Project proposal: Document Characterisation

Background & Motivation

At Adarga, we ingest raw, unstructured data from a wide range of sources. This data is passed through a suite of algorithms enriching and extracting information through its progression. Such enrichments include:

- Named Entity Recognition (NER)
- Event Extraction (EE)
- Event Coreference Resolution (ECR)
- Relationship Extraction (ER)

One of the problems we face in the data team is answering the question of: *How can we efficiently calibrate our algorithms with respect to the document being ingested?*

Which leads to the question: *What are the types of document we ingest and how do we discern between them?*

Through multiple iterations with our Subject Matter Experts (SME) we define the documents we ingest as belonging to one of the following categories:

- Situation Report
- Incident Report
- Profile Report
- Analytical Report

These categories will be introduced in the below sections.

Clarifying Notes (from meeting on 25/11/20):

Input Data

- Data will be provided as a CSV file containing the URL for each report document and a class label
- Students will have responsibility for how they ingest/scrape the data. Could use something like BeautifulSoup, or even manually extracting into a .txt file.
- It's advised that students don't rely on using website metadata as features for their classifiers (this is likely to be a strong feature – but would mask linguistic features).

Reporting Results

- Please report accuracy, recall and F1 for each class
- Provide a confusion matrix showing classification into the 4 classes
- Any other metrics you think will be interesting!

Project Aims

Exploration

Given a small sample of labelled data (equal number of samples per category), perform a data exploration focusing primarily (but not limited to) the syntactical structure of each of the categories. Are there any distinct patterns/properties observed unique to each of the document categories?

Resulting findings should be compiled into a presentable format (such as a Jupyter notebook) fit with details/visualisations deemed informative.

Modelling

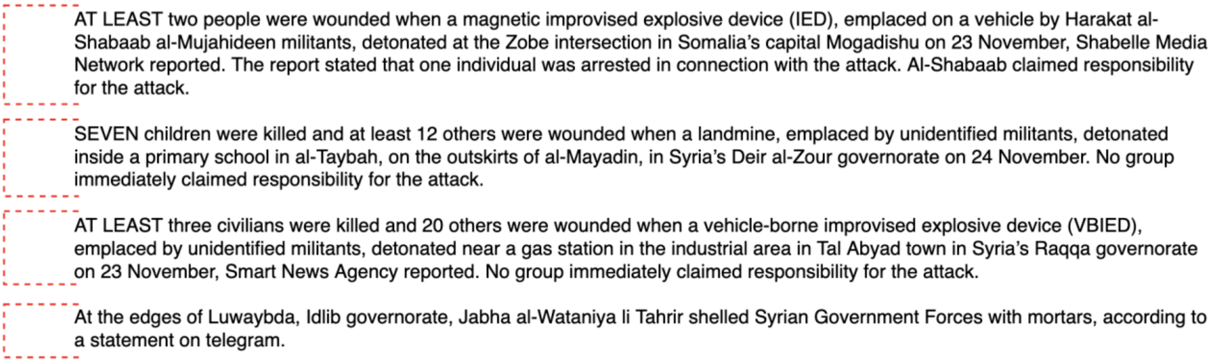
Given a larger sample of labelled data (approx. 400), formulate a supervised-learning problem tasked with document-type classification. Feature engineering efforts will be directly informed by your findings in the exploration portion of this project.

Document Types

Situation Report

These documents are made up of multiple short paragraphs, each describing unique event instances. The document in its entirety are based around a common focal point (domains, geolocations).

Let's look at an example situation report:



AT LEAST two people were wounded when a magnetic improvised explosive device (IED), emplaced on a vehicle by Harakat al-Shabaab al-Mujahideen militants, detonated at the Zobe intersection in Somalia's capital Mogadishu on 23 November, Shabelle Media Network reported. The report stated that one individual was arrested in connection with the attack. Al-Shabaab claimed responsibility for the attack.

