Unit 7

Arjun Chandrasekaran
Jouella Fabe
Jonathan Tran
Saurav Sharma
Peter Sulucz

Driving Question: What is the best way to summarize with indicative sentences?

1. **What sentences should be the input to the clustering?**

During our initial investigations, we were curious to see whether words that appear in similar contexts would be grouped in the same cluster. As features, we fed words to mahout k-means clustering algorithm. There were a few clusters with related words grouped together, however, they were not the Top words – the top words had a few good words along with some irrelevant words. Thus, overall, as expected, we could not obtain satisfactory results from this naïve approach.

A better approach is to use some of the sentences that we deemed important after cleaning up all of our files for our corpora - ClassEvent, YourSmall and YourBig. These input sentences are then broken down by mahout into vectors using tf-idf – which are our features. We placed these sentences in separate files, which is our input to mahout k-means algorithm. Also, we added another stage of selecting documents that were going to be the input to mahout k-means, as explained in 'cleaning the collection' section.

2. **How to cluster the sentences?**

We are using mahout's k-means to cluster our sentences. We mulled over the option of choosing the number of clusters manually (-k option) instead of specifying a list of centroids to mahout. But we decided against this because a predefined set of clusters is useful only in cases where the output is known to belong to one of K classes. In our case, while we wanted words that were relevant to our collection, but without an inherent method for tf-idf feature vectors to distinguish them, we decided not to specify the number of clusters.

Mahout's k-means algorithm automatically picks centroids for our input parameters that determine the size of a cluster and distance between two clusters, or roughly put, the overlap between two clusters. These two are the parameters -t1 and -t2 in mahout canopy command. We experimented extensively with different configurations of t1 and t2 to obtain the optimum cluster size and number of clusters.

For the Class Event, our total input is 288 points/sentences (split as 1 sentence per document).
With t1= 100 and t2= 50, we get an output with 209 clusters with almost all clusters having a single point assigned to them. Thus we see that it is not desirable to have small sized clusters since it resulted in a large number of individual clusters with too few points in them.
Output file : ClassEvent/CESsentDump_t1_100_t2_50.txt

We then tried with a bigger cluster size -t1 500 and -t2 350. Using this, we got about 83 clusters, which was a good number. We also found that when we eliminated clusters with too few points, we ended with a good representative set of clusters.
Output file : /ClassEvent/CESsentDump_t1_500_t2_350.txt
We parsed this file using our code /parseSentDump.py and obtained a raw list of indicative sentences.
Output file: Output file : ClassEvent/parsedSentDumpScript.txt

We stress that this is just a raw list of sentences, and there is much scope for further processing and refinement. We believe especially that using the Template in Unit8 will help us objectively handpick sentences from this pool of 'good' sentences to fit our template.

From our experiments on cluster size, we see that by having:

a) Larger t1 and t2, we have fewer clusters with big cluster sizes and (in general) a large number of points assigned to a few clusters. Using these parameters, we have many related sentences grouped together, which otherwise might have been scattered over a number of clusters.
b) Smaller t1 and t2, we have many clusters that are of small sizes. Using this, we don't have a Single cluster that can be indicative of our corpus, but have a number of clusters that contain indicative sentences. We need to weight these sentences from different clusters.
Due to hadoop memory issues, we used an approach with smaller t1 and t2 with our Small and Big collections. Using bigger t1 and t2 returned GC overhead errors.

3. Cleaning the collection:

In order to ensure that our data was relevant, we started out collection based only on files that had been previously classified as positive by our classifier. These files were all cleaned up of any extraneous spaces and newline characters that could ruin our data. On this, we ran a script that split each sentence from each file into its own file. So we ended up with one sentence per file. During this procedure, we did a bit of extra cleaning, to make sure that we didn't include any sentences that had less than 4 words. We also excluded any sentence that had a word in it which was longer than 16 characters, because in our case, these were usually irrelevant. We also created a specific list of stopwords to our corpus, which was gathered from previously found top words, and our LDA output. This list was focused on excluding any sentence which contained web related words. In our case "www.reddit.com" was a huge offender.

4. From a cluster, how can one find the best sentence to characterize the entire cluster?

We find the distances of sentence vectors from the centroid of the cluster using the script that the TA's gave us, and our own script. The sentences that have the least distance, are the sentences with the best matches with the centroid. In our script, we have implemented this as one of the parameters to choose the best sentences, even from within a cluster (please see section 5).
We will use words that we have found from past units as indicative and frequent as features for the cluster. Other possible features will be the topics that separate the sentences from the LDA. This was done in the background during the execution of our parse scirpts.

