2016

Topic Analysis of Tweets on the European Refugee Crisis Using Non-negative Matrix Factorization

Chong Shen
Claremont McKenna College

Claremont McKenna College

Topic Analysis of Tweets on the European Refugee Crisis Using Non-negative Matrix Factorization

submitted to
Professor Blake Hunter

by
Chong Shen

for
Senior Thesis
Spring 2016
April 25

Table of Contents

Abstract

The ongoing European Refugee Crisis has been one of the most popular trending topics on Twitter in the past 8 months. This paper applies topic modeling on bulks of tweets to discover the hidden patterns within these social media discussions. In particular, we perform topic analysis through solving Non-negative Matrix Factorization (NMF) as an Inexact Alternating Least Squares problem. We accelerate the computation using techniques including tweet sampling and augmented NMF, compare NMF results with different ranks and visualize the outputs through topic representation and frequency plots. We observe that supportive sentiments maintained a strong presence while negative sentiments such as safety concerns have emerged over time.

I. Introduction

Twitter exploded on September 2, 2015, when the picture of a drowned 3-year-old Syrian, Aylan Kurdi, was published, shocking people with the cruel reality of senseless deaths of refugees fleeing from Syria to Europe [1]. This set off the extensive discussions about the European Refugee Crisis on Twitter that has continued to today.

According to the United Nations High Commissioner for Refugees (UNHCR), "Globally, one in every 122 humans is now either a refugee, internally displaced, or seeking asylum [2]." We are interested in the dialogues about the Refugee Crisis because we hope to extract representations of the most concerning issues of the crisis from these dialogues.

This paper uses a mathematical topic model to discover and analyze hidden themes that pervade these discussions on social media. Topic modeling is a text mining method to extract patterns in a corpus of texts [3]. In mathematics, a long line of research has been conducted using topic modeling, but not on the topic of the Refugee Crisis, given that it started gaining attention very recently and is still ongoing. This paper formulates topic modeling as an alternating least squares problem, where the basic goal is to identify the best low rank approximation of a given set of document vectors.

We collected about 200,000 time-stamped and geo-tagged English tweets about refugees from across the world for a week each in September 2015, and February 2016. We cleaned the tweets by abstracting stop words and redressing minor misspellings, and then applied Non-negative Matrix Factorization to conduct the automated discovery of the hidden topics within this corpse of tweets.

II. Background on the European Refugee Crisis

Since the summer of 2015, an increasing number of refugees and migrants have begun fleeing to Europe from war, conflict and poverty in Syria, Afghanistan, Iraq and regions of North Africa [4].

According to the UNHCR, as of March 26, 2016, there are currently 4.8 million Syrian refugees, more than 10 percent of which have traveled to Europe [5]. Many of them are paying smugglers on overcrowded boats or buses, but thousands have died while trying to make it to the EU [6].

Refugees and migrants are fleeing to Europe because the refugee camps in the developing countries that are currently hosting them are impoverished and overcrowded [7]. For example, Lebanon, Jordan and Turkey, where refugees make up as much as 25% of the national population, are shelter to 3.6 million Syrian asylum-seekers and are reaching breaking point without nearly enough international humanitarian support [7]. In addition, global media coverage of the applause, flowers and teddy bears that greeted refugees when they made it to Germany and Austria has encouraged more people to embark on the difficult journeys through the Balkans [6].

According to Eurostat, "In the third quarter of 2015, EU countries received over 400,000 first time asylum applications, one third of which were Syrians (33%), followed by Afghans (14%) and Iraqis (11%) [8]." This surging number of migrants and refugees seeking shelter has created tensions among ill-prepared European countries on how to respond to the crisis [7].

So far, Germany has shown the most welcoming gestures with a commitment to take in 500,000 refugees annually [9]. France and the U.K. are much less enthusiastic about hosting refugees, each committing to taking 20,000 Syrian refugees over five years and 120,000 over two years respectively [9]. Central European countries like Czech Republic, Slovakia, Poland and Hungary called for Europe's borders to be sealed off to prevent refugees from flooding in [10].

On the other hand, despite the geographic proximity to Syria and Iraq, the Persian Gulf states like Saudi Arabia have faced international pressure for not taking in any refugees because none of them participated in the 1951 U.N. Refugee Convention [9]. The U.S. has also been pressured to increase its annual refugee admission quota from 70,000 to 85,000 for the year of 2016 [9].

In September, 2015, the European Commission proposed national quotas to relocate 160,000 asylum-seekers from Greece, Hungary and Italy across Europe [11]. But by January, 2016, only about 300 refugees were successfully relocated, which many news reports attributed to rising anti-refugee sentiment that has undermined many countries' commitments [12].

Anti-refugee sentiment has become a legitimate concern since the terrorist attacks on Paris in November and the alleged sexual assaults on women during New Year's Eve in Germany [13]. These incidents have significantly increased a sense of insecurity among the European citizens due to suspicions connecting the Paris attacks to the inflow of refugees [13]. In addition, a national poll in Germany revealed that 70 percent of respondents expected more crimes as refugees flooded in [14].

III. Topic Modeling

This section describes the process of building a topic model and illustrates the concepts of Non-negative Matrix Factorization through a toy example created from synthetic data.

According to David Blei of Columbia University, "Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents [15]." Through topic modeling, we can discover the hidden structure behind these documents including the topics, per-document topic distributions, and the per-document per-word topic assignments [15].

When applying our topic model, we adopt three main assumptions to simplify the processing of texts. First, we adopt the bag of words model that disregards the grammar and word order of a text and convert each tweet into an unordered histogram [16]. Without this assumption, we would need to apply n-gram models to consider the significance of the order of words. Secondly, we assume that the order of documents, i.e., tweets, does not matter. We could apply time series models to explore how discussions on Twitter have changed over a period of time, but this is not done here. Thirdly, we must assume a fixed number of topics. In this case, we vary the number of fixed topics from the tweet database, and compare their ability to extract significant topics.

Toy Example

To experiment with the methods used in topic modeling, we create a toy example using a synthetic data set created by taking random matrices, stacking them into a larger, low rank $M \times N$ matrix and adding noise. To mimic probability distributions, the entries are set to nonnegative values that allow us to apply Non-negative Matrix Factorization.

Non-negative Matrix Factorization

Nonnegative matrix factorization stands for the problem of approximating a nonnegative matrix by a product of two nonnegative matrices. Similar to the case of using real data sets, we suppose we have N tweets and M words in our vocabulary, then we have a $M \times N$ matrix V. Each entry V_{ij} would be the idf value of the i^{th} word in the vocabulary appearing in the j^{th} tweet, which this paper will explain later. For now, we can interpret V_{ij} as a frequency measurement of the i^{th} word's appearances in the j^{th} tweet.

