NLP Assessment

The purpose of this project is to analyze a dataset of news articles in two ways:

1. Parsing names and dates from each article in a useable format.
2. Applying text analytics to identify 5 news articles from the dataset that are most similar to a given sample article.

This project is an assessment given by Deloitte for potential Natural Language Processing work with the Centers for Disease Control and Prevention.

Summary of Results

After analyzing and processing the data, I determined that a Sentence Embedding model was the most accurate at predicting document similarity. A TF-IDF model and a model utilizing Spacy's internal similarity tools were far less performant.

For the final output of this project, I have created two files:

- 'texts_processed.csv': a final csv file of all relevant information, including the columns of names, distinct names, dates and datetime objects.
- 'closest_matches.csv': a sampling of the top five articles in similarity to the sample article, using the Sentence Embedding method.

Future Recommendations

With more time to devote to this project, there are several changes that could improve accuracy:

- Experimenting with various Spacy models, especially for vectorization/similarity testing.
- Training a custom model/class to detect document similarity utilizing the spacy or flair library.
- Utilizing Spark to increase compute power and decrease run time for text/date classification.

Ingesting and Analyzing Dataset

The source dataset was provided by Deloitte, as a csv file of unstructured text data from several news articles.

In [2]:
```
# Importing basic libraries
import pandas as pd
import numpy as np
import datetime

# Importing flair library for name recognition
from flair.data import Sentence, build_spacy_tokenizer
from flair.models import SequenceTagger

# Importing segtok segmenter
from segtok.segmenter import split_single

# Importing spacy for date recognition
import spacy
from spacy.tokenizer import Tokenizer

# Importing ctparse and timefhuman for converting date text into datetim
e objects
from ctparse import ctparse
from timefhuman import timefhuman

# Importing stopword, tokenizer and stemmer from nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
import nltk

# Importing tensorflow hub for Sentence Encoder
import tensorflow as tf
import tensorflow_hub as hub
```

```
/Users/eddiekirkland/anaconda3/lib/python3.7/site-packages/sklearn/bas
e.py:318: UserWarning: Trying to unpickle estimator CountVectorizer fro
m version 0.20.4 when using version 0.22.1. This might lead to breaking
code or invalid results. Use at your own risk.
  UserWarning)
/Users/eddiekirkland/anaconda3/lib/python3.7/site-packages/sklearn/bas
e.py:318: UserWarning: Trying to unpickle estimator MultinomialNB from
version 0.20.4 when using version 0.22.1. This might lead to breaking c
ode or invalid results. Use at your own risk.
  UserWarning)
/Users/eddiekirkland/anaconda3/lib/python3.7/site-packages/sklearn/bas
e.py:318: UserWarning: Trying to unpickle estimator Pipeline from versi
on 0.20.4 when using version 0.22.1. This might lead to breaking code o
r invalid results. Use at your own risk.
  UserWarning)
```

In [3]:
```
# Increasing max column width to display text
pd.set_option('max_colwidth', 10000)
```

In [4]:
```
# Reading data into dataframe
text_df = pd.read_csv('News-article-wikipedia-DFE.csv')
```

In [4]:
```
text_df.info(verbose=True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 14 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   _unit_id
3000 non-null   int64
 1   _unit_state
3000 non-null   object
 2   _trusted_judgments
3000 non-null   int64
 3   _last_judgment_at
3000 non-null   object
 4   given_the_news_article_headline_and_description_above_please_selec
t_the_most_relevant_wikipedia_page_from_the_following_options
3000 non-null   object
 5   given_the_news_article_headline_and_description_above_please_selec
t_the_most_relevant_wikipedia_page_from_the_following_options:confidenc
e  3000 non-null   float64
 6   wikipedia_page2__
3000 non-null   object
 7   article
3000 non-null   object
 8   gurl
3000 non-null   object
 9   id
3000 non-null   int64
 10  newdescp
3000 non-null   object
 11  nil
1 non-null      object
 12  option3
3000 non-null   object
 13  oururl
3000 non-null   object
dtypes: float64(1), int64(3), object(10)
memory usage: 328.2+ KB
```

It looks as though there are no null values in the dataset, and 3000 entries for each feature.

In [5]:
```
# Exploring dataframe
text_df.head()
```

Out[5]:

	_unit_id	_unit_state	_trusted_judgments	_last_judgment_at	given_the_news_article_headline_
0	691201838	finalized	5	3/19/2015 19:59	

	_unit_id	_unit_state	_trusted_judgments	_last_judgment_at	given_the_news_article_headline_
1	691201839	finalized	5	3/19/2015 20:34	

	_unit_id	_unit_state	_trusted_judgments	_last_judgment_at	given_the_news_article_headline_
2	691201840	finalized	5	3/19/2015 3:01	

_unit_id	_unit_state	_trusted_judgments	_last_judgment_at	given_the_news_article_headline_

_unit_id	_unit_state	_trusted_judgments	_last_judgment_at	given_the_news_article_headline_

	_unit_id	_unit_state	_trusted_judgments	_last_judgment_at	given_the_news_article_headline_
3	691201841	finalized	5	3/20/2015 6:27	
4	691201842	finalized	5	3/20/2015 6:51	

_unit_id	_unit_state	_trusted_judgments	_last_judgment_at	given_the_news_article_headline_

_unit_id	_unit_state	_trusted_judgments	_last_judgment_at	given_the_news_article_headline_

It seems as though much of the information in the dataset is not useful for the purpose of this project. I will simplify the dataframe through selecting relevant columns:

- 'article': the title of each article
- 'oururl': the url of the article's text
- 'newdescp': the basic text of each article

In [5]:
```
# Keeping only relevant columns
keepcols = ['article', 'oururl', 'newdescp']
text_simple = text_df[keepcols]
```

Part 1: Name and Date Recognition

Name Recognition

To identify names within the text of each article, I will utilize two external libraries:

- Spacy: Used to tokenize the data into sentences. Selected for fast and accurate performance.
- Flair: Used to tokenize spans within each sentence. Selected for its highly intuitive understanding of the context of tokens and token spans.

After identifying a list of names, I will create an additional column containing only distinct names.

Note: I considered removing all single-word names such as "Ghadafi," since many are duplicates of the first/last names located within the article. However, simply removing these names could eliminate important information such as people referred to only by last name. For this reason, I left the list as-is.

In [12]:
```
# Downloading named entity recognition model
tagger = SequenceTagger.load('ner')

# Creating spacy tokenizer for parsing sentences
nlp = spacy.load('en')
tokenizer = build_spacy_tokenizer(nlp)

# Defining function for tagging named entities in text
def find_names(text):
    # Creating empty list for name tokens
    names = []
    sentences = [Sentence(sent, use_tokenizer=tokenizer) for sent in spl
it_single(text)]
    tagger.predict(sentences)
    for sent in sentences:
        for entity in sent.get_spans('ner'):
            if entity.tag=='PER':
                names.append(entity.text)
            else:
                pass
    return names
```

```
2020-03-20 15:52:39,619 loading file /Users/eddiekirkland/.flair/model
s/en-ner-conll03-v0.4.pt
```