5. Which of the clusters should be considered when selecting sentences for the summary?

Selection of clusters:

1. We look at the number of points in a cluster – the higher the number of points, the more likely it is to be a cluster that is 'good' and contains relevant sentences. It is safe to make this assumption because we have cleaned our collection by filtering out regularly occurring noise, so we expect the highest cluster of related sentences to be directly from our articles. We find the clusters with highest points in our implementation – parseSentDump.py.

2. In addition, we don't want to take all the sentences from a cluster – we need just the ones that are closer to the cluster centroid. So we use a distance from centroid implementation provided by the

TAs (in parseClusteringOutput.py) in our own code parseSentDump.py to obtain the closest sentences to the centroid .

3. We combine ideas from 1 and 2 for weighted selection of sentences based on cluster size and distance from centroid. This, we feel is a further area for improvement based on the corpus characteristics. There are two parameters that we can choose to create a pool of good sentences : a) Size of cluster chosen, b) Distances from centroid of sentences chosen

6. How should the selected sentences be ordered?

For our data we decided that date order is not very useful. One of the reasons is that we could not find much date data. The main reason is that our shootings do not really span over a long time. So all of the articles would have been written around the same time, about an event that took place over 1 day. So due to our data constraints, we don't really have a good way to order sentences.

7. What criteria you consider when selecting the best sentences and best order, why you chose those, and how your solution optimizes for those criteria, to find a BestSummary.

Our scripts find a set of sentences that we deem are representative of our corpus articles.
We plan to choose from our pool of 'good' sentences by doing a combination of :

1. Look for YourWords - frequent words that occurred in our collection
2. Named entities from Stanford NER
3. Topics from LDA
4. Looking for answers to our template questions.

Selected results:

Class Event :
The wettest single month on record there is 14.07 inches, set in October 2005, according to weather.com.',
'TOTAL 24 HOUR PRECIPITATION AT ISLIP AS OF 9:30 AM WAS 13.57 INCHES.'
'THIS BREAKS THE PREVIOUS RECORD OF 13.78 INCHES SET BACK IN AUGUST OF 1990.',
'This broke the previous record of 11.6 inches, set at Tannserville, New York in August of 2011 during Hurricane Irene.',
'The all-time daily rainfall record there rains the 7.62 inches of rain that fell during the Great Chesapeake Potomac Hurricane of 1933, the National Weather Service said.',
'It is immediately apparent that the deepest convective clouds were not responsible for the heavy rains: cloud-top IR brightness temperatures over Islip were only near -30 C (per the 1200 UTC OKX sounding, that was around 300 hPa; the tropopause was closer to 150 hPa) and cloud-to-ground lightning was not detected.',
'THE 24 HOUR PERIOD FOR THIS RAINFALL EVENT WILL END AT 11PM ON AUGUST 13.'

We see that our method performed pretty well in selecting important sentences for the ClassEvent. It has, however included some noise, which is one point for improvement. Also, if we have a template, with the Points to cover, we need to verify if we really do have all the information about the rainfall that happened.

Small Event:

'A 28th person, believed to be Nancy Lanza, found dead in a house in town, was also believed to have been shot by Adam Lanza.'
'Twenty-six people were shot dead, including twenty children, after a gunman identified as Adam Lanza opened fire at Sandy Hook Elementary School.'
The shooter's mom was also a kindergarten teacherthere and was the target.",
'Authorities say gunman Adam Lanza killed his mother at their home on Friday and then opened fire inside the Sandy Hook Elementary School in Newtown, killing 26 people, including 20 children, before taking his own life.'
'The gunman also died at the scene, and a 28th body was found elsewhere.',
'The gunman also killed his mother and himself.Sen.',
'Many schools are already designated as "Gun free zones".Same can be said about a massive amount of drugs.'
Thats when the second officer shot and killed the man.'
'It was 9:25 in the morning when katelin roig began class.',
'Among his first victims, principal dawn hochsprung.',
['* Incident is among worst mass shootings in U.S. history.'
'A gunman at Virginia Tech University killed 33, including himself, in 2007.'

From these selected sentences, we can see that our method performed well in extracting sentences that have to do with our event of Newtown shootings, but not fantastically on keeping them strictly related to our event.We have a lot of sentences relating to gunmen and shootings, for example, the Virginia Tech shootings. This can be changed by changing our parameters for sentence extraction.

Big Event

'I am praying for Gabrielle and all of those who love her.'
4Suspect in attack on congresswoman acted alone911 call: "I do believe Gabby Giffords was hit.'
'"I was in shock," he said, describing his reaction to the shooting.',
'EDC Gear 1What Holster to Buy?'
'"Still, LeMole stresses that the danger for Giffords\' is far from over.'
'Diana Lopez from the Tucson Police Department.'