In order to find the hidden topics, we find the best rank K approximation of the document matrix through factorizing$V \approx WH^T$, by minimizing the objective function:

$$J(W,H) = \|V - WH^T\|_F \text{ [17][18].}$$

Word × Tweet Word × Topic Topic × Tweet

$$\begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ V_{m1} & V_{m2} & \cdots & V_{mn} \end{bmatrix} \approx \begin{bmatrix} W_{11} & \cdots & W_{1k} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ W_{m1} & \cdots & W_{mk} \end{bmatrix} \begin{bmatrix} H_{11} & \cdots & \cdots & H_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ H_{k1} & \cdots & \cdots & H_{kn} \end{bmatrix}$$

Figure 1 Approximation of $V \approx WH^T$

As shown in Figure 1, each row of W represents a word in our vocabulary and each column of W is a vector representation of a topic we extract from the tweets. Hence we can compose ranked lists of words related to the topic by sorting the entries of the row of W. We can interpret each of the K rows of H^T as a topic vector and each of the N columns as an encoding that is in one-to-one correspondence with one of the N tweets in V [18]. We can interpret each entry of a row of H^T as the fraction of the document that conforms to a given topic [18].

Alternating Least Squares

To optimize the NMF objective function J (W,H) = $\|V - WH^T\|_{\mathrm{F}}$, we consider a Least Squares problem introduced in Linear Algebra.

The Least Squares Theorem [19]:

Let A be an M×N matrix and let b be in R^{n}. Then $\mathrm{Ax} = \mathrm{b}$ always has at least one least squares solution $\bar{\mathrm{x}}$. Moreover:
a. $\bar{\mathrm{x}}$ is a least squares solution of Ax = b if and only if $\bar{\mathrm{x}}$ is a solution of the normal equations $\mathrm{A^TA}\,\bar{\mathrm{x}} = \mathrm{A^Tb}$.
b. A has linearly independent columns if and only if $\mathrm{A^TA}$ is invertible. In this case, the least squares solution of $\mathrm{Ax} = \mathrm{b}$ is unique and given by $\bar{\mathrm{x}} = (\mathrm{A^TA})^{-1}\mathrm{A^Tb}$.

By fixing W or H, the NMF problem can be solved using least squares methods with non-negativity constraints. Here we choose to use Alternating Least Squares (ALS). The algorithm for ALS is the following:

Algorithm (Alternating Least Squares) [20]**:**

1. Initialize W and H
2. **Repeat**
3. Solve: $min_{H\geq 0}\frac{1}{2}\|V - WH^T\|_F^2$
4. Solve: $min_{W\geq 0}\frac{1}{2}\|V^T - HW^T\|_F^2$
5. **Until** Stopping condition

Essentially, the algorithm first takes a random guess of W, $w_{ij} \geq 0$. It then takes a column of V (call it v) and a column of H (call it h) to find the best approximation of a for $v = Wh$. It does so solving $W^Tv = W^TWv$ with the constraint $v \geq 0$. The algorithm repeats this process for every column of H and produces a best rank k approximation of H matrix for this initial guess of W.

The Algorithm then implements a similar process for the approximation of W for a random guess of H. It takes a row of V and a row of W to find the nonnegative topic representation of V's row $v = wH$. It repeats this process for every row of W and produces the best-approximated Wmatrix for the current estimate of H. Through alternating between these two processes, we can optimize the approximation of matrices W and H since they may converge to a local minimum and we can compare each local minimum to find the best approximation of the solution.

If we find exact and unique solutions to the sub-problems (3) and (4), then we can consider every limit point of the ALS algorithm a stationary point of the NMF problem. To improve the running time of this method, we replace an exact solution of the nonnegative least squares problem by projecting the solution of the unrestricted least squares problem into a nonnegative quadrant [20].

Algorithm (Inexact Alternating Least Squares) [20]**:**

1. Initialize W and H
2. **Repeat**
3. Solve for W in equation: $WH^TH = VH$
4. $W = [W]_+$
5. Solve for H in equation: $HW^TW = V^TW$
6. $H = [H]_+$
7. **Until** Stopping condition

It is important to note that each alternating step at (3) and (5) is convex, which means that each step has a global solution, but together these steps are not convex and we cannot fund a global minimum to the least squares approximation. Instead we can find many local minimums by alternating the initial W, H and taking the optimum that gives the best approximation.

Back to the Toy Example

Through the method of IALS, we are able to factorize our data matrix V into topic matrices W and H. Since the algorithm ranks the topics in W, we plot the first three columns of W, which represent the top three topics in this corpse of tweets. Take W2's graph as an example, the x-axis denotes the index of tweets and the y-axis denotes the frequency measurement values of that tweet. We can observe that the peaks are concentrated in the range from 80 to 100 on the x-axis, which means that many of the last 20 tweets share a common topic 2.

If we compare $W2$ to $W1$, $W1$'s frequency has a less consistent peak and more "dives" in the non-peak area, which means that $W1$ picks up more noise among the tweets when forming its topic and hence is probably less meaningful than $W2$.

Next, let's look at the Image of W, which has 1 to 10 on its horizontal axis and 0 to 100 on its vertical axis. If we look at the second column or vertical block of the graph, that echoes $W2$'s graph, having many highlighted lines in the last 20 rows, which means that the last 20 tweets have a high concentration of topic 2. In contrast, the first column has the highlights spread out across vertically, which agrees with the fact that topic 1 is modeling noise. True hidden topics are like $W2$ and $W3$, which have more structure and much less noise.

Finally let's consider the graph of '$W1 \times W2 \times W3$', which plots the distribution of the top three topics in a 3-dimensional space. This allows us to see how frequently these three topics appear in the same tweet. It gives an idea as to how closely-related these topics are.

