A.L.I.C.E. Midterm Report

Cornelius Yap, Samuel Lye, Tiffany Goh

I. Executive Summary

Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information and to consume relevant information faster.

Analyzing Language Interface Created for Everyone (A.L.I.C.E.) is purposed to summarize text documents and output informative visualization displays that quickly and easily communicate the contents of the text to the user.

The key components of A.L.I.C.E. are the Frontend (React, D3), Backend (Flask, Machine Learning models), as well as DevOps (Docker, Openshift).

II. Objectives (Learning/Project goals)

a. To study and research into the most optimal natural language processing (NLP) and visualization tools to create an informative text summarizer.
b. To conduct Proof of Concepts (POC) testing through building of various NLP models using open source tools.
c. To develop greater understanding and knowledge of fundamental machine learning concepts and NLP tools to create a functioning end product.

III. Background

NLP technology is rapidly advancing due to an increased interest in human-to-machine communications coupled with an availability of big data, powerful computing and enhanced algorithms. While Deep Learning Models like Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) were leading NLP models, attention modelling has shown to produce state-of-the-art results in machine translation and other NLP tasks when combined with neural word embeddings. In addition, Transformers use attention mechanisms to gather information about the relevant context of words and encode that context in vectors that represent the words to provide the enabling of contextual information essential to NLP enhancement.

As a starter, the scope of our project will first utilize existing NLP libraries, such as gensim, spacy, and nltk, before exploring the use of attention modelling during the MP phase.

During the EP phase, we will focus on training our models on one domain – COVID-19. In other words, we would train our machine learning models using COVID-19 news articles.

The end product of our text summarizer would be an input form for users to upload a document/documents from a single category. The output would be in the form of informative visualization displays that quickly and easily summarize the contents of the text to the user.

IV. Application of A.L.I.C.E.

a. Quick summary of documents from different domains.
b. Comprehensive visualization tool that maps out the relation between different entities in the text.
c. Applicable to the security domain, where our model if trained on security-related domain, would be able to extract key information from sensitive documents and allow users to quickly understand the security concerns with regards to the documents.

V. Key Components

a. Topic Modelling

- Topic Modelling is the process of using unsupervised learning to extract main topics, represented as a collection of words, found in a document.
- We have tested 3 unsupervised learning approaches as follows:
    1. Latent Dirichlet Allocation (LDA)
        - LDA is a probabilistic model. It models each document as a Dirichlet distribution of topics and each topic as a Dirichlet distribution of words. Through an iterative process, it searches for the best topic mix and best word mix.
    2. Non-Negative Matrix Decomposition (NMF)
        - NMF decomposes high dimensional vectors into lower dimensional vectors. In this context, NMF decomposes the article-word matrix to article-topic matrix and topic-word matrix. The two lower dimensional matrices are optimised over an objective function for example Euclidean distance and updated iteratively until convergence.
    3. Principal Component Analysis (PCA), followed by Term Frequency Inverse Document Frequency (TF-IDF)
        - Documents are represented as word embeddings with a length of 7168. PCA is used to reduce the dimensionality of the word embedding to length 768, speeding up agglomerative clustering. Subsequently, we find words

or phrases in each topic with the highest TF-IDF scores to make sense of each topic.

- The approach adopted is NMF. This is because the topics could be more accurately deciphered from the list of words outputted from the NMF model. In other words, the list of words provided by the NMF model consists of more significant words. Moreover, given a small corpus, the NMF model is less likely to generate a repeated list of words for different topics.
- Libraries used include Gensim and Sklearn.

b. Named Entity Recognition (NER)

- A named entity is a word or a phrase that clearly identifies one item from a set of other items that have similar attributes. Examples of named entities are organization, person, and location names in general domain; gene, protein, drug and disease names in biomedical domain. NER is the process of locating and classifying named entities in text into predefined entity categories.
- The main datasets that NER models are trained and evaluated upon are the CoNLL03 and OntoNotes5.0 datasets.
- There are 4 main streams of NER:
    1) Rule-based approaches, which do not need annotated data as they rely on hand-crafted rules;
    2) Unsupervised learning approaches, which rely on unsupervised algorithms without hand-labeled training examples;
    3) Feature-based supervised learning approaches, which rely on supervised learning algorithms with careful feature engineering;
    4) Deep-learning based approaches, which automatically discover representations needed for the classification and/or detection from raw input in an end-to end manner.
- A list of existing NER libraries and comparison of the performance of different NER models can be found in Appendix I.
- Therefore, we explored the use of deep-learning based approaches, such as BERT and Transformers. We decided to use Flair, which is a new framework for sequence labelling (NER and parts of speech tagging). By learning to predict the next character on the basis of previous characters, Flair is able to leverage the internal states of a trained character language model to produce a novel type of word embedding which its developers refer to as contextual string embeddings. These embeddings have the distinct properties that they (a) are trained without any explicit notion of words and thus fundamentally model words as sequences of characters, and (b) are contextualized by their surrounding text, meaning that the same word will have different embeddings depending on its contextual use. Flair

achieved one of the higher performance (89.3%) on OntoNotes5.0 and the state-of-the-art performance (93.18%) on CoNLL03.

Task	Language	Dataset	Flair	Previous best
Named Entity Recognition	English	Conll-03	**93.18** (F1)	*92.22 (Peters et al., 2018)*
Named Entity Recognition	English	Ontonotes	**89.3** (F1)	*86.28 (Chiu et al., 2016)*
Emerging Entity Detection	English	WNUT-17	**49.49** (F1)	*45.55 (Aguilar et al., 2018)*
Part-of-Speech tagging	English	WSJ	**97.85**	*97.64 (Choi, 2016)*
Chunking	English	Conll-2000	**96.72** (F1)	*96.36 (Peters et al., 2017)*
Named Entity Recognition	German	Conll-03	**88.27** (F1)	*78.76 (Lample et al., 2016)*
Named Entity Recognition	German	Germeval	**84.65** (F1)	*79.08 (Hänig et al, 2014)*
Named Entity Recognition	Dutch	Conll-03	**90.44** (F1)	*81.74 (Lample et al., 2016)*
Named Entity Recognition	Polish	PolEval-2018	**86.6** (F1) (Borchmann et al., 2018)	*85.1 (PolDeepNer)*

c. Relation Extraction

- Relation extraction is the process of extracting semantic relationship from text, between two or more entities. Extracted relations usually occur between two or more entities of a certain type (e.g. Person, Organisation, Location) and fall into a number of semantic categories (e.g. married to, employed by, lives in).
- Initially, we utilized a rule-based relation extraction algorithm, i.e. Stanford Open Information Extraction. Stanford Open IE identifies relations between two entities in a sentence based on the contextual clues given in a single sentence. This is done through parts-of-speech (POS) tagging in the given sentence and using the tags to identify certain patterns that might indicate a relation. For example, the "Noun-Verb-Noun" POS-tag pattern might indicate a verbal relation such as "John went out with Tom". However, a limitation of this approach is that if the sentence is too complicated, the basic relation extraction feature might not be able to identify the relation even if it is present in the sentence.
- For the MP stage, we looked at supervised relation extraction. The models used are trained using the training and validation sets of FewRel (Han et al. 2018). The dataset contains 80 relations defined in Wikidata (Refer to Appendix II) and the corpus of totally 56,000 sentences comes from Wikipedia. Our model adopts a best-of-breed approach - Convolutional Neural Networks (CNN) work by first extracting semantic features from input sentences, obtaining the word embeddings and position embeddings of the sentences, and then sending them to the CNN to get the sentence representations. Finally those representations are fed to a fully-connected layer to calculate the probabilities for each relation; Bidirectional Encoder Representations from Transformers (BERT) (Devlin

et al. 2018) is a self-attention-based text encoder that achieves state-of-the-arts on several NLP benchmarks.

