DS5999 Final Project Report

Jingnan Yang – jy4fch

Introduction

The objective of this project is to keep track of the attention shift of digital media towards the political affairs in Middle East. Specifically, the project tries to investigate the trends of news coverage by New York Times across an eight-year time frame from 2011 to 2018. The data of this project includes 11772 news articles under the Middle East category of nytimes.com, 11665 of which are valid to use. The methods to investigate the news content include TFIDF, PCA, and topic modeling.

Text Representation

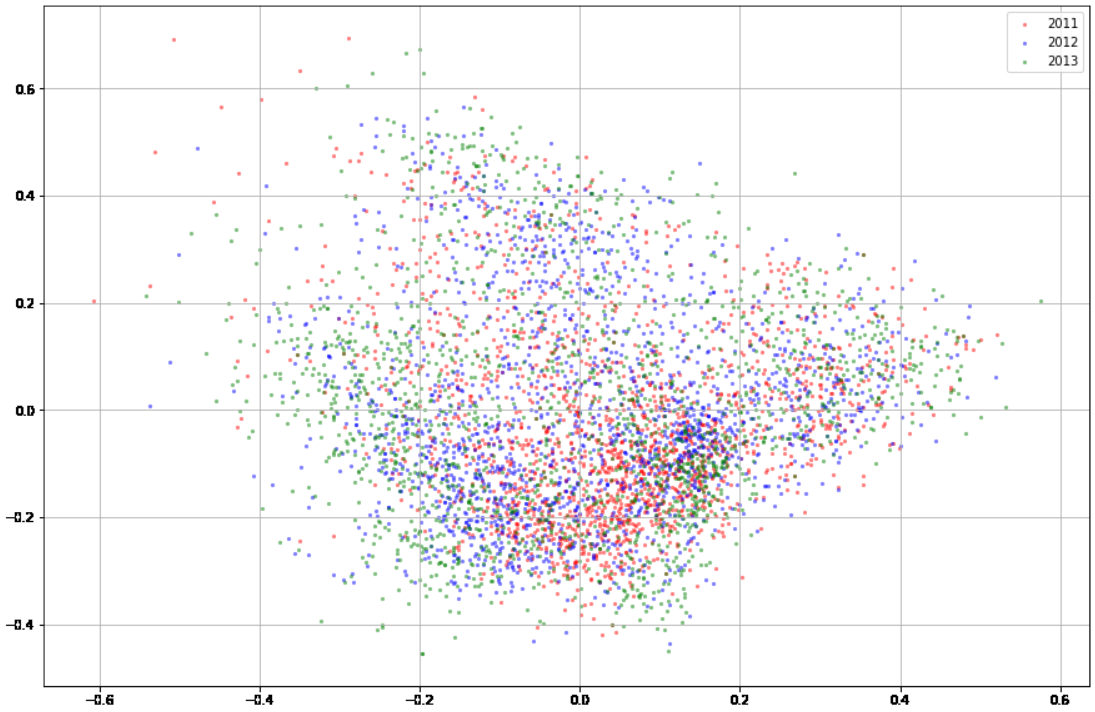
The first step of this project is to preprocess the 11665 news articles into machine learning formats. The preprocessing produces two tables, the tokens table, and the vocabulary table. The tokens table divides each news document into individual tokens, and marks the properties of each token, such as its part-of-speech tag. The vocabulary table is made of all unique tokens written in the 11665 news documents. Each token in the vocabulary table is assigned a term id to connect to the tokens table. The stems, term frequencies, mean TFIDF scores, and several other features are also calculated for each token in the vocabulary table.

Principle Component Analysis

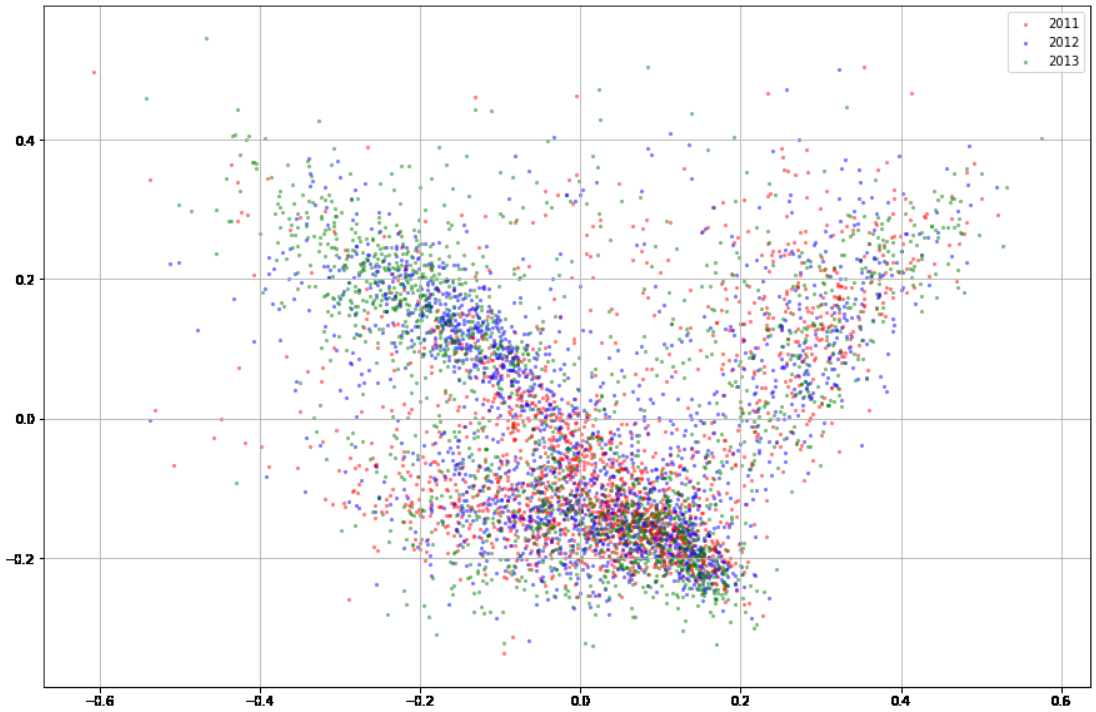
The PCA method reduces the dimensions of the corpus, so that the differences between documents can be observed. This project employs the 5000 tokens with the highest

mean TFIDF scores to generate 10 principal components for each news document, because tokens with higher TFIDF scores are more significant in differentiating document contents.

The 11665 news documents are divided into 8 groups by years. Every time 2 principal components of news documents in 3 consecutive years are compared to find if there exists any significant shift of news content across the consecutive years. For example, the PC1 and PC2 of documents in 2011, 2012, and 2013 are compared first, followed by PC1 and PC2 of documents in 2012, 2013, and 2014, and so on. Then, the PC1/PC3 and PC2/PC3 pairs are also compared for each 3 consecutive years.



As the scatter plot above shows, the PC1/PC2 distribution of 2011, 2012, and 2013 are almost the same. The PC1/PC2 distributions of other 3 consecutive years are also similar. It means the first and second principal components cannot show any shift in news coverage across consecutive years.



However, the PC1/PC3 distribution like the scatter plot above show 2011 has significantly different cluster from 2012 and 2013. The principle component analysis can tell if there exists a shift in document content, but cannot specify its reason. The topic modeling method is needed to classify the content of each document.

Topic Modeling

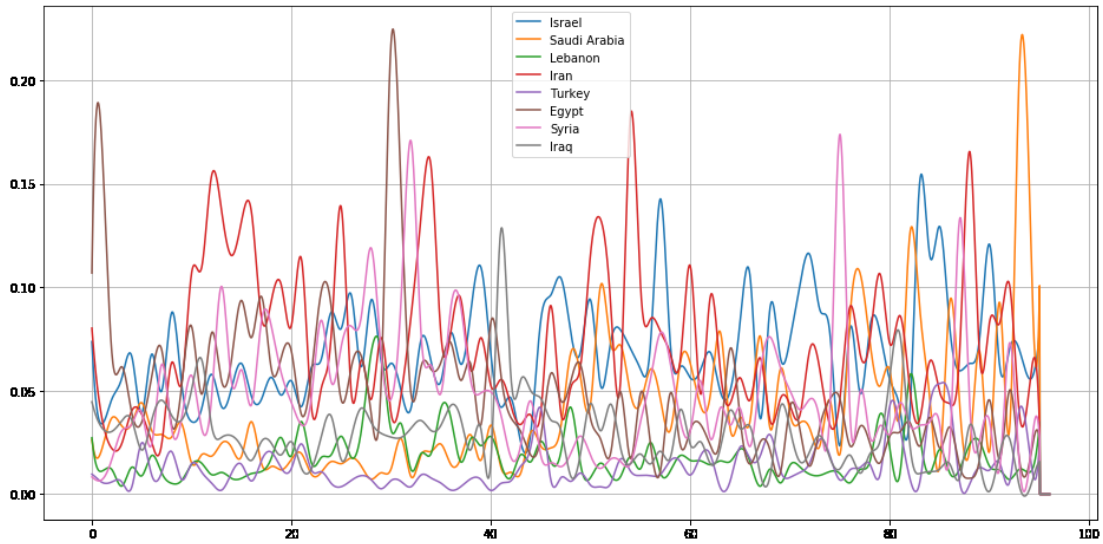
The topics modeling is based on the assumption that each document has a distribution over all topics, and each topic has a distribution over all tokens. The tokens selected for each document can be reversely engineered to generate the topic distribution of that document. The topic modeling identifies 20 most frequent topics across all new documents, each of which has 10 most used tokens as the table below shows.

ID	Topic	High-Frequency Tokens
0	America	american states officials united military would Obama administration intelligence official
1	Court	court news case prison family media two charges video authorities

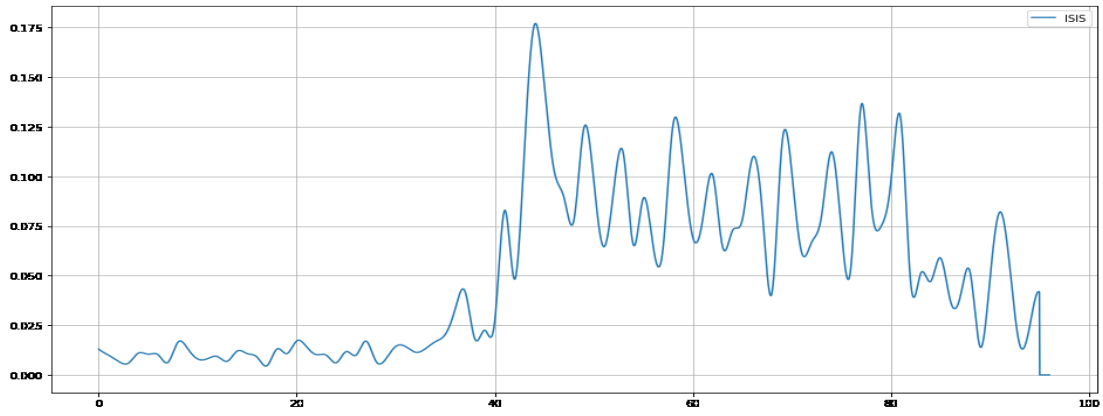
2	Israel	israel israeli palestinian netanyahu palestinians Jerusalem bank west jewish israelis
3	Military	killed attack military attacks security people two forces soldiers officials
4	Saudi	saudi arabia yemen arab yemeni united coalition country qaeda led
5	President	obama would kerry trump president united states american administration deal
6	Lebanon	hezbollah syria lebanon lebanese syrian jordan refugees war border shiite
7	Iran	iran iranian nuclear sanctions united tehran states would irans program
8	Protest	protesters police protests people security government forces friday activists violence
9	Turkey	turkey kurdish turkish border syrian syria region united states officials
10	Egypt	egypt egyptian military brotherhood mubarak morsi president cairo muslim government
11	ISIS	islamic state group isis fighters syria iraq militants forces also
12	Rights	rights human women report group groups law international government activists
13	Syria	syria syrian assad weapons government russia chemical opposition russian would
14	UN	united nations council security states aid international general people countries
15	People	one people many like years year even family would two
16	Rebel	government syrian rebel rebels syria aleppo fighters city assad damascus
17	Gaza	gaza hamas israel israeli palestinian military fire palestinians border west
18	Politics	government political minister would new party president prime parliament power
19	Iraq	iraq iraqi shiite sunni baghdad government american forces troops security

The 20 topics have covered the major nations and organizations involved in Middle East affairs and the major issues in that area such as military conflicts and human rights. The next step is to visualize the distribution or weights of the topics for each documents or each time period. This project generates three tables, the doc-topic table for topic

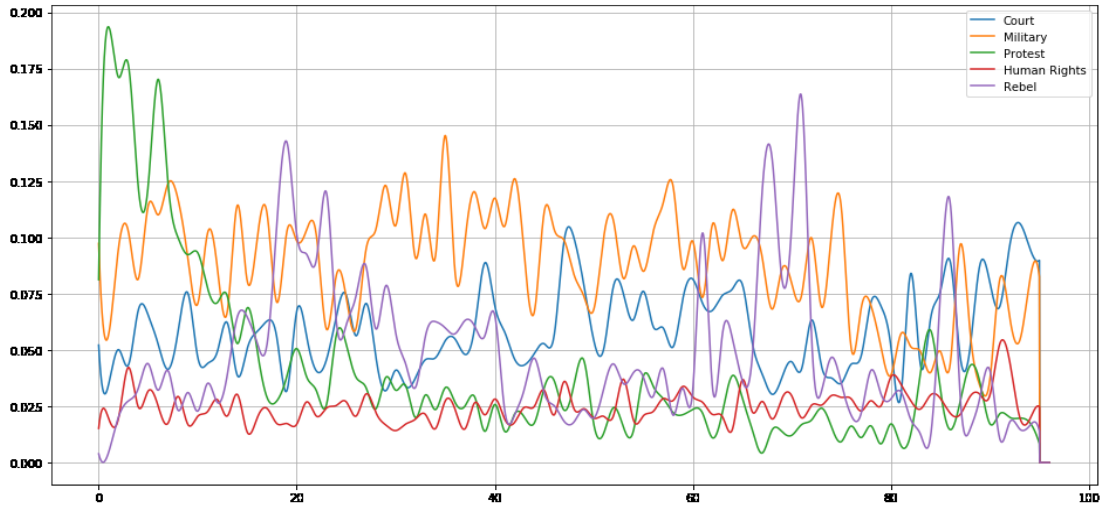
distribution of each document, the date-topic table for average topic distribution of all documents in each day, and the month-topic table for average topic distribution of all documents in each month. The month-topics is smoothed and visualized as follows.



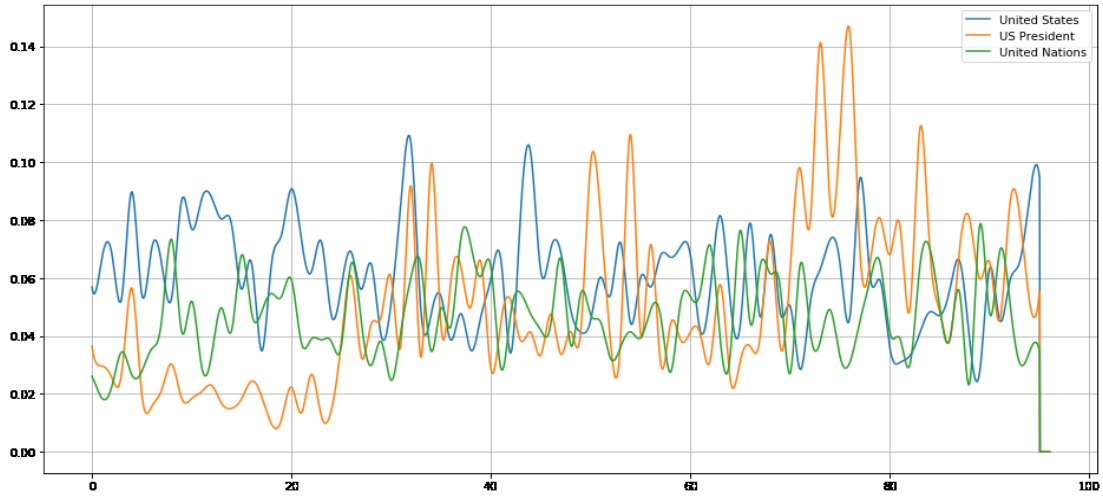
This graph demonstrates the weights for news coverage of 8 Middle East nations in the 96 months from 2011/01 to 2018/12. Egypt (brown) took the lead in Arab Spring and hence jumps at the beginning of the graph. Saudi Arab (orange) cached a lot of attention in 2018 probably because of her military operations in Yemen. Major players like Iran, Syria, and Israel have high weights across the entire time frame. Turkey and Lebanon are covered much less though. It is noteworthy that Iraq does not show up too much.



Another interesting news topic is the Islamic State. The graph above shows the rise and fall of the infamous “state”. ISIS began to catch attention in 2013, reached its peak in 2014 and 2015, and then gradually disappeared from news in 2018.



This graph shows the 5 popular issues reported by news. At the beginning of Arab Spring in 2011, a huge wave of protests (green) happened in Middle East countries. Then military conflicts (orange) and rebels showed up. Human rights (red) were violated frequently in this period, but they were not the focus point of the news media.



This graphs shows a rise in interest in US presidents (orange) over Middle East affairs during the presidential election period in 2016.