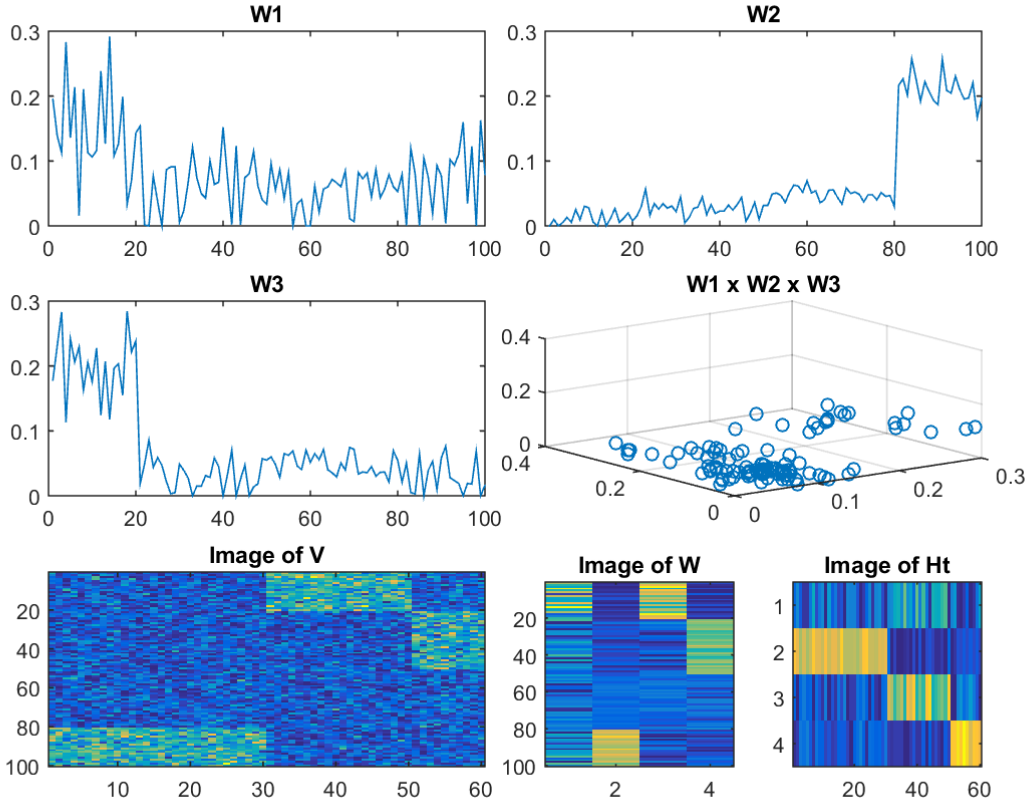

Figure 2 Rank 4 NMF Results of a Toy Example

Since we set a low noise level for the toy example, we can clearly identify from the image of V the light-colored clusters of "tweets" and "words." These highlighted clusters are decomposed into specific blocks of word-topic correspondence in the Image of W and blocks of topic-tweet correspondence in the Image of H^T. We keep in mind these simplified clusters and look for them in the analysis of the real data sets.

IV. Application

Having understood topic modeling and the algorithm to solve Non-negative Matrix Factorization through a toy example, we now apply these methods on the set of raw data that we collected.

Data Collection

We collected a total of 194,834 time-stamped and geo-tagged English tweets from across the world for a 10-day period in September 2015 and a week in February 2016. We downloaded tweets for Jan. 31 – Feb. 6, 2016 by using TAGS,[1] a free service that has streamlined connection to Twitter and allows me to download recent tweets with my designated hash tags and keywords. We used "refugee" as the keyword search to minimize restrictions and to maximize the amount of tweets that we can pull down. We collected tweets for Sept. 9 – 19, 2015 through Sifter,[2] a priced platform that provides retrieve access to historical tweets, since

[1] https://tags.hawksey.info/get-tags/
[2] http://sifter.texifter.com

media coverage and social media discussions about the Refugee Crisis broke out during September. Overall, we were able to pull down 100,000 tweets in September and 94,834 tweets in February.

We first tested our topic model on the September data set and then applied topic modeling to each month's data and compared the themes extracted from each month to reflect the changes in topics over time. We plotted the number of tweets versus the time intervals for each batch of data. The September tweets from Sifter were evenly spread out across each two-day interval, while the February tweets from TAGS were mostly from Feb. 3 – 4, 2016. We want to keep this in mind during the interpretation of our results later.

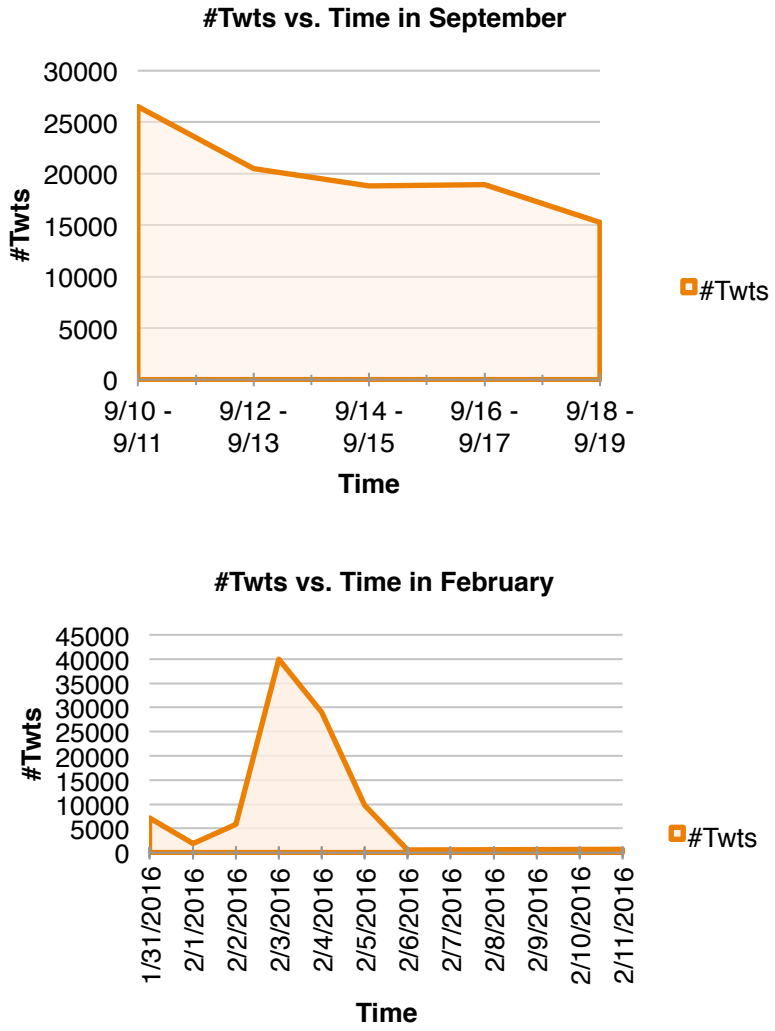

Figure 3 Number of Tweets in each time interval

Pre-Processing

We were able to adopt pre-existing code written by the 2014 & 2015 REU Groups to pre-process the raw data.[3] [3b]

Pre-processing is important because it allows us to apply efficient analysis of the sizable corpse that we collected. We followed the steps in Hunter et al to pre-process our raw data: "1. We encode the text into ASCII, discarding any Unicode characters. 2. We replace all double quotes with the empty string. 3. We extract all user references and all hash tags, denoted respectively with @ or # at the beginning of a token. 4. We attempt to remove any urls, specifically anything prefixed with "http." 5. We remove many non-alphanumeric characters, with the important exception of $ and @, with the latter only in the case that it is the only character in the token. 6. We change all characters to lowercase. 7. We remove any token on our Stop Words list, including a Twitter specific stop words list of the 50 most common words observed in our dataset. 8. We remove any token observed less than 10 times [18]."