- A limitation of the supervised relation extraction approach is that the models will have to run for every combination of two entities within a sentence, i.e. if there are N entities in a sentence, the time complexity of our model is 2^N

d. Sentiment Analysis

- Sentiment analysis, also known as opinion analysis/mining, is used to extract subjective and opinion related information, such as emotions, attitudes, and moods. A polarity weight is given to the text, based on whether it expresses a positive, negative, or a neutral sentiment.
- There are two major techniques for sentiment analysis: supervised machine learning and unsupervised lexicon-based learning. Lexicons refer to dictionaries or vocabularies specially constructed to be used for sentiment analysis and compute sentiment without any supervised techniques. Some examples are AFINN lexicon, Bing Liu's lexicon, MPQA subjectivity lexicon, SentiWordNet, VADER lexicon, and Pattern lexicon.
- Our current approach is to use unsupervised lexicon-based learning. We have tried using current available APIs provided by NLTK which allows us to perform sensitivity analysis using the Pattern, SentiWordNet and Vader lexicons. We also used a separate python library, TextBlob, to perform sentiment analysis, and compared the performance between the 4 different types used. For our current domain of Covid-19 news articles, all 4 methods have approximately the same performance and are fairly accurate in determining the polarity and objectivity of the text given.
- For the future versions of ALICE, a possible approach will be to allow the user to select the domain of the text and ALICE will use the method that is the most accurate for that specific domain. A separate approach will be to create our own sentiment analysis tool using the supervised machine learning method.

e. Classifier

- Text Classification refers to categorising documents into either of the 6 categories: Crime, Technology, Health, Finance, Terrorism, Politics.
- Below are models we have tested and their respective accuracy:

	Machine Learning Models	Deep Learning Models
1.	Logistic Regression (0.84)	Word Embedding + Bidirectional LSTM (0.75)

2.	Linear SVC (0.82)	XLNET (0.78)
3.	Naive Baye (0.88)	

- Naive Baye was trained as a binary classification model as it achieved a higher accuracy as compared to a multi-class classification model. This means that a single document needs to be passed into 6 binary classification models, one per domain to determine which domain the document belongs to. However, since the Naive Baye model implemented using Sklearn directly outputs whether a document falls in a certain domain rather than the probability of the document belonging to a domain, we were unable to determine the best category for the document. Through this experience, we understood the need for multi-label models.
- Searching for large samples of labelled data was challenging. Although external datasets such as AG News Topic Classification Datasets were available, there were typically merely a short extraction of the news and covered only a few domains out of the six domains mentioned above. Hence, it was difficult to improve the accuracy of deep learning models which learn from large datasets. We have made several attempts to combine various external datasets to create a larger and more comprehensive dataset. Unfortunately, since the news corpus were obtained from different news sources, the varied writing style could be a plausible reason for the minimal increase in accuracy of the models. An alternative solution was to use pre-trained models and fine tune them as fine tuning requires a smaller dataset as compared to training a model from scratch.
- We also looked into combining deep learning models to classify documents of variable length. We combined BERT with LSTM. We split each document into smaller text, 200 words each, with 50 words overlapped. Subsequent, we fed the smaller text into BERT to encode them. Before passing to LSTM, our classifier, we padded the shorter sequence with a special value to be masked. This ensures accurate encoding of documents with variable length.

f. Clustering

- Given a corpus, we would cluster documents with similar content together.
- The clustering algorithm used is K-Means. Upon deciding the desired number of clusters, k, K-Means starts with a group of k randomly selected centroids. They are used as the beginning points of each cluster and then perform iterative calculation to optimise the position of the centroids.
- The optimal k value was determined by performing K means with different k values and finding the k value that has the highest Silhouette value.
- We tried two ways to encode the document, namely using TF-IDF and Doc2Vec. Doc2Vec is an extension of the Word2Vec approach. It adds an additional input node

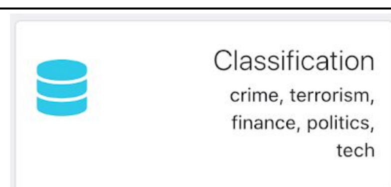
representing the document as an additional context in predicting the next word in a sentence.
- The clustering model chosen is Doc2Vec, along with K means.

g. Docker

- Docker is an open platform for developing, shipping, and running applications. Due to using a variety of ML modules which require different Python modules, we adopted the use of Docker containers to create isolated file systems for our ML modules. A Docker image is a file, comprised of multiple layers, that is used to execute code in a Docker container. An image is essentially built from the instructions for a complete and executable version of an application, which relies on the host OS kernel. A Docker container is a runnable instance of an image. Containers are lightweight because they don't need the extra load of a hypervisor, but run directly within the host machine's kernel.
- A.L.I.C.E. is comprised of the following Docker containers: Summarizer, Classifier, Sentiment, Topics, Wordcloud, NER, Relation Extraction, Clustering, Backend Flask server, and our Frontend React and Node server.
- For future plans, we aim to utilize Red Hat's Openshift platform that is powered by Kubernetes. Kubernetes provides the ability to run containers on various machines, auto-scale containers, distribute load between containers, manage storage required by containers, and provide resiliency of containers in case of failure. Kubernetes helps simplify container runtime by managing Docker-based applications that are placed on an underlying assures system that maintains several replicas of running applications. Kubernetes has enabled developers to accelerate the development of cloud-native applications and created an ecosystem of services that are self-driven and reusable.

h. Visualization

- Visualization is the key component of the frontend of A.L.I.C.E. This enables the user to quickly and easily gain an overview of the content of the text document.
- We utilized the following visualization libraries for A.L.I.C.E.: Nivo and React Force Graph (both built on top of D3).

Classifier a. Classifies the document into any of the following six categories: Health, Crime, Terrorism, Finance, Politics and Tech	

<table>
<tr>
<td>

Text summarizer

a. The text summarizer takes the most important sentences of the document as the summary. Currently the default settings displays 4-5 lines as the summary

</td>
<td>

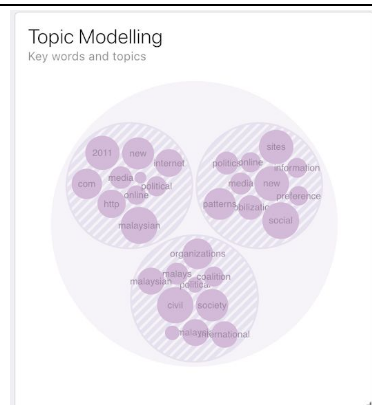
</td>
</tr>
<tr>
<td>

Topic Modelling

a. The larger bubble represents a topic and the smaller bubbles within it represents the words that make up the topic

b. The user has to infer what the topic is based on the words that are given

</td>
<td>



</td>
</tr>
<tr>
<td>

Wordcloud

a. Wordcloud is an image that is made up of the most frequent words in the document.

b. The larger the words, the higher the frequency count in the document.

c. Gives the user a brief idea on what the document is about

</td>
<td>

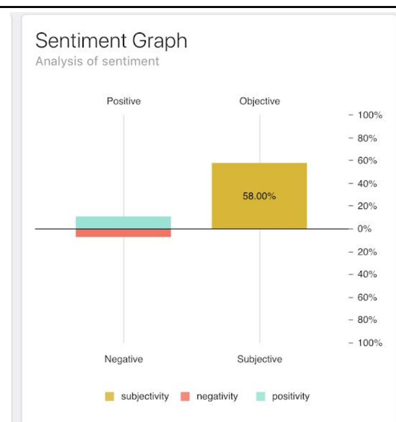
</td>
</tr>
<tr>
<td>

Sentiment Graph

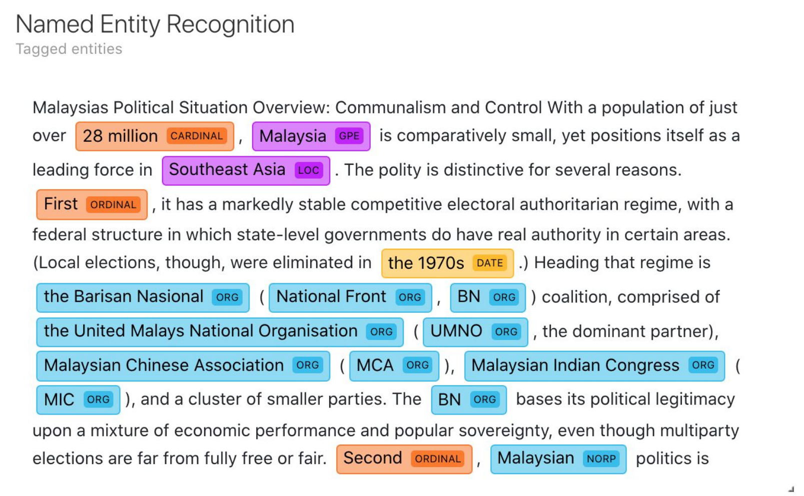
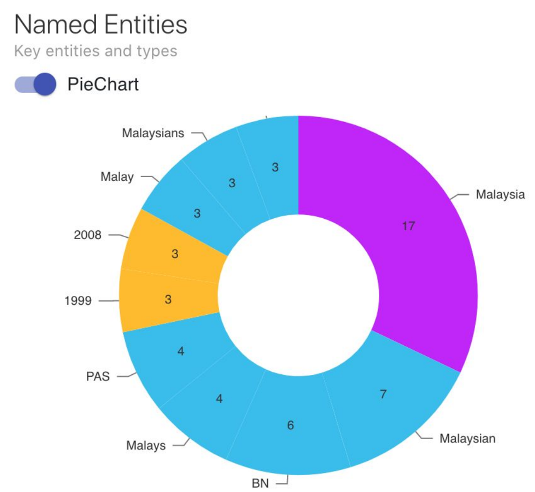
a. The sentiment graph provides more details about the positivity and subjectivity of the document.

