', 'but', 'some', 'experts', 'worry', 'that', 't
he', 'chemical', 'industry',
  'may', 'be', 'headed', 'for', 'trouble', 'if', 'companies', 'conti
nue', 'turning', 'their',
  'back', 'on', 'the', 'manufacturing', 'of', 'staple', 'petrochemic
al', 'commodities', ',', 'such',
  'as', 'ethylene', ',', 'in', 'favor', 'of', 'more', 'profitable',
'specialty', 'chemicals',
  'that', 'are', 'custom', '-', 'designed', 'for', 'a', 'small', 'gr
oup', 'of', 'buyers', '.', '"',
  'companies', 'like', 'dupont', '&', 'lt', ';', 'dd', '>', 'and',
'monsanto', 'co', '&', 'lt', ';',
  'mtc', '>', 'spent', 'the', 'past', 'two', 'or', 'three', 'years',
'trying', 'to', 'get', 'out',
  'of', 'the', 'commodity', 'chemical', 'business', 'in', 'reactio
n', 'to', 'how', 'badly', 'the',
  'market', 'had', 'deteriorated', ,"', 'dosher', 'said', '.', '"',
'but', 'i', 'think', 'they',
  'will', 'eventually', 'kill', 'the', 'margins', 'on', 'the', 'prof
itable', 'chemicals', 'in',
  'the', 'niche', 'market', '."', 'some', 'top', 'chemical', 'execut
ives', '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', 'no
t', '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', 'i
ndustry', 'trend', '.',
  'chairman', 'gordon', 'cain', ',', 'who', 'previously', 'led',
'a', 'leveraged', 'buyout', 'of',
  'dupont', "'", 's', 'conoco', 'inc', "'", 's', 'chemical', 'busine
ss', ',', 'has', 'spent', '1',
  '.', '1', 'billion', 'dlrs', 'since', 'january', 'to', 'buy', 'sev
en', 'petrochemical', 'plants',
  'along', 'the', 'texas', 'gulf', 'coast', '.', 'the', 'plants', 'p
roduce', 'only', 'basic',
  'commodity', 'petrochemicals', 'that', 'are', 'the', 'building',
'blocks', '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', 'd
ean', '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', 'upt
urn', 'in', 'the', 'chemical',
  'business', 'could', 'last', 'as', 'long', 'as', 'four', 'or', 'fi
ve', 'years', ',', 'provided',
  'the', 'u', '.', 's', '.', 'economy', 'continues', 'its', 'modes
t', 'rate', 'of', 'growth', '.',
  '<END>'],
 ['<START>', 'turkey', 'calls', 'for', 'dialogue', 'to', 'solve', 'd
ispute', 'turkey', 'said',
  'today', 'its', 'disputes', 'with', 'greece', ',', 'including', 'r
ights', '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', 'ca
n', 'only', 'be', 'found', 'by',
  'bilateral', 'negotiations', ',"', 'the', 'statement', 'said',
'.', 'greece', 'has', 'repeatedly',
  'said', 'the', 'issue', 'was', 'legal', 'and', 'could', 'be', 'sol
ved', 'at', 'the',
  'international', 'court', 'of', 'justice', '.', 'the', 'two', 'cou
ntries', 'approached', 'armed',
  'confrontation', 'last', 'month', 'after', 'greece', 'announced',
'it', 'planned', 'oil',
  'exploration', 'work', 'in', 'the', 'aegean', 'and', 'turkey', 'sa
id', 'it', 'would', 'also',
  'search', 'for', 'oil', '.', 'a', 'face', '-', 'off', 'was', 'aver
ted', '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', 'st
atement', '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', 'mini
ster', 'turgut', 'ozal', '.',
  'the', 'contents', 'of', 'the', 'message', 'were', 'not', 'disclos
ed', '.', '<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) 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).