Next we constructed a sorted vocabulary and built term-frequency vectors from the tweets. We concatenated them to generate a data matrix V' such that each row denoted a tweet and each column represented a unique word in our vocabulary. Then we re-weighed the V' through TF-IDF, also named term frequency-inverse document frequency, a weighting element that reflects the significance of a word to a document in a corpus [18].

Since the focus of this paper is topic modeling, we won't go into details about the methodology of TF-IDF. We borrowed the definition from William et al [21] that if S and T are two word sets, then the TF-IDF can be defined as

$$TFIDF(S,T) = \sum_{S \cap T} V(w,s)V(w,T)$$

where TF is the frequency of word in S, N is the size of the corpus. IDF is the inverse of the fraction of names in the corpus that contain w,

$$V'(w,S) = \log\left(TF_{w,s} + 1\right)\log(IDF_w)$$

$$V(w,s) = \frac{V'(w,S)}{\sqrt{\sum_w V'(w,S)^2}}.$$

According to Rajaraman and Ullman in *Mining of Massive Datasets*, "the TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general [22]."

[3] Research Experience for Undergraduates (REU) 2015: Katie Khuu, Daniel Balague-Guardia, Michael Boggess, Bo Jones, Eric Lai and Blake Hunter.

[3b] Research Experience for Undergraduates (REU) 2014: Baichuan Yuan, Daniel Moyer, Cristina Lopez, Eric Lai, Zhaoyi Meng, Xiyang Luo, Alex Robicquet and Blake Hunter.

Having calculated the value for each V_{ij}, we now have a $M \times N$ matrix such that N is the number of tweets and M is the number of words in our vocabulary. Since V is a $N \times M$ matrix of frequency counts, all entries in V have non-negative values.

Choosing Ranks

Following the analysis of Hunter et al. [18], our methods also assume that the rows of V are approximately the additive combination of K non-negative topic vectors, where $K \ll N$. Essentially we are assuming that "V is approximately of rank K, with the constraint that the subspace spanned by V has a set of non-negative coordinates in that basis [18]." The rank K of the factorization is chosen so that a number of K top topics are extracted from V.

We approximate V using rank 10 and rank 25, which allows us to retrieve 10 and 25 topics from the corpus respectively. From a mathematical perspective, the rank we pick gives a k-dimensional representation of V. Take entry V_j for example, $V_j = a_1 w_1 + a_2 w_2 + \cdots + a_k w_k$. Suppose V_j is in $\mathbb{R}_+^d$, then rank k will allow us to use k vectors to describe V_j as a subspace of $\mathbb{R}_+^d$. The larger the k, the more basic vectors, hence the more dimensions we can use to describe V_j, and the more accurate representation. The smaller the k, the more succinct our summary as we force topics that may be connected or partially similar to join into one topic.

Out of curiosity, we compare the topics generated from rank 10 and rank 25 approximations. We identify the most similar pair of topics from each batch by taking each W matrix and multiplying them. Consider the inner product of two unit vectors $< v, u >$, the closer to 1 the entry is, the more similar the two vectors are (since $< v, u > = 1$ when $v = u$). Hence the larger value an entry is, the more similar the two corresponding topics are. We will later use $\max_j (W^{10})^T W^{25}$ to find the pair or combination of topics that match each other the most closely.

Computing NMF

We were able to adopt pre-existing code written by the 2014 & 2015 REU Groups to apply NMF to our data.[4] [4b] But we had some difficulty computing NMF for matrices of size 100,000 × 10,000 since our computer has limited RAM and MATLAB consistently ran out of memory for these commands. We utilized the following technique to simplify the process, decrease the memory space required and reduce the computation time.

We first took a random sample containing 10,000 tweets, the largest amount of tweets that the lab's computer can handle in this case, and applied NMF to compute W and H matrices that approximate the hidden topics in this small corpus. Since tweets represented in the V matrix is distributed by time sequence, we think that a random sample will give us a fair rendition of the overall corpus, and that the W matrix computed from this sample can be used as our word by topic representation. Since $V \approx WH^T$, we can take a small block of V, as highlighted in Figure 3 (let it be V_1), and use simple matrix division to find the corresponding

[4] Research Experience for Undergraduates (REU) 2015: Katie Khuu, Daniel Balague-Guardia, Michael Boggess, Bo Jones, Eric Lai and Blake Hunter.
[4b] Research Experience for Undergraduates (REU) 2014: Baichuan Yuan, Daniel Moyer, Cristina Lopez, Eric Lai, Zhaoyi Meng, Xiyang Luo, Alex Robicquet and Blake Hunter.

block of H (let it be H_1) through $H_1^T = \frac{W}{V_1}$. We can continue this process until we finish processing each block of the V matrix and concatenate all the blocks of H to compose the full H matrix.

$$\begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ V_{m1} & V_{m2} & \cdots & V_{mn} \end{bmatrix} \approx \begin{bmatrix} W_{11} & \cdots & W_{1k} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ W_{m1} & \cdots & W_{mk} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ H_{k1} & H_{k2} & \cdots & H_{kn} \end{bmatrix}$$

Figure 4 Augmented Approximation of $V \approx WH^T$

This technique is suitable for computations using datasets that may be too large for the machine to process. One can easily process matrices with up to one billion rows by breaking down the matrix by blocks and applying NMF individually.

V. Research Questions & Results

We set two major research questions that we hope to answer with our analysis.

1) What issues triggered people's interest in discussing the European Refugees Crisis on twitter in September?
2) Have these issues of interest changed over time? If so, how?

Top 10 Topics for September

After analyzing the 100,000 tweets on refugees in September, we found the following top 10 topics. Each topic is illustrated with its most frequent 10 words, and each word is sized to represent its significance in its topic.