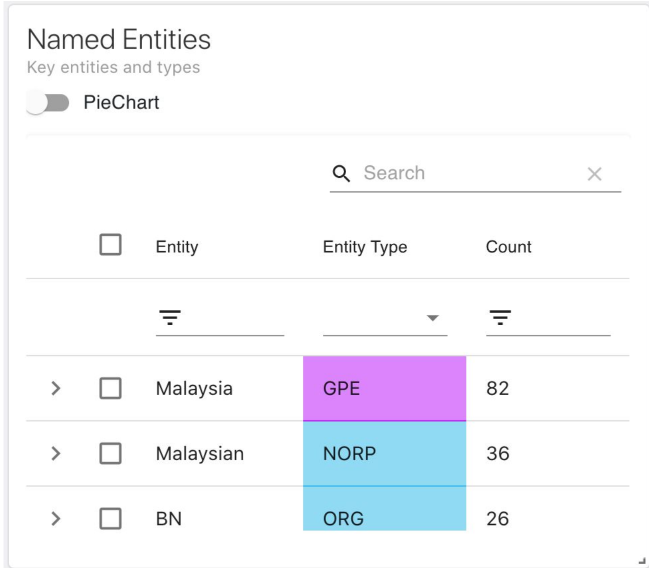
b. The left bar represents the positivity. A positive article will have the bar above the x-axis while a negative article will have the bar below the x-axis. The length of the bar represents the magnitude of the sentiment.

</td>
<td>



</td>
</tr>
</table>

c. The right bar represents the subjectivity. An objective article will have the bar above the x-axis while a subjectivity article will have the bar below the x-axis. Likewise, the length of the bar represents the magnitude.	
Named Entity Recognition a. The entity visualizer highlights named entities and their labels in a text. b. The entity pie chart displays the top 10 entities within the document, so that users can focus on the most frequent entities. c. The entity table allows the user to filter and sort the various entities within the document. The entity table differs between the Overview and individual document dashboards. - In the Overview dashboard, the NER table allows the user to filter entities in the network graph. The NER Table also allows the user to know which document the entity appears in. - In the individual document dashboard, the filtering of entities in the network graph is done by the relation table because of the smaller scale of relations in the individual documents. Also, the NER table allows the user to edit the entity type and delete non-entities.	 



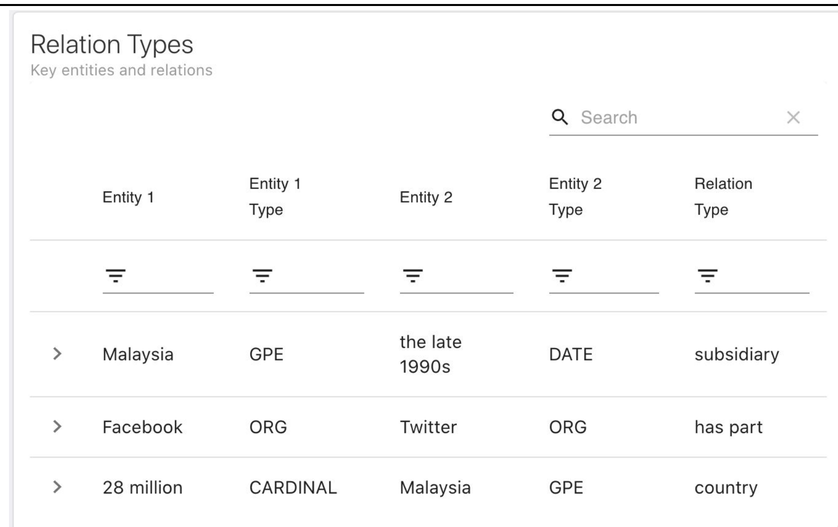
Named Entities
Key entities and types
PieChart

Entity	Entity Type	Count
Malaysia	GPE	82
Malaysian	NORP	36
BN	ORG	26

Relation Extraction

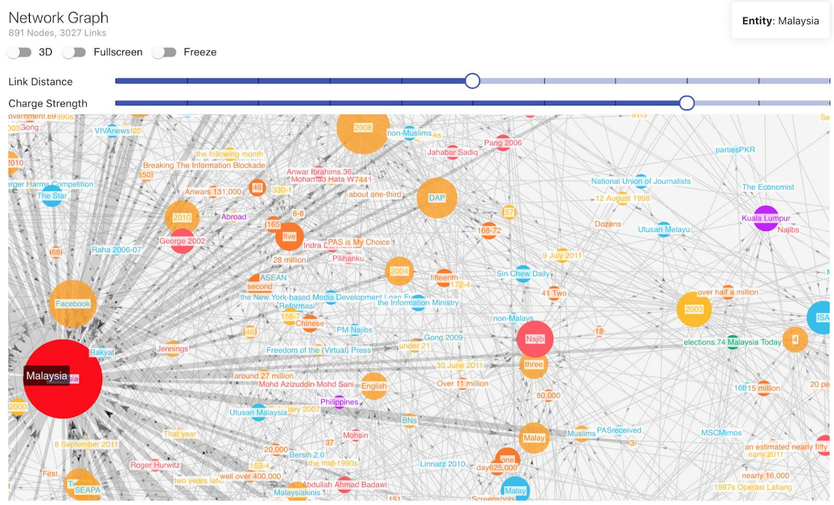
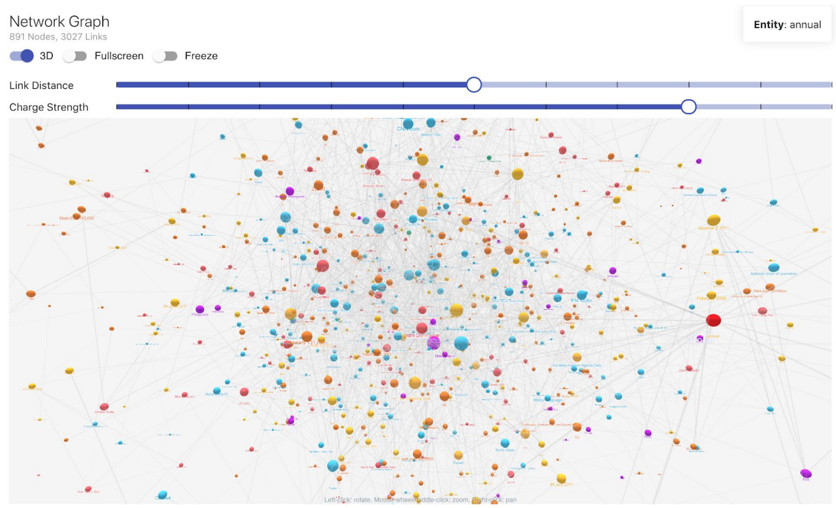
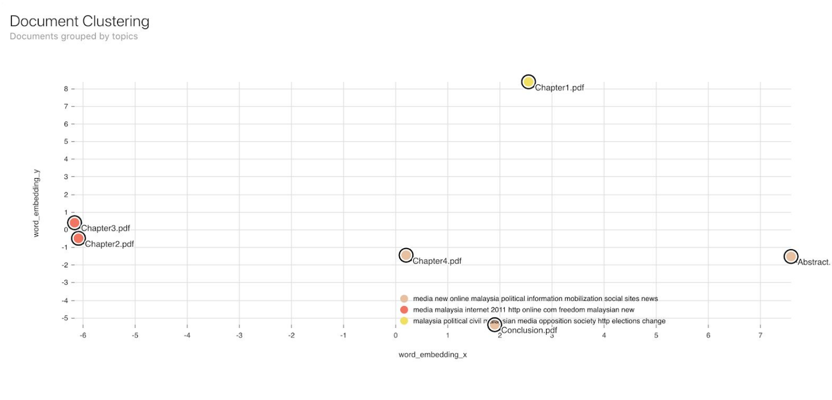
a. Similar to the NER table, the relation table has different functionalities between the Overview and individual document dashboards. Both tables do allow for filtering and searching different relations.
    - In the Overview dashboard, the relation table allows the user to know which document the relation appears in.
    - In the individual document dashboard, as mentioned above, the filtering of entities in the network graph is done by the relation table because of the smaller scale of relations in the individual documents. Also, the relation table allows the user to delete the relation if it is irrelevant.

