Meeting Notes

Daily Stand Up Meeting Notes

TUE JULY 01

* Research on the various options for technical methods/which is best-suited for our purpose
* Prof. Satadisha will get back to us about the locked pdf
* What are topics/generate these from articles (Nicole’s question; not sure what she asked)/what kind of topic classifications do we want for grievances
* Find keywords/phrases that are constantly re-occurring in complaints and use these to train for their categorization
* Cleaning/Stop-word ‘removal’; word frequency
* Text topic classification/ Specific topic classification model
* Generate a distribution e.g this grievance/the entire article is like 60% topic 1, 20% topic 2, 10% topic 3 etc (spread/variation of data-type amongst the classifications)
* Satadisha will provide some resources on topic modeling (Pass a few links on to us)

EXPECTED TIME-LINE??

* Division of Labour; a system to pass data on from a sub-section of us to another sub-section who’s actively working with the data
* Teams: Industry-specific; Technical research (Code dissection; how to separate the topics); Dataset exploration (visualize and key word analysis)

-Identify the different keywords within the data

-Manually read some of the entries; additional details provide additional context

-Topic classification usually unsupervised

WED JULY 02

* All the models can separate into topics and classify new data, so no need to combine
* Try both LDA and LSA and compare the outcomes, they are unsupervised so don’t worry about the smaller dataset concerns
* Try word cloud

THUR JULY 03

* Document intermediate steps (like us checking words)
* Try building function to filter out topics, assign a list of documents to a topic (visa versa)
* When sure about topics, visualize the distribution of words per topic. Should help us give manual labels to each topic
* We want to look into phrases (anagrams?)
* We can do LSA and LDA same steps with anagrams, we set our size to trigrams? (n-grams)
* Satadisha will send some n-gram resources
* Try without removing any words
* Run with TF-IDF to handle stop words
* We should be doing LSA and LDA with TF-IDF
* Review the NLTK packages we are using

**Check Out:**

— Progress your group made on To-Do items discussed in stand up

— Challenges you faced today and how you overcame them or are still struggling

— Remaining To-Do's you will keep working on going forward

**Progress:**

* Started with a pLSA model and were able to improve it to give us solid differentiable topics, whilst also modifying it to accept the stopwords we had identified using BOW and TF-IDF scores.
* Began an organized way to document output from each model.
* Created two CSVs which display all summaries associated with the top 25 words from BOW and TF-IDF.
* Continuation of Background research.
* Explored the possibility of using coherence scores on models to decide on the optimum number of topics.

**Challenges:**

* We are trying a LLM problem as well, and are having some embedding issues with openAI.
* The pLSA implementation we found had great visualizations that we are trying to replicate for the other models, but since these were built-in for the *psla* library, we are struggling to do similar visualizations for the other models.

**TO-DO’s:**

* Coherence scores??
* Complete “When We Lost the Forest, We Lost Everything” notes
* Explore n-grams
* Finish LLM-powered topic modeling
* Properly compare topic modeling results and decide which to continue with
* Start slides for lightning talk

MON JULY 07

* LLMs thing with Helena. Not extremely sure. It seemed sad :(
* (Uriel )pLSA model; trying to replicate sth for another model (LSA\_TFIDF)
* Our final product should be able to take in a new document and without shifting everything into new categories, either sort the new grievances in the pre-existing categories or create new categories(if the new grievances are so distinct) in which only the new grievances are sorted
* Play around with number of topics specificity

**Check Out:**

We are currently reviewing and upgrading all five of our models. So far, we’ve implemented TF-IDF across each model and begun generating visualizations and calculating coherence scores for comparison. Below is a summary of our progress:

pLSA: It’s been challenging to find coherence scores due to the lack of documentation.

LLM model: Helena is working on this and is looking into how to calculate coherence for it.

Literature review: Nicole is conducting additional research to better understand the nature of the grievances and how they relate to the topics.

N-grams: Kamare has begun working on n-grams and will continue tomorrow. Theron started building a model with n-grams, completed a working example, and plans to apply it to our dataset tomorrow.

So far, both the LSA and pLSA models have partial visualizations, but they need refinement to improve readability and allow for meaningful comparisons.

Uriel did a great job with improving the LSA model and getting the classification for each grievance.