1	2	3	4	5
rt	europe	agreed	border	refugees
//t	refugees	aid	hungary	welcome
children	crisis	vulnerable	croatia	via
refugee	across	support	serbia	syria
today	migrants	sum	migrants	muslim
come	thousands	additional	refugees	germany
photos	rallies	shouldnt	germany	take
http	isis	game	closes	eu
women	refugees	funds	austria	country
crisis	via	supporting	cross	countries

6	7	8	9	10
saudi	police	000	help	syrian
arabia	tear	10	match	refugees
million	gas	obama	need	u
100	water	accept	migrants	s
3	hungarian	us	donate	help
tents	cannons	s	google	us
take	riot	u	donation	lebanon
000	used	year	urgent	take
house	refugees	refugees	refugees	states
people	cannon	next	million	accept

Figure 5 Top 10 Topics Discovered related to Refugees in Sept. 9 – 20, 2015
Note: Each word in a topic is sized to represent its significance in the topic.

To interpret these topics without searching for each of these keywords in the corpus of tweets, we sorted each row of H^T (i.e., we sorted all tweets' frequency values corresponding to each topic) and found the tweet that matches the topic the most closely. For example, we learned that Topic 8 is best described by the tweet "Obama administration preparing for at least 10,000 Syrian refugees next year."

From these 10 topics and their closest matches of tweets, we observed strong interests in supporting and accepting refugees through keywords like "welcome", "aid", "funds", "accept", "help" and "donate." Europe, the center of the refugee crisis, also had a strong presence in keywords like "EU", "Germany", "Austria", "Hungary" and "Croatia." In addition, "Saudi Arabia" and the "U.S." received a lot of traction in their policies regarding admitting and assisting refugees.

To understand the noise level and significance of each keyword in a topic, we plotted the topic distributions like we did with the toy example. $W1$ is topic 1 "rt", $W2$ is topic 2 "europe", and $W3$ is Topic 3 "agreed." From the Image of $W3$, we see that Topic 3 picked up about six keywords that were equally significant at the frequency level of 0.3, which means that the keywords in Topic 3 carry more significance.

Image of H shows that rows 1 and 5 have a lot of highlighted lines, which means Topic 1 and 5 pick up a lot of noise across all tweets. We suspect that it is because "rt" in Topic 1 picked up many retweets, and that "refugees" in Topic 5 picked up the common keyword "refugee" that we originally searched for in downloading the data.

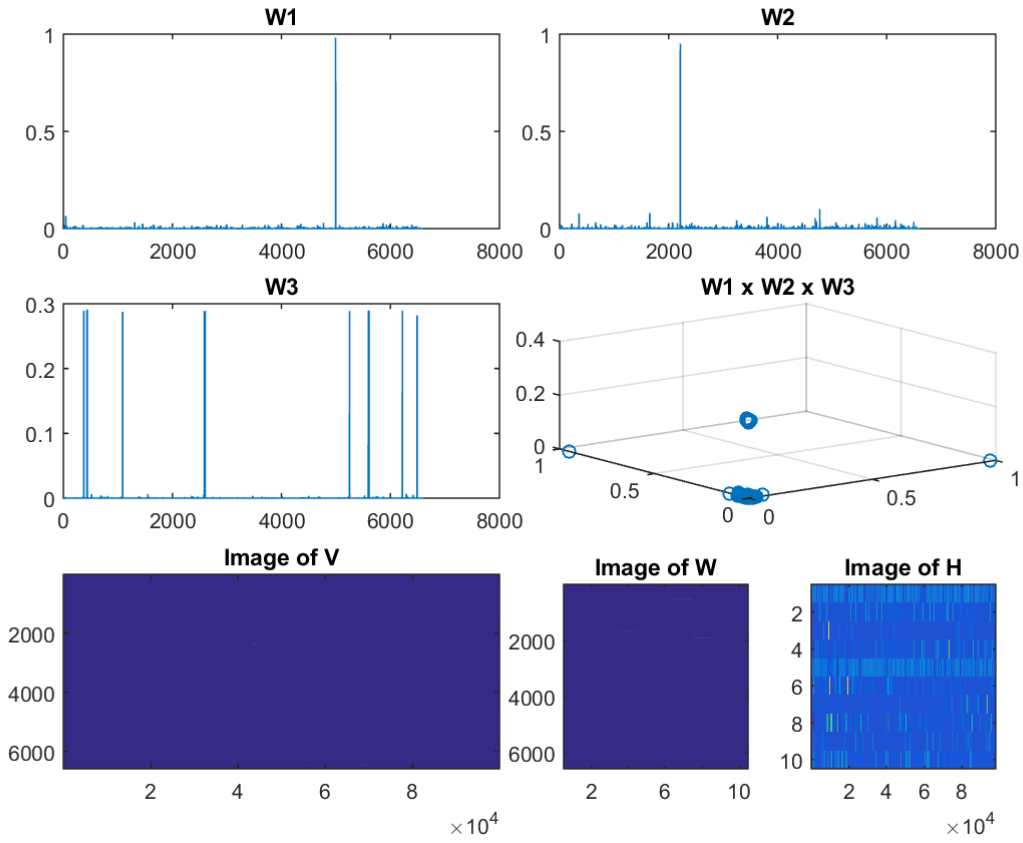

Figure 6 Rank 10 NMF Results of Sep Tweets on Refugees

To understand the popularity of each topic over the 10-day period, we plotted their frequencies from September 9 – 19, 2015 in Figure 6.

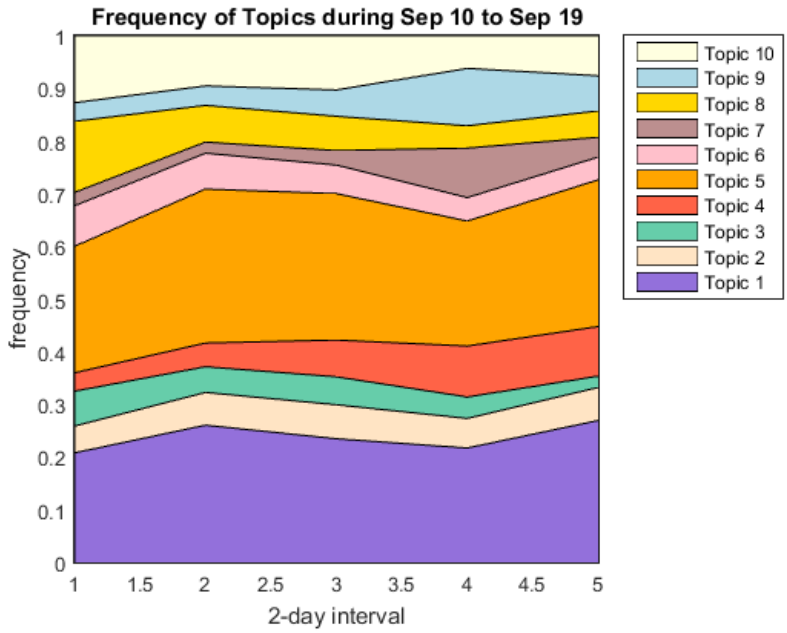

Figure 7 Frequencies of Topics in Sept. 9 – 19, 2015

Interestingly, topics 1 and 5 dominated the discussions in the 10-day period. We suspected that topic 1 was popular because Sept. 9th was a week after Aylan Kurdi's photo went viral, and many more photos became widely spread across Twitter for recording the suffering of child and woman refugees [23]. We guessed that "rt" was the top word for Topic 1 because people could be mostly retweeting the pictures on Twitter. We conjectured that Topic 5 trended because many local German communities were welcoming refugees during their arrivals. In addition, Topic 7 peaked during the fourth period, i.e. Sept. 16 – 17, which we guessed to be a result of increased news coverage of the Hungarian police deploying tear gas and water cannons against refugees.