b. The network graph displays all the entities and relations via nodes and links respectively. The network graph has



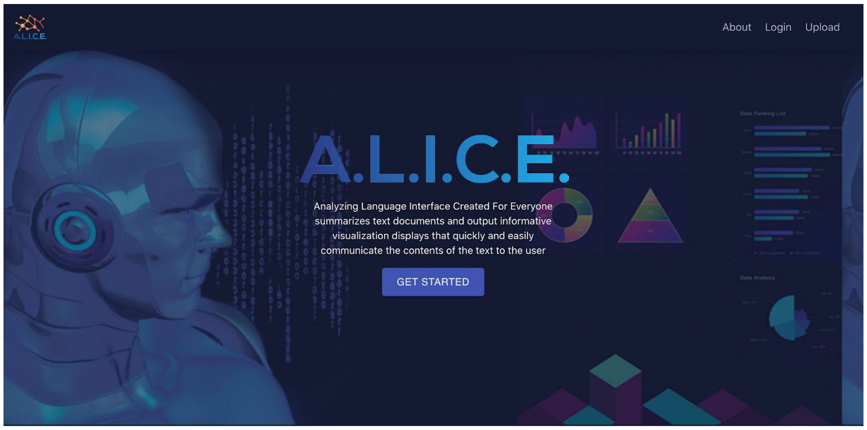
Relation Types
Key entities and relations

Entity 1	Entity 1 Type	Entity 2	Entity 2 Type	Relation Type
Malaysia	GPE	the late 1990s	DATE	subsidiary
Facebook	ORG	Twitter	ORG	has part
28 million	CARDINAL	Malaysia	GPE	country

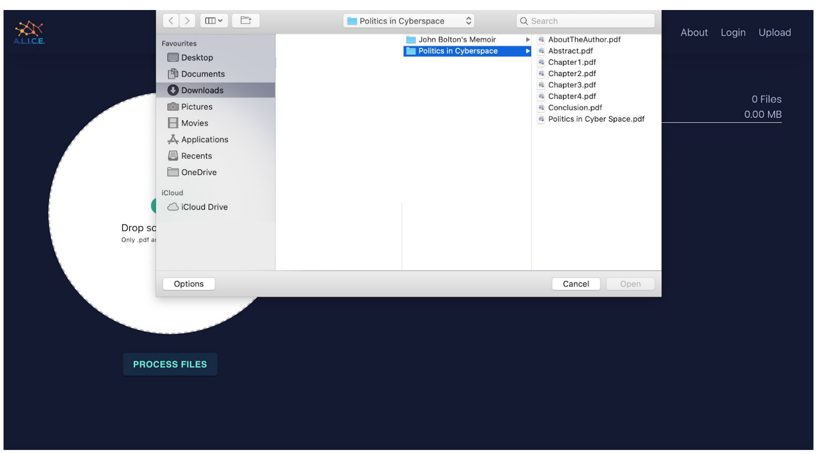
several options available to the user - 3D mode, fullscreen mode, freezing the graph so the nodes do not move, and changing the link distance and charge strength between nodes. The graph also informs the user on how many nodes and links there are in total.	
Document Clustering a. Shows how closely related each document is to each other. The closer the distance between the document node in the graph, the more similar the documents are.	

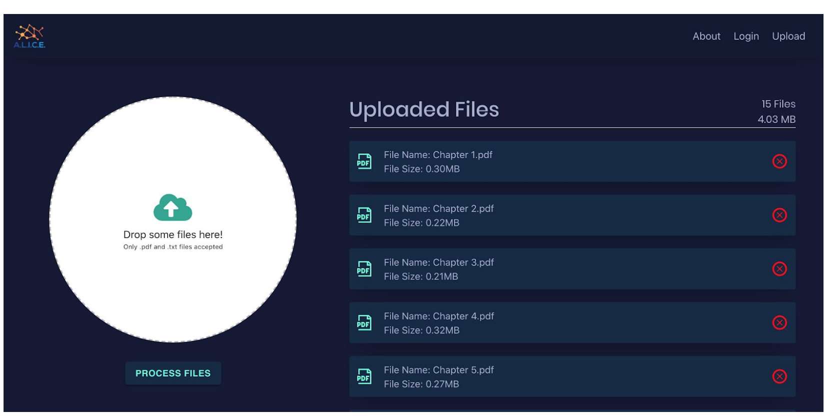
VI. Walkthrough

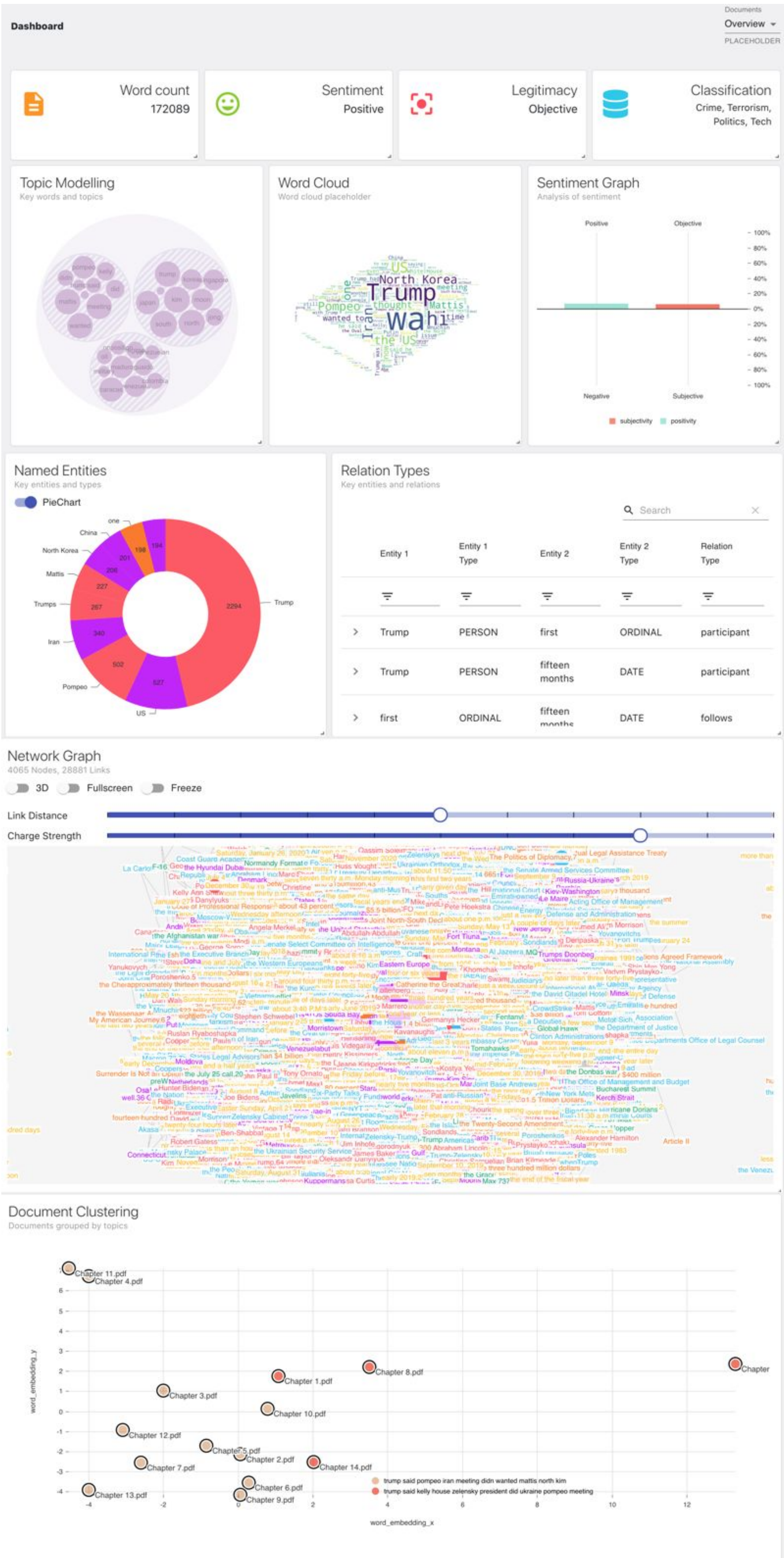
The first page users will see is the about page, which provides a short description of the different components of A.L.I.C.E.