Tomorrow, we’ll continue refining the models, finalizing visualizations, and ensuring coherence scores are calculated for all five models.

| pLSA | LSA | NMF | LDA | LLM-BERT |
| --- | --- | --- | --- | --- |
| TF-IDF✅ | TF-IDF✅ | TF-IDF✅ | TF-IDF✅ | TF-IDF✅  (INTERNAL) |
| VISUALIZATION | VISUALIZATION (somewhat done) | VISUALIZATION | VISUALIZATION | VISUALIZATION |
| COHERENCE SCORES | COHERENCE SCORES ✅ | COHERENCE SCORES ✅ | COHERENCE SCORES ✅ | COHERENCE SCORES |

TUE JULY 08

* Either everyone uses lemma or everyone drops it
* Concern about TF IDF bias towards rare words
* Coherence score changing in LDA TFIDF
* Setting a seed
* Outlier topic (-1) that didn't fit into anything in LDA\_TFIDF
* Ask for more data to test
* Theron using a supervised model for bigram; trying to find model for unsupervised
* Theron discovers that you can apply tf idf vectorizers to every model (with varying results)
* Yuliia complains about the recurrence of “land” throwing off our results

WED JULY 09

* Rather than just banishing words to “stopwords”, we can vary the weights assigned to certain words

FRI JULY 11

July 11

* Final presentation is 15min length
* Add a slide briefly explaining why we need to do topic modeling, because we don’t have labels
* Think about the story we want to tell with the slides
* Try out notion or onenote
* Name some of the topics officially
* Mount shared google drive??

David said:

* Mention the visualizations we have when we talk, the picture of deforestation before and after was important and should be highlighted
* Emphasize the size of the dataset and remind the audience

FUTURE SUGGESTIONS

* Follow the story of a single example farmer to demonstrate what life is like pre- and post- our project
* Push code to github (perhaps link it on LinkedIn for online professional presence)

MON JULY 14

* Bertopic documentation, try using k nearest neighbors in the model
* Check satadisha link to different clustering options
* Initialize a specific clustering technique instead of hdb scan
* We could also not do any dimensionality reduction or try pca/svd (but don’t touch this idea until we figure out the clustering)
* Satadisha likes Nicole’s idea about documents having more than one topic label, and we should look into how we could add this to the code
* … add topics to a doc until we reach like 50% of the total explanation of the document
* If a doc has more than 50% of the total topics at that point, then it is probably an outlier, this will help us identify outliers
* Logic, if dominant topic is larger than 50% then assign to 1, if not assign to 0 something like that

TUE JULY 15

Uriel & Helena mentioned the thing about the column with no values that does weird things to the model running in LLM

Worry about lack of probabilities interfering with visualizations; no distribution just getting 0 and 100

Dimensionality reduction; U-map??

Clustering; Agglumeration?

Parameters in BERTtopic

WED JULY 16

* Try soft kmeans, like [Gaussian Mixture Model](https://en.wikipedia.org/wiki/Mixture_model) (megan put in chat), bert has a couple requires a clustering algorithm that has fit and predict

MON JULY 21

Begin abstract

Have all experimental related work finished by August 1

TSNE

Virtual Check Outs

Virtual Check Outs

CO 7/3

**7/3/25**

Progress:

* Started with a pLSA model and were able to improve it to give us solid differentiable topics, whilst also modifying it to accept the stopwords we had identified using BOW and TF-IDF scores.
* Began an organized way to document output from each model.
* Created two CSVs which display all summaries associated with the top 25 words from BOW and TF-IDF.
* Continuation of Background research.
* Explored the possibility of using coherence scores on models to decide on the optimum number of topics.

Challenges:

* We are trying a LLM problem as well, and are having some embedding issues with openAI.
* The pLSA implementation we found had great visualizations that we are trying to replicate for the other models, but since these were built-in for the psla library, we are struggling to do similar visualizations for the other models.

TO-DO’s:

* Coherence scores??
* Complete “When We Lost the Forest, We Lost Everything” notes
* Explore n-grams
* Finish LLM-powered topic modeling
* Properly compare topic modeling results and decide which to continue with
* Start slides for lightening talk

CO 7/10

**7/10/25**

Progress:

* Decided presentation split for talk
* Finishing up the slides for our lightning talk and practicing
* Fine tuning visualizations for presentation and future use
* Created plan of action graphic for presentation and future use

Challenges:

* We struggled a bit to cut down on what we talk about tomorrow, we have a lot of information and work to show at the moment

TO-DO’s:

* Presentation
* Watch videos provided by Satadisha to learn more about each model
* Finish up any last tweeks to the models, visualizations, and coherence scores before progressing to n-grams on Monday