Rank 10 versus Rank 25 Topic Comparison

Given that we were analyzing 100,000 tweets, we thought 10 topics might have over-condensed the similar sub-topics and brushed over nuances that could be significant. To understand these sub-topics in detail, we took a rank 25 approximation of our V matrix and broke the tweets from 10 topics down to 25 sub-topics. We then looked for $max_j(W^{10})^T W^{25}$ to identify the sub-topics corresponding to each major topic from our rank 10 approximation.

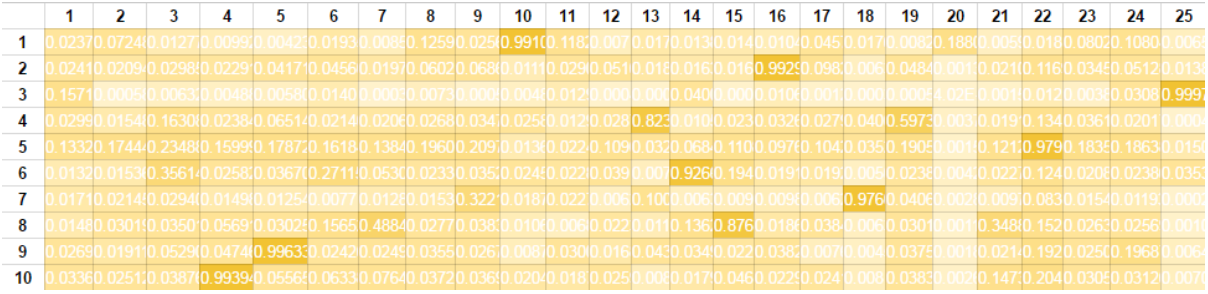

Figure 8 Heat map of $\max_j(W^{10})^T W^{25}$ for September
Note: The darker color represents the higher value of $(W^{10})^T W^{25}$

As an example, the greatest frequency value in column 25 is in row 3, meaning that Topic 25 in rank 25 approximation is most similar to Topic 3 in rank 10 approximation. As the side-by-side comparison below shows, the two have 9 out of 10 keywords in common. Both topics also share the same closest tweet: "Supporting the most vulnerable shouldn't be a zero sum game. Support refugees with additional funds to agreed aid budget."

3	25
agreed	agreed
aid	vulnerable
vulnerable	support
support	sum
sum	additional
additional	shouldnt
shouldnt	funds
game	supporting
funds	game
supporting	budget

Figure 9 Sub-topic (25) for Topic 3

Following this logic, we found that the closest match for Topic 5 in rank 10 approximation are a set of sub-topics in rank 25 approximation, including sub-topics 2, 8, 9, 12, 17, 22 and 23. This was not a surprise, given that it dominated other topics in frequency. We noticed that these sub-topics revealed a more complicated mix of concerns besides the welcoming and supportive sentiments we previously observed. Not only were religious identities including Christian and Muslim discussed comparatively in Topic 2, the terrorist group ISIS also came up in Topic 9, indicating safety concerns with the inflow of refugees. But supportive sentiments also had a strong presence in topics 8, 22 and especially 23, which talked about barbers giving haircuts to refugees on the streets.

Figure 10 Sub-topics (2, 8, 9, 12, 17, 22, 23) for Topic 5

September versus February Topic Comparison

To understand how discussions have changed five months after the outbreak in September, we carried out the same analysis on 94,834 tweets about refugees during Jan. 31 – Feb. 6, 2016. Since many more refugees have reached Europe over the interval of five months, we predicted a stronger presence of negative sentiments due to overcrowding European countries and problems in the integration of refugees into local communities. In addition, since the Paris Attack took place in November 2015, we expected to see topics more related to safety concerns in Europe given the inflow of refugees. We found the following top 10 topics for the February batch of tweets.

1	2	3	4	5
greek	german	refugees	s	children
prize	mayor	syrian	u	europe
nobel	don't	million	syrian	migrants
peace	harassment	us	kerry	3
helping	provoke	syria	announce	move
nominated	sexual	4	house	1
fisherman	walk	000	boost	refugees
grandmother	avoid	2	significant	000
refugees	fighters	eu	white	says
rt	agency	via	refugees	rt

6	7	8	9	10
many	time	=	memorial	rt
nations	show	allowing	might	refugees
poland	next	bull	swedish	muslim
budapest	come	drop	offend	high
1968	risking	slowly	worker	go
czechoslovakia	someone	driving	migrant	today
1981	everything	insanity	aid	germany
1956	wonders	suicide	murdered	commisioner
quickly	europe	deaths	young	filippo
forget	refugees	shit	bans	grandi

Figure 11 Top 10 Topics Discovered related to Refugees in Jan 31 – Feb 6, 2016
Note: Colored keywords appeared in both September and February.

We highlighted the keywords in these topics that overlapped with those in our September analysis, including the "U.S.", "Germany", "help", "children", "Europe" and "Budapest." It was interesting to observe whether these keywords were discussed in the same context. For example, "Budapest," the capital of Hungary, was still the center of condemnation for not accepting refugees. On the other hand, "German" came up in Topic 2 about a statement from a German mayor telling schoolgirls to not walk near refugees to avoid sexual harassment.

In fact, "harassment" along with "deaths" and "murder" came up as completely new keywords in the top 10 topics. We observed from the new set of topics higher levels of frustration from supporters who believed that more needed to be done for refugees and from cynics who found the inflow of refugees problematic.

Recall that the tweets collected in February were unevenly distributed across each time interval, so we decided that plotting the frequency of topics over this time period would not be meaningful. Instead, we repeated the process of taking a rank 25 approximation of our V matrix and breaking the tweets from 10 topics down to 25 sub-topics.