The upload page allows the user to select and upload files or drag and drop files. The number of files and the total filesize will also be displayed.





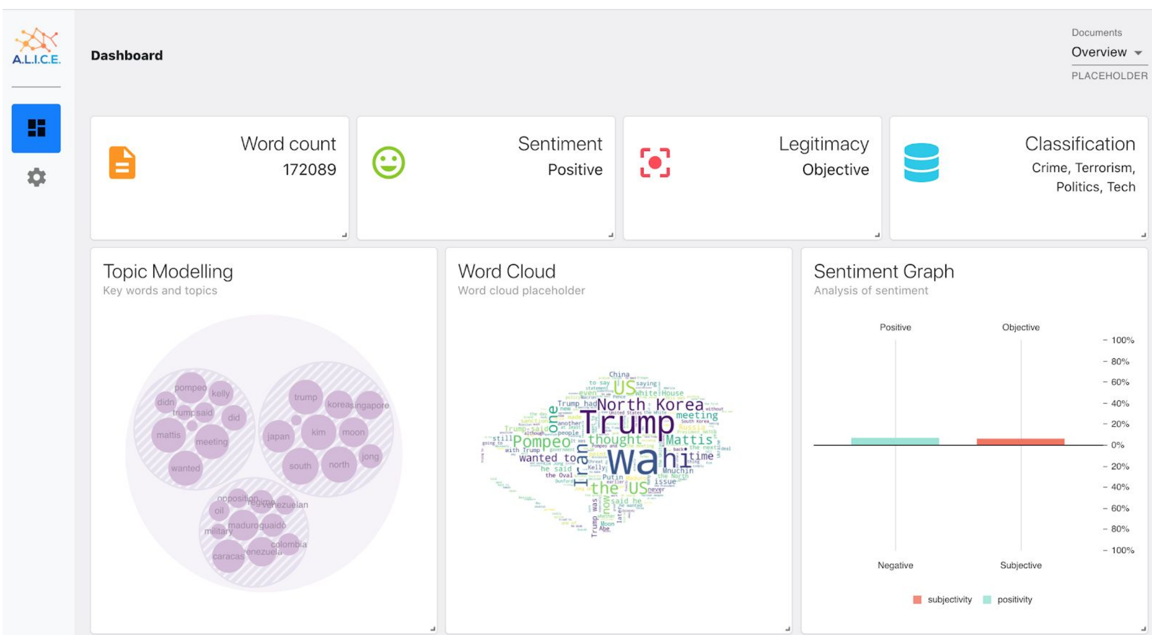

Dashboard
Documents
Overview
PLACEHOLDER
Word count
172089
Sentiment
Positive
Legitimacy
Objective
Classification
Crime, Terrorism, Politics, Tech
Topic Modelling
Key words and topics
Word Cloud
Word cloud placeholder
Sentiment Graph
Analysis of sentiment
Positive
Objective
Negative
Subjective
subjectivity
positivity
Named Entities
Key entities and types
PieChart
one
China
North Korea
Mattis
Trumps
Iran
Pompeo
US
Trump
Relation Types
Key entities and relations
Search
Entity 1
Entity 1 Type
Entity 2
Entity 2 Type
Relation Type
Trump PERSON first ORDINAL participant
Trump PERSON fifteen months DATE participant
first ORDINAL fifteen months DATE follows
Network Graph
4065 Nodes, 28881 Links
3D
Fullscreen
Freeze
Link Distance
Charge Strength
Document Clustering
Documents grouped by topics
Chapter 11.pdf
Chapter 4.pdf
Chapter 8.pdf
Chapter 1.pdf
Chapter 3.pdf
Chapter 10.pdf
Chapter 12.pdf
Chapter 5.pdf
Chapter 2.pdf
Chapter 7.pdf
Chapter 14.pdf
Chapter 6.pdf
Chapter 13.pdf
Chapter 9.pdf
trump said pompeo iran meeting didn wanted mattis north kim
trump said kelly house zelensky president did ukraine pompeo meeting
word_embedding_y
word_embedding_x

From the Overview dashboard, we can tell that the entire pdf document has 172089 words, has positive sentiment and is written objectively. The documents talks about crime, terrorism, politics and technology related issues.

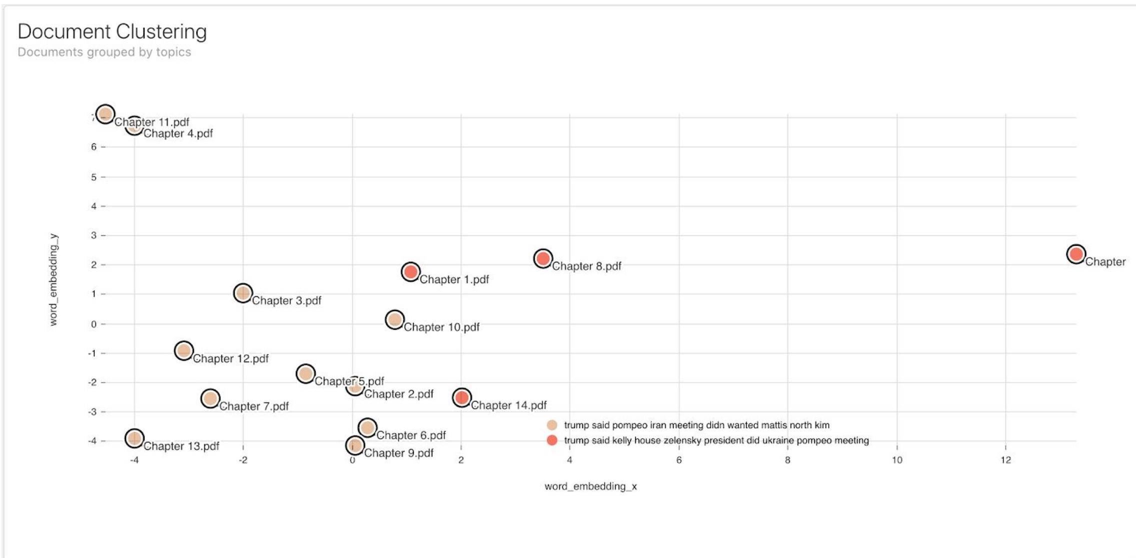
From the topic modelling chart, the first topic (top left bubble) tells us there is a meeting between Kelly, Pompeo and Mattis. The second topic (top right bubble) is related to USA involvement in Asia, in particular the USA-North Korea summit between Kim Jong Un and Donald Trump. The last topic (bottom bubble) is related to USA relations with its South American neighbors, such as Venezuela and Colombia, and the topics of oil and military.

Looking at the word cloud, we could infer that there were probably international meetings involving mainly US, North Korea and Iran. Political figures involved include Trump, Mattis and Pompeo. Washington (WA) is a significant location.