We see results that are not good for our chosen parameter values and cluster sizes. This is because the ratio of relevant content to noise is very skewed in our big collection. So, the parameters, cluster sizes and even cleaning methods that we use for our collections do not perform good enough for our big collection. These results can be improved by specifically taking into account the noisy nature of our big collection and making assumptions while cleaning and sizing our clusters.

Some statistics of our runs:

Collection	Size of all positive files(bytes)	Size of selected sentences(bytes)	Number of Selected Sentences	Size of Parsed Output (bytes)
ClassEvent	4096	12288	289	1297
Small	32768	1064960	14069	239592
Large	212922	491520	15001	79117

Custom Stopwords

facebook	like
twitter	post
reddit	account
todayilearned	subscribe
comment	tweet
comments	email
link	e-mail
askreddit	karma
google	upvote
bing	downvote
password	reddit
username	point
search	points
perma	permalink
minecraft	points1
points	points2
point2	point1

Best Summary Questions:

- Who is in the target audience?
    - What would each such person want to find in a summary?
    - How can involvement of such people help in finding the best summary?
        - Should you interview several such people? Ask them to critique your summaries?

Our target audience is those that are interested in information about certain shootings. These events are breaking news, so it is people that are interested in current events. Each person would want to know the who, what, where, when and why, along with several more shooting specific information. Involvement of people would help us refine our methods and change our template to include/look for the specific kind of information that they may be interested in, that we have overlooked.

- What are the successful scenarios for producing good summaries?
    - What disciplines (using what approaches and methods?) teach how to make good summaries?
        - Those related to: Communication? Literature? Linguistics?
        - Those related to: Science? Social science? Humanities? Engineering? Arts?

Discipline wise we believe that it is important to have a good balance between all of the disciplines. Of course, sometimes it depends on the topic. For a more scientific article, it would help to have more scientific skills to be able to provide a good summary. In our case, our events are general, and well known shootings. It is not really a specific discipline that is required to provide a good summary, but a good understanding of the type of event itself. In our case: Shootings, it helps to have knowledge of the important points of a shooting.

    - How well does it work to focus on answering questions like:
        - Who? What? When? Where? Why? How?

In our case, these are the most important questions to be answered. Our events are all breaking news. That's all anybody wants to know. Who is the shooter? Who was killed? What happened? When did it happen? Does anybody know why it happened? Where did it take place? How did he get away with it? Why did he do it?

    - How well does it work to be concerned about:
        - Avoiding bias, or at least identifying biased statements as to expressing a particular viewpoint
        - Having facts, or at least presenting alternative statements of facts along with their sources
        - Including discussion of emotions, sentiments, and other human considerations

In our case, bias really isn't an issue. There's not too much to be biased on related to a shooting, unless you get into the age old "gun control" debate, which we are not focused on. We could take a look at emotions, but we don't believe they have a significant impact on summarizing our corpus.

- How and to what degree are spatial (covering all types of mathematical spaces) concerns important?

    - How does geography fit, and how can that be expressed?
    - Is the phenomenon local or global? Centralized or distributed? Concentrated or spreading?
        - the phenomenon is local.
    - Are boundaries based on considerations of: nature, politics, culture, language, economics, etc.
    - Is consideration of probability relevant: uncertainty, odds, randomness, variance, etc.
    - Are there applicable feature spaces?

Our phenomenon our mostly concentrated, unless we look at the big picture. We can broaden our perspective and look at the relationships between shootings, or try to find connections between shooters and the areas in which they enacted violence. In some cases we could also take a look at connections of religion, language, economics, and politics. Our large collection is the Gabriel Giffords shooting. She was a political figure, so we could take a look at the relation or motive with politics. But this approach does not generalize well to our other events, so we chose to stay away for now.

- How well is structure addressed?
    - Of the domain? Event-type?
    - Of the organization of the knowledge about the event?
    - Of the summary?

We believe that the structure of our event type is addressed pretty well. Shootings have a simple structure, and we can summarize them quite well. We will be taking a more in depth look at this later with templates.

- How well are stream-related considerations addressed?
    - Is the sequence of (sub)events covered so there is an accurate historical record?
    - Is the flow / ordering correct?
    - When there are multiple streams involved, are they each identified and covered appropriately?

We do not need to worry too much about stream events, since our events take place in such a short period of time. Usually within a small matter of hours. The articles that are released about these events usually cover the entire thing, rather than any subsets of time.