In [13]:
```
# Applying function to text and creating column of names
text_simple['names'] = text_df['newdescp'].apply(find_names)
```

```
2020-03-20 15:59:32,664 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 15:59:32,666 Ignore 1 sentence(s) with no tokens.
2020-03-20 16:55:47,542 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 16:55:47,549 Ignore 1 sentence(s) with no tokens.
2020-03-20 17:19:57,990 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 17:20:30,706 Ignore 1 sentence(s) with no tokens.
2020-03-20 17:41:22,967 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 17:41:22,969 Ignore 1 sentence(s) with no tokens.
2020-03-20 18:45:08,357 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 18:45:08,360 Ignore 1 sentence(s) with no tokens.
2020-03-20 19:18:21,727 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 19:18:21,729 Ignore 1 sentence(s) with no tokens.
2020-03-20 19:24:44,244 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 19:24:44,245 Ignore 1 sentence(s) with no tokens.
2020-03-20 20:58:59,907 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 20:58:59,909 Ignore 1 sentence(s) with no tokens.
2020-03-20 21:01:50,005 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 21:01:50,007 Ignore 1 sentence(s) with no tokens.
2020-03-20 21:10:04,642 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 21:10:04,643 Ignore 1 sentence(s) with no tokens.
2020-03-20 22:35:51,588 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 22:35:51,590 Ignore 1 sentence(s) with no tokens.
2020-03-20 22:45:53,566 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 22:45:53,567 Ignore 1 sentence(s) with no tokens.
2020-03-20 22:52:30,500 ACHTUNG: An empty Sentence was created! Are the
re empty strings in your dataset?
2020-03-20 22:52:30,502 Ignore 1 sentence(s) with no tokens.

/Users/eddiekirkland/anaconda3/lib/python3.7/site-packages/ipykernel_la
uncher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

In [14]:
```
# Saving analyzed results to file in case of needed restart
text_simple.to_csv('text_simple_names.csv')
```

In [15]:
```
text_simple.head()
```

Out[15]:

	article	oururl	newdescp
0	Gaza aid ship to dock in Egypt after Israel pressure	http://en.wikipedia.org/wiki/Gaza_flotilla_raid	A ship with supplies for Gaza will dock at el-Arish in Egypt, officials say, after Israeli pressure to stop the vessel breaking its Gaza blockade. The Moldovan-flagged ship chartered by a charity run by the son of Libyan leader Col Muammar Gaddafi, left a Greek port on Saturday. Israel asked for help from the UN, and had talks with Greece and Moldova. But organisers insist they will go to Gaza. An Israeli raid on a Gaza-bound ship in May killed nine Turkish activists. Israel insisted its troops were defending themselves but the raid sparked international condemnation. Israel recently eased its blockade, allowing in almost all consumer goods but maintaining a "blacklist" of some items. Israel says its blockade of the Palestinian territory is needed to prevent the supply of weapons to the Hamas militant group which controls

	article	oururl	newdescp
			Gaza. The Amalthea, renamed Hope for the mission, set off from the Greek port of Lavrio, loaded with about 2,000 tonnes of food, cooking oil, medicines and pre-fabricated houses. It has been chartered by the Gaddafi International Charity and Development Foundation. Its chairman is Saif al-Islam Gaddafi. The organisation said the 92m (302ft) vessel would also carry "a number of supporters who are keen on expressing solidarity with the Palestinian people".
1	Mel Gibson	http://en.wikipedia.org/wiki/Mel_Gibson_filmography	Often acts and directs stories involving an individual who is persecuted, and fights for justice Has often portrayed a widower, in films such as Mad Max (the sequels), Lethal Weapon film series, Braveheart, The Patriot, Signs, and Edge of Darkness. Often portrays men who seek revenge for the murder of family or friends Ranked #12 in Empire (UK) magazine's "The Top 100 Movie Stars of All Time" list.

	article	oururl	newdescp	
			[October 1997] Chosen by People (USA) magazine as one of the "50 Most Beautiful People" in the world. Educated at University of New South Wales, Australia. Chosen by People magazine as one of the "50 Most Beautiful People" in the world. Chosen by People magazine as one of the "50 Most Beautiful People" in the world. Awarded the AO (Officer of the Order of Australia), Australia's highest honor, in mid-1997. He took up acting only because his sister submitted an application behind his back. The night before an audition, he got into a fight, and his face was badly beaten, an accident that won him the role.	
2	Talent Agency WME drops Mel Gibson	http://en.wikipedia.org/wiki/Lethal_Weapon_(film_series)	Cast member Mel Gibson (R) and Oksana Grigorieva attend the premiere of the film ''Edge of Darkness'' in Los Angeles January 26, 2010. Earlier this week, the agency's Patrick Whitesell informed the actors'	[

article	oururl	newdescp
		representatives that he would no longer be represented by the agency. Gibson's longtime agent, Ed Limato, died July 3, and a funeral will take place in New York next week. William Morris Endeavor (WME) partner Ari Emanuel had previously expressed hostility toward Gibson after the actor made anti-Semitic remarks and made remarks implying skepticism about the Holocaust. An agency source said the only reason the agency had represented Gibson in the first place was his association with Limato. "Mel was really important to Ed," an agency source said. "He was with him for 32 years and I think Ed saw him as a son." But he added, "The world knows how Ari feels and he has never changed that opinion." Gibson's troubles have only increased in recent weeks with allegations of bigoted tirades and reports that he is under investigation for assaulting his ex-

article	oururl	newdescp
		girlfriend. Several studio executives have said in the wake of these disclosures that they consider the troubled actor too untouchable in the industry. "I'd rather get engaged to Lindsay Lohan than have anything to do with him," one studio chief said. A spokesman for Gibson could not be reached for comment.

	article	oururl	newdescp
3	Suicide bomber killed in Tehran-Fars	http://en.wikipedia.org/wiki/Timeline_of_the_2009_Iranian_election_protests	(Adds details) TEHRAN, June 20 (Reuters) - A suicide bomber was killed and two people were wounded in Tehran on Saturday, near the shrine of Iran's revolutionary founder, Ayatollah Ruhollah Khomeini, Iran's semi-official Fars news agency reported. "A suicide bomber was killed at the northern wing of Imam Khomeini's shrine. Two people were injured," Fars said. It did not explain the exact circumstances. Iranian riot police used teargas elsewhere in Tehran to disperse demonstrators protesting against a disputed presidential election, a witness said. (Editing by Jon Boyle)
4	Iran's 10% ballot boxes to be recounted	http://en.wikipedia.org/wiki/Timeline_of_the_2009_Iranian_election_protests	Tehran - Iran's Guardian Council is ready to recount up to 10 percent of the ballot boxes randomly in last week's presidential election, state television reported on Saturday. "The Guardian Council is

article	oururl	newdescp
		ready to recount randomly up to 10 percent of ballot boxes in last week's disputed presidential election," the council's spokesman Abbas Ali Kadkhodai was quoted as saying. "The Guardian Council is not legally obliged," Kadkhodai said, "we will recount the votes in the presence of the three (defeated) candidates." Whenever the examination and the recount is finished the council will announce its final decision, he added. He also said that Mir-Hossein Mousavi and Mehdi Karroubi still have time to express their opinions until Wednesday. Only Iran's former Revolutionary Guards Chief Mohsen Rezaei attended a special meeting of the Guardians Council with presidential candidates on Saturday, the official IRNA news agency reported. Iran's former Prime Minister Mir-Hossein Mousavi and former Parliament

	article	oururl	newdescp
			Speaker Mehdi Karroubi failed to attend the meeting without giving any reason. The Spokesman of the Guardian Council Abbas-Ali Kadkhodaei said on Wednesday that candidates of Iran's recent presidential election were invited to its upcoming meeting session which is to be held within the next few days. Iran's Supreme Leader Ayatollah Seyyed Ali Khamenei has ordered Iran's Guardian Council, the top legislative body, to investigate the claims of "fraud" in the recent presidential election.

In [68]:
```
# Removing duplicates from name list
text_simple['names_distinct'] = text_simple['names'].apply(lambda x: lis
t(set(x))).copy()
```

Results

Overall, the results of the text recognition seem very reliable. This process took a significant amount of compute power and processing time. In the future, I would experiment with implementing this process on a Spark cluster to increase compute power.

Date Recognition

Next, I will parse each article to remove any relevant date information. I will do this in two steps:

- Use Spacy's tagging function to identify any date-related words, placing those words in a list within a new column.
- Converting all date-related words in the column to datetime objects, if possible. This should help with future useability of the date information.

In [10]:
```
# Defining function to detect dates in text
def date_detect(string):
    # Creating tagged doc from string
    doc = nlp(string)
    # Create empty list of dates
    date_list = []
    # Detecting date using spacy libraray
    for ent in doc.ents:
        if ent.label_=='DATE':
            # append date to list
            date_list.append(ent.text)
        else:
            pass
    return date_list
```

In [11]:
```
# Applying function to text and creating column of dates
text_simple['dates'] = text_simple['newdescp'].apply(date_detect).copy()
```

In [12]:
```
# Define function to apply timefhuman for readable timestamp objects
def timestamp(timelist):
    output = []
    for i in timelist:
        try:
            output.append(timefhuman(i))
        # Exception handling for non readable dates
        except Exception:
            pass
    # Removing empty items
    output = list(filter(None, output))
    return output
```

In [13]:
```
# Applying function to text and creating column of datetime objects
text_simple['datetimes'] = text_simple['dates'].apply(timestamp).copy()
```

In [14]:
```
text_simple.head()
```

Out[14]:

	Unnamed: 0	article	oururl	r
0	0	Gaza aid ship to dock in Egypt after Israel pressure	http://en.wikipedia.org/wiki/Gaza_flotilla_raid	A s Gaz at Egy Israe t b bloc l fla cha cha Lib Col Gac Gre Israe hel UI G Mc insi go t Israe Ga s activ i tr them in conc Isra ma "b sc Isra b the I p w t mili

	Unnamed: 0	article	oururl	r
				whic rena miss from por lo al tonn c med pre- hou been by th In C De Four c Sa Ga or sai (30 s wh on e soli the I
1	1	Mel Gibson	http://en.wikipedia.org/wiki/Mel_Gibson_filmography	Ofte dire in indiv is p an j p v filn Ma Leth f E T : Ofte men rever frien #12 n "Th

	Unnamed: 0	article	oururl	r
				Mov All [Oct (Pe ma one Mos Pec E U N (ma one Mos Pec worl ma one Mos Pec Av AO th high in H a b sul a night a got i ar ac w
2	2	Talent Agency WME drops Mel Gibson	http://en.wikipedia.org/wiki/Lethal_Weapon_(film_series)	Cas Mel ar prem filn Da Lo J 20 this

Unnamed: 0	article	oururl	r
			inf repre tha nc repre tl Li Ju tal New wee (WM A had hosti Gi the a ar re mac s Hol ager sa a re Gib first his a w "Mel im Ed," sc "H y thi him But ' know fe cha trou only w alle bigo a

Unnamed: 0	article	oururl	r
			inv for Sev (ha th c co trou unto th "I'd e Lind anyt with s spok Gib not b for

	Unnamed: 0	article	oururl	r
				(Ad TEH 20 bo kille pe w Satu th rev
3	3	Suicide bomber killed in Tehran-Fars	http://en.wikipedia.org/wiki/Timeline_of_the_2009_Iranian_election_protests	Ir o ne re bo k nor k s pe inju said e circu p els dem p wit (Edit
4	4	Iran's 10% ballot boxes to be recounted	http://en.wikipedia.org/wiki/Timeline_of_the_2009_Iranian_election_protests	Teh rec 10 ra l p elec re Satu

Unnamed: 0	article	oururl	r
			rand 10 ballc l p ele s I was sa Cou I sai r v p ca Wh ex fi c an fina he als M Mc Ka ha ex opi W (Rev Gu a mee Cc p can Sa of ne repo for M Mc

Unnamed: 0	article	oururl	r
			Spea Karr to with a Spo the K W car Ira p ele in ses is with Iran's Kha ord C top inve "fra p

Results

After the data processing, we have several new columns of relevant data:

- 'names': a list containing all personal names detected in the article
- 'names_distinct': a list of all distinct names detected
- 'dates': a list of all date-type strings detected
- 'datetimes': a list of datetime objects for any detectable timestamps

Note: The datetime objects could be more accurate if each article had a corresponding date of publication. This could allow the library to index those dates such as "last week" or "next Tuesday" to specific timestamps. From the given data, we might have been able to use the `'last_judgement_at'` column, but without documentation we cannot tell if this is an appopriate reference date to use.

Part 2: Check for Document Similarity

For part two of this assessment, we will scan the news articles to find those most similar to a provided sample. To do this, we will try a few machine learning techniques:

- "Term Frequency - Inverse Document Freqency" (TF-IDF) with a cosine similarity test
- Sentence Emedding with Google Sentence Encoder
- Spacy Vectorization - using Spacy's internal vectorization and similariy function

Preprocessing

To begin, we need to preprocess the data in order to:

- Lowercase all words
- Remove stop words
- Remove puntuation
- Remove single characters (unneeeded for our analysis)
- Stem each word
- Lemmatize each word
- Convert numbers to words

Creating Preprocessing Functions

In [143]:
```
# Function to lowercase all words
def convert_lower_case(text):
    return np.char.lower(text)
```

In [144]:
```
# Function to remove stop words
def remove_stop_words(text):
    stop_words = stopwords.words('english')
    words = word_tokenize(str(text))
    new_text = ""
    for w in words:
        if w not in stop_words and len(w) > 1:
            new_text = new_text + " " + w
    return new_text
```

In [145]:
```
# Function to remove punctuation
def remove_punctuation(text):
    symbols = "!\"#$%&()*+-./:;<=>?@[\]^_`{|}~\n"
    for i in range(len(symbols)):
        text = np.char.replace(text, symbols[i], ' ')
        text = np.char.replace(text, "  ", " ")
    text = np.char.replace(text, ',', '')
    return text
```

In [146]:
```
# Function to remove apostrophes
# We call this function separately to avoid erros in possessive words
def remove_apostrophe(text):
    return np.char.replace(text, "'", '')
```

In [147]:
```
# Function to stem words
def stemming(text):
    stemmer = PorterStemmer()
    tokens = word_tokenize(str(text))
    new_text = ""
    for w in tokens:
        new_text = new_text + " " + stemmer.stem(w)
    return new_text
```

In [153]:
```
# Function to convert numbers to text
def convert_numbers(text):
    tokens = word_tokenize(str(text))
    new_text = ""
    for w in tokens:
        try:
            w = num2words(int(w))
        except:
            a = 0
        new_text = new_text + " " + w
    new_text = np.char.replace(new_text, "-", " ")
    return new_text
```

In [158]:
```
# Function to preprocess data using previously defined functions
def preprocess(text):
    text = convert_lower_case(text)
    text = remove_punctuation(text)
    text = remove_apostrophe(text)
    text = remove_stop_words(text)
    text = convert_numbers(text)
    text = stemming(text)
    text = remove_punctuation(text)
    text = convert_numbers(text)
    text = stemming(text)
    text = remove_punctuation(text) # needed again as num2word adds some
hyphens and commas
    text = remove_stop_words(text) # needed again as num2word adds some
 stop words
    return text
```

In [161]:
```
# Applying preprocessing function to all articles
text_simple['processed_text'] = text_simple['newdescp'].apply(preprocess
).copy()
```

In [162]:
```
text_simple.head()
```

Out[162]:

	article	oururl	newdescp
0	Gaza aid ship to dock in Egypt after Israel pressure	http://en.wikipedia.org/wiki/Gaza_flotilla_raid	A ship with supplies for Gaza will dock at el-Arish in Egypt, officials say, after Israeli pressure to stop the vessel breaking its Gaza blockade. The Moldovan-flagged ship chartered by a charity run by the son of Libyan leader Col Muammar Gaddafi, left a Greek port on Saturday. Israel asked for help from the UN, and had talks with Greece and Moldova. But organisers insist they will go to Gaza. An Israeli raid on a Gaza-bound ship in May killed nine Turkish activists. Israel insisted its troops were defending themselves but the raid sparked international condemnation. Israel recently eased its blockade, allowing in almost all consumer goods but maintaining a "blacklist" of some items. Israel says its blockade of the Palestinian territory is needed to prevent the supply of weapons to the Hamas militant group which controls

	article	oururl	newdescp
			Gaza. The Amalthea, renamed Hope for the mission, set off from the Greek port of Lavrio, loaded with about 2,000 tonnes of food, cooking oil, medicines and pre-fabricated houses. It has been chartered by the Gaddafi International Charity and Development Foundation. Its chairman is Saif al-Islam Gaddafi. The organisation said the 92m (302ft) vessel would also carry "a number of supporters who are keen on expressing solidarity with the Palestinian people".
1	Mel Gibson	http://en.wikipedia.org/wiki/Mel_Gibson_filmography	Often acts and directs stories involving an individual who is persecuted, and fights for justice Has often portrayed a widower, in films such as Mad Max (the sequels), Lethal Weapon film series, Braveheart, The Patriot, Signs, and Edge of Darkness. Often portrays men who seek revenge for the murder of family or friends Ranked #12 in Empire (UK) magazine's "The Top 100 Movie Stars of All Time" list.

	article	oururl	newdescp	
			[October 1997] Chosen by People (USA) magazine as one of the "50 Most Beautiful People" in the world. Educated at University of New South Wales, Australia. Chosen by People magazine as one of the "50 Most Beautiful People" in the world. Chosen by People magazine as one of the "50 Most Beautiful People" in the world. Awarded the AO (Officer of the Order of Australia), Australia's highest honor, in mid-1997. He took up acting only because his sister submitted an application behind his back. The night before an audition, he got into a fight, and his face was badly beaten, an accident that won him the role.	
2	Talent Agency WME drops Mel Gibson	http://en.wikipedia.org/wiki/Lethal_Weapon_(film_series)	Cast member Mel Gibson (R) and Oksana Grigorieva attend the premiere of the film ''Edge of Darkness'' in Los Angeles January 26, 2010. Earlier this week, the agency's Patrick Whitesell informed the actors'	[

article	oururl	newdescp
		representatives that he would no longer be represented by the agency. Gibson's longtime agent, Ed Limato, died July 3, and a funeral will take place in New York next week. William Morris Endeavor (WME) partner Ari Emanuel had previously expressed hostility toward Gibson after the actor made anti-Semitic remarks and made remarks implying skepticism about the Holocaust. An agency source said the only reason the agency had represented Gibson in the first place was his association with Limato. "Mel was really important to Ed," an agency source said. "He was with him for 32 years and I think Ed saw him as a son." But he added, "The world knows how Ari feels and he has never changed that opinion." Gibson's troubles have only increased in recent weeks with allegations of bigoted tirades and reports that he is under investigation for assaulting his ex-

article	oururl	newdescp
		girlfriend. Several studio executives have said in the wake of these disclosures that they consider the troubled actor too untouchable in the industry. "I'd rather get engaged to Lindsay Lohan than have anything to do with him," one studio chief said. A spokesman for Gibson could not be reached for comment.

	article	oururl	newdescp
3	Suicide bomber killed in Tehran-Fars	http://en.wikipedia.org/wiki/Timeline_of_the_2009_Iranian_election_protests	(Adds details) TEHRAN, June 20 (Reuters) - A suicide bomber was killed and two people were wounded in Tehran on Saturday, near the shrine of Iran's revolutionary founder, Ayatollah Ruhollah Khomeini, Iran's semi-official Fars news agency reported. "A suicide bomber was killed at the northern wing of Imam Khomeini's shrine. Two people were injured," Fars said. It did not explain the exact circumstances. Iranian riot police used teargas elsewhere in Tehran to disperse demonstrators protesting against a disputed presidential election, a witness said. (Editing by Jon Boyle)
4	Iran's 10% ballot boxes to be recounted	http://en.wikipedia.org/wiki/Timeline_of_the_2009_Iranian_election_protests	Tehran - Iran's Guardian Council is ready to recount up to 10 percent of the ballot boxes randomly in last week's presidential election, state television reported on Saturday. "The Guardian Council is

article	oururl	newdescp
		ready to recount randomly up to 10 percent of ballot boxes in last week's disputed presidential election," the council's spokesman Abbas Ali Kadkhodai was quoted as saying. "The Guardian Council is not legally obliged," Kadkhodai said, "we will recount the votes in the presence of the three (defeated) candidates." Whenever the examination and the recount is finished the council will announce its final decision, he added. He also said that Mir-Hossein Mousavi and Mehdi Karroubi still have time to express their opinions until Wednesday. Only Iran's former Revolutionary Guards Chief Mohsen Rezaei attended a special meeting of the Guardians Council with presidential candidates on Saturday, the official IRNA news agency reported. Iran's former Prime Minister Mir-Hossein Mousavi and former Parliament

article	oururl	newdescp
		Speaker Mehdi Karroubi failed to attend the meeting without giving any reason. The Spokesman of the Guardian Council Abbas-Ali Kadkhodaei said on Wednesday that candidates of Iran's recent presidential election were invited to its upcoming meeting session which is to be held within the next few days. Iran's Supreme Leader Ayatollah Seyyed Ali Khamenei has ordered Iran's Guardian Council, the top legislative body, to investigate the claims of "fraud" in the recent presidential election.

Treating Sample Text

In [164]:
```
# Importing sample text
sample = '''
        Human Rights Watch says government-controlled health services i
n Egypt have been pressured into playing down the number of casualties d
uring anti-government protests. The group has documented the deaths of 2
97 people, but says the final toll is likely to be significantly higher.
Human Rights Watch says the vast majority of the deaths in Cairo, Alexan
dria and Suez were on January 28 and 29 as a result of live gunfire as r
iot police fought running battles with protesters. A significant proport
ion came as a result of rubber bullets fired at too close a range and fr
om teargas canisters fired into the crowds at very close range. Human Ri
ghts Watch says the actual number of deaths is likely to be an underesti
mate because the organisation had only included those deaths it had veri
fied itself at key hospitals in the three major cities.
        '''
# Preprocessing sample to match dataset
sample = preprocess(sample)
```

In [165]:
```
sample
```

Out[165]:
```
' human right watch say govern control health servic egypt pressur play
number casualti anti govern protest group document death two hundr nine
ti seven peopl say final toll like significantli higher human right wat
ch say vast major death cairo alexandria suez januari twenti eight twen
ti nine result live gunfir riot polic fought run battl protest signif p
roport came result rubber bullet fire close rang tearga canist fire cro
wd close rang human right watch say actual number death like underestim
organi includ death verifi key hospit three major citi'
```

Test 1: TF-IDF Cosine Similarity

In [166]:
```
from collections import Counter
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
```

In [167]:
```
# defining vectorization function for texts
def get_vectors(strings):
    text = [t for t in strings]
    vectorizer = CountVectorizer(text)
    vectorizer.fit(text)
    return vectorizer.transform(text).toarray()
```

In [178]:
```
# defining cosine similarity function
def get_cosine_sim(strings):
    vectors = [t for t in get_vectors(strings)]
    return cosine_similarity(vectors)[1,0]
```

In [182]:
```
# defining function to combine sample with each item
def cosine_test(item):
    sample = '''
         Human Rights Watch says government-controlled health services i
n Egypt have been pressured into playing down the number of casualties d
uring anti-government protests. The group has documented the deaths of 2
97 people, but says the final toll is likely to be significantly higher.
Human Rights Watch says the vast majority of the deaths in Cairo, Alexan
dria and Suez were on January 28 and 29 as a result of live gunfire as r
iot police fought running battles with protesters. A significant proport
ion came as a result of rubber bullets fired at too close a range and fr
om teargas canisters fired into the crowds at very close range. Human Ri
ghts Watch says the actual number of deaths is likely to be an underesti
mate because the organisation had only included those deaths it had veri
fied itself at key hospitals in the three major cities.
         '''
    test_items = [item, sample]
    return get_cosine_sim(test_items)
```

In [194]:
```
# Applying cosine test to all articles
text_simple['cosine_test'] = text_simple['processed_text'].apply(cosine_
test).copy()
```

In [195]:
```
text_simple = text_simple.sort_values(by='cosine_test', ascending=False)
```

In [196]:
```
text_simple.head()
```

Out[196]:

	article	oururl	newdescp	n
76	Egypt hospitals 'told to downplay protest deaths'	http://en.wikipedia.org/wiki/Egyptian_Revolution_of_2011	Human Rights Watch says government-controlled health services in Egypt have been pressured into playing down the number of casualties during anti-government protests. The group has documented the deaths of 297 people, but says the final toll is likely to be significantly higher. Human Rights Watch says the vast majority of the deaths in Cairo, Alexandria and Suez were on January 28 and 29 as a result of live gunfire as riot police fought running battles with protesters. A significant proportion came as a result of rubber bullets fired at too close a range and from teargas canisters fired into the crowds at very close range. Human Rights Watch says the actual number of deaths is likely to be an underestimate because the organisation had only included those deaths it had verified itself at key hospitals in the three major cities.	
2361	Libya: Security Forces Kill 84	http://en.wikipedia.org/wiki/Libyan_Civil_War_(2011)	Muammar Gaddafi's security forces	[Mua Ga Mua

article	oururl	newdescp	n
Over Three Days		are firing on Libyan citizens and killing scores simply because they're demanding change and accountability. Libyan authorities should allow peaceful protesters to have their say. (New York) - Government security forces have killed at least 84 people in three days of protests in several cities in Libya, Human Rights Watch said today, based on telephone interviews with local hospital staff and witnesses. The Libyan authorities should immediately end attacks on peaceful protesters and protect them from assault by pro-government armed groups, Human Rights Watch said. Thousands of demonstrators gathered in the eastern Libyan cities of Benghazi, Baida, Ajdabiya, Zawiya, and Derna on February 18, 2011, following violent attacks against peaceful protests the day before that killed 20 people in Benghazi, 23 in Baida, three in Ajdabiya, and three in Derna. Hospital sources told	Ga Joe

article	oururl	newdescp	n
		Human Rights Watch that security forces killed 35 people in Benghazi on February 18, almost all with live ammunition. "Muammar Gaddafi's security forces are firing on Libyan citizens and killing scores simply because they're demanding change and accountability," said Joe Stork, deputy Middle East and North Africa director at Human Rights Watch. "Libyan authorities should allow peaceful protesters to have their say." The protests in Benghazi on February 18 began during funerals for the 20 demonstrators killed by security forces the day before. Eyewitnesses told Human Rights Watch that security forces with distinctive yellow uniforms opened fire on protesters near the Fadil Bu Omar Katiba, a security force base in the center of Benghazi. One protester told Human Rights Watch he witnessed four men shot dead. By 11 p.m. on February 18, Al Jalaa Hospital in Benghazi had received the bodies of 35	

article	oururl	newdescp	n
		people killed that day, a senior hospital official told Human Rights Watch. He said the deaths had been caused by gunshot wounds to the chest, neck, and head. Two sources at the hospital confirmed to Human Rights Watch that the death toll for February 17 was 20, and that at least 45 people had been wounded by bullets. The senior hospital official told Human Rights Watch, "We put out a call to all the doctors in Benghazi to come to the hospital and for everyone to contribute blood because I've never seen anything like this before." Witnesses said that after the February 18 shootings, protesters in Benghazi continued on to the courthouse and gathered there throughout the evening, the crowd swelling to thousands. In Baida, further to the east, protesters on February 18 buried the 23 people who had been shot dead the day before. One protester told Human Rights Watch that police were patrolling the streets but he	

	article	oururl	newdescp	n
			had seen no further clashes.	
2624	Libya: Security Forces Fire on 'Day of Anger' Demonstrations	http://en.wikipedia.org/wiki/Libyan_Civil_War_(2011)	The security forces' vicious attacks on peaceful demonstrators lay bare the reality of Muammar Gaddafi's brutality when faced with any internal dissent. Libyans should not have to risk their lives to make a stand for their rights as human beings. (New York) - The Libyan security forces killed at least 24 protesters and wounded many others in a crackdown on peaceful demonstrations across the country, Human Rights Watch said today. The authorities should cease the use of lethal force unless absolutely necessary to protect lives and open an independent investigation into the lethal shootings, Human Rights Watch said. Hundreds of peaceful protesters took to the streets on February 17, 2011, in Baida, Benghazy, Zenten, Derna, and Ajdabiya. According to multiple witnesses, Libyan security forces shot and killed the demonstrators in efforts to disperse the	[Mua Ga Mua Ga Wh S Nas Ju Ahn Mua Ga H H So Nas Ra

article	oururl	newdescp	n
		protests. "The security forces' vicious attacks on peaceful demonstrators lay bare the reality of Muammar Gaddafi's brutality when faced with any internal dissent," said Sarah Leah Whitson, Middle East and North Africa director at Human Rights Watch. "Libyans should not have to risk their lives to make a stand for their rights as human beings." Some of the worst violence was in the eastern city of Baida. At around 1 p.m. on February 17, according to sources in Libya, hospital staff put out a call for additional medical supplies, as they became overwhelmed with the medical needs of 70 injured protesters, half of them in critical condition due to gunshot wounds. On the night of February 16, security forces had attacked peaceful protesters with teargas and live ammunition, shooting dead two protesters, according to protesters who spoke to Human Rights Watch. Geneva-based Libya Human Rights	

article	oururl	newdescp	n
		Solidarity has confirmed three of the names of those shot dead so far: Safwan Attiya, Nasser Al Juweigi, and Ahmad El Qabili. One protester told Human Rights Watch that a new protest started on February 17, after noon prayers and the funerals of those killed on February 16. Joined by hundreds of other protesters, families marched toward the Internal Security office, chanting, "Down with the regime" and "Get out Muammar Gaddafi." Some protesters filmed the protests with mobile phones and posted them online. One injured protester in a hospital in Baida told Human Rights Watch that he was sitting near the intensive care unit there and had confirmed that security forces had shot dead 16 people and wounded dozens of others. He said that Special Forces and armed men in street clothes fired live ammunition to deter protesters. A protester in Benghazi told	

article	oururl	newdescp	n
		Human Rights Watch that hundreds of lawyers, activists, and other protesters gathered on the steps of the Benghazi Court calling for a constitution and respect for the rule of law. Early in the day, sources in Libya told Human Rights Watch that security forces had arrested a Benghazi journalist, Hind El Houny, and Salem Souidan, a family member of a group that has been seeking justice for the massacre of inmates in Abu Salim prison in 1996. Security forces also arrested a former political prisoner, Abdel Nasser al-Rabbasi, in Bani Walid. The protester said he saw groups of men in street clothes armed with knives, later joined by Internal Security forces, charging the protesters to disperse them. The protester told Human Rights Watch that he believed security forces had shot dead at least 17 protesters during the day, mostly near Abdel Nasser Street. Human Rights Watch was able to confirm eight of those deaths. It appears that	

	article	oururl	newdescp	n
			the government also has coordinated pro-government supporters to confront the demonstrations. On February 16, subscribers to Libyana, one of two Libyan mobile phone networks, received a text message calling upon "nationalist youth" to go out and "defend national symbols." At around 11:30 p.m. on February 17, a protester in Tripoli told Human Rights Watch that anti-government protests had started in Tripoli also.	
1434	Clue to slow human bird flu jump	http://en.wikipedia.org/wiki/Singapore_2006	Flu viruses which target man tend to attach to cells further up the airway - maximising their chances of being passed on by coughing or sneezing. Researchers found the bird flu virus attaches itself to cells deep down in the human airways. The University of Wisconsin research is published in the journal Nature. But it still cannot jump easily from human to human. Scientists fear that if it mutates and gains that ability, it could result in a human flu pandemic, with	

	article	oururl	newdescp	n
			millions of deaths world-wide. The Wisconsin team investigated why the virus could not spread easily between humans despite the fact that it could replicate efficiently in human lungs. Flu viruses infecting humans and birds are known to home in on slightly different versions of the same molecule, found on the surface of cells which line the respiratory tract. The latest study found the version of the molecule targeted by human viruses was more prevalent on cells higher up in the airway. The molecule targeted by bird viruses, on the other hand, tended to be found on cells deep within the lungs, in structures called alveoli. Thus the bird flu virus tended to be buried so deep in the lungs that it was unlikely to be spread by coughing or sneezing.	
1332	Mystery China bug toll reaches 17	http://en.wikipedia.org/wiki/Influenza_A_virus_subtype_H5N1	The indications are that the disease is a bacterial infection spread by contact with dead pigs, and not a virus, officials in Sichuan province said. At least 58	H

article	oururl	newdescp	n
		people showed symptoms, which include high fever, nausea and vomiting, during June and July. The World Health Organization has urged calm, saying the disease is unable to spread from human to human. "I can assure you that the disease is absolutely not Sars, anthrax or bird flu," Zeng Huajin, a Sichuan health official, told the China Daily newspaper. The number of people infected with the illness has risen steadily as health officials searched through remote villages in the province for people with symptoms. A total of 17 people have died, with just two discharged from hospital. Twelve people remain in a critical condition while 27 are described as "stable", doctors said. Health officials said the illness could be a variant of the streptococcus bacteria, often found in pigs. The symptoms cannot be spread from human to human, and those most at risk from animal carcasses are people with	

article	oururl	newdescp	n
		vulnerable, low immune systems, officials said. Experts had expressed fears that pigs, which can also carry human influenza, could accelerate mutation of the bird flu virus into a form which can be transmitted between people.	

From this test, we can tell that the sample text is still represented in the dataset. For the purposes of our testing, we will leave the sample text in the dataset to ensure that our similarity testing is working appropriately. In each case, this should be our top result. For now, let's explore the remaining top 5 articles in similarity.

In [205]:
```
cosine_samples = text_simple.sort_values(by='cosine_test', ascending=False).head(6)
cosine_samples = cosine_samples[1:]
```

In [206]:
```
cosine_samples.head()
```

Out[206]:

	article	oururl	newdescp	r
2361	Libya: Security Forces Kill 84 Over Three Days	http://en.wikipedia.org/wiki/Libyan_Civil_War_(2011)	Muammar Gaddafi's security forces are firing on Libyan citizens and killing scores simply because they're demanding change and accountability. Libyan authorities should allow peaceful protesters to have their say. (New York) - Government security forces have killed at least 84 people in three days of protests in several cities in Libya, Human Rights Watch said today, based on telephone interviews with local hospital staff and witnesses. The Libyan authorities should immediately end attacks on peaceful protesters and protect them from assault by pro-government armed groups, Human Rights Watch said. Thousands of demonstrators gathered in the eastern Libyan cities of Benghazi, Baida, Ajdabiya, Zawiya, and Derna on February 18, 2011, following violent attacks against peaceful protests the day before that killed 20 people in Benghazi, 23 in Baida, three in Ajdabiya, and	[Mua Ga Mua Ga Joe

article	oururl	newdescp	r
		three in Derna. Hospital sources told Human Rights Watch that security forces killed 35 people in Benghazi on February 18, almost all with live ammunition. "Muammar Gaddafi's security forces are firing on Libyan citizens and killing scores simply because they're demanding change and accountability," said Joe Stork, deputy Middle East and North Africa director at Human Rights Watch. "Libyan authorities should allow peaceful protesters to have their say." The protests in Benghazi on February 18 began during funerals for the 20 demonstrators killed by security forces the day before. Eyewitnesses told Human Rights Watch that security forces with distinctive yellow uniforms opened fire on protesters near the Fadil Bu Omar Katiba, a security force base in the center of Benghazi. One protester told Human Rights Watch he witnessed four men shot dead. By 11 p.m. on February 18, Al Jalaa Hospital in Benghazi had received the	

article	oururl	newdescp	r
		bodies of 35 people killed that day, a senior hospital official told Human Rights Watch. He said the deaths had been caused by gunshot wounds to the chest, neck, and head. Two sources at the hospital confirmed to Human Rights Watch that the death toll for February 17 was 20, and that at least 45 people had been wounded by bullets. The senior hospital official told Human Rights Watch, "We put out a call to all the doctors in Benghazi to come to the hospital and for everyone to contribute blood because I've never seen anything like this before." Witnesses said that after the February 18 shootings, protesters in Benghazi continued on to the courthouse and gathered there throughout the evening, the crowd swelling to thousands. In Baida, further to the east, protesters on February 18 buried the 23 people who had been shot dead the day before. One protester told Human Rights Watch that police were patrolling the streets but he	

	article	oururl	newdescp	r
			had seen no further clashes.	
2624	Libya: Security Forces Fire on 'Day of Anger' Demonstrations	http://en.wikipedia.org/wiki/Libyan_Civil_War_(2011)	The security forces' vicious attacks on peaceful demonstrators lay bare the reality of Muammar Gaddafi's brutality when faced with any internal dissent. Libyans should not have to risk their lives to make a stand for their rights as human beings. (New York) - The Libyan security forces killed at least 24 protesters and wounded many others in a crackdown on peaceful demonstrations across the country, Human Rights Watch said today. The authorities should cease the use of lethal force unless absolutely necessary to protect lives and open an independent investigation into the lethal shootings, Human Rights Watch said. Hundreds of peaceful protesters took to the streets on February 17, 2011, in Baida, Benghazy, Zenten, Derna, and Ajdabiya. According to multiple witnesses, Libyan security forces shot and killed the demonstrators in efforts to disperse the	[Mua Ga Mua Ga WI S Nas Ju Ahr Mua Ga H H So Nas Ra

article	oururl	newdescp	r
		protests. "The security forces' vicious attacks on peaceful demonstrators lay bare the reality of Muammar Gaddafi's brutality when faced with any internal dissent," said Sarah Leah Whitson, Middle East and North Africa director at Human Rights Watch. "Libyans should not have to risk their lives to make a stand for their rights as human beings." Some of the worst violence was in the eastern city of Baida. At around 1 p.m. on February 17, according to sources in Libya, hospital staff put out a call for additional medical supplies, as they became overwhelmed with the medical needs of 70 injured protesters, half of them in critical condition due to gunshot wounds. On the night of February 16, security forces had attacked peaceful protesters with teargas and live ammunition, shooting dead two protesters, according to protesters who spoke to Human Rights Watch. Geneva-based Libya Human Rights Solidarity has confirmed three of the	

article	oururl	newdescp	r
		names of those shot dead so far: Safwan Attiya, Nasser Al Juweigi, and Ahmad El Qabili. One protester told Human Rights Watch that a new protest started on February 17, after noon prayers and the funerals of those killed on February 16. Joined by hundreds of other protesters, families marched toward the Internal Security office, chanting, "Down with the regime" and "Get out Muammar Gaddafi." Some protesters filmed the protests with mobile phones and posted them online. One injured protester in a hospital in Baida told Human Rights Watch that he was sitting near the intensive care unit there and had confirmed that security forces had shot dead 16 people and wounded dozens of others. He said that Special Forces and armed men in street clothes fired live ammunition to deter protesters. A protester in Benghazi told Human Rights Watch that hundreds of lawyers, activists, and other protesters	

article	oururl	newdescp	r
		gathered on the steps of the Benghazi Court calling for a constitution and respect for the rule of law. Early in the day, sources in Libya told Human Rights Watch that security forces had arrested a Benghazi journalist, Hind El Houny, and Salem Souidan, a family member of a group that has been seeking justice for the massacre of inmates in Abu Salim prison in 1996. Security forces also arrested a former political prisoner, Abdel Nasser al-Rabbasi, in Bani Walid. The protester said he saw groups of men in street clothes armed with knives, later joined by Internal Security forces, charging the protesters to disperse them. The protester told Human Rights Watch that he believed security forces had shot dead at least 17 protesters during the day, mostly near Abdel Nasser Street. Human Rights Watch was able to confirm eight of those deaths. It appears that the government also has coordinated pro-government supporters to confront the demonstrations.	

	article	oururl	newdescp	r
			On February 16, subscribers to Libyana, one of two Libyan mobile phone networks, received a text message calling upon "nationalist youth" to go out and "defend national symbols." At around 11:30 p.m. on February 17, a protester in Tripoli told Human Rights Watch that anti-government protests had started in Tripoli also.	
1434	Clue to slow human bird flu jump	http://en.wikipedia.org/wiki/Singapore_2006	Flu viruses which target man tend to attach to cells further up the airway - maximising their chances of being passed on by coughing or sneezing. Researchers found the bird flu virus attaches itself to cells deep down in the human airways. The University of Wisconsin research is published in the journal Nature. But it still cannot jump easily from human to human. Scientists fear that if it mutates and gains that ability, it could result in a human flu pandemic, with millions of deaths world-wide. The Wisconsin team investigated why the virus could not	

	article	oururl	newdescp	r
			spread easily between humans despite the fact that it could replicate efficiently in human lungs. Flu viruses infecting humans and birds are known to home in on slightly different versions of the same molecule, found on the surface of cells which line the respiratory tract. The latest study found the version of the molecule targeted by human viruses was more prevalent on cells higher up in the airway. The molecule targeted by bird viruses, on the other hand, tended to be found on cells deep within the lungs, in structures called alveoli. Thus the bird flu virus tended to be buried so deep in the lungs that it was unlikely to be spread by coughing or sneezing.	
1332	Mystery China bug toll reaches 17	http://en.wikipedia.org/wiki/Influenza_A_virus_subtype_H5N1	The indications are that the disease is a bacterial infection spread by contact with dead pigs, and not a virus, officials in Sichuan province said. At least 58 people showed symptoms, which include high fever, nausea and vomiting, during June and July. The World	I

article	oururl	newdescp	r
		Health Organization has urged calm, saying the disease is unable to spread from human to human. "I can assure you that the disease is absolutely not Sars, anthrax or bird flu," Zeng Huajin, a Sichuan health official, told the China Daily newspaper. The number of people infected with the illness has risen steadily as health officials searched through remote villages in the province for people with symptoms. A total of 17 people have died, with just two discharged from hospital. Twelve people remain in a critical condition while 27 are described as "stable", doctors said. Health officials said the illness could be a variant of the streptococcus bacteria, often found in pigs. The symptoms cannot be spread from human to human, and those most at risk from animal carcasses are people with vulnerable, low immune systems, officials said. Experts had expressed fears that pigs, which can also carry human	

	article	oururl	newdescp	r
			influenza, could accelerate mutation of the bird flu virus into a form which can be transmitted between people.	
1070	U.N. rights council condemns Syrian abuses	http://en.wikipedia.org/wiki/Syrian_Civil_War	The U.N. Human Rights Council decried a wide range of human rights violations in Syria on Friday and called for U.N. bodies to consider a recent report detailing the abuses and take" appropriate action." The council passed a resolution that "strongly condemns the continued widespread, systematic and gross violations of human rights and fundamental freedoms by the Syrian authorities, such as arbitrary executions, excessive use of force and the killing and persecution of protesters, human rights defenders and journalists, arbitrary detention, enforced disappearances, torture and ill-treatment, including against children." There were 37 yes votes, four against and six abstentions at the meeting in Geneva, Switzerland. The group convened to	[E Navi

article	oururl	newdescp	r
		consider action against Syria after a troubling report issued Monday by the Independent International Commission of Inquiry, a body appointed by the council. That report concluded security and military forces "committed crimes against humanity" against civilians. The resolution recommends that U.N. bodies "urgently consider" the commission report and "take appropriate action." The group decided to send the Commission of Inquiry report to U.N. Secretary-General Ban Ki-moon "for appropriate action and transmission to all U.N. relevant bodies." It backs "efforts to protect the population of the Syrian Arab Republic and to bring an immediate end to gross human rights violations." And, it urged Syria "to protect its population" and "to immediately put an end to all human rights violations." The resolution also decided "to establish a mandate of a Special Rapporteur on the situation of human rights in Syria and urges Syria to cooperate with	

article	oururl	newdescp	r
		it. Before the resolution was adopted, U.N. High Commissioner for Human Rights Navi Pillay told the council Syria faces a "full-fledged civil war" if the regime's "continual ruthless repression" against peaceful demonstrators and civilians isn't stopped now. She noted with concern the reports of "increased armed attacks by the opposition forces, including the so-called Free Syrian army, against the Syrian military and security apparatus." "In light of the manifest failure of the Syrian authorities to protect their citizens, the international community needs to take urgent and effective measures to protect the Syrian people," Pillay said.	

Test 1: Results

From exploring the top 5 articles by cosine similarity, it seems that articles 1, 2 and 5 have a good deal of contextual overlap. These artciles refer to protests and human rights abuses in Libya and Syria, while the sample text is about human rights abuses from protests in Egypt.

However, articles 3 and 4 are related mainly to viruses and have nothing to do protests, human rights, or the middle east. Further similarity testing is needed to refine results.

Test 2: Sentence Embedding

For this test, we will utilize the Google Sentence Encoder. This encoder uses embeddings for words within larger sentence structures which are averaged together and given an embedding for the overall sentence. These larger embeddings are compared for correlation.

Because this model utilizes sentence structure, we will use the raw text from each article rather than its cleaned and tokenized version.

In [10]:
```
# Import module from Universal Sentence Encoder
module_url = "https://tfhub.dev/google/universal-sentence-encoder/4" #@p
aram ["https://tfhub.dev/google/universal-sentence-encoder/4", "https://
tfhub.dev/google/universal-sentence-encoder-large/5"]
model = hub.load(module_url)
print ("module %s loaded" % module_url)

# Define function to embed sentences.
def embed(input):
  return model(input)
```

```
INFO:absl:Using /var/folders/1_/yqjpttyx6yvg73s_v4rn7_gc0000gn/T/tfhub_
modules to cache modules.
INFO:absl:Downloading TF-Hub Module 'https://tfhub.dev/google/universal
-sentence-encoder/4'.
INFO:absl:Downloading https://tfhub.dev/google/universal-sentence-encod
er/4: 180.00MB
INFO:absl:Downloading https://tfhub.dev/google/universal-sentence-encod
er/4: 400.00MB
INFO:absl:Downloading https://tfhub.dev/google/universal-sentence-encod
er/4: 620.00MB
INFO:absl:Downloading https://tfhub.dev/google/universal-sentence-encod
er/4: 850.00MB
INFO:absl:Downloaded https://tfhub.dev/google/universal-sentence-encode
r/4, Total size: 987.47MB
INFO:absl:Downloaded TF-Hub Module 'https://tfhub.dev/google/universal-
sentence-encoder/4'.

module https://tfhub.dev/google/universal-sentence-encoder/4 loaded
```

In [11]:
```
# Define a function to score similarity based on sentence enbeddings
def embedding_corr(test_items):
  features = embed(test_items)
  # Return correlation between sentence embeddings
  # Using absolute value to standardize correlations
  corr = np.abs(np.inner(features, features))
  return corr[1,0]

# defining function to combine sample with each item
def embedding_test(item):
    sample = '''
         Human Rights Watch says government-controlled health services i
n Egypt have been pressured into playing down the number of casualties d
uring anti-government protests. The group has documented the deaths of 2
97 people, but says the final toll is likely to be significantly higher.
Human Rights Watch says the vast majority of the deaths in Cairo, Alexan
dria and Suez were on January 28 and 29 as a result of live gunfire as r
iot police fought running battles with protesters. A significant proport
ion came as a result of rubber bullets fired at too close a range and fr
om teargas canisters fired into the crowds at very close range. Human Ri
ghts Watch says the actual number of deaths is likely to be an underesti
mate because the organisation had only included those deaths it had veri
fied itself at key hospitals in the three major cities.
         '''
    test_items = [item, sample]
    return embedding_corr(test_items)
```

In [12]:
```
# Applying sentence embedding test to all articles
text_simple['sentence_test'] = text_simple['newdescp'].apply(embedding_t
est).copy()
```

In [13]:
```
sentence_samples = text_simple.sort_values(by='sentence_test', ascending
=False).head(6)
sentence_samples = sentence_samples[1:]
sentence_samples.head()
```

Out[13]:

	Unnamed: 0	article	oururl	newdescp	na
89	585	Deaths in Egypt's Suez after Port Said football unrest	http://en.wikipedia.org/wiki/2012?13_Egyptian_protests	Two people have been killed and more than 400 injured in protests across Egypt sparked by the deaths of 74 people after a football match. The two killed were shot by police trying to disperse angry crowds in the city of Suez, medical officials said. In the capital Cairo, thousands of protesters remained on the streets following a day of clashes with police. Thousands marched to the interior ministry, where security forces fired tear gas to keep them back. Earlier, the Egyptian prime minister announced the sackings of several senior officials. Funerals of some of the 74 victims took place in Port Said, where the football match had taken place on Wednesday. The deaths came when fans invaded the pitch after a fixture between top Cairo club al-Ahly and the Port Said side al-Masry. As night fell in Cairo, several thousand demonstrators remained in the streets around the interior ministry, witnesses said. In Suez, health	['M 'Mohan Lash

	Unnamed: 0	article	oururl	newdescp	na
				official Mohammed Lasheen said two people had been shot dead early on Friday. A witness quoted by Reuters said: "Protesters are trying to break into the Suez police station and police are now firing live ammunition."	
1	2361	Libya: Security Forces Kill 84 Over Three Days	http://en.wikipedia.org/wiki/Libyan_Civil_War_(2011)	Muammar Gaddafi's security forces are firing on Libyan citizens and killing scores simply because they're demanding change and accountability. Libyan authorities should allow peaceful protesters to have their say. (New York) - Government security forces have killed at least 84 people in three days of protests in several cities in Libya, Human Rights Watch said today, based on telephone interviews with local hospital staff and witnesses. The Libyan authorities should immediately end attacks on peaceful protesters and protect them from assault by pro-government armed groups, Human Rights Watch said. Thousands of demonstrators gathered in the	['Muar Gac 'Muar Gac 'Joe S

Unnamed: 0	article	oururl	newdescp	na
			eastern Libyan cities of Benghazi, Baida, Ajdabiya, Zawiya, and Derna on February 18, 2011, following violent attacks against peaceful protests the day before that killed 20 people in Benghazi, 23 in Baida, three in Ajdabiya, and three in Derna. Hospital sources told Human Rights Watch that security forces killed 35 people in Benghazi on February 18, almost all with live ammunition. "Muammar Gaddafi's security forces are firing on Libyan citizens and killing scores simply because they're demanding change and accountability," said Joe Stork, deputy Middle East and North Africa director at Human Rights Watch. "Libyan authorities should allow peaceful protesters to have their say." The protests in Benghazi on February 18 began during funerals for the 20 demonstrators killed by security forces the day before. Eyewitnesses told Human Rights Watch that security forces with distinctive yellow uniforms opened fire on	

Unnamed: 0	article	oururl	newdescp	na
			protesters near the Fadil Bu Omar Katiba, a security force base in the center of Benghazi. One protester told Human Rights Watch he witnessed four men shot dead. By 11 p.m. on February 18, Al Jalaa Hospital in Benghazi had received the bodies of 35 people killed that day, a senior hospital official told Human Rights Watch. He said the deaths had been caused by gunshot wounds to the chest, neck, and head. Two sources at the hospital confirmed to Human Rights Watch that the death toll for February 17 was 20, and that at least 45 people had been wounded by bullets. The senior hospital official told Human Rights Watch, "We put out a call to all the doctors in Benghazi to come to the hospital and for everyone to contribute blood because I've never seen anything like this before." Witnesses said that after the February 18 shootings, protesters in Benghazi continued on to the courthouse and gathered there	

	Unnamed: 0	article	oururl	newdescp	na
				throughout the evening, the crowd swelling to thousands. In Baida, further to the east, protesters on February 18 buried the 23 people who had been shot dead the day before. One protester told Human Rights Watch that police were patrolling the streets but he had seen no further clashes.	
29	2625	Libya protests leave 24 dead, says rights group	http://en.wikipedia.org/wiki/Libyan_Civil_War_(2011)	At least 24 people have been killed in anti-government protests in Libya in recent days, rights activists say. Many others were wounded in the clashes between security forces and protesters, the US-based Human Rights Watch said. Protests continued overnight with thousands on the streets of the eastern city of Benghazi, where there is now a heavy military presence, witnesses said. Large protests are uncommon in Libya, where dissent is rarely allowed. Pro-democracy protests have recently swept through several Arab nations, with the presidents of Tunisia and Egypt forced from power amid growing unrest. The	Le

	Unnamed: 0	article	oururl	newdescp	na
				BBC's Jon Leyne in Cairo says violent confrontations are reported to have spread to five Libyan cities in demonstrations so far, but not yet to Tripoli, the capital, in any large numbers. Our correspondent says the reports reflect an extremely tough government response, including the use of gunfire and even denying supplies to hospitals. Funerals of some of those killed are expected to be held on Friday in Benghazi and al-Bayda, which correspondents say could spur more protests. Activists set up camps in al-Bayda after Thursday's "Day of Rage" protest against the government, witnesses said. Eyewitnesses believe that the death toll could be even higher, our correspondent says.	
873	638	Hundreds hurt, 6 killed in Yemen violence	http://en.wikipedia.org/wiki/Yemeni_Revolution	(CNN) -- Yemeni protesters and military and pro-government gangs clashed in several areas Tuesday, with at least six killed and hundreds more injured, as the future of President Ali Abdullah Saleh remained uncertain. The	Abd S ' Mo 'Abd Sa 'S 'Abd Qi 'S

Unnamed: 0	article	oururl	newdescp	na
			United States has no intention of stopping its military aid to Yemen, despite the unrest, Pentagon spokesman Geoff Morrell said Tuesday. The aid, in support of Yemeni counterterrorism efforts, continues to be essential because of the "real threat" from al Qaeda in the country, he said. In Sanaa, the capital, eyewitnesses and field medical teams told CNN that security forces and anti-riot police used batons to attack protesters among 40,000 people marching on Zubairy Street Tuesday evening. In addition, pro-government gangs attacked protesters on Tuesday near a military base. Four people were killed -- three pro-government demonstrators and one anti-government demonstrator. Windows were shattered on an ambulance carrying some of the 56 injured protesters to a hospital, witnesses said. "The government forces are killing us," said Abdullah Salem, a youth activist who was at the	

Unnamed: 0	article	oururl	newdescp	na
			protest. "Saleh and his militia will not succeed, and every blood spilt will be accounted for in international courts." In the city of Taiz, meanwhile, at least two anti-government protesters were killed when security forces and Republican Guards fired on protesters, according to medical teams. Hundreds of people were injured, 55 of them from gunshot wounds. The security chief in Taiz denied his forces fired on demonstrators. "Security forces did not attack protesters," said Abdullah Qiaran. "We were dispersing pro and anti-government protesters after we saw that both sides were clashing." An estimated 30,000 demonstrators marched near the presidential palace in the port city of Hodeida Tuesday evening, witnesses said. The violence comes as the United States is helping to mediate a transition out of office for Saleh, who has been facing popular protests for weeks, according to		

	Unnamed: 0	article	oururl	newdescp	na
				two Yemeni officials.	
161	804	Yemen toll rises as U.S. seen pressing Saleh to go	http://en.wikipedia.org/wiki/Yemeni_Revolution	1 of 13. Anti-government protesters run after police fired tear gas during a demonstration in the southern Yemeni city of Taiz April 4, 2011. The attempt to suppress mounting protests inspired by uprisings in Egypt and Tunisia came amid signs that the United States is seeking an end to Saleh's 32-year rule, long seen as a rampart against Yemen-based al Qaeda in the Arabian Peninsula. In Taiz, south of the capital Sanaa, police shot at protesters trying to storm the provincial government building, killing at least 15 and wounding 30, hospital doctors said. "The regime has surprised us with this extent of killing. I don't think the people will do anything other than come out with bare chests to drain the government of all its ammunition," parliamentarian Mohammed Muqbil al-Hamiri told Al Jazeera TV. The television showed a row of men, apparent tear gas victims,	['S 'Mohan Muqb 'Ha 'S 'S

Unnamed: 0	article	oururl	newdescp	na
			lying motionless and being tended by medics on the carpeted floor of a makeshift hospital in Taiz. In the Red Sea port of Hudaida, police and armed men in civilian clothes attacked a march toward a presidential palace. Three people were hit by bullets, around 30 were stabbed with knifes, and 270 were hurt from inhaling tear gas, doctors said. Later on Monday, doctors said at least six demonstrators were shot dead and several wounded during evening rallies, and that the toll was likely to rise. In Washington, the U.S. State Department called the latest violence in Yemen "appalling." Yemen's opposition coalition appealed in a statement to the United Nations, human rights groups and other international bodies "to intervene quickly to stop President Saleh and his entourage from shedding more blood." As opposition forces stepped up their actions, Saleh again appeared defiant.	

Test 2: Results

The results of the sentence encoder seem much more promising. Overall, all five articles share many similarities. They all refer to violent protests, and all are located somewhere in the Middle East/North Africa region (Egypt, Libya, Yemen). Several refer to human rights issues and anti-government protests, much like the sample article.

The results of this test seem far more accurate than the cosine similarity test.

Test 3: Spacy Vectorization

Finally, we will run a similarity test using Spacy's internal similarity function. This model vectorizes the sample and article texts using spacy's nlp vectorizer utilized above. It then compares similarity in both directions (sample->test, test->sample) and averages the similarity score.

For this test, we will utilize the tokenized and cleaned version of the article text.

In [5]:
```
# Loading spacy vectorized model - en_core_web_md
nlp = spacy.load("en_core_web_md")
```

In [6]:
```
# Defining function to use spacy's vectorizer
def spacy_test(string):
    # Checking to make sure each value contains information
    if string:
        # Storing sample article and string, vectorizing using spacy nlp
        sample = nlp('''
             Human Rights Watch says government-controlled health servic
es in Egypt have been pressured into playing down the number of casualti
es during anti-government protests. The group has documented the deaths
 of 297 people, but says the final toll is likely to be significantly hi
gher. Human Rights Watch says the vast majority of the deaths in Cairo,
 Alexandria and Suez were on January 28 and 29 as a result of live gunfi
re as riot police fought running battles with protesters. A significant
 proportion came as a result of rubber bullets fired at too close a rang
e and from teargas canisters fired into the crowds at very close range.
 Human Rights Watch says the actual number of deaths is likely to be an
 underestimate because the organisation had only included those deaths i
t had verified itself at key hospitals in the three major cities.
            ''')
        testdoc = nlp(string)
        # Testing for similarity in both directions
        similarity = sample.similarity(testdoc)
        similarity_rev = testdoc.similarity(sample)
        # Averaging both similarity scores
        sim_score = (similarity + similarity_rev) / 2
    else:
        pass
    return sim_score
```

In [7]:
```
# Applying spacy vectorization test to all articles
text_simple['spacy_test'] = text_simple['processed_text'].apply(spacy_te
st).copy()
```

In [8]:
```
spacy_samples = text_simple.sort_values(by='spacy_test', ascending=False
).head(6)
spacy_samples = spacy_samples[1:]
spacy_samples.head()
```

Out[8]:

	Unnamed: 0	article	oururl	
35	1188	UPDATE 1-Libya said to use cluster arms, Tripoli denies it	http://en.wikipedia.org/wiki/Timeline_of_the_2011_Libyan_Civil_War	Cluster statem to civili munitior beca Times re m de ma bombs soc

	Unnamed: 0	article	oururl	
53	8	US admits Afghan airstrike errors	http://en.wikipedia.org/wiki/Drone_strikes_in_Pakistan	A row prov die stri conclu what r was a fa a grea ir est casu accomp strikes c all of atta recomr be intr US

Unnamed: 0	article	oururl

Unnamed: 0	article	oururl

	Unnamed: 0	article	oururl	
657	2473	Report on sex abuse 'to be worse than Ferns'	http://en.wikipedia.org/wiki/Sexual_abuse_in_Cloyne_diocese	THE Go horrific in detai alloc night. T

	Unnamed: 0	article	oururl	
418	2815	Iraq death toll 'soared post-war'	http://en.wikipedia.org/wiki/Casualties_of_the_Iraq_War	A stu invasi small s modern h databa The L
1986	2508	U.S. military scales	http://en.wikipedia.org/wiki/Typhoon_Haiyan	Typh

Unnamed: 0	article	oururl
	down aid efforts in Philippines	274*mil*)*worth* . *TheU* . *Navy* – *pow* *theUS* *butstil* – 130*a* . *Lastu* *theUn* . *Jeren* *direct* . *S* . *Agen* (*USAI* *saidth* . *milita* . " *Wha* *partic* *isnow* – *secto* " *Kon* . " *Youd* *theyar* *butinth* *thatre* . " *Kony* – *hitL* . " *Fooa* *shelte* . *Ithini* . " *Aidd* *theU*. . *huma* . *How* *major* . *Kony* – *term* *giving* . *TheU* millior

	Unnamed: 0	article	oururl

Test 3 Results

This test required a significant amount more compute power than the others, and its results are similar to those of Test 1. Two articles seem to be related to civil unrest in the middle east region. The others, however, relate to a child sex abuse scandal in Cork, an estimate of civilian death toll in Iraq, and an article on Typhoon Haian in the Philippines.

On the whole, this model seems to be the least accurate.

Part 2: Results

After processing the text using three separate tools, it seems that the Sentence Encoding test provides the most accurate result.

Outputting Final Results to Files

Finally, I will output two files:

- 'texts_processed.csv': a final csv file of all relevant information, including the columns of names, distinct names, dates and datetime objects.
- 'closest_matches.csv': a sampling of the top five articles in similarity to the sample article, using the Sentence Embedding method.

In [14]:
```
text_simple.to_csv('texts_processed.csv')
sentence_samples.to_csv('closest_matches.csv')
```