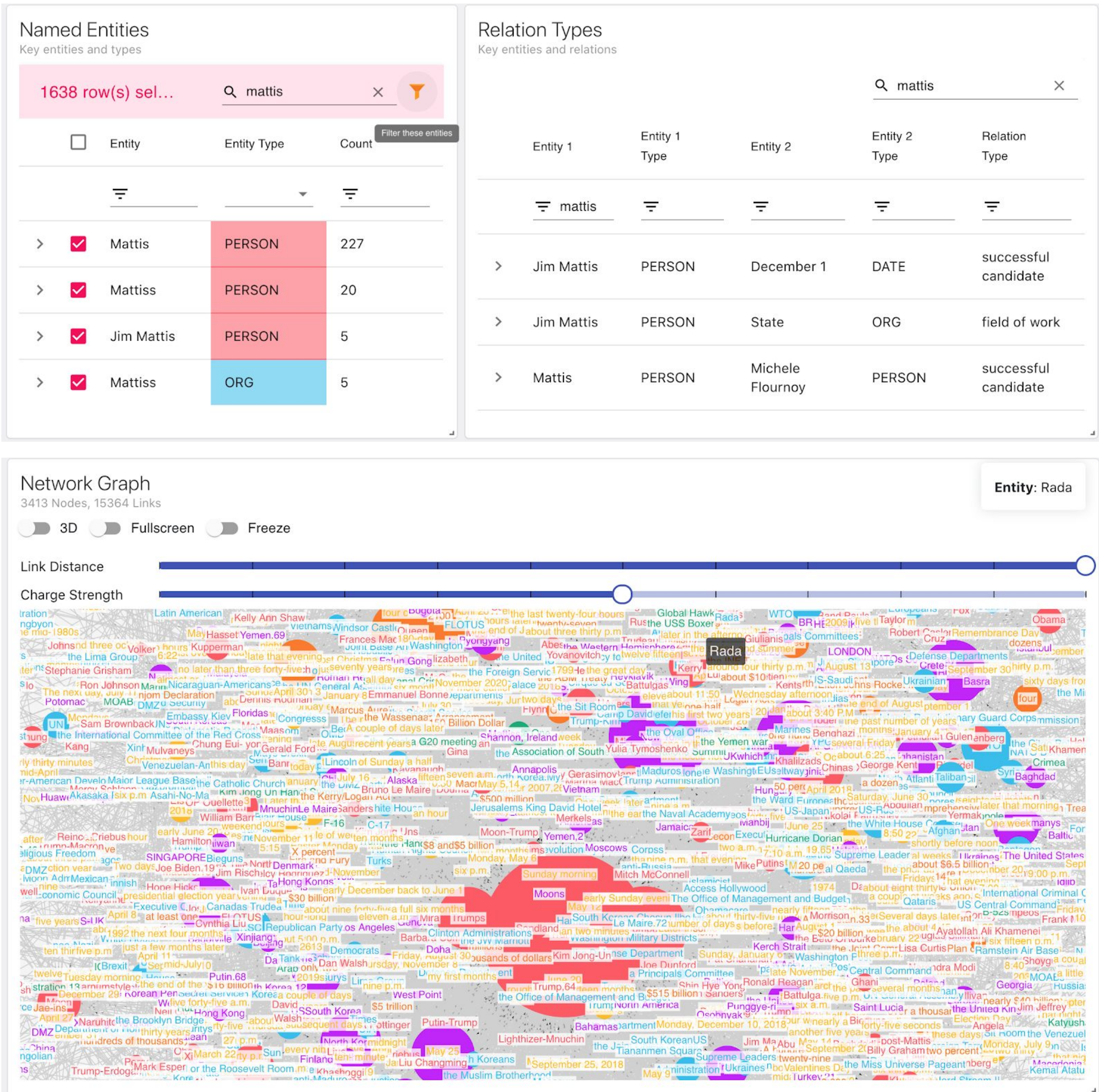
Overall, the sentiment of the entire document is slightly positive and slightly objective.



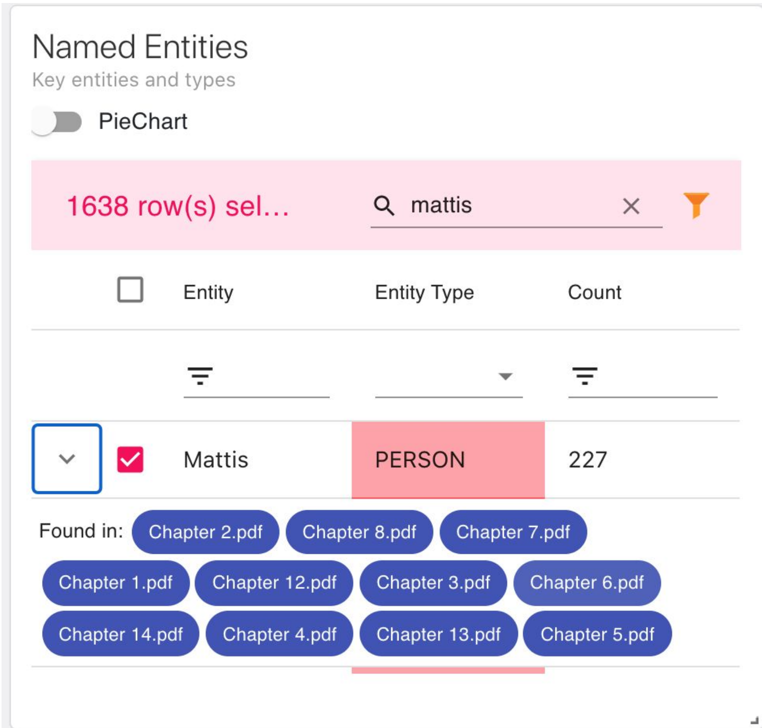
The document clustering scatter plot groups chapters with similar content together. Chapter 1, 8 14 and 15 form a cluster, talking about a meeting between Trump, Pompeo, Mattis and Kim Jong Un. The remaining chapter talks about another meeting between Kelly, Zelensky, Trump and Pompeo. In addition, chapters that are closer together are more similar in content. Hence, although Chapter 6, 9 and 14 are in the same cluster, perhaps Chapter 6 and 9 mentions more similar aspects of the meeting as compared to Chapter 9 and 4. This would provide users with a better idea of which chapter to further delve into.



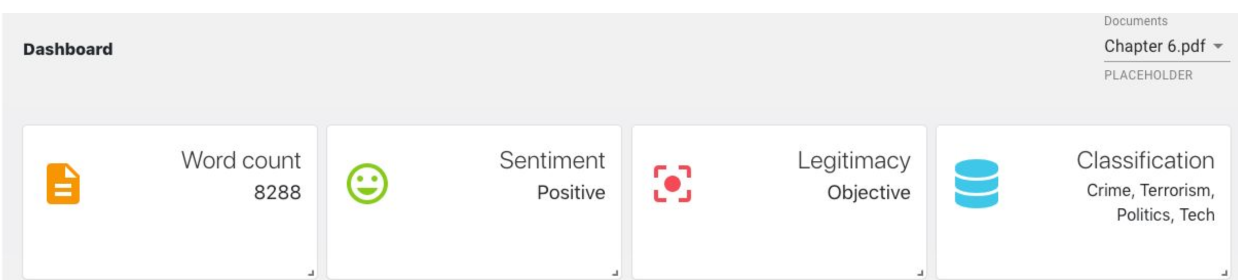
In the Overview dashboard, let's say we want to find out more about the entity Mattis. We can first filter the entity Mattis using the NER table as well as the Relation table. The NER table tells us the possible entity types of Mattis and the number of times the entity appears throughout the entire document while the Relation table tells us what kind of relations there are between Mattis and other entities.



However, there are still 3,413 nodes and 15,364 links in the network graph after filtering. To narrow down the search, we can zoom in on one of the chapters where the entity Mattis appears as depicted below. Clearly, the entity appears in many chapters of the document. Let's choose a chapter to focus on, for example Chapter 6.



This section of the dashboard provides a brief breakdown of chapter 6. We can see that the chapter has 8288 words, a positive sentiment and is written in an objective manner. It is also classified under the Crime, Terrorisim, Politics and Tech category, giving us a rough idea on what the chapter will be focusing on.



From the Sentiment graph, we can get more details about the positivity and subjectivity of the chapter. We can see that the article is slightly positive and slightly objective as well.

Inferring from the bubble chart, we can tell that Trump and Putin were in a meeting discussing a treaty called INF. They might be talking about the withdrawal from the treaty. A person called Mattis is involved, exactly how is still unsure.

The word clouds displays an image of the most frequent words that appear in the document. This helps us to further understand and analyse what the chapter is about. We can see that the US, Trump and Russia are being mentioned a number of times. We can also see that another person called Mattis is involved as well as the INF treaty.

Word Cloud

Word cloud placeholder

Based on what we gather from the topic modelling and the word cloud, we can tell that the chapter will be about talks between Trump and Putin about a possible withdrawal from the INF treaty. Now we can read the summary of the document to get a better understanding on what the INF treaty talks might be about.

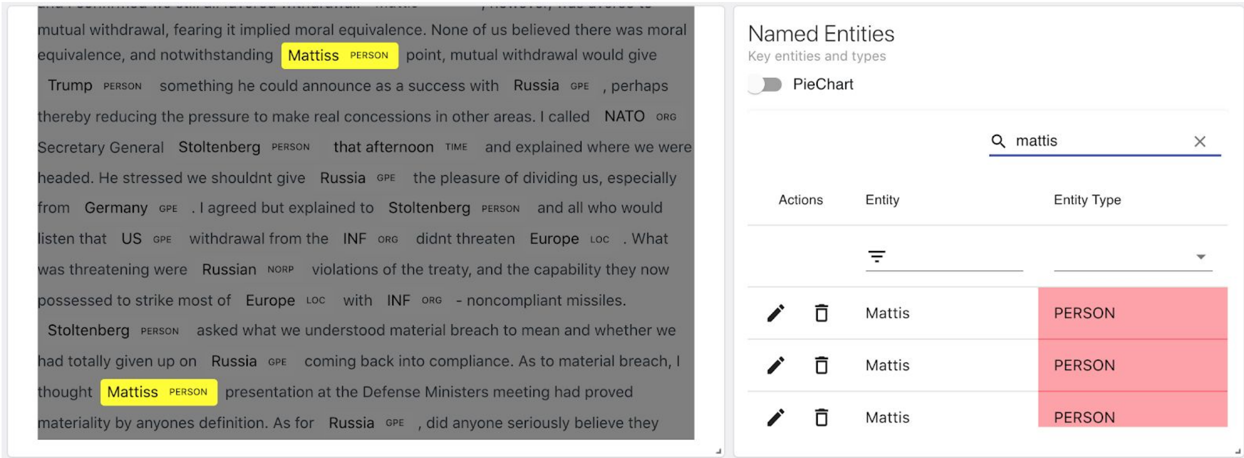
From the summary we can tell that Putin does not wish to pull out from the INF treaty, however, Trump will be terminating the INF treaty at the request of the NATO allies. We can also see that Trump does not wish to pull out from New START but at the same time cannot allow it to be extended for another five more years.

Summary of document

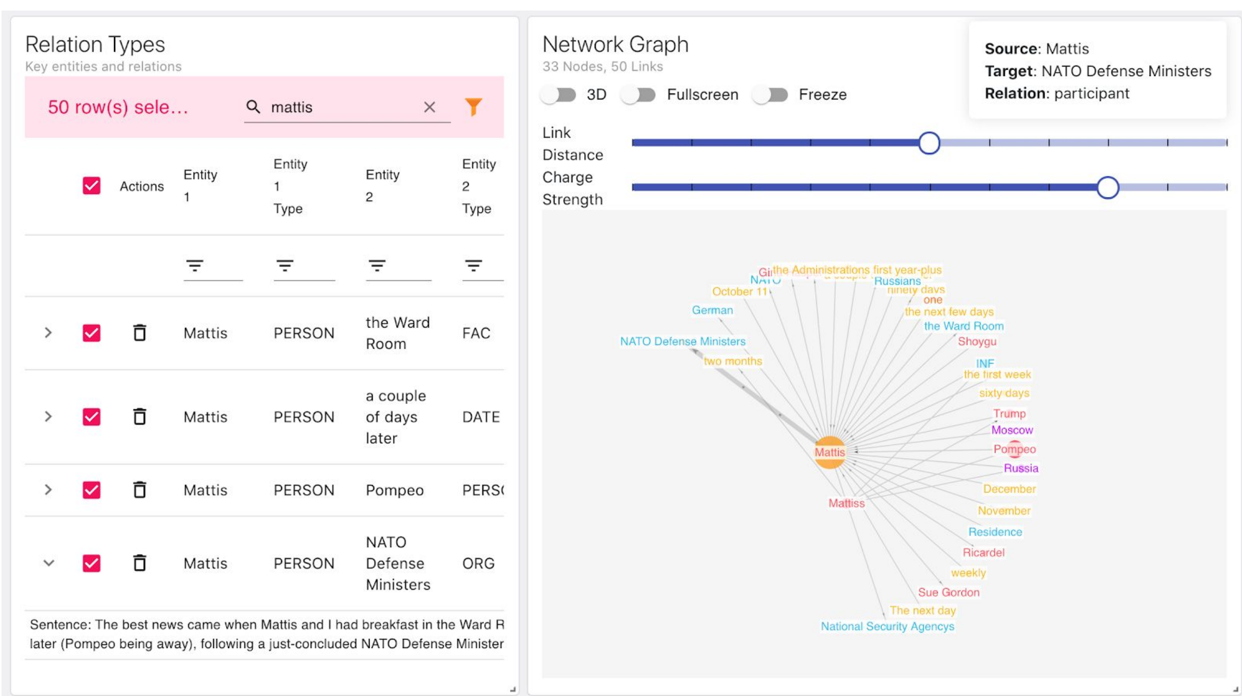
PLACEHOLDER

Putin agreed that he had acknowledged the China issue but said he had not mentioned wishing to pull out of the INF, agreeing with my point that Russia and the United States were the only two countries bound by the treaty. Why go through the agony of renegotiating New START, adding, for example, reductions or limitations on tactical nuclear weapons, which were of major importance to the US given the large number of such weapons Russia had?5 In response to Putin's questions, I said we had no intention of withdrawing from New START, but we were also essentially certain not to allow it simply to be extended for five years as Russia was asking (along with almost all Senate Democrats). We discussed explaining this to NATO allies, and Trump proposed saying, "At the request of Germany and others, we will terminate the INF Treaty sixty days from now." It struck me he still didn't appreciate the 180-day clock that had to run before withdrawal actually took effect, but it was too late by then to reopen the discussion.

Narrowing our analysis of the chapter to the entity Mattis, we can see that the entity appears 33 times in the chapter. We can simply do a search of the entity in the NER table and the NER text will highlight where exactly the entity appears in the text.



The relation table can also be used to filter the relations where the entity Mattis appears and display them in the network graph as shown below. From the graph, we can possibly infer that Mattis could have participated as a NATO Defense Minister and have been involved in meetings about Russia. We can also click on a row in the relation table to see what the exact sentence is where the relation appeared. Similar to the NER table, we can also delete that particular relation if it is inaccurate.



APPENDIX I. NER Models

Existing NER libraries

NER System	URL
StanfordCoreNLP	https://stanfordnlp.github.io/CoreNLP/
OSU Twitter NLP	https://github.com/aritter/twitter_nlp
Illinois NLP	http://cogcomp.org/page/software/
NeuroNER	http://neuroner.com/
NERsuite	http://nersuite.nlplab.org/
Polyglot	https://polyglot.readthedocs.io
Gimli	http://bioinformatics.ua.pt/gimli
spaCy	https://spacy.io/api/entityrecognizer
NLTK	https://www.nltk.org
OpenNLP	https://opennlp.apache.org/
LingPipe	http://alias-i.com/lingpipe-3.9.3/
AllenNLP	https://demo.allennlp.org/
IBM Watson	https://natural-language-understanding-demo.ng.bluemix.net
FG-NER	https://fgner.alt.ai/extractor/
Intellexer	http://demo.intellexer.com/
Repustate	https://repustate.com/named-entity-recognition-api-demo
AYLIEN	https://developer.aylien.com/text-api-demo
Dandelion API	https://dandelion.eu/semantic-text/entity-extraction-demo
displaCy	https://explosion.ai/demos/displacy-ent
ParallelDots	https://www.paralleldots.com/named-entity-recognition
TextRazor	https://www.textrazor.com/named_entity_recognition

Performance of existing NER models

Model	F1	Paper / Source
Flair embeddings (Akbik et al., 2018)	89.71	Contextual String Embeddings for Sequence Labeling
CVT + Multi-Task (Clark et al., 2018)	88.81	Semi-Supervised Sequence Modeling with Cross-View Training
Bi-LSTM-CRF + Lexical Features (Ghaddar and Langlais 2018)	87.95	Robust Lexical Features for Improved Neural Network Named-Entity Recognition
BiLSTM-CRF (Strubell et al, 2017)	86.99	Fast and Accurate Entity Recognition with Iterated Dilated Convolutions
Iterated Dilated CNN (Strubell et al, 2017)	86.84	Fast and Accurate Entity Recognition with Iterated Dilated Convolutions
AllenNLP Bert Implementation (Shi et al, 2019)	86.49	Semantic Role Labelling
Chiu and Nichols (2016)	86.28	Named entity recognition with bidirectional LSTM-CNNs
spaCy en_core_web_lg v2.0.0a3	85.85	NER Accuracy (OntoNotes5.0)
Joint Model (Durrett and Klein 2014)	84.04	A Joint Model for Entity Analysis: Coreference, Typing, and Linking
Averaged Perceptron (Ratinov and Roth 2009)	83.45	Design Challenges and Misconceptions in Named Entity Recognition (These scores reported in (Durrett and Klein 2014))

APPENDIX II. Relation Table (Wikidata)

QID	Relation Name	Description
P931	place served by transport hub	territorial entity or entities served by this transport hub (airport, train station, etc.)
P4552	mountain range	range or subrange to which the geographical item belongs
P140	religion	religion of a person, organization or religious building, or associated with this subject
P1923	participating team	Like 'Participant' (P710) but for teams. For an event like a cycle race or a football match you can use this property to list the teams and P710 to list the individuals (with 'member of sports team' (P54)' as a qualifier for the individuals)
P150	contains administrative territorial entity	(list of) direct subdivisions of an administrative territorial entity
P6	head of government	head of the executive power of this town, city, municipality, state, country, or other governmental body
P27	country of citizenship	the object is a country that recognizes the subject as its citizen
P449	original network	network(s) the radio or television show was originally aired on, including
P1435	heritage designation	heritage designation of a cultural or natural site
P175	performer	performer involved in the performance or the recording of a musical work
P1344	participant of	event a person or an organization was/is a participant in, inverse of P710 or P1923
P39	position held	subject currently or formerly holds the object position or public office
P527	has part	part of this subject; inverse property of 'part of' (P361). See also 'has parts of the class' (P2670).
P740	location of formation	location where a group or organization was formed

P706	located on terrain feature	located on the specified landform. Should not be used when the value is only political/administrative (P131) or a mountain range (P4552).
P84	architect	person or architectural firm that designed this building
P495	country of origin	country of origin of this item (creative work, food, phrase, product, etc.)
P123	publisher	organization or person responsible for publishing books, periodicals, games or software
P57	director	director(s) of film, TV-series, stageplay, video game or similar
P22	father	male parent of the subject. For stepfather, use 'stepparent' (P3448)
P178	developer	organisation or person that developed the item
P241	military branch	branch to which this military unit, award, office, or person belongs, e.g. Royal Navy
P403	mouth of the watercourse	the body of water to which the watercourse drains
P1411	nominated for	award nomination received by a person, organisation or creative work (inspired from 'award received' (Property:P166))
P135	movement	literary, artistic, scientific or philosophical movement associated with this person or work
P991	successful candidate	person(s) elected after the election
P156	followed by	immediately following item in a series of which the subject is a part [if the subject has been replaced, e.g. political offices, use 'replaced by' (P1366)]
P176	manufacturer	manufacturer or producer of this product
P31	instance of	that class of which this subject is a particular example and member (subject typically an individual member with a proper name label); different from P279; using this property as a qualifier is deprecated—use P2868 or P3831 instead
P1877	after a work by	artist whose work strongly inspired/ was copied in this item
P102	member of political party	the political party of which this politician is or has been a member

P1408	licensed to broadcast to	place that a radio/TV station is licensed/required to broadcast to
P159	headquarters location	specific location where an organization's headquarters is or has been situated. Inverse property of 'occupant' (P466).
P3373	sibling	the subject has the object as their sibling (brother, sister, etc.). Use 'relative' (P1038) for siblings-in-law (brother-in-law, sister-in-law, etc.) and step-siblings (step-brothers, step-sisters, etc.)
P1303	instrument	musical instrument that a person plays
P17	country	sovereign state of this item; don't use on humans
P106	occupation	occupation of a person; see also 'field of work' (Property:P101), 'position held' (Property:P39)
P551	residence	the place where the person is or has been, resident
P937	work location	location where persons were active
P355	subsidiary	subsidiary of a company or organization, opposite of parent organization (P749)
P710	participant	person, group of people or organization (object) that actively takes/took part in an event or process (subject). Preferably qualify with 'object has role' (P3831). Use P1923 for participants that are teams.
P137	operator	person, profession, or organization that operates the equipment, facility, or service; use country for diplomatic missions
P674	characters	characters which appear in this item (like plays, operas, operettas, books, comics, films, TV series, video games)
P466	occupant	a person or organization occupying property
P136	genre	creative work's genre or an artist's field of work (P101). Use main subject (P921) to relate creative works to their topic
P306	operating system	operating system (OS) on which a software works or the OS installed on hardware
P127	owned by	owner of the subject
P400	platform	platform for which a work was developed or released, or the specific platform version of a software product

P974	tributary	stream or river that flows into this main stem (or parent) river
P1346	winner	winner of an event - do not use for awards (use P166 instead), nor for wars or battles
P460	said to be the same as	this item is said to be the same as that item, but the statement is disputed
P86	composer	person(s) who wrote the music [for lyricist, use 'lyrics by' (P676)]
P118	league	league in which team or player plays or has played in
P264	record label	brand and trademark associated with the marketing of subject music recordings and music videos
P750	distributor	distributor of a creative work; distributor for a record label
P58	screenwriter	person(s) who wrote the script for subject item
P3450	sports season of league or competition	property that shows the competition of which the item is a season. Use P5138 for 'season of club or team'.
P105	taxon rank	level in a taxonomic hierarchy
P276	location	location of the item, physical object or event is within. In case of an administrative entity use P131. In case of a distinct terrain feature use P706.
P101	field of work	specialization of a person or organization; see P106 for the occupation
P407	language of work or name	language associated with this creative work (such as books, shows, songs, or websites) or a name (for persons use P103 and P1412)
P1001	applies to jurisdiction	the item (an institution, law, public office ...) or statement belongs to or has power over or applies to the value (a territorial jurisdiction: a country, state, municipality, ...)
P800	notable work	notable scientific, artistic or literary work, or other work of significance among subject's works
P131	located in the administrative territorial entity	the item is located on the territory of the following administrative entity. Use P276 (location) for specifying the location of non-administrative places and for items about events

P177	crosses	obstacle (body of water, road, ...) which this bridge crosses over or this tunnel goes under
P364	original language of film or TV show	language in which a film or a performance work was originally created. Deprecated for written works; use P407 ('language of work or name') instead.
P2094	competition class	official classification by a regulating body under which the subject (events, teams, participants, or equipment) qualifies for inclusion
P361	part of	object of which the subject is a part (it's not useful to link objects which are themselves parts of other objects already listed as parts of the subject). Inverse property of 'has part' (P527, see also 'has parts of the class' (P2670)).
P641	sport	sport in which the subject participates or belongs to
P59	constellation	the area of the celestial sphere of which the subject is a part (from a scientific standpoint, not an astrological one)
P413	position played on team / speciality	position or specialism of a player on a team, e.g. Small Forward
P206	located in or next to body of water	sea, lake or river
P412	voice type	person's voice type. expected values: soprano, mezzo-soprano, contralto, countertenor, tenor, baritone, bass (and derivatives)
P155	follows	immediately prior item in a series of which the subject is a part [if the subject has replaced the preceding item, e.g. political offices, use 'replaces' (P1365)]
P26	spouse	the subject has the object as their spouse (husband, wife, partner, etc.). Use 'unmarried partner' (P451) for non-married companions
P410	military rank	military rank achieved by a person (should usually have a 'start date' qualifier), or military rank associated with a position
P25	mother	female parent of the subject. For stepmother, use 'stepparent' (P3448)
P463	member of	organization or club to which the subject belongs. Do not use for membership in ethnic or social groups, nor for holding a position

		such as a member of parliament (use P39 for that).
P40	child	subject has the object in their family as their offspring son or daughter (independently of their age)
P921	main subject	primary topic of a work (see also P180: depicts)