SEVEN children were killed and at least 12 others were wounded when a landmine, emplaced by unidentified militants, detonated inside a primary school in al-Taybah, on the outskirts of al-Mayadin, in Syria's Deir al-Zour governorate on 24 November. No group immediately claimed responsibility for the attack.

AT LEAST three civilians were killed and 20 others were wounded when a vehicle-borne improvised explosive device (VBIED), emplaced by unidentified militants, detonated near a gas station in the industrial area in Tal Abyad town in Syria's Raqqa governorate on 23 November, Smart News Agency reported. No group immediately claimed responsibility for the attack.

At the edges of Luwaybda, Idlib governorate, Jabha al-Wataniya li Tahrir shelled Syrian Government Forces with mortars, according to a statement on telegram.

This report is based around the common focal point of attacks in the Eastern world. We've used dashed red boxes to highlight the individual event instance components.

In this type of document, we are more likely to see:

- Repetition in syntactical patterns. Each of the above components talk about different event instances but follow a similar structure in how the sentences are constructed.
- Entity mentions to be tightly distributed around specific regions of the document. Given that each component talks about a unique event instance, it is unlikely that we will see the same entities consistently mentioned in the first and last components.

- A larger number of entities with fewer mentions of each.

Incident Report

These reports typically focus around a single event instance. These reports contain multiple event mention instances describing various meta-information associated with a single (global) event instance.

These types of reports are like that of a **Profile Report** in the sense that they are specific, but incident reports tend to focus primarily on an event instance whereas a **Profile Report** is focused on an entity.

Example incident report:

> Armed officers swooped on Folkestone earlier today following a shooting. A man is now in custody following the incident, which happened close to Morrisons superstore, in Cheriton Road. Officers were called to the scene at 10.30am after reports a man was carrying a weapon. Shots were reportedly fired and a car was damaged, with 'two indents' left in its bodywork.

This example incident report describes a single event instance involving police officers and a shooting.

In this type of document, we are more likely to see:

- A fewer number of entities compared to that of a **Situation Report** (with respect to document size).
- Entity mentions more evenly distributed across the document.
- Contain a larger number of event mentions than that of a **Profile Report**.
- Less syntactical repetition than that of a **Situation Report**.

Profile Report

Profile report are typically focused around a single entity. The style of writing is descriptive in nature and is likely to contain a small number of entities.

Example profile report:

> Cemil Bayik is an Executive Committee Member, founding member and senior leader of the PKK. Bayik is also designated by the Department of the Treasury.
>
> The Kurdistan Workers Party (PKK), also known as Kongra-Gel is a regionally active terrorist organization and a U.S. designated Foreign Terrorist Organization (FTO).

In this type of document, we are more likely to see:

- Small set of entities.
- Small number of events.
- Mentions of the root entity (the entity being profiled) is likely to be distributed across the document.

- Broad date ranges than that of an **Incident/Situation Report.** (Profiles are not anchored at a given time whereas events are).

Analytical Report

These reports tend to be more detailed than an **Incident Report** in that these reports tends to answer the question **why** as opposed to **what.** The reports are typically focused around facts and attempt to describe causalities of events.

Their choice of words tends to follow a more cautious tense. We are likely to observe words such as could, should, may. These reports are differentiable from **Incident/Situation Reports** easily by humans as they are more analytical and factual.

Example analytical report:

> The Islamic State continues to show very significant resilience inside Iraq, undertaking a surge in attack activities in the second half of 2019 and the first quarter of 2020. According to the authors' dataset, the number of reported Islamic State attacks increased from 1,470 in 2018 to 1,669 in 2019, with 566 reported attacks in the first quarter of 2020 alone.[1]
>
> These national-level figures, supported by detailed qualitative and province-by-province breakdowns in the following sections, paint a picture of a militant organization that is reestablishing itself in Iraq, possibly drawing (in the authors' assessment) on a cadre of experienced tactical leaders and bomb makers that returned from the Syrian battlefields in 2019.

In this style of document, we are more likely to see:

- High frequency of words such as "could", "should", "may".
- A larger use of vocabulary than that of a **Situation/Incident Report.**
- Statistics.