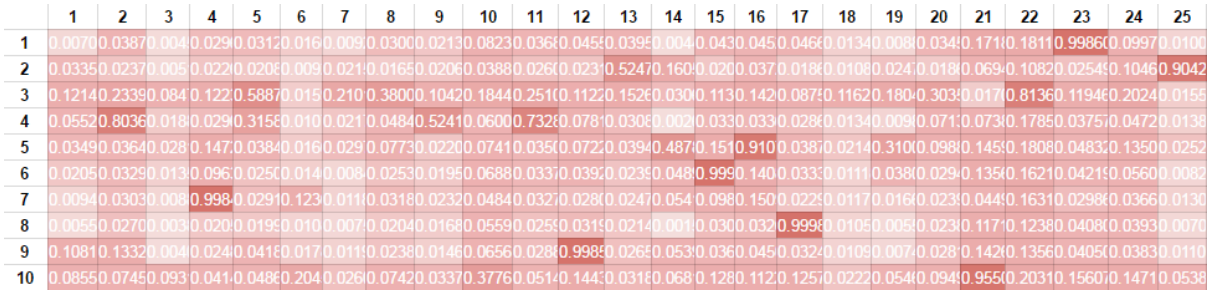

Figure 12 Heat map of $\max_j(\mathbf{W^{10}})^T\mathbf{W^{25}}$ for February

Note: The darker color represents the higher value of $(W^{10})^T W^{25}$

Figure 13 Sub-topics (3, 6, 10, 21) for Topic 10

Note: Each box at the bottom of the figure contains the closest match of tweet for each sub-topic.

We decided to highlight Topic 10 and its closest match of sub-topics including sub-topics 3, 6, 10 and 21 because Topic 10 revolved around Filippo Grandi from UNHCR, the first time an international organization became relevant in our topics. In particular, Topic 10 stemmed from retweets of Serbian Prime Minister's tweet "With @UN High Commissioner for Refugees Filippo Grandi on possible solutions for #migrantcrisis. #SupportSyrians." We found it interesting that all of these sub-topics showed support for refugees.

Figure 14 Serbian Prime Minister's Tweet
Source: https://twitter.com/SerbianPM/status/695312243307761664

While we observed many meaningful topics in the dialogues surrounding refugees, we also found it meaningful to learn what issues didn't come up as any keyword. For example, only refugees from Syria were the center of the discussions while Iraqi, Afghan and North African refugees didn't draw any significant interests on Twitter. In addition, we didn't see much presence of international organizations besides UNHCR, such as Doctors Without Borders whom have been providing medical care to refugees across Europe [24]. Last but not least, it was interesting to see Lebanon in the topics but not Turkey or Jordan, despite the fact that all three countries are hosting the largest amount of refugees from Syria.

VI. Future Work

We found a number of areas for improvement in our analysis. Firstly, we can eliminate "rt", "refugees" and other keywords during pre-processing which may not carry significance to our analysis so that we can observe and interpret topics that pick up less noise.

Secondly, we observed from our topic analysis that many of the topics discovered were heavily correlated with news reports and other forms of media coverage. We looked up the top news items in both periods of our analysis, and easily located most of the key words in these news items. This inspired the question regarding the relationship between media coverage and twitter discussions regarding the refugee crisis. Given more time and resources, we would like to pull down batches of news articles and apply NMF to discover how closely topics and tweets relate to a news article. We envisioned that a time series topic analysis of public tweets and the news articles for at least one month might yield meaningful results.

How much longer can Germany keep its doors open to **refu**...
The Guardian - Sep 12, 2015
On Thursday, 500 **refugees** moved into another sports hall, next door to the athletics complex. Its giant playing area for handball and volleyball is now decked ...

Croatia opens borders, welcomes sudden stream of **refugees**
CNN - Sep 16, 2015

Obama Directs Administration to Accept at Least 10000 Syr...
ABC News - Sep 10, 2015
President Obama has directed his administration to accept at least 10,000 more Syrian **refugees** starting on Oct. 1, White House press secretary Josh Earnest ...

Background Briefing on the United States **Refugee** Admissions **...**
Highly Cited - US Department of State (press release) - Sep 9, 2015

Figure 15 Top News Items on Google search results for "refugee" during Sept. 9 - 19, 2015
Source: Google

Thirdly, we also thought that the topics generated didn't seem to reflect personal experience or attitudes toward refugees because we took tweets worldwide in the English language. We conjectured that the people who posted these tweets may not have had personal experience with the refugee crisis, which is why their discussions were driven by what news outlets chose to report on. We predicted tweets in languages such as German and Turkish and from geographic locations in Europe would yield more meaningful reflections of how the local communities in Europe have interacted with the inflow of refugees. In addition, we thought it would be meaningful to conduct a time series topic analysis with these "European" tweets and match the changes in topics with the shifts in European governments' policies regarding refugees and migrants.

VII. Conclusion

In this paper, we applied topic modeling on a total of 194,834 tweets to discover the hidden patterns within these social media discussions surrounding refugees. The core of our topic model was solving Non-negative Matrix Factorization as an Inexact Alternating Least Squares problem. We introduced techniques including tweet sampling and NMF by batches to accelerate the computation. To interpret our outputs, we visualized our analysis through topic representation and frequency plots in addition to comparing NMF results with different ranks.

We compiled the following answers to our research questions:

1) What triggered people's interest in discussing the Syrian Refugees Crisis on twitter in September?

In September, people showed major interests in two areas: humanitarian support and admissions of refugees. We discovered people's interests in providing support for refugees through topics with keywords including "welcome", "aid", "funds", "accept", "help" and "donate." We observe interests in where refugees are admitted through topics related to "Germany", "Hungary", the "U.S.", "border", "cross" and "Saudi Arabia."

2) Have issues of interest changed from September to February? If so, how?

Issues of interest have changed over the span of 5 months. In particular, new topics like "suicides" and "sexual harassment" have emerged, indicating rising interests in safety concerns. However, we were also able to observe subtle consistency between the two time frames through rank 10 versus rank 25 comparisons. For example, we detected early signs of safety concerns in September when sub-topics from rank 25 approximations involved the terrorist group ISIS. We also noticed consistent presence of support for refugees in February when sub-topics expressed concerns over the well-being of child and woman refugees.

Our finding is significant on two levels. As a mathematical model, this method can be used to uncover topics for any other corpus of documents and is scalable for analyzing a massive volume of documents. As a sociological research, our results unveil many specific issues to be explored, such as the role of religious identities in the European Refugee Crisis and the crimes perceived related to refugees.

Most recently, EU and Turkey has reached an agreement in March to send all refugees and migrants arriving in Europe to Turkey in an attempt to reduce the overwhelming migration to Europe [25]. Migration expert, Patrick Kingsley, has suggested that despite the closure of the Aegean smuggling route, another route is likely going to open since refugees are still aiming for Europe as their destinations [25]. We think this indicates that dialogues surrounding refugees will remain relevant on social media and there will likely be more new topics emerging from these dialogues.

VIII. Acknowledgement

This work was supported in part by the Mgrublian Center for Human Rights at Claremont McKenna College. First and foremost, the author would like to sincerely thank Blake Hunter, her thesis reader and greatest supporter, because he has helped and taught her immensely throughout this process. The author would also like to extend her gratitude to Jennifer Taw, her Academic advisor, for her wisdom on the Refugee Crisis as well as her invaluable guidance in the past four years. In addition, the author wants to acknowledge Heather Ferguson from the History department for sharing her expertise on the Refugee Crisis, and Wendy Lower from the Mgrublian Center for Human Rights for generously sponsoring the purchase of the Twitter data. Lastly, the author would like to appreciate her amazing family members and friends for being her greatest cheerleaders in life.

IX. Bibliography

[1] Laurent, Olivier. "What the Image of Aylan Kurdi Says About the Power of Photography." *Time*. September 4, 2015. Accessed March 26, 2016. http://time.com/4022765/aylan-kurdi-photo/

[2] "Worldwide displacement hits all-time high as war and persecution increase." *UNHCR.org*. June 18, 2015. Accessed April 22, 2016. http://www.unhcr.org/558193896.html

[3] Brett, Megan R. “Topic Modeling: A Basic Introduction.” *Journal of Digital Humanities*. Vol.2, No. 1 Winter 2012. December 12, 2012. Accessed April 16, 2016. http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

[4] “The 2015 Immigration Crisis.” *Infoplease.* @2000-2015 Sandbox Networks, Inc. publishing as Infoplease. 23 April, 2016. http://www.infoplease.com/world/statistics/immigration-crisis.html

[5] "Syria Regional Refugee Response." *UNHCR.org*. March 16, 2016. Accessed March 26, 2016. http://data.unhcr.org/syrianrefugees/regional.php

[6] Fleming, Melissa. "Six Reasons Why Syrians Are Fleeing to Europe in Increasing Numbers." *The Guardian*. October 25, 2015. Accessed March 26, 2016. http://www.theguardian.com/global-development-professionals-network/2015/oct/25/six-reasons-why-syrians-are-fleeing-to-europe-in-increasing-numbers

[7] Bajekal, Naina. “The 5 Big Questions About Europe’s Migrant Crisis.” *Time*. September 9, 2015. Accessed April 21, 2016. http://time.com/4026380/europe-migrant-crisis-questions-refugees/

[8] “More than 410,000 first time asylum seekers registered in the third quarter of 2015.” *Eurostat*. 10 December 2015. Accessed April 23, 2016. http://ec.europa.eu/eurostat/documents/2995521/7105334/3-10122015-AP-EN.pdf/04886524-58f2-40e9-995d-d97520e62a0e

[9] Martinez, Michael. “Syrian Refugees: Which countries welcome them, which ones don’t.” *CNN*. September 10, 2015. Accessed April 10, 2016. http://www.cnn.com/2015/09/09/world/welcome-syrian-refugees-countries/

[10] Fenton, Siobhan. “Czech Republic, Slovakia, Hungary and Poland call for Europe to block main route used by refugees” *The Independent*. Feb 16, 2016. Accessed April 23, 2016. http://www.independent.co.uk/news/world/europe/czech-republic-slovakia-hungary-and-poland-call-for-europe-to-block-the-main-route-used-by-refugees-a6876846.html

[11] “Refugee Crisis: European Commission takes decisive action – Questions and answers.” *European Commission Press Release Database*. September 9, 2015. Accessed April 23, 2016. http://europa.eu/rapid/press-release_MEMO-15-5597_en.htm

[12] “Member States’ Support to Emergency Relocation Mechanism.” *Eurostats*. April 21, 2016. Accessed April 23, 2016. http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/european-agenda-migration/press-material/docs/state_of_play_-_relocation_en.pdf

[13] Sazonov, Andrey. "Paris Attacks and Europe's Response to the Refugee Crisis." *The Huffington Post*. December 2, 2015. Accessed April 2, 2016. http://www.huffingtonpost.com/european-horizons/paris-attacks-and-europes_b_8684636.html

[14] "Poll: Germans Increasingly Skeptical of Refugees." *DW*. January 15, 2016. Accessed April 2, 2016. http://www.dw.com/en/poll-germans-increasingly-skeptical-of-refugees/a-18982943

[15] Blei, David. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (April 2012): 77-84.

[16] Kosala, Raymond and Hendrik Blockeel. Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1):1–15, 2000.

[17] Lee, Daniel D. and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature*, volume 401 (21 October 1999): 790.

[18] Lai, Eric; Moyer, Daniel; Yuan, Baichuan; Fox, Eric; Hunter, Blake; Bertozzi, Andrea L.; and Jeffery Brantingham. "Topic Time Series Analysis of Microblogs." Accessed April 10, 2016. ftp://ftp.math.ucla.edu/pub/camreport/cam14-76.pdf

[19] Strang, Gilbert. *Introduction to Linear Algebra, 5th Edition*. Wellesley-Cambridge Press. 2016.

[20] Ho, Ngoc-Diep. Nonnegative Matrix Factorization Algorithms and Applications. UNIVERSITÉ CATHOLIQUE DE LOUVAIN. June 2008.

[21] Cohen, William W.; Ravikumar, Pradeep; and Stephen E. Fienberg. "A Comparison of String Distance Metrics for Name-Matching Tasks." *American Association for Artificial Intelligence* (2003).

[22] Rajaraman, A.; Ullman, J. D. (2011). "Data Mining". *Mining of Massive Datasets*. pp. 1-17.

[23] Pensiero, Karen Miller. "Aylan Kurdi and the Photos That Change History." *The Wall Street Journal*. September 11, 2015. Accessed April 15, 2016. http://www.wsj.com/articles/aylan-kurdi-and-the-photos-that-change-history-1442002594

[24] Goldberg, Eleanor. "How to Help Refugees and Migrants In Dire Need of Aid." *The Huffington Post.* September 3, 2015. Accessed April 23, 2016. http://www.huffingtonpost.com/entry/syrian-boy-death-how-to-help_us_55e869d5e4b0c818f61af36a

[25] Kingsley, Patrick. "Refugee Crisis: What does the EU's deal with Turkey mean?" *The Guardian.* March 18, 2016. Accessed April 22, 2016. http://www.theguardian.com/world/2016/mar/18/eu-deal-turkey-migrants-refugees-q-and-a