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

In [0]:

```
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 co
rpus, sorted (using python 'sorted' function)
            num_corpus_words (integer): number of distinct words across the corp
us
    """
    corpus_words = []
    num_corpus_words = -1

### SOLUTION BEGIN
    corpus_words = set([ w for d in corpus for w in d ])
    num_corpus_words = len(corpus_words)

### SOLUTION END

    return sorted(corpus_words), num_corpus_words
```

In [0]:

```
# ---------------------
# 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.\nC
orrect: {}\nYours:   {}".format(str(ans_test_corpus_words), str(test_corpus_word
s))

# 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/).

In [0]:

```
def compute_co_occurrence_matrix(corpus, window_size=4):
    """ Compute co-occurrence matrix for the given corpus and window_size (defau
lt of 4).

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

              For example, if we take the document "START All that glitters is n
ot gold END" with window size of 4,
              "All" will co-occur with "START", "that", "glitters", "is", and "n
ot".

        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 unique words in the corpus , num
ber of unique words in the corpus):
                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 = {}

### SOLUTION BEGIN
    M = np.zeros((num_words, num_words))
    word2Ind = {w:i for i,w in enumerate(words)}
    for d in corpus:
      for i, w in enumerate(d):
        center = word2Ind[w]
        for j in range(i - window_size, i + window_size + 1 ):
          if -1 < j < len(d) and j != i:
            M[center][word2Ind[d[j]]] += 1
### SOLUTION END
    return M, word2Ind
```

In [0]:

```
# ---------------------
# 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, window_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, 'glitter
s': 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 shape.\nCorre
ct: {}\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, idx2, 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).

In [0]:

```
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.decomposi
tion.TruncatedSVD.html

        Params:
            M (numpy matrix of shape (number of unique words in the corpus , num
ber of unique words in the corpus)): 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)): matri
x of k-dimensional 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]))

    ### SOLUTION BEGIN
    svd = TruncatedSVD(k, n_iter=100)
    M_reduced = svd.fit_transform(M)
    ### SOLUTION END
    print("Done.")
    return M_reduced
```

In [0]:

```
# ---------------------
# 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/). In the future, a good way to make a plot is to look at the Matplotlib gallery (https://matplotlib.org/gallery/index.html), find a plot that looks somewhat like what you want, and adapt the code they give.

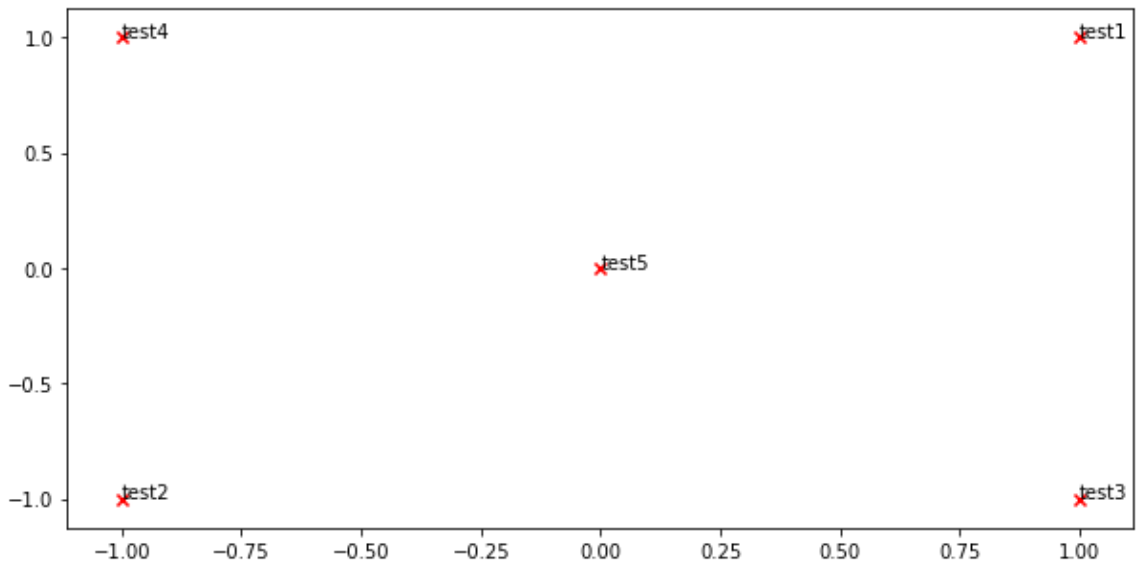
In [0]:

```
def plot_embeddings(M_reduced, word2Ind, words):
    """ Plot in a scatterplot the embeddings of the words specified in the list
 "words".
        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 corp
us , 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 to visualize
    """


### SOLUTION BEGIN
    for (x,y),w in zip(M_reduced, words):
      plt.scatter(x,y, c='r', marker = 'x')
      plt.text(x,y, w)



### SOLUTION END
```

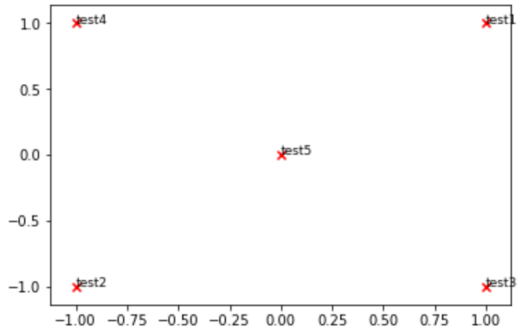
In [0]:

```
# ---------------------
# 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': 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).

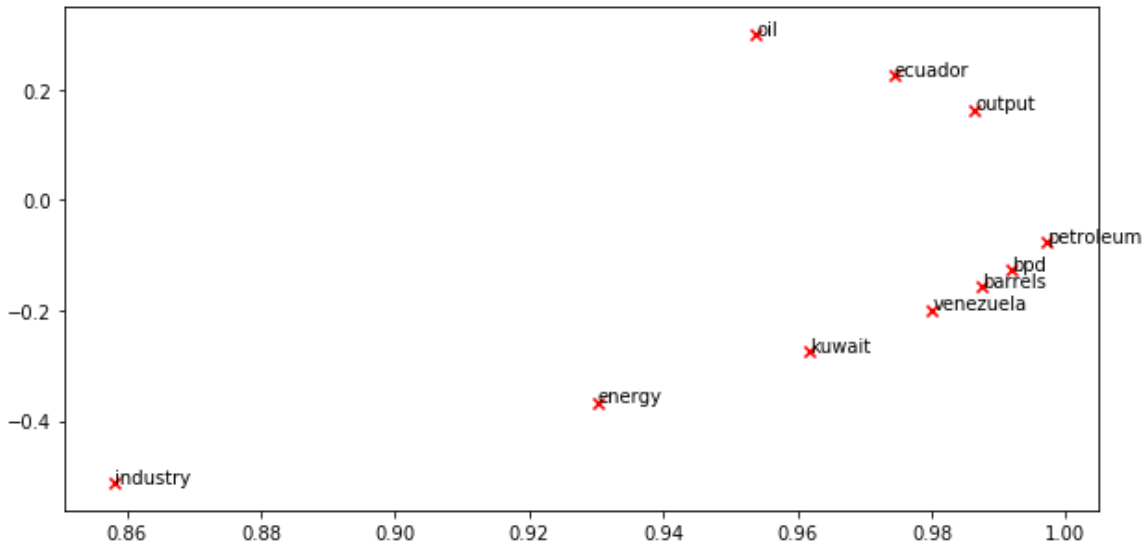
Run the below cell to produce the plot. It'll probably take a few seconds to run.

Written Question: What clusters together in 2-dimensional embedding space (in the given plot)? 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 [0]:

```
# ----------------------------
# Run This Cell to Produce Your Plot
# ----------------------------
reuters_corpus = read_corpus()
M_co_occurrence, word2Ind_co_occurrence = compute_co_occurrence_matrix(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] # broadcasting

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.
```



In [0]:

```
len(word2Ind_co_occurrence)
```

Out[0]:

```
8185
```

Write your answer here.

