STOCK ANALYSIS AND MOVEMENT PREDICTION USING NEWS HEADLINES

INDEX

PACKAGES USED

PANDAS

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way toward this goal.

Pandas are well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure

The two primary data structures of pandas are:

- Series (1-dimensional)
- Data Frame (2-dimensional)

They handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. For R users, Data Frame provides everything that R's data. Frame provides and much more.

Pandas are built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.

Here are just a few of the things that pandas does well:

Easy handling of missing data (represented as Nan) in floating point as well as non-floating point data

Size mutability: columns can be **inserted and deleted** from Data Frame and higher dimensional objects

Automatic and explicit **data alignment**: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, Data Frame, etc. automatically align the data for you in computations

Powerful, flexible **group by** functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data.

Make it **easy to convert** ragged, differently-indexed data in other Python and NumPy data structures into Data Frame objects

Intelligent label-based **slicing**, **fancy indexing**, and **subletting** of large data sets

Intuitive **merging** and **joining** data sets

Flexible **reshaping** and pivoting of data sets

Hierarchical labeling of axes (possible to have multiple labels per tick)

Robust IO tools for loading data from **flat files** (CSV and delimited), Excel files, databases, and saving / loading data from the ultrafast **HDF5 format**

Time series-specific functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging, etc.

Many of these principles are here to address the shortcomings frequently experienced using other languages / scientific research environments. For data scientists, working with data is typically divided into multiple stages: munging and cleaning data, analyzing / modeling it, then organizing the results of the analysis into a form suitable for plotting or tabular display. Pandas is the ideal tool for all of these tasks.

NUMPY

NumPy is an open source library available in Python that aids in mathematical, scientific, engineering, and data science programming. NumPy is an incredible library to perform mathematical and statistical operations. It works perfectly well for multi-dimensional arrays and matrices multiplication

For any scientific project, NumPy is the tool to know. It has been built to work with the N-dimensional array, linear algebra, random number, Fourier transform, etc. It can be integrated to C/C++ and FORTRAN.

NumPy is a programming language that deals with multi-dimensional arrays and matrices. On top of the arrays and matrices, NumPy supports a large number of mathematical operations

Why use NumPy?

NumPy is memory efficiency, meaning it can handle the vast amount of data more accessible than any other library. Besides, NumPy is very convenient to work with, especially for matrix multiplication and reshaping. On top of that, NumPy is fast. In fact, Tensor Flow and Scikit learn to use NumPy array to compute the matrix multiplication in the back end.

TENSORFLOW

Tensorflow's name is directly derived from its core framework: **Tensor**. In Tensor flow, all the computations involve tensors. A tensor is a **vector** or **matrix** of n-dimensions that represents all types of data. All values in a tensor hold identical data type with a known (or partially known) **shape**. The shape of the data is the dimensionality of the matrix or array.

A tensor can be originated from the input data or the result of a computation. In Tensor Flow, all the operations are conducted inside a **graph**. The graph is a set of computation that takes place successively. Each operation is called an **op node** and are connected to each other.

The graph outlines the ops and connections between the nodes. However, it does not display the values. The edge of the nodes is the tensor, i.e., a way to populate the operation with data.

In Machine Learning, models are feed with a list of objects called **feature vectors**. A feature vector can be of any data type. The feature vector will usually be the primary input to populate a tensor.

These values will flow into an op node through the tensor and the result of this operation/computation will create a new tensor which in turn will be used in a new operation. All these operations can be viewed in the graph.

KERAS

Keras is a high-level neural networks API, written in Python and capable of running on top of Tensor Flow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. *Being able to go from idea to result with the least possible delay is key to doing good research.*

Use Keras if you need a deep learning library that:

- Allows for easy and fast prototyping (through user friendliness, modularity, and extensibility).
- Supports both convolution networks and recurrent networks, as well as combinations of the two.
- Runs seamlessly on CPU and GPU.

RE

A regular expression in a programming language is a special text string used for describing a search pattern. It is extremely useful for extracting information from text such as code, files, log, spreadsheets or even documents.

While using the regular expression the first thing is to recognize is that everything is essentially a character, and we are writing patterns to match a specific sequence of characters also referred as string. Ascii or Latin letters are those that are on your keyboards and Unicode is used to match the foreign text. It includes digits and punctuation and all special characters like $#@! %, etc.

For instance, a regular expression could tell a program to search for specific text from the string and then to print out the result accordingly. Expression can include

Text matching

Repetition

Branching

Pattern-composition etc.

In Python, a regular expression is denoted as RE (REs, regexes or regex pattern) are imported through **re module**. Python supports regular expression through libraries. In Python regular expression supports various things like **Modifiers, Identifiers, and White space characters**.

Identifiers	**Modifiers**	**White space characters**	**Escape required**

\d= any number (a digit)	\d represents a digit.Ex: \d{1,5} it will declare digit between 1,5 like 424,444,545 etc.	\n = new line	. + * ? [] $ ^ () {} | \
\D= anything but a number (a non-digit)	+ = matches 1 or more	\s= space	
\s = space (tab,space,new line etc.)	? = matches 0 or 1	\t =tab	
\S= anything but a space	* = 0 or more	\e = escape	
\w = letters (Match alphanumeric character, including "_")	$ match end of a string	\r = carriage return	
\W =anything but letters (Matches a non-	^ match start of a string	\f= form feed	

alphanumeric character excluding "_")			
. = anything but letters (periods)	| matches either or x/y	----------	
\b = any character except for new line	[] = range or "variance"	----------	
\.	{x} = this amount of preceding code	----------	

SKLEARN

Is a library in Python that provides many unsupervised and supervised learning algorithms. It's built upon some of the technology you might already be familiar with, like NumPy, pandas, and Matplotlib!

The functionality that scikit-learn provides include:

- **Regression**, including Linear and Logistic Regression
- **Classification**, including K-Nearest Neighbors
- **Clustering**, including K-Means and K-Means++

- **Model selection**
- **Preprocessing**, including Min-Max Normalization

ABOUT GLOVE

Introduction

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Highlights

Nearest Neighbors

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

a. *frog*
b. frogs
c. toad
d. litoria
e. leptodactylidae
f. rana
g. lizard

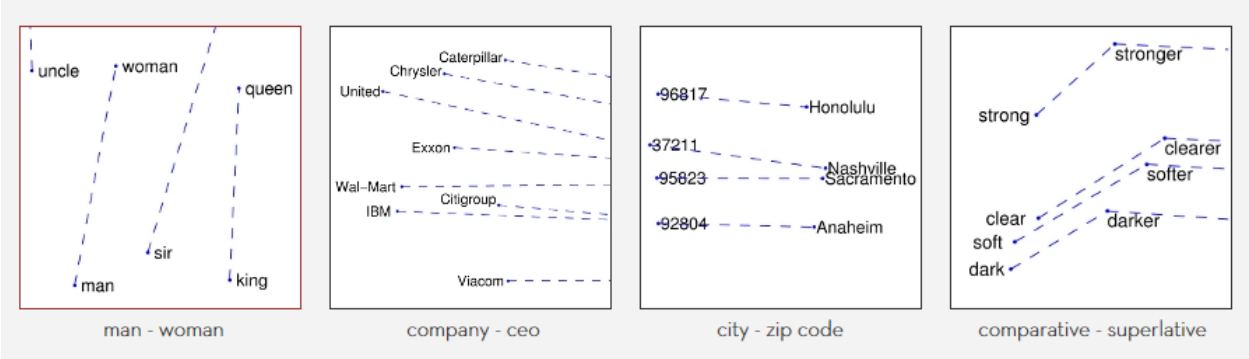
h. eleutherodactylus

3. litoria 4. leptodactylidae 5. rana 7. eleutherodactylus

Linear substructures

The similarity metrics used for nearest neighbor evaluations produce a single scalar that quantifies the relatedness of two words. This simplicity can be problematic since two given words almost always exhibit more intricate relationships than can be captured by a single number. For example, *man* may be regarded as similar to *woman* in that both words describe human beings; on the other hand, the two words are often considered opposites since they highlight a primary axis along which humans differ from one another.

In order to capture in a quantitative way the nuance necessary to distinguish *man* from *woman*, it is necessary for a model to associate more than a single number to the word pair. A natural and simple candidate for an enlarged set of discriminative numbers is the vector difference between the two word vectors. GloVe is designed in order that such vector differences capture as much as possible the meaning specified by the juxtaposition of two words.
The underlying concept that distinguishes man from woman, i.e. sex or gender, may be equivalently specified by various other word pairs, such as king and queen or brother and sister. To state this observation mathematically, we might expect that the vector differences man - woman, king - queen, and brother - sister might all be roughly equal. This property and other interesting patterns can be observed in the above set of visualizations.



Training

The GloVe model is trained on the non-zero entries of a global word-word co-occurrence matrix, which tabulates how frequently words co-occur with one another in a given corpus. Populating this matrix requires a single pass through the entire corpus to collect the statistics. For large corpora, this pass can be computationally expensive, but it is a one-time up-front cost. Subsequent training iterations are much faster because the number of non-zero matrix entries is typically much smaller than the total number of words in the corpus.

The tools provided in this package automate the collection and preparation of co-occurrence statistics for input into the model. The core training code is separated from these preprocessing steps and can be executed independently.

Model Overview

GloVe is essentially a log-bilinear model with a weighted least-squares objective. The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning. For example, consider the co-occurrence probabilities for target words *ice* and *steam* with various probe words from the vocabulary. Here are some actual probabilities from a 6 billion word corpus:

As one might expect, *ice* co-occurs more frequently with *solid* than it does with *gas*, whereas *steam* co-occurs more frequently with *gas* than it does with *solid*. Both words co-occur with their shared property *water* frequently, and both co-occur with the unrelated word *fashion* infrequently.
Only in the ratio of probabilities does noise from non-discriminative words like *water* and *fashion* cancel out, so that large values (much greater than 1) correlate well with properties specific to ice, and small values (much less than 1) correlate well with properties specific of steam. In this way, the ratio of probabilities encodes some crude form of meaning associated with the abstract concept of thermodynamic phase.

The training objective of GloVe is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence. Owing to the fact that the logarithm of a ratio equals the difference of logarithms, this objective associates (the logarithm of) ratios of co-occurrence probabilities with vector differences in the word vector space. Because these ratios can encode some form of meaning, this information gets encoded as vector differences as well. For this reason, the resulting word vectors perform very well on word analogy tasks, such as those examined in the word2vec package.

WORD EMBEDDINGS

What are word embeddings?

"Word embeddings" are a family of natural language processing techniques aiming at mapping semantic meaning into a geometric space. This is done by associating a numeric vector to every word in a dictionary, such that the distance (e.g. L2 distance or more commonly cosine distance) between any two vectors would capture part of the semantic relationship between the two associated words. The geometric space formed by these vectors is called an *embedding space*.
For instance, "coconut" and "polar bear" are words that are semantically quite different, so a reasonable embedding space would represent them as vectors that would be very far apart. But "kitchen" and "dinner" are related words, so they should be embedded close to each other.

Ideally, in a good embeddings space, the "path" (a vector) to go from "kitchen" and "dinner" would capture precisely the semantic relationship between these two concepts. In this case the relationship is "where x occurs", so you would expect the vector kitchen - dinner (difference of the two embedding vectors, i.e. path to go from dinner to kitchen) to capture this "where x occurs" relationship. Basically, we should have the vectorial identity: dinner + (where x occurs) = kitchen (at least approximately). If that's indeed the case, then we can use such a relationship vector to answer questions. For instance, starting from a new vector, e.g. "work", and applying this relationship vector, we should get sometime meaningful, e.g. work + (where x occurs) = office, answering "where does work occur?".
Word embeddings are computed by applying dimensionality reduction techniques to datasets of co-occurence statistics between words in a corpus of text. This can be done via neural networks (the "word2vec" technique), or via matrix factorization

.

GloVe word embeddings

We will be using GloVe embeddings, which you can read about here. GloVe stands for "Global Vectors for Word Representation". It's a somewhat popular embedding technique based on factorizing a matrix of word co-occurence statistics.
Specifically, we will use the 300-dimensional GloVe embeddings of 400k words computed on a 2014 dump of English Wikipedia.

Approach

Here's how we will solve the classification problem:

- convert all text samples in the dataset into sequences of word indices. A "word index" would simply be an integer ID for the word. We will only consider the top 20,000 most commonly occurring words in the dataset, and we will truncate the sequences to a maximum length of 1000 words.
- prepare an "embedding matrix" which will contain at index I the embedding vector for the word of index I in our word index.
- load this embedding matrix into a Keras Embedding layer, set to be frozen (its weights, the embedding vectors, will not be updated during training).
- Build on top of it a convolution neural network, ending in a softmax output over our 20 categories.

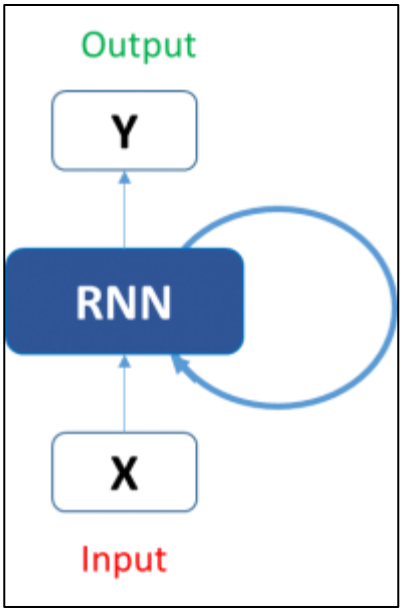
LSTM ARCHITECTURE

Flashback: A look into Recurrent Neural Networks(RNN)

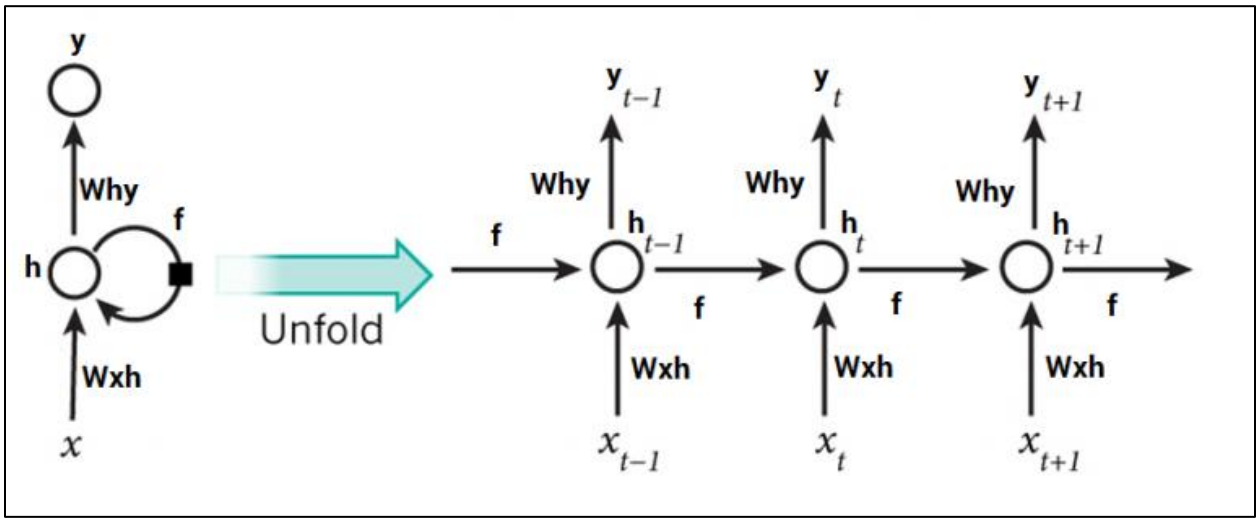
Take an example of sequential data, which can be the stock market's data for a particular stock. A simple machine learning model or an Artificial Neural Network may learn to predict the stock prices based on a number of features: the volume of the stock, the opening value etc. While the price of the stock depends on these features, it is also largely dependent on the stock values in the previous days. In fact for a trader, these values in the previous days (or the trend) is one major deciding factor for predictions.

In the conventional feed-forward neural networks, all test cases are considered to be independent. That is when fitting the model for a particular day, there is no consideration for the stock prices on the previous days.

This dependency on time is achieved via Recurrent Neural Networks. A typical RNN looks like:



This may be intimidating at first sight, but once unfolded, it looks a lot simpler:



Now it is easier for us to visualize how these networks are considering the trend of stock prices, before predicting the stock prices for today. Here every prediction at time t (h_t) is dependent on all previous predictions and the information learned from them.

RNNs can solve our purpose of sequence handling to a great extent but not entirely.

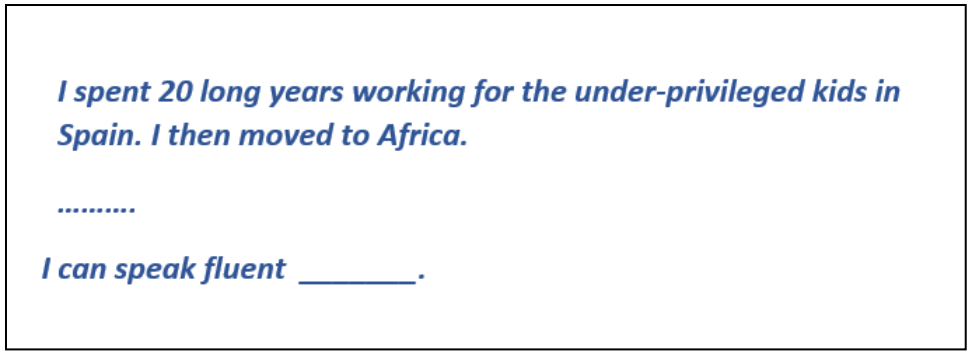
We want our computers to be good enough to write Shakespearean sonnets. Now RNNs are great when it comes to short contexts, but in order to be able to build a story and remember it, we need our models to be able to understand and remember the context behind the sequences, just like a human brain. This is not possible with a simple RNN.

Limitations of RNNs

Recurrent Neural Networks work just fine when we are dealing with short-term dependencies. That is when applied to problems like:

RNNs turn out to be quite effective. This is because this problem has nothing to do with the context of the statement. The RNN need not remember what was said before this, or what was its meaning, all they need to know is that in most cases the sky is blue. Thus the prediction would be:

However, vanilla RNNs fail to understand the context behind an input. Something that was said long before, cannot be recalled when making predictions in the present. Let’s understand this as an example:



Here, we can understand that since the author has worked in Spain for 20 years, it is very likely that he may possess a good command over Spanish. But, to make a proper prediction, the RNN needs to remember this context. The relevant information may be separated from the point where it is needed, by a huge load of irrelevant data. This is where a Recurrent Neural Network fails!

The reason behind this is the problem of **Vanishing Gradient.** In order to understand this, you'll need to have some knowledge about how a feed-forward neural network learns. We know that for a conventional feed-forward neural network, the weight updating that is applied on a particular layer is a multiple of the learning rate, the error term from the previous layer and the input to that layer. Thus, the error term for a particular layer is somewhere a product of all previous layers' errors. When dealing with activation functions like the sigmoid function, the small values of its derivatives (occurring in the error function) gets multiplied multiple times as we move towards the starting layers. As a result of this, the gradient almost vanishes as we move towards the starting layers, and it becomes difficult to train these layers.

A similar case is observed in Recurrent Neural Networks. RNN remembers things for just small durations of time, i.e. if we need the information after a small time it may be reproducible, but once a lot of words are fed in, this information gets lost somewhere. This issue can be resolved by applying a slightly tweaked version of RNNs – the Long Short-Term Memory Networks.

Improvement over RNN: LSTM (Long Short-Term Memory) Networks

When we arrange our calendar for the day, we prioritize our appointments right? If in case we need to make some space for anything important we know which meeting could be canceled to accommodate a possible meeting.

Turns out that an RNN doesn't do so. In order to add a new information, it transforms the existing information completely by applying a function. Because of this, the entire information is modified, on the whole, I. e. there is no consideration for *'important'* information and *'not so important'* information.

LSTMs on the other hand, make small modifications to the information by multiplications and additions. With LSTMs, the information flows through a mechanism known as cell states. This way, LSTMs can selectively remember or forget things. The information at a particular cell state has three different dependencies.

We'll visualize this with an example. Let's take the example of predicting stock prices for a particular stock. The stock price of today will depend upon:

1. The trend that the stock has been following in the previous days, maybe a downtrend or an uptrend.
2. The price of the stock on the previous day, because many traders compare the stock's previous day price before buying it.
3. The factors that can affect the price of the stock for today. This can be a new company policy that is being criticized widely, or a drop in the company's profit, or maybe an unexpected change in the senior leadership of the company.

These dependencies can be generalized to any problem as:

1. The previous cell state *(i.e. the information that was present in the memory after the previous time step)*
2. The previous hidden state *(i.e. this is the same as the output of the previous cell)*
3. The input at the current time step *(i.e. the new information that is being fed in at that moment)*

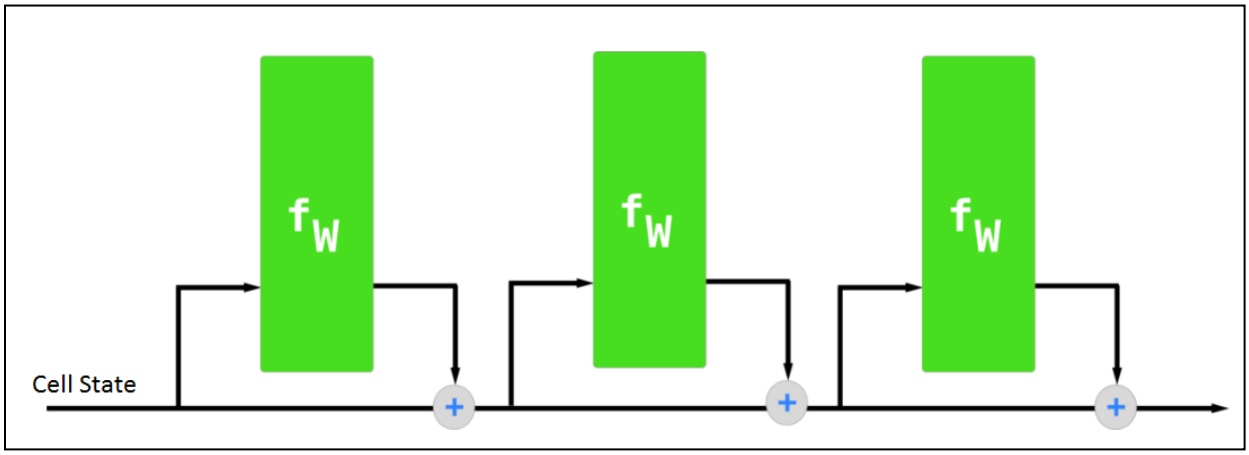
Another important feature of LSTM is its analogy with conveyor belts!

That's right!

Industries use them to move products around for different processes. LSTMs use this mechanism to move information around.

We may have some addition, modification or removal of information as it flows through the different layers, just like a product may be molded, painted or packed while it is on a conveyor belt.

The following diagram explains the close relationship of LSTMs and conveyor belts.



Source

Although this diagram is not even close to the actual architecture of an LSTM, it solves our purpose for now.

Just because of this property of LSTMs, where they do not manipulate the entire information but rather modify them slightly, they are able to *forget* and *remember* things selectively. How do they do so, is what we are going to learn in the next section?

Architecture of LSTMs

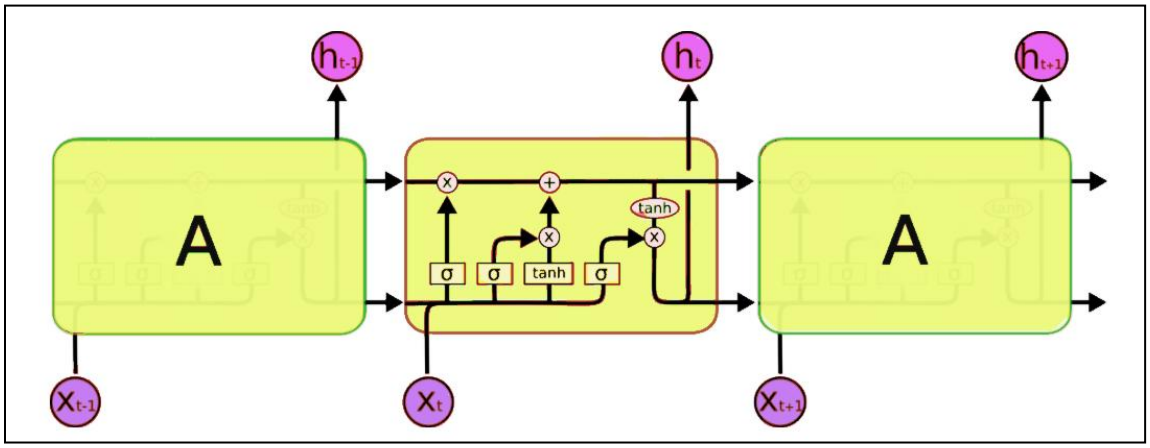
The functioning of LSTM can be visualized by understanding the functioning of a news channel's team covering a murder story. Now, a news story is built around facts, evidence and statements of many people. Whenever a new event occurs you take either of the three steps.

Let's say, we were assuming that the murder was done by 'poisoning' the victim, but the autopsy report that just came in said that the cause of death was 'an impact on the head'. Being a part of this news team what do you do? You immediately **forget** the previous cause of death and all stories that were woven around this fact.

What, if an entirely new suspect is introduced into the picture. A person who had grudges with the victim and could be the murderer?
You **input** this information into your news feed, right?

Now all these broken pieces of information cannot be served on mainstream media. So, after a certain time interval, you need to summarize this information and **output** the relevant things to your audience. Maybe in the form of "*XYZ turns out to be the prime suspect.*".

Now let's get into the details of the architecture of LSTM network:



Source

Now, this is nowhere close to the simplified version which we saw before, but let me walk you through it. A typical LSTM network is comprised of different memory blocks called **cells** (the rectangles that we see in the image)**.** There are two states that are being transferred to the next cell; the **cell state** and the **hidden state**. The memory blocks are responsible for remembering things and manipulations to this memory is done through three major mechanisms, called **gates.** Each of them is being discussed below.

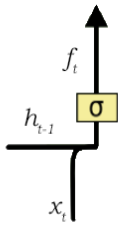
Forget Gate

Taking the example of a text prediction problem. Let's assume an LSTM is fed in, the following sentence:

> ***Bob is a nice person. Dan on the other hand is evil.***

As soon as the first full stop after "*person"* is encountered, the forget gate realizes that there may be a change of context in the next sentence.

As a result of this, the *subject* of the sentence is *forgotten* and the place for the subject is vacated. And when we start speaking about "*Dan"* this position of the subject is allocated to "*Dan"*. This process of forgetting the subject is brought about by the forget gate.



A forget gate is responsible for removing information from the cell state. The information that is no longer required for the LSTM to understand things or the information that is of less importance is removed via multiplication of a filter. This is required for optimizing the performance of the LSTM network.

This gate takes in two inputs; h_t-1 and x_t.

h_t-1 is the hidden state from the previous cell or the output of the previous cell and x_t is the input at that particular time step. The given inputs are multiplied by the weight matrices and a bias is added. Following this, the sigmoid function is applied to this value. The sigmoid function outputs a vector, with values ranging from 0 to 1, corresponding to each number in the cell state. Basically, the sigmoid function is responsible for deciding which values to keep and which to discard. If a '0' is output for a particular value in the cell state, it means that the forget gate wants the cell state to forget that piece of information completely. Similarly, a '1' means that the forget gate wants to remember that entire piece of information. This vector output from the sigmoid function is multiplied to the cell state.

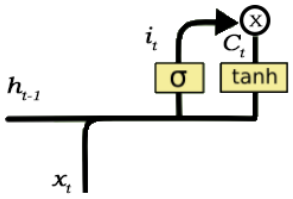
Input Gate

Okay, let's take another example where the LSTM is analyzing a sentence:

> ***Bob knows swimming. He told me over the phone that he had served the navy for 4 long years.***

Now the important information here is that "Bob" knows swimming and that he has served the Navy for four years. This can be added to the cell state, however, the fact that he told all this over the phone is a less important fact and can be ignored. This process of adding some new information can be done via the **input** gate.

Here is its structure:



The input gate is responsible for the addition of information to the cell state. This addition of information is basically three-step process as seen from the diagram above.

1. Regulating what values need to be added to the cell state by involving a sigmoid function. This is basically very similar to the forget gate and acts as a filter for all the information from h_t-1 and x_t.

2. Creating a vector containing all possible values that can be added (as perceived from h_t-1 and x_t) to the cell state. This is done using the **tanh** function, which outputs values from -1 to +1.
3. Multiplying the value of the regulatory filter (the sigmoid gate) to the created vector (the tanh function) and then adding this useful information to the cell state via addition operation.

Once this three-step process is done with, we ensure that only that information is added to the cell state that is *important* and is not *redundant*.

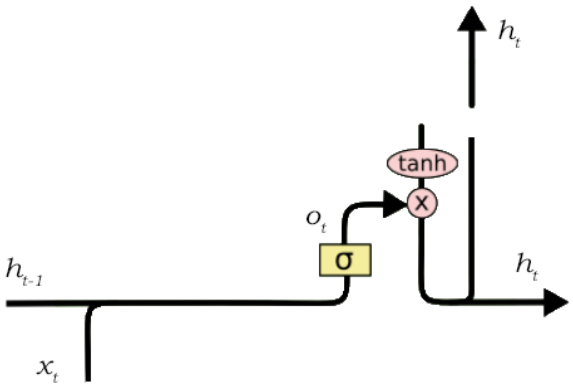
Output Gate

Not all information that runs along the cell state, is fit for being output at a certain time. We'll visualize this with an example:

> ***Bob fought single handedly with the enemy and died for his country. For his contributions brave _____.***

In this phrase, there could be a number of options for the empty space. But we know that the current input of *'brave'*, is an adjective that is used to describe a noun. Thus, whatever word follows, has a strong tendency of being a noun. And thus, Bob could be an apt output.

This job of selecting useful information from the current cell state and showing it out as an output is done via the output gate. Here is its structure:



The functioning of an output gate can again be broken down to three steps:

1. Creating a vector after applying **tanh** function to the cell state, thereby scaling the values to the range -1 to +1.
2. Making a filter using the values of h_t-1 and x_t, such that it can regulate the values that need to be output from the vector created above. This filter again employs a sigmoid function.
3. Multiplying the value of this regulatory filter to the vector created in step 1, and sending it out as a output and also to the hidden state of the next cell.

The filter in the above example will make sure that it diminishes all other values but *'Bob'*. Thus the filter needs to be built on the input and hidden state values and be applied on the cell state vector.

CODE

In [80]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [81]:

```
ST = pd.read_csv("AMZN.csv")
HL = pd.read_csv("Headlines.csv")
```

Data Description¶

In [82]:

```
ST.head(20)
#hello
```

Out[82]:

	Date	Open	High	Low	Close	Adj Close	Volume
0	2016-07-01	717.320007	728.000000	716.539978	725.679993	725.679993	2920400
1	2016-06-30	717.200012	719.369995	712.539978	715.619995	715.619995	2855100
2	2016-06-29	715.750000	719.500000	713.539978	715.599976	715.599976	3070100
3	2016-06-28	700.000000	708.000000	698.169983	707.950012	707.950012	4037000
4	2016-06-27	692.010010	696.820007	682.119995	691.359985	691.359985	5568000
5	2016-06-24	693.000000	712.530029	692.200012	698.960022	698.960022	7632500
6	2016-06-23	715.500000	722.119995	712.500000	722.080017	722.080017	2825000
7	2016-06-22	716.580017	717.000000	707.570007	710.599976	710.599976	2260500
8	2016-06-21	715.719971	718.400024	712.719971	715.820007	715.820007	2137500

	Date	Open	High	Low	Close	Adj Close	Volume
9	2016-06-20	713.500000	721.309998	710.809998	714.010010	714.010010	3677200
10	2016-06-17	718.190002	718.200012	699.179993	706.390015	706.390015	5897800
11	2016-06-16	712.049988	718.000000	705.299988	717.510010	717.510010	3098000
12	2016-06-15	722.000000	722.559998	713.349976	714.260010	714.260010	2709400
13	2016-06-14	712.330017	720.809998	712.270020	719.299988	719.299988	2506900
14	2016-06-13	714.010010	721.989990	711.159973	715.239990	715.239990	3352200
15	2016-06-10	722.349976	724.979980	714.210022	717.909973	717.909973	3425700
16	2016-06-09	723.099976	728.909973	722.299988	727.650024	727.650024	2170300
17	2016-06-08	726.400024	729.419983	721.599976	726.640015	726.640015	2223400
18	2016-06-07	729.890015	730.000000	720.549988	723.739990	723.739990	2732500
19	2016-06-06	726.500000	731.500000	724.419983	726.729980	726.729980	2704800

In [83]:

ST.tail(10)
Out[83]:

	Date	Open	High	Low	Close	Adj Close	Volume
1980	2008-08-20	82.000000	83.250000	81.199997	82.129997	82.129997	5950600
1981	2008-08-19	83.089996	83.510002	81.059998	81.290001	81.290001	6630400
1982	2008-08-18	86.089996	86.279999	83.040001	83.110001	83.110001	6547400
1983	2008-08-15	88.279999	89.529999	86.260002	86.400002	86.400002	6871600
1984	2008-08-14	85.709999	88.750000	85.220001	88.029999	88.029999	6901700
1985	2008-08-13	86.279999	88.250000	84.540001	86.690002	86.690002	7208800

	Date	Open	High	Low	Close	Adj Close	Volume
1986	2008-08-12	87.320000	88.480003	86.099998	87.250000	87.250000	8026500
1987	2008-08-11	80.180000	91.750000	79.779999	88.089996	88.089996	25070200
1988	2008-08-08	76.779999	81.209999	76.290001	80.510002	80.510002	9162700
1989	2008-08-07	77.010002	78.050003	76.000000	76.949997	76.949997	5444800

In [84]:

HL.head(10)
Out[84]:

	Date	News
0	2016-07-01	A 117-year-old woman in Mexico City finally re...
1	2016-07-01	IMF chief backs Athens as permanent Olympic host
2	2016-07-01	The president of France says if Brexit won, so...
3	2016-07-01	British Man Who Must Give Police 24 Hours' Not...
4	2016-07-01	100+ Nobel laureates urge Greenpeace to stop o...
5	2016-07-01	Brazil: Huge spike in number of police killing...
6	2016-07-01	Austria's highest court annuls presidential el...
7	2016-07-01	Facebook wins privacy case, can track any Belg...
8	2016-07-01	Switzerland denies Muslim girls citizenship af...
9	2016-07-01	China kills millions of innocent meditators fo...

In [85]:

HL.tail(10)
Out[85]:

	Date	News
73598	2008-06-08	b"S. Korean protesters, police clash in beef r...
73599	2008-06-08	b"Oil reserves 'will last decades' - a BBC Sco...
73600	2008-06-08	b'Cameras designed to detect terrorist facial ...
73601	2008-06-08	b'Israeli peace activists protest 41 years of ...
73602	2008-06-08	b"A 5.1 earthquake hits China's Southern Qingh...
73603	2008-06-08	b'Man goes berzerk in Akihabara and stabs ever...
73604	2008-06-08	b'Threat of world AIDS pandemic among heterose...
73605	2008-06-08	b'Angst in Ankara: Turkey Steers into a Danger...
73606	2008-06-08	b"UK: Identity cards 'could be used to spy on ...
73607	2008-06-08	b'Marriage, they said, was reduced to the stat...

In [86]:

ST.describe()
Out[86]:

	Open	High	Low	Close	Adj Close	Volume
count	1990.000000	1990.000000	1990.000000	1990.000000	1990.000000	1.990000e+03
mean	264.210412	267.391015	260.883854	264.311880	264.311880	5.480421e+06
std	161.474786	162.888409	159.816264	161.429345	161.429345	3.969781e+06
min	35.290001	39.000000	34.680000	35.029999	35.029999	9.844000e+05
25%	136.977500	139.012501	135.152501	137.130005	137.130005	3.018325e+06
50%	230.279999	234.354996	227.424996	230.529999	230.529999	4.443150e+06
75%	335.232506	337.474998	331.612503	333.922493	333.922493	6.644050e+06

	Open	High	Low	Close	Adj Close	Volume
max	729.890015	731.500000	724.419983	728.239990	728.239990	5.830580e+07

In [87]:

HL.describe()
Out[87]:

	Date	News
Count	73608	73608
unique	2943	73537
Top	2008-10-26	Iceland Declares Independence from Internation...
Freq	50	3

In [88]:

HL.isnull().sum()
Out[88]:

```
Date    0
News    0
dtype: int64
```

In [89]:

ST.isnull().sum()
Out[89]:

```
Date         0
Open         0
High         0
Low          0
Close        0
Adj Close    0
Volume       0
dtype: int64
```

In [90]:

```
print(ST.shape)
print(HL.shape)
(1990, 7)
```

(73608, 2)
In [91]:

```
print(len(set(ST.Date)))
print(len(set(HL.Date)))
```

1990
2943
In [92]:

```
HL = HL[HL.Date.isin(ST.Date)]
```

In [93]:

```
HL
```

Out[93]:

	Date	News
0	2016-07-01	A 117-year-old woman in Mexico City finally re...
1	2016-07-01	IMF chief backs Athens as permanent Olympic host
2	2016-07-01	The president of France says if Brexit won, so...
3	2016-07-01	British Man Who Must Give Police 24 Hours' Not...
4	2016-07-01	100+ Nobel laureates urge Greenpeace to stop o...
...	...	...
72103	2008-08-07	b'Pininfarina, Italian Car Designer, Dies in T...
72104	2008-08-07	b'Venezuelas Chavez Pushes for New World Finan...
72105	2008-08-07	b'Stalinism Was Just as Bad as Nazism'
72106	2008-08-07	b'Anti-War Website Operator Threatened By Arme...
72107	2008-08-07	b''Police fear Maddie 'stolen to order'"

49743 rows Ã—2 columns

In [94]:

print(len(set(ST.Date)))
print(len(set(HL.Date)))
1990
1990
In [95]:

ST = ST.set_index('Date').diff(periods=1)
In [96]:

ST
Out[96]:

	Open	High	Low	Close	Adj Close	Volume
Date						
2016-07-01	Nan	Nan	Nan	Nan	Nan	Nan
2016-06-30	-0.119995	-8.630005	-4.000000	-10.059998	-10.059998	-65300.0
2016-06-29	-1.450012	0.130005	1.000000	-0.020019	-0.020019	215000.0
2016-06-28	-15.750000	-11.500000	-15.369995	-7.649964	-7.649964	966900.0
2016-06-27	-7.989990	-11.179993	-16.049988	-16.590027	-16.590027	1531000.0
...	...	...	...	...	...	...
2008-08-13	0.570000	-0.500000	-0.680000	-1.339997	-1.339997	307100.0
2008-08-12	1.040001	0.230003	1.559997	0.559998	0.559998	817700.0
2008-08-11	-7.140000	3.269997	-6.319999	0.839996	0.839996	17043700.0
2008-08-08	-3.400001	-10.540001	-3.489998	-7.579994	-7.579994	-15907500.0
2008-08-07	0.230003	-3.159996	-0.290001	-3.560005	-3.560005	-3717900.0

1990 rows Ã—6 columns

In [97]:

```
ST['Date'] = ST.index
```

In [98]:

```
ST
```

Out[98]:

	Open	High	Low	Close	Adj Close	Volume	Date
Date							
2016-07-01	Nan	Nan	Nan	Nan	Nan	Nan	2016-07-01
2016-06-30	-0.119995	-8.630005	-4.000000	-10.059998	-10.059998	-65300.0	2016-06-30
2016-06-29	-1.450012	0.130005	1.000000	-0.020019	-0.020019	215000.0	2016-06-29
2016-06-28	-15.750000	-11.500000	-15.369995	-7.649964	-7.649964	966900.0	2016-06-28
2016-06-27	-7.989990	-11.179993	-16.049988	-16.590027	-16.590027	1531000.0	2016-06-27
...	...	...	...	...	...	...	...
2008-08-13	0.570000	-0.500000	-0.680000	-1.339997	-1.339997	307100.0	2008-08-13
2008-08-12	1.040001	0.230003	1.559997	0.559998	0.559998	817700.0	2008-08-12
2008-08-11	-7.140000	3.269997	-6.319999	0.839996	0.839996	17043700.0	2008-08-11
2008-08-08	-3.400001	-10.540001	-3.489998	-7.579994	-7.579994	-15907500.0	2008-08-08
2008-08-	0.230003	-3.159996	-0.290001	-3.560005	-3.560005	-3717900.0	2008-08-

	Open	High	Low	Close	Adj Close	Volume	Date
Date							
07							07

1990 rows Ã—7 columns

In [99]:

ST = ST.reset_index(drop=True)
In [100]:

ST
Out[100]:

	Open	High	Low	Close	Adj Close	Volume	Date
0	Nan	Nan	Nan	Nan	Nan	Nan	2016-07-01
1	-0.119995	-8.630005	-4.000000	-10.059998	-10.059998	-65300.0	2016-06-30
2	-1.450012	0.130005	1.000000	-0.020019	-0.020019	215000.0	2016-06-29
3	-15.750000	-11.500000	-15.369995	-7.649964	-7.649964	966900.0	2016-06-28
4	-7.989990	-11.179993	-16.049988	-16.590027	-16.590027	1531000.0	2016-06-27
...	...	...	...	...	...	...	...
1985	0.570000	-0.500000	-0.680000	-1.339997	-1.339997	307100.0	2008-08-13
1986	1.040001	0.230003	1.559997	0.559998	0.559998	817700.0	2008-08-12
1987	-7.140000	3.269997	-6.319999	0.839996	0.839996	17043700.0	2008-08-11
1988	-3.400001	-10.540001	-3.489998	-7.579994	-7.579994	-15907500.0	2008-08-08
1989	0.230003	-3.159996	-0.290001	-3.560005	-3.560005	-3717900.0	2008-08-07

1990 rows Ã—7 columns

Data Dropping¶

In [101]:

```
# Remove unneeded features
ST =ST.drop(['High','Low','Close','Volume','Adj Close'], 1)
```

In [102]:

```
ST
```

Out[102]:

	Open	Date
0	Nan	2016-07-01
1	-0.119995	2016-06-30
2	-1.450012	2016-06-29
3	-15.750000	2016-06-28
4	-7.989990	2016-06-27
...	...	...
1985	0.570000	2008-08-13
1986	1.040001	2008-08-12
1987	-7.140000	2008-08-11
1988	-3.400001	2008-08-08
1989	0.230003	2008-08-07

1990 rows Ã—2 columns

In [103]:

```
ST = ST[ST.Open.notnull()]
```

In [104]:

```
ST
```

Out[104]:

	Open	Date
1	-0.119995	2016-06-30
2	-1.450012	2016-06-29
3	-15.750000	2016-06-28
4	-7.989990	2016-06-27
5	0.989990	2016-06-24
...	...	...
1985	0.570000	2008-08-13
1986	1.040001	2008-08-12
1987	-7.140000	2008-08-11
1988	-3.400001	2008-08-08
1989	0.230003	2008-08-07

1989 rows Ã—2 columns

In [105]:

ST.isnull().sum()
Out[105]:

Open 0
Date 0
dtype: int64

In [106]:

List of headlines and corresponding stock prices

In [107]:

```
price = []
headlines = []

for I in ST.iterrows():
    daily_headlines = []
    date = I[1]['Date']
    price.append(I[1]['Open'])
    for j in HL[HL.Date==date].iterrows():
        daily_headlines.append(j[1]['News'])
    headlines.append(daily_headlines)
```

In [31]:

```
price
```

Out[31]:

```
[3.400001000000003,
 7.139999999999986,
 -1.0400009999999895,
 -0.5700000000000074,
 2.5700000000000074,
 -2.1900030000000044,
 -3.0,
 -1.0899959999999993,
 -0.5899959999999993,
 2.859993000000003,
 0.35999999999999943,
 -1.889999000000003,
 -1.299995999999993,
 0.7699969999999894,
 0.6900030000000044,
 0.26000200000000007,
 -1.760002,
 -0.8000040000000013,
 -3.260002,
 4.910004000000001,
 -1.489998,
 -0.75,
 -4.7400049999999965,
 3.5400009999999895,
 -1.9499969999999962,
 -0.05999799999999311,
 0.879997000000003,
 -4.599997999999999,
```

7.0400009999999895,
1.009994000000006,
-5.339995999999999,
-3.489998,
-1.639999000000003,
-1.140006999999997,
-1.109993000000003,
-2.5700080000000014,
5.940003000000004,
-4.150002000000001,
1.7900009999999895,
-5.359999999999999,
1.7300030000000106,
-10.150002000000008,
6.3300020000000075,
-8.979999,
5.969996999999992,
4.540001000000004,
-9.040001000000004,
-8.059996999999996,
2.4299999999999997,
3.3399959999999993,
-0.30999699999999564,
-1.8100020000000043,
-6.68,
1.380001,
4.25,
2.5,
4.029998999999997,
3.2299989999999994,
-2.75,
0.3400000000000034,
0.8100019999999972,
0.31000100000000685,
-7.670002000000004,
-2.0400010000000037,
2.220002000000001,
-2.8800019999999975,
-3.1099959999999953,
-2.5900000000000034,
2.2099990000000034,
-3.7000010000000074,
-0.17999999999999972,
-1.7599989999999934,
-2.6799999999999997,
1.0999979999999923,

2.4000020000000077,
3.2999989999999997,
-1.2200010000000034,
3.1599999999999966,
-2.0299989999999966,
-0.610001000000004,
-1.2299989999999994,
5.779999000000004,
-0.05000000000000426,
3.2600030000000046,
0.6599989999999991,
1.6499979999999965,
-2.129996999999996,
-1.8100019999999972,
3.1300019999999975,
-1.0300030000000007,
2.639998999999996,
0.740002000000004,
-1.4099999999999966,
0.009997999999995955,
-1.4799989999999994,
1.5400009999999966,
2.130001000000007,
-2.360001000000004,
-1.9200020000000038,
1.230004000000008,
0.6099959999999953,
4.3800019999999975,
-1.1800009999999972,
1.740002000000004,
-1.2999989999999997,
1.9299959999999956,
-2.7999989999999997,
-3.1600000000000037,
-0.8600009999999969,
-1.5399969999999996,
3.239998,
-1.0499989999999997,
-1.4399990000000003,
0.1099969999999999,
-0.519995999999999,
1.279997999999999,
-0.880001,
0.42000200000000376,
0.23999799999999283,
7.400002000000008,

1.2099989999999963,
2.2999989999999997,
2.510002,
-2.2299989999999994,
2.029997999999999,
3.3799979999999934,
-0.4099959999999925,
-3.0500040000000013,
0.15000200000000063,
0.7200010000000034,
-2.300003000000004,
0.6000020000000035,
0.5700000000000003,
-1.7700000000000031,
3.0900040000000004,
-2.1900029999999973,
2.930000999999997,
-0.7799990000000037,
-2.860005000000001,
2.6800010000000043,
-1.1899990000000003,
0.04999899999999968,
1.3200039999999973,
1.0799939999999992,
-2.999995999999996,
0.29000100000000373,
3.749995999999996,
2.1099999999999994,
1.2900010000000037,
-1.08000100000001,
-1.1899949999999961,
3.5999979999999994,
0.4300010000000043,
-1.0100029999999975,
0.9700019999999938,
3.4899970000000025,
-1.760002,
0.38000499999999704,
-1.8599999999999994,
-1.2099989999999963,
2.2099989999999963,
0.4099960000000067,
0.6099999999999994,
2.7900009999999895,
0.8400040000000075,
-0.29000100000000373,

-1.0400009999999895,
1.5800019999999932,
2.3299939999999992,
-1.8399959999999993,
-2.5,
-0.16999799999999254,
1.4499969999999962,
1.6600030000000032,
-1.1099999999999994,
0.8399959999999851,
3.160004000000015,
0.6999969999999962,
1.8499979999999994,
-1.1999969999999962,
0.3099979999999931,
-2.059997999999993,
-0.5500030000000038,
-0.11999500000000296,
0.75,
1.869995000000003,
-0.14999400000000662,
-3.0400009999999895,
-2.889999000000003,
1.8999939999999924,
-2.449996999999996,
-2.4599989999999963,
0.75,
-0.5800020000000075,
1.4700010000000105,
3.069999999999993,
-1.75,
-0.6500020000000006,
-1.0699989999999957,
3.4800029999999964,
-0.8100050000000039,
0.020004000000000133,
0.48999799999999993,
4.209998999999996,
0.8900000000000006,
2.1099999999999994,
0.8700030000000112,
0.45999899999999627,
0.18000000000000682,
0.6500020000000006,
-1.2799990000000037,
-1.2700040000000001,

-2.199996999999996,
0.8699949999999887,
-1.1999969999999962,
0.3099979999999931,
-0.6599959999999925,
0.26000200000000007,
-3.4400030000000044,
-0.9700009999999963,
1.0999979999999994,
2.8000040000000013,
1.9700009999999963,
-0.25,
0.7999949999999956,
-3.3099969999999956,
-2.6600040000000007,
0.08000200000000746,
-2.0700000000000074,
1.7900010000000037,
-0.7300029999999964,
0.5500029999999896,
3.4100030000000032,
1.5199970000000036,
1.4199979999999925,
1.3800050000000113,
0.4699939999999998,
2.25,
0.13000499999999704,
1.1899939999999987,
-2.239998,
-1.3599999999999994,
-2.4000020000000006,
0.630004999999997,
1.1699980000000068,
0.12000299999999697,
0.799995999999993,
0.9200050000000033,
-1.8400039999999933,
-1.33000100000001,
-0.20999999999999375,
0.9100040000000007,
-1.370002999999997,
-0.22000100000001055,
2.300003000000004,
-1.6099999999999994,
-2.949996999999996,
0.25,

-0.5,
2.0699989999999957,
1.7999960000000073,
0.3899989999999889,
-0.5099939999999918,
-0.5600060000000013,
-0.0899959999999993,
0.7699970000000036,
-2.849998999999997,
-1.1900020000000069,
-1.6999969999999962,
-0.6500020000000006,
-0.1200019999999995,
1.5900039999999933,
0.7399970000000025,
1.730004000000008,
2.1099999999999994,
-0.6300040000000138,
0.35999999999999943,
1.7400060000000082,
4.839995999999999,
-0.010002000000000066,
-1.049995999999993,
1.7699969999999894,
1.360000999999997,
-0.8199999999999932,
-0.5599979999999931,
-0.40000100000000316,
0.9199979999999925,
0.3000030000000038,
0.23999799999999993,
-2.449996999999996,
0.1999969999999962,
-0.9199979999999925,
2.1699979999999925,
3.300003000000004,
0.1999969999999962,
1.1699979999999925,
-2.339995999999985,
2.389998999999989,
-1.0900039999999933,
0.17000600000000077,
0.04999499999999557,
0.5600060000000013,
-0.6400069999999971,
-1.609993000000003,

17.389999000000003,
8.159995999999992,
3.7200010000000105,
-1.3600000000000136,
2.3300020000000075,
-1.9300010000000043,
-3.3099969999999956,
-0.9900060000000082,
1.3300020000000075,
-1.5400010000000037,
5.540001000000004,
4.110000999999997,
-0.3099979999999931,
4.279999000000004,
-1.1000060000000076,
1.180008000000015,
0.9599909999999738,
-0.7200009999999963,
-0.5,
-0.3600010000000111,
-2.7799909999999812,
3.2900010000000037,
2.520004,
-0.2600089999999966,
-3.0099950000000035,
1.8899989999999889,
4.75,
2.2099919999999997,
4.470000999999996,
-0.1999969999999962,
-5.4199979999999925,
-3.699996999999996,
0.3000030000000038,
-2.1900019999999927,
3.660002999999989,
-3.570007000000004,
-1.7400049999999965,
0.16999799999999254,
-1.569991999999985,
-1.4499970000000104,
2.569991999999999,
3.2799990000000037,
1.0400080000000003,
4.399993999999992,
0.5500030000000038,
1.5399929999999813,

-2.889998999999989,
-1.3099980000000073,
-0.8399959999999851,
-2.820007000000004,
1.1700130000000115,
-2.590011000000004,
-1.4499969999999962,
2.0599969999999814,
-3.6299899999999923,
-1.0900029999999958,
1.2399969999999882,
0.03999400000000719,
-2.869995000000003,
0.8199990000000099,
0.13000499999999704,
-1.6600040000000007,
-3.5,
-1.5400000000000063,
0.47000100000001055,
3.400001000000003,
5.340003999999993,
-6.590003999999993,
-4.389999000000003,
-1.6699980000000068,
1.5199960000000061,
-2.760002,
3.5,
-1.1800000000000068,
-0.1999969999999962,
-0.7900010000000037,
1.7799990000000037,
1.0699999999999932,
-2.989998,
-1.2300039999999939,
2.0700080000000014,
-0.5400010000000037,
0.6399990000000031,
-0.050003000000003794,
0.20999899999999627,
-0.2900009999999895,
0.8199999999999932,
6.310005000000004,
0.39000000000000057,
0.5800009999999958,
3.150002000000015,
-0.8300020000000075,

1.2899929999999813,
-0.47999499999997397,
1.339995999999985,
3.75,
-2.5,
-0.45999199999999973,
1.1699990000000184,
-1.3900000000000148,
2.6900030000000186,
-3.5100100000000225,
0.6900019999999927,
-2.25,
0.5,
5.7599950000000035,
0.47000099999999634,
0.3700100000000077,
0.25999500000000353,
-0.1999969999999962,
-2.949996999999996,
-1.6200100000000077,
4.730011000000019,
-1.25,
6.009993999999978,
-0.7200009999999963,
1.2299959999999999,
-0.8900000000000148,
4.210007000000019,
0.33000200000000746,
-2.5299990000000037,
1.4799959999999999,
1.339995999999985,
1.839997000000011,
-1.6299899999999923,
-2.180008000000015,
2.3500060000000076,
-2.9600070000000187,
-2.5,
1.3099980000000073,
-4.199996999999996,
-1.5800020000000075,
-7.619994999999989,
2.0,
-2.0299990000000037,
1.7599950000000035,
0.22000099999999634,
1.4600070000000187,

2.519988999999981,
-3.569991999999985,
-2.1199960000000146,
1.8999939999999924,
-4.629996999999989,
-2.870002999999997,
-4.7399970000000025,
4.6699979999999925,
-4.0299989999999895,
6.510002,
-0.07000000000000739,
1.0899969999999968,
-1.0999989999999968,
-0.9500039999999927,
2.2300029999999964,
0.08000200000000746,
-0.49000600000000816,
-3.8399959999999993,
-1.6900020000000069,
-0.3099979999999931,
1.389999000000003,
2.849998999999997,
-1.0400010000000037,
2.190002000000007,
1.3499989999999968,
-0.25999500000000353,
0.3099980000000073,
-4.139999000000003,
-0.5400010000000037,
-1.5,
-2.470001999999994,
0.7099989999999963,
-2.5899959999999993,
-7.679999999999993,
0.3199999999999932,
2.019995999999992,
-0.2699959999999919,
-0.8100060000000013,
5.180001000000004,
1.5300060000000002,
1.2599949999999893,
2.8800040000000138,
2.3399969999999968,
-2.9000020000000006,
1.1500020000000006,
-2.9000020000000006,

2.230003999999994,
0.010002000000000066,
-1.9100040000000007,
-12.77999899999999,
12.330001999999993,
0.16999800000000675,
-1.4300000000000068,
0.9899979999999999,
-2.4599989999999963,
3.620002999999997,
0.8499979999999994,
3.059997999999993,
3.7099990000000105,
-0.04999600000000726,
1.7400060000000224,
-0.5100100000000225,
0.15000900000001138,
-4.3400040000000075,
2.299995999999993,
-2.449996999999996,
3.7299950000000024,
1.9100040000000007,
-0.02999900000000366,
-2.020004,
0.6600040000000007,
-2.449996999999996,
-1.5600060000000013,
3.3000040000000013,
-1.6900030000000044,
0.5700000000000074,
-3.1800010000000043,
3.5100029999999975,
5.809996999999996,
4.720000999999996,
0.6699990000000184,
0.36999499999998875,
2.460005999999993,
0.3600010000000111,
3.320007000000004,
0.4299929999999961,
0.38000500000001125,
0.5199889999999812,
3.5,
-0.1999969999999962,
2.0599980000000073,
-0.9199990000000184,

1.3600010000000111,
4.229996,
4.790008,
-0.38000500000001125,
-0.8099969999999814,
0.9799959999999999,
-2.929992999999996,
-3.1300050000000113,
3.1300050000000113,
3.520004,
-4.090011000000004,
-1.3300020000000075,
-0.2899940000000072,
-2.3800039999999854,
4.180006999999989,
-1.5299979999999778,
3.2599939999999776,
6.5800020000000075,
-4.320007000000004,
-1.8999939999999924,
3.889998999999989,
-0.22000099999999634,
9.120010000000008,
-4.0,
1.339997000000011,
-0.6000060000000076,
-2.5099950000000035,
-1.3500060000000076,
-0.6999969999999962,
1.6499939999999924,
4.460007000000019,
-0.5099950000000035,
1.4899899999999775,
1.8300020000000075,
-2.0800020000000075,
0.4100040000000149,
-0.8800050000000113,
-4.959990999999974,
-6.419999000000018,
-0.9000090000000114,
2.9000090000000114,
3.2099919999999997,
1.1500090000000114,
3.5099950000000035,
2.9199979999999925,
5.8300020000000075,

2.6300039999999854,
-3.0400080000000003,
2.2100070000000187,
-2.300003000000004,
-1.3600010000000111,
0.020004000000000133,
4.979996,
-3.0099950000000035,
0.27999900000000366,
-2.889998999999989,
1.4499969999999962,
-2.050003000000004,
-0.5599980000000073,
1.8600010000000111,
2.8300020000000075,
0.8599999999999852,
4.610001000000011,
1.1199949999999887,
-0.6600040000000149,
-2.4400019999999927,
0.20001200000001518,
-0.3000030000000038,
2.1199949999999887,
-1.9599909999999738,
-0.59001200000003,
4.779999000000004,
-2.049987999999985,
2.3999939999999924,
1.3800050000000113,
-2.84001200000003,
0.38000500000001125,
-0.059996999999981426,
-1.7599950000000035,
1.8999939999999924,
3.160004000000015,
2.2399899999999775,
-5.610001000000011,
-2.2899929999999813,
-5.050003000000004,
-2.449996999999996,
2.0099950000000035,
-0.02999900000000366,
-6.029999000000004,
-1.2899929999999813,
0.3599999999999852,
0.8999939999999924,

2.0800020000000075,
0.5,
2.1499939999999924,
0.5100100000000225,
6.4899899999999775,
1.2400049999999965,
1.1699990000000184,
3.6900019999999927,
-0.5200040000000001,
1.0400080000000003,
-4.0,
1.7699889999999812,
-3.859999999999985,
-3.429992999999996,
-3.389998999999989,
2.089995999999985,
-5.039992999999981,
-0.38000500000001125,
-4.440003000000019,
4.620011000000034,
-1.09001200000003,
-0.6999969999999962,
-2.5299990000000037,
-2.7200009999999963,
0.4000090000000114,
-1.570007000000004,
1.1000060000000076,
-5.210007000000019,
3.3099980000000073,
1.2100070000000187,
-4.720002000000022,
2.179992999999996,
0.7000120000000152,
-1.7700040000000001,
5.910004000000015,
3.42999199999997,
0.1600040000000149,
-1.070007000000004,
7.050003000000004,
1.5299990000000037,
2.270004,
-0.6900030000000186,
1.2100070000000187,
4.049987999999985,
-3.3699949999999887,
2.479996,

-0.3999939999999924,
-1.8000030000000038,
-2.229996,
0.5599969999999814,
-0.3899989999999889,
-2.6199949999999887,
-0.02999900000000366,
3.269988999999981,
2.9400030000000186,
1.089995999999985,
0.6200100000000077,
-3.070007000000004,
12.760010000000023,
-1.5800020000000075,
2.1900019999999927,
4.429992999999996,
-2.75,
0.4100040000000149,
0.4400019999999927,
-0.7600100000000225,
3.6000060000000076,
1.179992999999996,
1.1000060000000076,
1.4799959999999999,
-5.160004000000015,
-8.719985999999977,
2.3099980000000073,
4.199996999999996,
-0.38000500000001125,
-2.389998999999989,
1.4400019999999927,
-3.429992999999996,
-2.3300020000000075,
3.519990000000007,
1.1800069999999891,
0.11999600000001465,
-3.7799990000000037,
-1.0500030000000038,
-3.2200009999999963,
-2.289994000000007,
1.7299959999999999,
2.2900080000000003,
-0.49000499999999647,
-2.4400019999999927,
2.180006999999989,
-0.9500120000000152,

-2.299987999999985,
0.7699900000000071,
-0.5499879999999848,
2.339995999999985,
5.660004000000015,
-4.460007000000019,
4.380005000000011,
0.6199949999999887,
7.4199979999999925,
0.75,
-1.8899989999999889,
4.770004,
3.2099919999999997,
3.3600000000000136,
2.970000999999968,
-0.7899930000000097,
2.4400019999999927,
-2.1000059999999507,
0.05999800000000732,
-1.1199950000000172,
-0.5,
-0.5500029999999754,
1.240004999999968,
6.279998999999975,
-3.3099980000000073,
-2.8800039999999854,
1.6300039999999854,
-0.5,
9.39999400000005,
-1.1199950000000172,
-1.9800109999999904,
3.7100069999999903,
-4.6799930000000245,
-7.350006000000008,
-6.240004999999968,
-2.0599980000000073,
-8.270004,
0.3000030000000038,
4.059998000000007,
-3.75,
3.270004,
1.7799990000000037,
-0.9199990000000184,
-2.609999999999985,
-7.319991999999985,
-10.920014000000037,

2.540009000000026,
-3.910004000000015,
14.970000999999996,
0.520005000000026,
-3.1699990000000184,
11.580002000000007,
2.9599919999999997,
6.4900049999999965,
3.009995000000032,
-6.339997000000039,
-4.1699979999999925,
14.029998999999975,
-0.5,
-3.25,
-6.300002999999975,
9.03999300000001,
2.430007999999958,
3.770004,
3.5800020000000075,
9.539994000000007,
3.6900019999999927,
-6.290007999999943,
-9.789994000000064,
-4.210005999999936,
6.970000999999968,
6.740004999999968,
-7.86999499999996,
7.819991999999985,
-15.979996,
-1.1800069999999607,
-7.390000000000043,
2.9100040000000433,
7.75,
2.199996999999968,
3.75,
4.370010000000008,
6.03999300000001,
0.3600009999999827,
3.869995000000017,
3.4199979999999925,
-1.9799950000000024,
-1.6400000000000148,
-8.53999300000001,
4.779999000000032,
-0.8900000000000148,
2.569991999999985,

-34.89999399999999,
0.5699930000000109,
2.270004,
9.259994000000006,
-7.679991999999999,
7.440001999999964,
0.75,
1.349991000000017,
-0.8099980000000073,
2.3600010000000395,
-4.25,
-1.4499970000000246,
-0.9799959999999999,
3.1299899999999923,
2.3500060000000076,
-1.7299959999999999,
-3.760008999999968,
-7.1799930000000245,
-12.040009000000026,
-6.339995999999985,
6.110001000000011,
-2.6499939999999924,
1.2399899999999775,
3.1300050000000113,
-0.020004000000000133,
-2.909988999999996,
5.220000999999996,
1.7899940000000072,
-2.8800050000000113,
-4.949996999999996,
2.5400080000000003,
-2.359999999999985,
-1.180008000000015,
-1.4700009999999963,
-9.559998000000007,
3.050003000000004,
0.36999499999998875,
-0.41999799999999254,
0.6900019999999927,
-0.7700040000000001,
-6.8300020000000075,
4.130005000000011,
-1.4900049999999965,
-1.339997000000011,
-6.770004,
3.7400060000000224,

2.5299979999999778,
3.32000800000003,
-3.270005000000026,
2.1300050000000113,
4.689988,
-1.659988999999996,
-1.4600070000000187,
-0.22000099999999634,
-3.609999999999985,
4.339995999999985,
1.7900080000000003,
8.940003000000019,
-0.16999799999999254,
0.07998599999996259,
-5.789992999999981,
1.9900049999999965,
2.3099980000000073,
3.7899929999999813,
0.589997000000011,
0.3200070000000039,
-20.190001999999993,
5.839995999999985,
3.180008000000015,
3.449996999999996,
-3.6300050000000113,
2.300003000000004,
-0.4499969999999962,
-1.0800020000000075,
3.75,
3.910004000000015,
0.20999099999997384,
-13.5,
2.300003000000004,
2.5599980000000073,
-0.6999969999999962,
-2.3099980000000073,
0.05999800000000732,
-2.160004000000015,
1.3600010000000111,
4.9900049999999965,
-4.0,
-0.410002999999989,
-0.47999599999999987,
-0.3200070000000039,
3.9700009999999963,
1.5200040000000001,

2.6199949999999887,
-2.65998799999997,
-0.2100070000000187,
-0.27000400000000013,
-1.6299899999999923,
1.2599950000000035,
0.16999799999999254,
1.430008000000015,
7.619994999999989,
-1.9600070000000187,
1.4700020000000222,
4.470000999999996,
7.109999999999985,
2.550003000000004,
-4.859999999999985,
3.7400049999999965,
-7.0,
0.22000099999999634,
-1.2900080000000003,
-3.3999939999999924,
-1.5299990000000037,
0.7299959999999999,
-3.1199949999999887,
-1.570007000000004,
1.839995999999985,
-0.8899989999999889,
-1.7999879999999848,
1.6099999999999852,
4.109985999999992,
-0.589997000000011,
-3.3499909999999886,
-0.3100120000000004,
2.9900049999999965,
1.9000090000000114,
31.259995000000004,
-0.8800049999999828,
5.449996999999968,
-1.579987000000017,
1.9199979999999925,
-1.9400019999999927,
-5.440001999999993,
0.7899930000000097,
-2.5599980000000073,
3.32000800000003,
2.0399930000000097,
-0.34999100000001704,

0.8999939999999924,
-1.5,
0.05000299999997537,
-5.6399989999999605,
-5.380004999999983,
4.279998999999975,
-3.599991000000017,
2.2699890000000096,
-1.9899910000000318,
-0.6900019999999927,
-2.160003999999958,
-2.6600030000000174,
-1.0399940000000072,
-1.0400080000000003,
6.450012000000015,
0.7799989999999752,
5.01998900000001,
-1.0399930000000097,
-1.1199960000000146,
0.15998900000002436,
-2.1299899999999923,
-0.2600089999999682,
0.029997999999977765,
1.9900060000000224,
5.979996,
1.25,
-0.6699990000000469,
-2.0099939999999776,
-1.529999000000032,
1.1499940000000493,
3.5599980000000073,
-1.0899970000000394,
0.7799990000000321,
4.600005999999951,
-0.16000399999995807,
-0.5200040000000001,
-2.2699890000000096,
-1.3500060000000076,
1.25,
-7.300002999999975,
-2.339996000000042,
-0.9799959999999999,
0.9700010000000248,
0.849991000000017,
-1.3000030000000322,
4.5800020000000075,

4.630004999999983,
-0.6400000000000148,
1.5500030000000322,
-4.270004,
-2.0,
5.25,
...]

In [108]:

headlines[0]
Out[108]:

['Jamaica proposes marijuana dispensers for tourists at airports following legalization: The kiosks and desks would give people a license to purchase up to 2 ounces of the drug to use during their stay',
"Stephen Hawking says pollution and 'stupidity' still biggest threats to mankind: we have certainly not become less greedy or less stupid in our treatment of the environment over the past decade",
'Boris Johnson says he will not run for Tory party leadership',
'Six gay men in Ivory Coast were abused and forced to flee their homes after they were pictured signing a condolence book for victims of the recent attack on a gay nightclub in Florida',
'Switzerland denies citizenship to Muslim immigrant girls who refused to swim with boys: report',
'Palestinian terrorist stabs Israeli teen girl to death in her bedroom',
'Puerto Rico will default on $1 billion of debt on Friday',
'Republic of Ireland fans to be awarded medal for sportsmanship by Paris mayor.',
"Afghan suicide bomber 'kills up to 40' - BBC News",
'US airstrikes kill at least 250 ISIS fighters in convoy outside Fallujah, official says',
'Turkish Cop Who Took Down Istanbul Gunman Hailed a Hero',
"Cannabis compounds could treat Alzheimer's by removing plaque-forming proteins from brain cells, research suggests",
"Japan's top court has approved blanket surveillance of the country's Muslims: 'They made us terrorist suspects, we never did anything wrong,' says Japanese Muslim, Mohammed Fujita",
'CIA Gave Romania Millions to Host Secret Prisons',
'Groups urge U.N. to suspend Saudi Arabia from rights council',
'Goggles free wifi at Indian railway stations is better than most of the countries paid services',
"Mounting evidence suggests 'hobbits' were wiped out by modern humans' ancestors 50,000 years ago.",
"The men who carried out Tuesday's terror attack at Istanbul's Ataturk Airport were from Russia, Uzbekistan and Kyrgyzstan, a Turkish official said.",
'Calls to suspend Saudi Arabia from UN Human Rights Council because of military aggression in Yemen',
'More Than 100 Nobel Laureates Call Out Greenpeace For Anti-GMO Obstruction In Developing World',

'British pedophile sentenced to 85 years in US for trafficking child abuse images: Domminich Shaw, a kingpin of sexual violence against children, sent dozens of images online and discussed plans to assault and kill a child while on probation',
'US permitted 1,200 offshore fracks in Gulf of Mexico between 2010 and 2014 and allowed 72 billion gallons of chemical discharge in 2014.',
'We will be swimming in ridicule - French beach police to carry guns while in swimming trunks: Police lifeguards on Frances busiest beaches will carry guns and bullet-proof vests for the first time this summer amid fears that terrorists could target holidaymakers.',
"UEFA says no minutes of silence for Istanbul victims at Euro 2016 because 'Turkey have already been eliminated'",
'Law Enforcement Sources: Gun Used in Paris Terrorist Attacks Came from Phoenix']

In [109]:

```
print(len(price))
print(len(headlines))
```

1989
1989

In [110]:

```
for I in range(0,10):
    print(headlines[I])
    print('\n')
```

['Jamaica proposes marijuana dispensers for tourists at airports following legalization: The kiosks and desks would give people a license to purchase up to 2 ounces of the drug to use during their stay', "Stephen Hawking says pollution and 'stupidity' still biggest threats to mankind: we have certainly not become less greedy or less stupid in our treatment of the environment over the past decade", 'Boris Johnson says he will not run for Tory party leadership', 'Six gay men in Ivory Coast were abused and forced to flee their homes after they were pictured signing a condolence book for victims of the recent attack on a gay nightclub in Florida', 'Switzerland denies citizenship to Muslim immigrant girls who refused to swim with boys: report', 'Palestinian terrorist stabs Israeli teen girl to death in her bedroom', 'Puerto Rico will default on $1 billion of debt on Friday', 'Republic of Ireland fans to be awarded medal for sportsmanship by Paris mayor.', "Afghan suicide bomber 'kills up to 40' - BBC News", 'US airstrikes kill at least 250 ISIS fighters in convoy outside Fallujah, official says', 'Turkish Cop Who Took Down Istanbul Gunman Hailed a Hero', "Cannabis compounds could treat Alzheimer's by removing plaque-forming proteins from brain cells, research suggests", "Japan's top court has approved blanket surveillance of the country's Muslims: 'They made us terrorist suspects, we never did anything wrong,' says Japanese Muslim, Mohammed Fujita", 'CIA Gave Romania Millions to Host Secret Prisons', 'Groups urge U.N. to suspend Saudi Arabia from rights council', 'Goggles free wifi at Indian railway stations is better than most of the countries paid services', "Mounting evidence suggests 'hobbits' were wiped out by modern humans' ancestors 50,000 years ago.", "The men who carried out Tuesday's terror attack at Istanbul's

Ataturk Airport were from Russia, Uzbekistan and Kyrgyzstan, a Turkish official said.", 'Calls to suspend Saudi Arabia from UN Human Rights Council because of military aggression in Yemen', 'More Than 100 Nobel Laureates Call Out Greenpeace For Anti-GMO Obstruction In Developing World', 'British pedophile sentenced to 85 years in US for trafficking child abuse images: Domminich Shaw, a kingpin of sexual violence against children, sent dozens of images online and discussed plans to assault and kill a child while on probation', 'US permitted 1,200 offshore fracks in Gulf of Mexico between 2010 and 2014 and allowed 72 billion gallons of chemical discharge in 2014.', 'We will be swimming in ridicule - French beach police to carry guns while in swimming trunks: Police lifeguards on Frances busiest beaches will carry guns and bullet-proof vests for the first time this summer amid fears that terrorists could target holidaymakers.', "UEFA says no minutes of silence for Istanbul victims at Euro 2016 because 'Turkey have already been eliminated'", 'Law Enforcement Sources: Gun Used in Paris Terrorist Attacks Came from Phoenix']

['Explosion At Airport In Istanbul', 'Yemeni former president: Terrorism is the offspring of Wahhabism of Al Saud regime', 'UK must accept freedom of movement to access EU Market', 'Devastated: scientists too late to captive breed mammal lost to climate change - Australian conservationists spent 5 months obtaining permissions & planning for a captive breeding program. But when they arrived on the rodents tiny island, they they were too late.', 'British Labor Party leader Jeremy Corbyn loses a no-confidence vote but refuses to resign', 'A Muslim Shop in the UK Was Just Firebombed While People Were Inside', 'Mexican Authorities Sexually Torture Women in Prison', 'UK shares and pound continue to recover', 'Iceland historian Johannesson wins presidential election', '99-Million-Yr-Old Bird Wings Found Encased in Amber - Finding things trapped in amber is far from rare. But when researchers in Burma found a pair of tiny bird-like wings frozen inside, they knew they had something special.', 'A Chabot programmed by a British teenager has successfully challenged 160,000 parking tickets since its launch last year.', 'The Philippine president-elect said Monday he would aggressively promote artificial birth control in the country even at the risk of getting in a fight with the dominant Catholic church, which staunchly opposes the use of contraceptives.', 'Former Belgian Prime Minister ridicules Nigel Farage and accuses Ukip leader of lying in EU referendum campaign', "Brexiteer Nigel Farage To EU: 'You're Not Laughing Now, Are You?'", 'Islamic State bombings in southern Yemen kill 38 people', 'Escape Tunnel, Dug by Hand, Is Found at Holocaust Massacre Site', 'The land under Beijing is sinking by as much as four inches per year because of the overconsumption of groundwater, according to new research.', 'Car bomb and Anti-Islamic attack on Mosque in Perth, Australia', 'Emaciated lions in Taiz Zoo are trapped in blood-soaked cages and left to starve for months due to the Yemeni civil war', "Rupert Murdoch describes Brexit as 'wonderful'. The media mogul likened leaving the EU to a prison break and shared his view of Donald Trump as a very able man.", 'More than 40 killed in Yemen suicide attacks', "Google Found Disastrous Symantec and Norton Vulnerabilities That Are 'As Bad As It Gets'", 'Extremist violence on the rise in Germany: Domestic intelligence agency says far-right, far-left and Islamist radical groups gaining membership in country', 'BBC News: Labor MPs pass Corbyn no-confidence motion', "Tiny New Zealand town with 'too many jobs' launches drive to recruit outsiders"]

['2,500 Scientists To Australia: If You Want To Save The Great Barrier Reef, Stop Supporting Coal', 'The personal details of 112,000 French police officers have been uploaded to Google Drive in a security breach just a fortnight after two officers were murdered at their home by a jihadist.', "Sample cuts United Kingdom sovereign credit rating to 'AA' from 'AAA'", 'Huge helium deposit found in Africa', 'CEO of the South African state broadcaster quits shortly after negative news about president is banned.', 'Brexit cost investors $2 trillion, the worst one day drop ever', 'Hong Kong democracy activists call for return to British rule as first step to independence', "Brexit: Iceland president says UK can join 'triangle' of non-EU countries", "UK's Osborne: 'Absolutely' going to have to cut spending, raise taxes", '"Do not let Scotland down now' : Scottish MEP Alyn Smith has urged members of the European Parliament to stand by his country following the UK referendum on EU membership.", 'British pound could hit history-making dollar parity by end of 2016', "Merkel vows to strengthen EU, tells UK no 'cherry-picking'", '"Ryanair will not deploy new aircraft on routes to and from the UK [United Kingdom] next year [2017], following the Brexit vote, and will instead focus on the European Union [EU]."', 'People, ever more greedy and stupid, destroy the world - Stephen Hawking to Larry King', 'Siemens freezes new UK wind power investment following Brexit vote', 'US, Canada and Mexico pledge 50% of power from clean energy by 2025', 'There is increasing evidence that Australia is torturing refugees, medical experts claim', 'Richard Branson, the founder of Virgin Group, said Tuesday that the company has lost about a third of its value since the U.K. voted to leave the European Union last week.', '37,000-yr-old skull from Borneo reveals surprise for scientists - Study of the "Deep Skull" - oldest modern human discovered in SE Asia - reveals this ancient person was not related to Indigenous Australians, as originally thought. "Our discovery is a game changer."', 'Palestinians stone Western Wall worshipers; police shut Temple Mount to non-Muslims', 'Jean-Claude Juncker asks Farage: Why are you here?', '"Romanians for Remainians" offering a new home to the 48% of Britons who voted to stay in the EU | Bucharest newspaper\'s app connects loving Romanian families with needy Brits, allowing people to offer to help would-be immigrants apply for a Romanian ID', 'Brexit: Gibraltar in talks with Scotland to stay in EU', '8 Suicide Bombers Strike Lebanon', "Mexico's security forces routinely use 'sexual torture' against women: Rights group Amnesty International has compiled testimonies of sexual violence used as torture by Mexican security forces. Despite thousands of complaints, only 15 probes have led to criminal convictions since 1991."]

['Barclays and RBS shares suspended from trading after tanking more than 8%', 'Pope says Church should ask forgiveness from gays for past treatment', "Poland 'shocked' by xenophobic abuse of Poles in UK", 'There will be no second referendum, cabinet agrees', 'Scotland welcome to join EU, Merkel ally says', "Sterling dips below Friday's 31-year low amid Brexit uncertainty", 'No negative news about South African President allowed on state broadcaster.', 'Surge in Hate Crimes in the U.K. Following U.K.s Brexit Vote', 'Weapons shipped into Jordan by the CIA and Saudi Arabia intended for Syrian rebels have been systematically stolen by Jordanian intelligence operatives and sold to arms merchants on the black market, according to American and Jordanian officials', 'Angela Merkel said the U.K. must file exit papers with the European Union before talks can begin', 'In a birth offering hope to a threatened species, an aquarium in Osaka, Japan, has succeeded in artificially breeding a southern rockhopper penguin for the first time in the world.', "Sky News Journalist Left Speechless As Leave MP Tells Him 'There Is No Plan'", 'Giant panda in Macau gives birth to twins', 'Get out now: EU leader tells

Britain it must invoke Article 50 on Tuesday', "Sea turtle 'beaten and left for dead' on beach by people taking selfies: Loggerhead sea turtle receiving treatment after it was beaten with sticks and stepped on in Lebanon", 'German lawyers to probe Erdogan over alleged war crimes', 'Boris Johnson says the UK will continue to "intensify" cooperation with the EU and tells his fellow Leave supporters they must accept the 52-48 referendum win was "not entirely overwhelming".', "Richard Branson is calling on the UK government to hold a second EU referendum to prevent 'irreversible damage' to the country.", "Turkey 'sorry for downing Russian jet'", 'Edward Snowden lawyer vows new push for pardon from Obama', 'Brexit opinion poll reveals majority don\'t want second EU referendum: "half (48%) of British adults say that they are happy with the result, with two in five (43%) saying they are unhappy with the outcome."', 'Conservative MP Leave Campaigner: "The leave campaign don\'t have a post-Brexit plan..."', 'Economists predict UK recession, further weakening of Pound following Brexit.', "New EU 'superstate plan by France, Germany: Creating a European superstate limiting the powers of individual members following Britains referendum decision to leave the EU", 'Pakistani clerics declare transgender marriages legal under Islamic law']

['David Cameron to Resign as PM After EU Referendum', 'BBC forecasts UK votes to Leave the European Union', 'Nicola Sturgeon says a second independence referendum for Scotland is "now highly likely"', "It's official. Britain votes to leave the European Union.", "World's Largest Tibetan Buddhist Institute Ordered to be Demolished by Chinese Government", 'Not a single place in Scotland voted to leave the EU...', 'Rich Getting Richer at the Expense of the Poor, Oxfam Warns', 'Spanish minister calls for Gibraltar to be returned to Spain on back of Brexit vote', 'British Pound drops nearly 5% in minutes following strong results for leave campaign in Newcastle', 'J.K. Rowling leads the charge for Scottish independence after UK votes for Brexit', "Buenos Aires zoo to close after 140 years: 'Captivity is degrading' - Mayor Horacio Rodrguez Larreta said that the zoos 2,500 animals will be gradually moved to nature reserves in Argentina which can provide a more suitable environment.", "Northern Ireland's Deputy First Minister calls for poll on united Ireland after Brexit", 'Polls close | Brexit polling day as it happened', "Brexit: Petition for second EU referendum so popular the government site's crashing | UK | News", 'North Korea printing massive amounts of fake Chinese currency, defectors say', 'Sinn Fein calls for a referendum on Irish reunification after Brexit', '$70 billion wiped off the Australian sharemarket as a result of Brexit.', "Nigel Farage disowns Vote Leave '350m for the NHS' pledge hours after result", 'Top EU leader: we want Britain out as soon as possible', 'Nigel Farage: 350 million pledge to fund the NHS was a \'big mistake\'.Captions such as "Let\'s give our NHS the 350 million the EU takes every week" and "We send the EU 350 million a week, let\'s fund our NHS instead" were stamped on Vote Leave campaign material.', 'Thousands of London banking job cuts to start next week', 'Google says there was a large spike in searches for Irish passport applications as news broke', 'EU referendum; Gibraltar backs Remain with 94% in first result of the night', "After Brexit, U.K. Residents Google 'What Is The EU?'", 'A Turkish man has been found guilty of insulting President Recep Tayyip Erdogan for depicting him as the Gollum character from the Lord of the Rings. A court in the south-western Antalya province gave Rifat Cetin a suspended one-year jail sentence and stripped him of parental custody rights.']

['Today The United Kingdom decides whether to remain in the European Union, or leave', "E-cigarettes should not be banned in public, medical experts warn: 'A ban on using e-cigarettes in public places could be damaging, as it may put off smokers from using e-cigarettes to help them quit,' says Rosanna O'Connor from Public Health England", 'Report: China is still harvesting organs from prisoners', 'Man opens fire at cinema complex in Germany, several people wounded report', 'Erdoan: Europe, you dont want us because were Muslim', 'Asian millionaires now control more wealth than those in North America, Europe and other regions', 'A Japanese porn industry association has apologised and promised reform, amid allegations women are being forced to perform sex acts on film.', 'University students are being warned when classes contains graphic or sensitive content, including sexual abuse, rape and transgenderism, to protect their mental health. Australian academics are issuing so-called "trigger warnings" for confronting material in classrooms.', "Afghan interpreters 'betrayed' by UK and US", 'Contagious cancer cells are spreading between different animals and even different species in the sea, according to new research which raises the prospect of the disease becoming infectious in humans.', '51 Killed in China by Powerful Tornado', 'Teacher Killings Ignite Calls for Revolution in Mexico: Police crackdown on rural educator protest spurs wide rebuke of government privatization and repression', 'Solar plane lands in Spain after three-day Atlantic crossing', 'Brexit supporters urged to take own pens to polling station amid fears of MI5 conspiracy', "Cities forge world's largest alliance to curb climate change | More than 7,100 cities in 119 countries formed the Global Covenant of Mayors for Climate and Energy, a network for helping exchange information on such goals as developing clean energy, organizers said.", "Colombia, FARC announce full ceasefire, 'last day of the war'", 'Gunmen kill Sufi devotional singer Amjad Sabri in Pakistan, Pakistani Taliban claim responsibility', 'India launches 20 satellites in single mission', 'F-16s to be manufactured soon in an assembly line in India', "Australia's gun laws stopped mass shootings and reduced homicides, study finds.", 'French cement company in Syria buys oil from ISIS Documents', "Pope to visit Armenia after irking Turkey with 'genocide' label", 'Merkel says NATO must be strengthened', 'China cracks down on online comments, click-bait stories, foreign TV content as Xi reshapes media landscape', 'The prime minister of India is set to get a brand new Air India One aircraft, will be more advanced than Air Force One.']

['German government agrees to ban fracking indefinitely', 'Teenage recruits were raped by staff and forced to rape each other as part of initiation practices in the Australian military going back to 1960, a public inquiry heard on Tuesday', 'Pakistan is selling nuclear materials to N Korea and China knows it, US sources say', 'Amazon jaguar shot dead at Olympic torch ceremony', "Mexican flags raised around Donald Trump's golf course - Ahead of Trump's visit to Scotland this week", 'EU smashes 2020 emissions target six years early | Greenhouse gas emissions in 2014 were 4.1% lower than the previous year and 24.4% below the 1990 baseline. The 2020 target is a 20% cut.', 'Police kill eight striking Mexican teachers as Activists Denounce Police Killings & Crackdowns on Teachers in Oaxaca', 'Pro-choice activists have delivered abortion pills to women in Northern Ireland using a drone.', 'A French football fan shoved an 18cm-long flare up his rectum in order to smuggle it into a Euro 2016 game. He then pulled the device out and set it off in the crowd, burning himself and two others in the process, before being charged over the incident by French police.', 'Indian space agency ISRO launches 20 satellites in a single launch, setting a new record for the agency', 'Japanese power company TEPCO admits it lied about meltdown after Fukushima', "Murdoch's News Corp buys 72 regional newspapers for

$37 million", '"Europe\'s growing army of robot workers could be classed as \'electronic persons\' and their owners liable to paying social security for them if the European Union adopts a draft plan to address the realities of a new industrial revolution."', 'Intel Fights Record $1.2 Billion Antitrust Fine at Top EU Court for bribing computer makers to not use AMD', 'Russian security service conducts raids on Church of Scientology Federal Security Service raid over a dozen locations in Moscow and St Petersburg as part of investigation into alleged illegal business dealings', "Turkish students and graduates from 370 schools issue statements against AKP government's Islamization of education, student protests continue", 'Japan Election Campaign Kicks off, Voting Age Lowered to 18', 'Moscow has signed an agreement with Los Angeles-based company Hyperloop One to explore building a futuristic, high-speed transportation system known as a Hyperloop in the Russian capital.', 'TEPCO admits meltdown cover-up - The president of Tokyo Electric Power Company has admitted the company concealed the reactor meltdowns at its Fukushima Daiichi nuclear plant immediately after the March 2011 earthquake and tsunami', 'Indian State Grants Jews Minority Status - India Real Time', 'Canadian Rescue Plane successfully reaches South Pole Research Outpost.', 'The Swedish parliament on Tuesday voted in favour of tough new asylum and residency laws.', 'French police teargas migrants trying to board trucks to the UK', 'Qawwali musician Amjad Sabri was shot dead in Pakistan. In 2014, the Islamabad Hight Court had issued a notice in a blasphemy case to two private TV channels. The show had mixed a mock wedding with a qawwali sung by Sabri related to religious figures, and was considered offensive.', 'N. Korea launches what appears to be Musudan mid-range missile from east coast']

['An Australian athlete who has competed in six Paralympic Games has been robbed at gunpoint in the Brazilian city of Rio de Janeiro. Liesl Tesch said a man brandishing a gun pushed her off her bicycle and stole it on Sunday.', 'Russian state television accidentally broadcasts evidence that Moscow uses cluster bombs in Syria', 'In 2015, 50 environmentalists were killed in Brazil more than a quarter of those killed worldwide and last year marked deadliest year for environmental activists, with 185 total murders across world. Out of the 10 deadliest countries in world for environmentalist, seven were in Latin America.', "China's plan to cut meat consumption by 50% cheered by climate campaigners: New dietary guidelines could reduce greenhouse gas emissions by 1bn tonnes by 2030, and could lessen countries problems with obesity and diabetes", "Coral bleaching event now biggest in history and about to get worse: 'Coral in every major reef region has already experienced severe bleaching. About 93% of the reefs on Australias Great Barrier Reef have been affected, and almost a quarter of the reef on the 2,300km stretch is now dead.'", 'Super-rich quaff champagne in Venezuela country club while middle classes scavenge for food', 'Hong Kong bookseller refuses to be silenced after harsh detention in China', 'Chinese prosecutors have successfully sued a county environmental agency for inadequately punishing a sewage firm that produced dye without appropriate safeguards the first such public interest case against a government department.', 'A London-based advocacy group says it has documented the killings of 185 environmental activists in 16 countries last year, nearly 60 percent more than in 2014.', 'Erdogan loses appeal against German media boss', "Chinese supercomputer is the world's fastest and without using US chips", '7-Eleven operator handed record penalty of more than $400,000 for systematically exploiting staff', 'A Honduran military unit trained by US was ordered to kill environmental activist who was slain in March', "African Union plans to introduce single passport to create EU-style 'continent without borders'",

'More refugees became citizens of Canada than any other country last year, UN says', "Turkey charges Reporters Without Borders press freedom representative with 'terrorist propaganda'", 'Brussels: Bomb alert at shopping centre sparks anti-terror operation | Europe | News', 'Australian Paralympian Liesl Tesch robbed at gunpoint in Rio de Janeiro', "China issues orders to demolish Buddhist 'towns' (religious communities) with intensive 'instruction' for leaders and members of religious orders. Only a tiny fraction of the existing communities will be allowed to remain at each site.", 'Gazans squeezed by triple taxes as Hamas replaces lost income', "US and Russian fighters in dramatic showdown over Syria: American pilots scramble to confront Putin's jets as they bomb Pentagon-backed Syrian rebels", "Rising Tide of 'Politically Acceptable' Killings Spells Danger for Environmentalists Worldwide: More than three people were slain each week in 2015 for 'protecting the land, forests, and rivers,' a new report reveals", 'Mexico teachers protest: Six people are dead and more than 100 are injured following a weekend of violence between members of a teachers union and police in southern Mexico.', 'Canada is set to launch a paid whistleblower program.', 'Russian football fan leader Alexander Shprygin has been detained in France, two days after being expelled from the country, French officials say.']

['A staggering 87 percent of Venezuelans say they do not have money to buy enough food', 'Two corporate whistleblowers may enter into plea bargain deal that would tie 30% of Brazilian lawmakers to corruption cases.', 'Poland, together with Russia, Iran, and several Gulf states successfully removed decriminalization of homosexuality from UN resolution.', 'Three environmental activists were killed per week last year, murdered defending land rights and the environment from mining, dam projects and logging', 'Ontario funeral business dissolves the dead, pours them into town sewers', 'New Declassified Documents Reveal How CIA Abused, Tortured Prisoners [Graphic]', "Tens of thousands of people gathered in sweltering heat on Japan's Okinawa island on Sunday in one of the biggest demonstrations in two decades against U.S. military bases, following the arrest of an American suspected of murdering a local woman.", "Japan's dementia crisis hits record levels as thousands go missing. National Police Agency reports more than 12,000 patients going missing in 2015, with hundreds of those later found dead", 'Icelands Hekla volcano, a very popular tourist destination, now ready to blow', 'Corbyn pledges to kill TTIP if elected', 'Venezuelans Ransack Stores as Hunger Grips the Nation - 4-year-old girl was shot to death as street gangs fought over food.', 'Rome elects first female mayor', "Saudi Arabia and Kuwait angry about Hillary Clintons claims that they 'fund terrorism': The two embassies in Canberra, Australia denounced the presidential candidate's remarks and said they also suffer from terrorism", "Professor Dismissed for Insulting Turkey's President", 'Russian soldier dies in Syria after preventing car bomb attack on aid distribution point', 'Three dead, 45 injured as labor union clashes with police in Mexico', 'Indonesia vows to stand firm after skirmishes with Chinese ships.China claims most of the South China Sea, through which $5 trillion in ship-borne trade passes every year. The Philippines, Vietnam, Malaysia, Taiwan and Brunei have overlapping claims.', 'A study of ocean plankton has shown that an increase in the water temperature of the worlds oceans of around 6C (10.8F) which some scientists predict could occur as soon as 2100 could stop oxygen production by phytoplankton by disrupting the process of photosynthesis.', 'Australia taxes foreign home buyers as affordability bites - Sydney is imposing new taxes on foreigners buying homes as concerns grow that a flood of mostly Chinese investors is crowding out locals and killing the "Great Australian Dream" of owning property', "Paris isn't happy about Amazon's one-hour delivery service",

'Australian man pleads guilty to making sexual threats on social media, in a landmark victory for opponents of online harassment. "When her friends defended her online, Alchin wrote more than fifty posts, including rape threats, and saying that women should \'never have been given rights\' "', 'Trudeau condemns killing of 14 Canadian Embassy security guards in Kabul', 'Vladimir Putin is considering selling part of Russias corporate crown jewels to China and India as the president struggles to meet spending commitments before his possible re-election bid in less than two years', 'An elephant has survived being shot in the head by suspected poachers in Zimbabwe. It was found by vets in Mana Pools national park and is believed to have had the bullet lodged in its head for up to six weeks.', 'Wikileaks founder Julian Assange marks 5 years holed up in Ecuadorean embassy']

['MP Jo Cox dead after shooting attack', 'Saudi Arabia upset after Hillary Clinton links oil kingdom to terrorism', 'India may be building an underwater wall of microphones to keep track of Chinas submarines, Hydrophones can record and listen to underwater sounds', 'Ex-Auschwitz guard Reinhold Hanning, 94, sentenced to 5 years for being accessory to murder in 170,000 deaths.', 'The annual Gay Pride Parade in Istanbul is under threat from ultra-nationalist and conservative groups who have pledged to do "what is necessary" to stop the event.', 'Russian athletes to remain banned from international competition', 'Israel seeks life plus 60 years for gay pride stabber', "Brazil's tourism minister resigns weeks before Olympics", 'A rare, risky mission is underway to rescue sick scientists from the South Pole', "China behaving like 'gangster' state with bookseller kidnap, say Hong Kong politicians", 'Putin: Today USA is the only superpower.', 'Report shows young people today in Australia are overqualified, underemployed and swimming in debt', 'Alleged killer of British MP was a longtime supporter of the neo-Nazi National Alliance', "Taliban use 'honey trap' boys to kill Afghan police - The Taliban are using child sex slaves to mount crippling insider attacks on police in southern Afghanistan", 'COGAT: Israel water supply to Palestinians increased, not decreased', 'Deaths, arrests as looting erupts in Venezuela', "Elderly Japanese, among the world's richest retirees, are flocking to inheritance advisers, tackling historical taboos on discussing death and providing a rare avenue of growth for the country's brokerages and banks.", 'Russian hooligans attack Spanish tourists outside cathedral - Independent.ie', 'Boko Haram shoot dead 18 women at funeral in northern Nigeria', 'ISIS Committed Genocide Against Yazidis in Syria and Iraq, U.N. Panel Says', 'Greece wants to send thousands of migrants back to Turkey in coming weeks', 'Missing Hong Kong bookseller: I was kidnapped - One of five Hong Kong booksellers who went missing in late 2015 said he was kidnapped after crossing the border into mainland China', "Kim Jong-un destroys model village in North Korea built by his 'despicable human scum' uncle: Sources claim the theme park made the North Korean leader uneasy", 'Turkey Bans Gay Pride Parade in Istanbul, Citing Security', 'One dead, others injured in mass stabbing at Calgary medical clinic.']

In [111]:

#max and min number of headlines for each day

In [112]:

```
print(max(len(I) for I in headlines))
print(min(len(I) for I in headlines))
```

25
22

In [113]:

```
## http://stackoverflow.com/questions/19790188/expanding-english-language-contractions-in-python
```

In [114]:

```
# A list of contractions from http://stackoverflow.com/questions/19790188/expanding-english-language-contractions-in-python
contractions = {
"ain't": "am not",
"aren't": "are not",
"can't": "cannot",
"can't've": "cannot have",
"'cause": "because",
"could've": "could have",
"couldn't": "could not",
"couldn't've": "could not have",
"didn't": "did not",
"doesn't": "does not",
"don't": "do not",
"hadn't": "had not",
"hadn't've": "had not have",
"hasn't": "has not",
"haven't": "have not",
"he'd": "he would",
"he'd've": "he would have",
"he'll": "he will",
"he's": "he is",
"how'd": "how did",
"how'll": "how will",
"how's": "how is",
"i'd": "I would",
"i'll": "I will",
"i'm": "I am",
"i've": "I have",
"isn't": "is not",
"it'd": "it would",
```

"it'll": "it will",
"it's": "it is",
"let's": "let us",
"ma'am": "madam",
"mayn't": "may not",
"might've": "might have",
"mightn't": "might not",
"must've": "must have",
"mustn't": "must not",
"needn't": "need not",
"oughtn't": "ought not",
"shan't": "shall not",
"sha'n't": "shall not",
"she'd": "she would",
"she'll": "she will",
"she's": "she is",
"should've": "should have",
"shouldn't": "should not",
"that'd": "that would",
"that's": "that is",
"there'd": "there had",
"there's": "there is",
"they'd": "they would",
"they'll": "they will",
"they're": "they are",
"they've": "they have",
"wasn't": "was not",
"we'd": "we would",
"we'll": "we will",
"we're": "we are",
"we've": "we have",
"weren't": "were not",
"what'll": "what will",
"what're": "what are",
"what's": "what is",
"what've": "what have",
"where'd": "where did",
"where's": "where is",
"who'll": "who will",
"who's": "who is",
"won't": "will not",
"wouldn't": "would not",
"you'd": "you would",
"you'll": "you will",
"you're": "you are"
}

In [115]:

```
def clean(text, remove_stopwords = True):


    # Convert words to lower case
    text = text.lower()

    # Replace contractions with actual words
    if True:
        text = text.split()
        new_text = []
        for word in text:
            if word in contractions:
                new_text.append(contractions[word])
            else:
                new_text.append(word)
        text = " ".join(new_text)

    # remove unwanted characters
    text = re.sub(r'&amp;amp;', '', text)
    text = re.sub(r'0,0', '00', text)
    text = re.sub(r'[_"\-;%()|.,+&=*%.,!?:#@\[\]]', ' ', text)
    text = re.sub(r'\'', ' ', text)
    text = re.sub(r'\$', ' $ ', text)
    text = re.sub(r'u s ', ' united states ', text)
    text = re.sub(r'u n ', ' united nations ', text)
    text = re.sub(r'u k ', ' united kingdom ', text)
    text = re.sub(r'j k ', ' jk ', text)
    text = re.sub(r' s ', ' ', text)
    text = re.sub(r' yr ', ' year ', text)
    text = re.sub(r' l g b t ', ' lgbt ', text)
    text = re.sub(r'0km ', '0 km ', text)

    # Optionally, remove stop words
    if remove_stopwords:
        text = text.split()
        stops = set(stopwords.words("english"))
        text = [w for w in text if not w in stops]
        text = " ".join(text)

    return text
```

In [116]:

```
clean_headlines = []

for daily_headlines in headlines:
    clean_daily_headlines = []
    for headline in daily_headlines:
        clean_daily_headlines.append(clean(headline))
    clean_headlines.append(clean_daily_headlines)
```

In [117]:

```
import re
from nltk.corpus import stopwords
```

In [118]:

```
#HEADLINES ARE CLEANED AND FREE FROM UNWANTED CHARACTERS
```

In [119]:

```
clean_headlines [0]
```

Out [119]:

```
['Jamaica proposes marijuana dispensers tourists airports following legalization kiosks desks would give people license purchase 2 ounces drug use stay',
 'Stephen hawking says pollution stupidity still biggest threats mankind certainly become less greedy less stupid treatment environment past decade',
 'Boris Johnson says run Tory party leadership',
 'Six gay men Ivory Coast abused forced flee homes pictured signing condolence book victims recent attack gay nightclub Florida',
 'Switzerland denies citizenship Muslim immigrant girls refused swim boys report',
 'Palestinian terrorist stabs Israeli teen girl death bedroom',
 'Puerto rice default $ 1 billion debt Friday',
 'Republic Ireland fans awarded medal sportsmanship Paris mayor',
 'Afghan suicide bomber kills 40 bib news',
 'We airstrikes kill least 250 Isis fighters convoy outside Fallujah official says',
 'Turkish cop took Istanbul gunman hailed hero',
 'Cannabis compounds could treat Alzheimer removing plaque forming proteins brain cells research suggests',
 'Japan top court approved blanket surveillance country Muslims made us terrorist suspects never anything wrong says Japanese Muslim Mohammed fajita',
 'CIA gave Romania millions host secret prisons',
 'Groups urge united nations suspend Saudi Arabia rights council',
```

'Goggles free wifi Indian railway stations better countries paid services',
'Mounting evidence suggests hobbits wiped modern human’s ancestors 50000 years ago',
'Men carried Tuesday terror attack Istanbul Ataturk airport Russia Uzbekistan Kyrgyzstan Turkish official said',
'Calls suspend Saudi Arabia un human rights council military aggression Yemen',
'100 Nobel laureates call Greenpeace anti gmo obstruction developing worBritishbritish pedophile sentenced 85 years us trafficking child abuse images domminich Shaw kingpin sexual violence children sent dozens images online discussed plans assault kill child probation',
'us permitted 1 200 offshore fracks gulf Mexico 2010 2014 allowed 72 billion gallons chemical discharge 2014',
'swimming ridicule French beach police carry guns swimming trunks police lifeguards frances busiest beaches carry guns bullet proof vests first time summer amid fears terrorists could target holidaymakers',
'Uefa says minutes silence Istanbul victims euro 2016 turkey already eliminated',
'law enforcement sources gun used Paris terrorist attacks came phoenix']

: