**Introduction:**

Sometimes your parents hide facts about your family but your neighbors know and speak about that. One way to know about the truth is to listen to them not to one but many of them. Then you can learn the truth by majority (though not always majority tells it)

It goes without saying that how a government manages its international behavior affects people life. However, in most cases, governments not only show no tendency to reveal this fact but also sometimes they misinform people. On the other hand, in countries like Iran there is no independent organization or media to enlighten the society about what is going on from international prospective. Some people achieve knowledge about this fact through browsing websites, being informed by others, and so on yet the knowledge is partial and occasionally biased.

Fortunately, there exists a large amount of text in book, magazine, and website format. However, one cannot process all of these data to come up with an idea about any issue yet machine can. I am addressing the above mentioned issue making use of data science. In this work I stat with country Kuwait for several reasons:

1. The country is really small country so the media including magazines, attitude is a perfect reflection of both its people and government.
2. It’s an Arabic country and the procedure can be applied to other Arab countries easily.
3. Kuwait attitude is not trivial as negatively Saudi Arabia and positively Iraq are. Therefore, we may achieve surprising results after the analysis.
4. **Phase 1: Data gathering, cleansing, and Visualization**

**Cleansing file names:**

In the Almojtama magazine archive files are named chronologically and each and categorized according to the published year. Based on these facts one can find the approximate time of the year which a given magazine has been published. Therefore, the file names contain information too. However, the file names are consisting of an Arabic term meaning number and in some cases some meaning subtitle. Furthermore, sometimes a file repeated twice.

In order to get rid of the above mention problems, I renamed all the file names programmically.

After year 2010 the pdf files are encrypted in such a way one can read them with pdf reader but they are split by machine, there is no output. To tackle this problem, I had to copy and paste whole pdf file and then split the txt file into several txt file.

**Splitting Pdf files into PNG files**

Most of the magazine files are pdfs of images. Therefore, in order to extract their text, we need to both split each file into a number of image files. On the other hand, this splitting will be useful when we want to count the number of articles in the magazine. To do so, I made use of PythonMagick and PyPDF2 libraries. After converting the pdf files we apply OCR‌ on the resulting PNG‌ to extract their text.

**Doing OCR:**

In order to extract pages text, I have used pytesseract module with Arabic tessdata. The accuracy is about 90 percent but I think it will work well for our purpose. There is problem needed to be addressed and it is the trade of between accuracy and speed of the ocr process. There were two options;

1. Ocr pages with low resolution but faster result
2. Ocr pages with higher dpi with slower operation

I decided to go with the first option; every page’s text ocr consumes about 16 seconds with a single thread. However, I used multi-threading (with five threads) which boosted the performance 3 times.

**Presentation:**

At first glance, one may not see a way to do sentiment analysis on this data. But domain knowledge will help here. Arab people are used to call Iranians as “Al-Ajam” (means mute) or “Al-Majoos” (means Zoroastrian) when they are angry about Iran; this is great because we found a feature to detect negative sentiment. So we can count these words. They also use “Aljomhoori Irani” (means Iranian republic), “Al-fars” (means Persia), and Iran when they refer to Iran. I first identify all the pages where any of these items are mentioned. Then adding up negative words (i.e. “Al-Ajam” , “Al-Majoos”) and scale them by multiplying the average number of pages in a year and dividing by the year pages; we do so because the yearly pages of the magazine have a large variation. However we can find the Iran-Importance factor too.

Therefore, I do the same scaling method for remaining words (i.e. “Aljomhoori Irani”, “Al-fars”, and “Iran”) to extract the Iran importance for Kuwait country.

After these modifications we are ready for Visualization part. I have shown some aspects of the resulted data in histogram, series plot, and scatter plot which can be viewed at git-hub: <https://github.com/afshinamiree/Kuwai_Iran_Temporal_Relation-through-Text-Mining/blob/master/Almojtama%20Plots.ipynb>

The results are informative and explain many situations in Iran-Kuwait relation and their corresponding sentiment toward Iran.

**Plot analysis:**

In the histogram we see that in some cases there is a little attention toward Iran from Kuwait. On the other hand, there is a focus on Iran by Kuwait. We also see in the time series plot that the attention toward Iran grows as the time passes. In the scatter plot I have used the negative-sentiment-indicator words to achieve the polarity of Kuwait attention toward Iran. These are points that can be mentioned:

a) Before Iran’s revolution, the attitude toward Iran was positive and then after the revolution it gets negative indicating that Kuwait has had some problem with the revolution from start.

b) Kuwait was an ally of Iraq. As can be seen, the end of war was not change the Kuwait attitude but the Invasion of Kuwait by Iraq was.

c) After Khatami presidency, the attitude gets more positive which shows he really did well in foreign policy.

d) After Syria war again the attitude again headed to negative side which is logical because Iran backed Assad’s regime and Kuwait and Saudi Arabia have serious problem with it.

1. **Phase 2: Providing labeled Data, Modelling**

We have achieved good results by now, but it will be more accurate to use some part of the data to label it and use it for supervised learning. There are some Arabic lexicon and corpus online but they are all prepared for short messages of social media. However, if we label around 2 to 5 percent of the magazine texts, not only we improve the results of the analysis but also the whole magazine texts can serve as a good Arabic corpus in social and political text analysis.

1. **Comparing the result with other countries changes**

We won’t persuade ourselves just with Kuwait country when we can repeat the same procedure for other countries. Therefore, I will do the same procedure for other Kuwait magazines and other countries. In this way, by comparing other magazines and number of their audience we can find the whole country attitude toward other countries (in this example Iran).

Another great thing is, by comparing the time series resulted from different countries we can learn much about them. For example using correlation coefficient we will found which country changes its attitude first and that would be the leading country.

In fact, when such data are provided, it can be used in many contexts that we are not able to think about.