It seems that there are 4 Clusters formed with one containining petroleum, barrel, bpd, and venezuela. Second with oil, ecuador, and output. Third with Kuwait and energy. Fourth with Industry.

Oil and petroleum should have been clustered together. Oil producting countries like Ecuador, Kuwait, and veneuzela should have been close based on semantics.

Part 2: Prediction-Based Word Vectors (14 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).

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

In [0]:

```
def load_word2vec():
    """ Load Word2Vec Vectors
        Return:
            wv_from_bin: 2.5 million of 3 million embeddings, each lengh 300
    """
    import gensim.downloader as api
    from gensim.models import KeyedVectors
    # let's load 2.5 million of the 3 million word embeddings so we don't run ou
t of memory on Colab
    wv_from_bin = KeyedVectors.load_word2vec_format(api.load("word2vec-google-ne
ws-300", return_path=True), limit=2500000, binary=True)
    vocab = list(wv_from_bin.vocab.keys())
    print("Loaded vocab size %i" % len(vocab))
    return wv_from_bin
```

In [0]:

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

```
[==================================================] 100.0% 1662.8/1
662.8MB downloaded

/usr/local/lib/python3.6/dist-packages/smart_open/smart_open_lib.py:
398: UserWarning: This function is deprecated, use smart_open.open i
nstead. See the migration notes for details: https://github.com/RaRe
-Technologies/smart_open/blob/master/README.rst#migrating-to-the-new
-open-function
  'See the migration notes for details: %s' % _MIGRATION_NOTES_URL

Loaded vocab size 2500000
```

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 2.5 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 [0]:

```
def get_matrix_of_vectors(wv_from_bin, required_words=['barrels', 'bpd', 'ecuador', 'energy', 'industry', 'kuwait', 'oil', 'output', 'petroleum', 'venezuela']):
    """ Put the word2vec vectors into a matrix M.
        Param:
            wv_from_bin: KeyedVectors object; the 2.5 million word2vec vectors loaded from file
        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.vocab.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.word_vec(w))
            word2Ind[w] = curInd
            curInd += 1
        except KeyError:
            continue
    for w in required_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 [0]:

```
# -----------------------------------------------------------------
# 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)
```

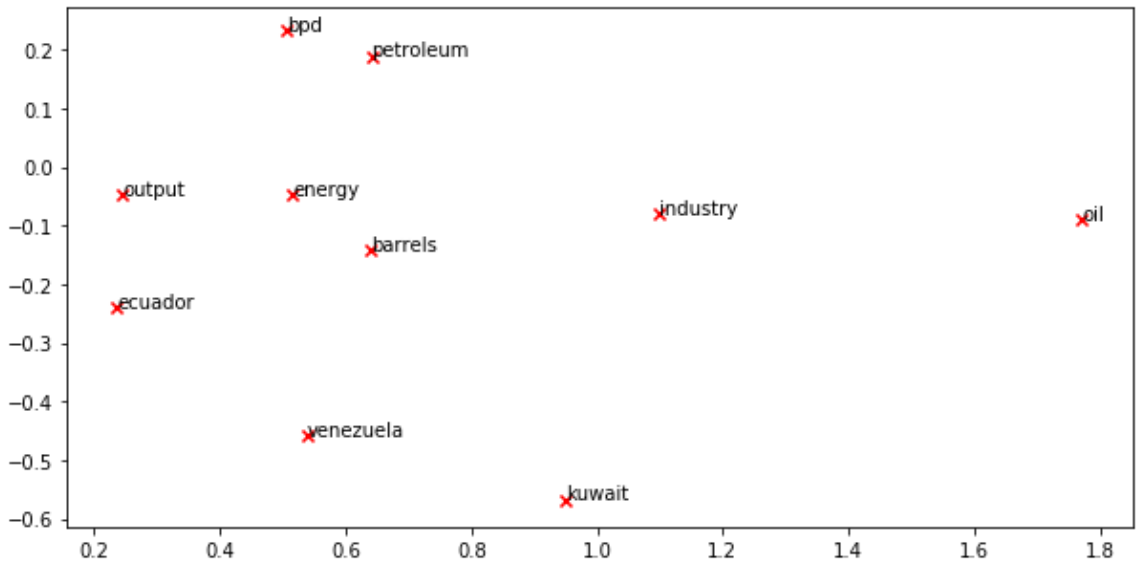
```
Shuffling words ...
Putting 10000 words into word2Ind and matrix M...
Done.
Running Truncated SVD over 10010 words...
Done.
```

Question 2.1: Word2Vec Plot Analysis [written] (2 points)

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

In [0]:

```
words = ['barrels', 'bpd', 'ecuador', 'energy', 'industry', 'kuwait', 'oil', 'ou
tput', 'petroleum', 'venezuela']
plot_embeddings(M_reduced, word2Ind, words)
```



Multiple Choice Question: Why aren't countries "venezuela", "ecuador" and "kuwait" clustered together in the Word2Vec plot while they were clustered together in the co-occurence plot? - State All That Apply

A) Word2Vec was trained on a larger dataset in which the countries did not always appear in the same context as the small dataset, used to compute the co-occurence matrix

B) The countries are not geographically close to each other

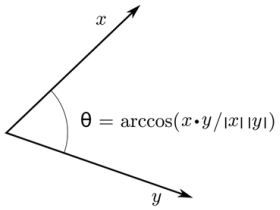
Write your answer here. A.

Other case might be the when using PCA, there is lot of information that is lost and that might be the cause of why these countries are not appearing to be clusetered.

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: Homophonous Words (2 points) [code + written]

Find a homophonous (https://en.wikipedia.org/wiki/Homophony) 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 homophonous words before you find one.**

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.FastTextK**

Written Question: State the homophonous word you discover and the multiple meanings that occur in the top 10. Why do you think many of the homophonous words you tried didn't work?

In [0]:

```
### SOLUTION BEGIN

wv_from_bin.most_similar("lead")

### SOLUTION END
```

```
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Futur
eWarning: Conversion of the second argument of issubdtype from `int`
to `np.signedinteger` is deprecated. In future, it will be treated a
s `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):
```

Out[0]:

```
[('advantage', 0.5628466606140137),
 ('trailed', 0.5312870740890503),
 ('cushion', 0.4993368983268738),
 ('led', 0.4984801411628723),
 ('midway_through', 0.483384370803833),
 ('thelead', 0.48188185691833496),
 ('leads', 0.47565916180610657),
 ('commanding', 0.4725237488746643),
 ('lead.The', 0.466875284910202),
 ('Lead', 0.4649958610534668)]
```

In [0]:

```
wv_from_bin.most_similar("record")
```

```
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Futur
eWarning: Conversion of the second argument of issubdtype from `int`
to `np.signedinteger` is deprecated. In future, it will be treated a
s `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):
```

Out[0]:

```
[('records', 0.6369920969009399),
 ('equaling', 0.5535237789154053),
 ('Record', 0.5457872152328491),
 ('equaled', 0.5234976410865784),
 ('eclipsing', 0.5121445655822754),
 ('surpassing', 0.5103195905685425),
 ('mark', 0.5074685215950012),
 ('#,###-###_.###', 0.5008059144020081),
 ('recod', 0.48859328031539917),
 ('shattering_Roger_Maris', 0.48826664686203003)]
```

Write your answer here.

"lead" included "advantage" which determines leading a score,"Lead" as a chemical element, "led" and "commanding" which means a leader.

Other homophobic like "record" was biased towards surpassing or equaling records of something. There was no similarity with video or audio, medical record, criminal record. This may be the fact that context words in Google news near word "record" are always asociated to making/breaking a record and less about particualr domain like health-care.

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". (1 point)

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.FastTextK** for further assistance.

Written Question: State which three words (w1, w2, w3) you found. What are some possible explanations for why this counterintuitive result happened?

In [0]:

```
### SOLUTION BEGIN

w1 = "large"
w2 = "largest"
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))



### SOLUTION END
```

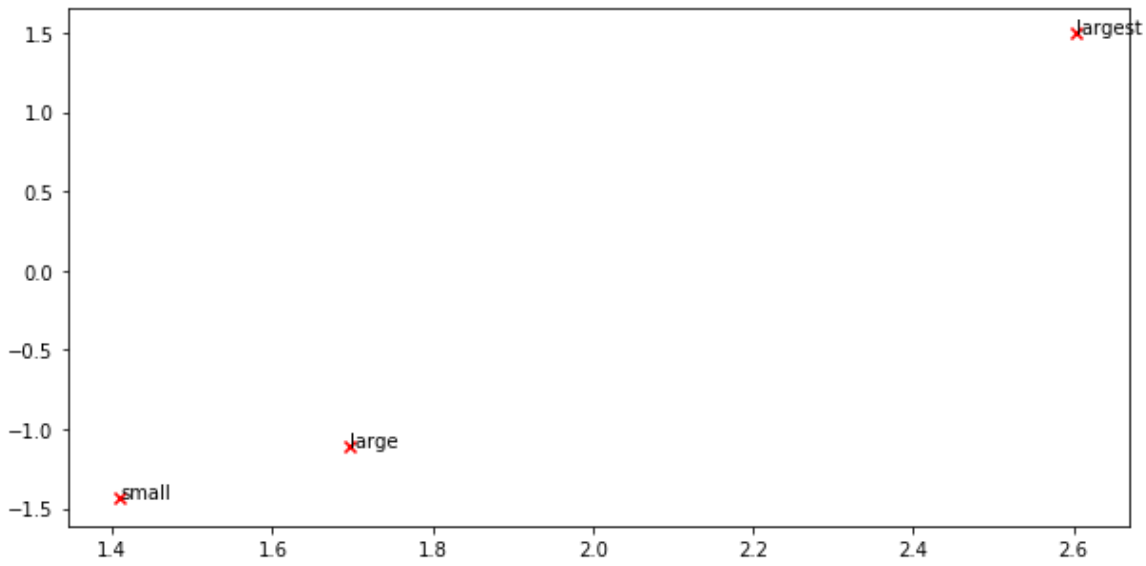
```
Synonyms large, largest have cosine distance: 0.5916160047054291
Antonyms large, small have cosine distance: 0.266884982585907

/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Futur
eWarning: Conversion of the second argument of issubdtype from `int`
to `np.signedinteger` is deprecated. In future, it will be treated a
s `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):
```

Write your answer here. Cosine similarity, althought helps in identifying the similarity of the two text vector, cannot handle the semantic meaning of the text.
Even in the below plot of reduced dimension, we see large and largest with further than large and small, with the cosine similarity value concuring with the results.

In [0]:

```
from sklearn.metrics.pairwise import cosine_similarity
w1 = "largest"
w2 = "large"
w3 = "small"
v1 = wv_from_bin.word_vec(w1)
v2 = wv_from_bin.word_vec(w2)
v3 = wv_from_bin.word_vec(w3)
M1 = np.stack([v1, v2, v3])
k = reduce_to_k_dim(M1, 2)
plot_embeddings(k,None, [w1,w2,w3])
print(f"Checking similarity of {w1} and {w2}", cosine_similarity([k[0]],[k[1]]))
print(f"Checking similarity of {w3} and {w2}", cosine_similarity([k[2]],[k[1]]))
```

```
Running Truncated SVD over 3 words...
Done.
Checking similarity of largest and large [[0.4531818]]
Checking similarity of small and large [[0.97732747]]
```



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

In [0]:

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

```
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Futur
eWarning: Conversion of the second argument of issubdtype from `int`
to `np.signedinteger` is deprecated. In future, it will be treated a
s `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):

[('queen', 0.7118192911148071),
 ('monarch', 0.6189674139022827),
 ('princess', 0.5902431011199951),
 ('crown_prince', 0.5499460697174072),
 ('prince', 0.5377321243286133),
 ('kings', 0.5236844420433044),
 ('Queen_Consort', 0.5235945582389832),
 ('queens', 0.518113374710083),
 ('sultan', 0.5098593235015869),
 ('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).
Note: You may have to try many analogies to find one that works!

Written Question: State the successful analogy you found in the form x:y :: a:b. If you believe the analogy is complicated, give a short explanation as to why it holds.

In [0]:

```
### SOLUTION BEGIN


pprint.pprint(wv_from_bin.most_similar(
    positive=['good','sad'], negative=['bad']))

# ### SOLUTION END
```

```
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Futur
eWarning: Conversion of the second argument of issubdtype from `int`
to `np.signedinteger` is deprecated. In future, it will be treated a
s `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):

[('wonderful', 0.6414927244186401),
 ('happy', 0.6154337525367737),
 ('great', 0.5803680419921875),
 ('nice', 0.5683972835540771),
 ('saddening', 0.5588892698287964),
 ('bittersweet', 0.5544660687446594),
 ('glad', 0.551203727722168),
 ('fantastic', 0.5471093654632568),
 ('proud', 0.5305150747299194),
 ('saddened', 0.5293528437614441)]
```

Write your answer here. In this example, I wanted to find analogy of sad in terms of bad:good. I got very high results of positive words like wonderful, happy, nice and

Question 2.5: Incorrect Analogy [code + written] (2 point)

Find an example of analogy that does **not** hold according to these vectors.

Written Question: State the intended analogy in the form x:y :: a:b, and state the (incorrect) value of b according to the word vectors.

In [0]:

```
### SOLUTION BEGIN

pprint.pprint(wv_from_bin.most_similar(positive=["wine", "pepsi"], negative=['be
er']))

pprint.pprint(wv_from_bin.most_similar(
    positive=['lock','goldsmith'], negative=['locksmith']))

### SOLUTION END
```

```
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Futur
eWarning: Conversion of the second argument of issubdtype from `int`
to `np.signedinteger` is deprecated. In future, it will be treated a
s `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):

[("s'il_vous_plaît", 0.5249711275100708),
 ('danielle', 0.4998072385787964),
 ('Loire_wines', 0.4923233687877655),
 ('Pinot_grigio', 0.4906075894832611),
 ('asti', 0.4899405539035797),
 ('gbr', 0.4868606925010681),
 ('Savennieres', 0.4867655634880066),
 ('brunello', 0.48486918210983276),
 ('Beyers_Truter', 0.4846540093421936),
 ('whitney', 0.4742027521133423)]
[('locking', 0.37111473083496094),
 ('locked', 0.3384663462638855),
 ('hilts', 0.33627286553382874),
 ('goldsmiths', 0.33505356311798096),
 ('blued_steel', 0.32900142669677734),
 ('silversmiths', 0.32708287239074707),
 ('bangle', 0.32086703181266785),
 ('exquisitely_carved', 0.31835252046585083),
 ('bangles', 0.3160739541053772),
 ('silver_gilt', 0.3150860667228699)]
```

Write your answer here. wine:beer::pepsi:coke I was expecting coke-cola or any softdrink that was possible, but it seems to be giving out collection of wines.

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".

Written Question: Point out one difference between the list of female-associated words and the list of male-associated words, and explain how it is reflecting a potential gender bias. Where the search has returned named entities, you may want to perform a quick search to understand what/who they refer to.

In [0]:

```
# Run this cell
# Here `positive` indicates the list of words to be similar to and `negative` in
dicates the list of words to be
# most dissimilar from.
pprint.pprint(wv_from_bin.most_similar(positive=['woman', 'boss'], negative=['ma
n']))
print()
pprint.pprint(wv_from_bin.most_similar(positive=['boss', 'man'], negative=['woma
n']))
```

```
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Futur
eWarning: Conversion of the second argument of issubdtype from `int`
to `np.signedinteger` is deprecated. In future, it will be treated a
s `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):

[('bosses', 0.5522644519805908),
 ('manageress', 0.49151360988616943),
 ('exec', 0.459408164024353),
 ('Manageress', 0.45598435401916504),
 ('receptionist', 0.4474116861820221),
 ('Jane_Danson', 0.44480547308921814),
 ('Fiz_Jennie_McAlpine', 0.44275766611099243),
 ('Coronation_Street_actress', 0.44275569915771484),
 ('supremo', 0.4409852921962738),
 ('coworker', 0.4398624897003174)]

[('supremo', 0.6097397804260254),
 ('MOTHERWELL_boss', 0.5489562153816223),
 ('CARETAKER_boss', 0.5375303626060486),
 ('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),
 ('WIGAN_boss', 0.5175146460533142)]
```

Write your answer here. a. Mangeress,Receptionist, actress seems like a gender bias

b. MOTHERWELL_boss, YEOVIL_Town Boss, BARNS_boss etc are soccer club manager which has been biased

to a Male boss

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.

Written Question: Briefly explain the additional example of bias that you discover.

In [0]:

```
### SOLUTION BEGIN

pprint.pprint(wv_from_bin.most_similar(positive=['latino', 'burger'], negative=[
'white']))
pprint.pprint(wv_from_bin.most_similar(positive=["female", "doctor"], negative=[
"male"]))
# print()
# pprint.pprint(wv_from_bin.most_similar(positive=[], negative=[]))

### SOLUTION END
```

```
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Futur
eWarning: Conversion of the second argument of issubdtype from `int`
to `np.signedinteger` is deprecated. In future, it will be treated a
s `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):

[('Burrito', 0.5746285319328308),
 ('burrito', 0.568935751914978),
 ('taco', 0.5676559209823608),
 ('Tacos', 0.5324654579162598),
 ('burritos', 0.5316759943962097),
 ('burgers', 0.5200715065002441),
 ('vegetarian_burrito', 0.5107964277267456),
 ('carnitas', 0.5081146955490112),
 ('taqueria', 0.5071629881858826),
 ('Tex_Mex', 0.5069916248321533)]
[('physician', 0.6744842529296875),
 ('doctors', 0.6639320850372314),
 ('nurse', 0.6262701153755188),
 ('gynecologist', 0.6154145002365112),
 ('surgeon', 0.61018967628479),
 ('pediatrician', 0.5857931971549988),
 ('dentist', 0.5751978158950806),
 ('nurse_practitioner', 0.5651593208312988),
 ('neurologist', 0.5592567920684814),
 ('oncologist', 0.5587260127067566)]
```

Write your answer here. In the first example, it shows an ethnic bias where in white:burger::latino shows burrito and all the mexican food.

Second example is related to occupation, where in doctor related female occupation shows nurse, nurse practicioner.

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

Multiple Choice Question: What factors might contribute to the biases observed in the word vectors? (State all that apply)

A) These biases may come from bias in the text (possibly held by people generating the text) that was used as training data

B) The training corpus may have many instances of sentences that relate certain races, genders, etc. to certain properties or behaviours

C) These biases come from the dictionary meaning of words

Write your answer here. A and B

Submission Instructions

1. Please make sure you have entered your name and SCPD XID Number above
2. Click the Save button at the top ("File > Save")
3. Select "Edit > Clear All Outputs". This will clear all the outputs from all cells (but will keep the content of all cells).
4. Select "Runtime > Run All". This will run all the cells in order, and will take several minutes.
5. Once you've rerun everything, select "File > Print > Save as PDF"
6. Look at the PDF file and make sure all your solutions are there, displayed correctly including the output cells. The PDF is the only thing your graders will see!
7. Submit your PDF via the Gradescope submission link in the Assignment 1 block of your SCPD learning portal. The system will ask you to take 1-2 minutes to tag pages of your PDF to the corresponding response item. You can see a demonstration of this submission process at this link - https://youtu.be/yocclo79qh4 (https://youtu.be/yocclo79qh4)

In [0]: