Topic Segmentation

Introduction to problem

- Segmenting a document into meaningful sections of text
- Example

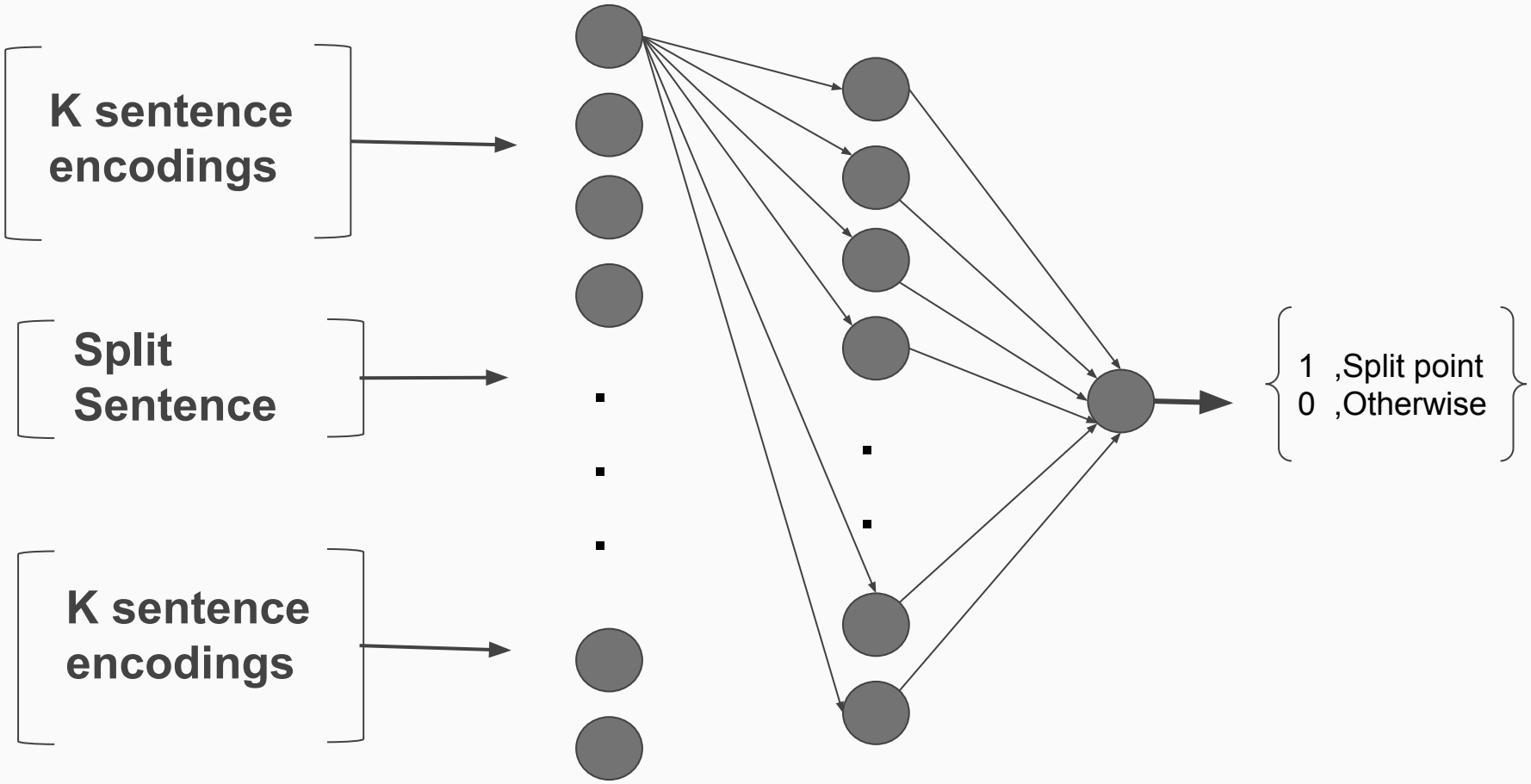
… The term deschooling was popularised by Ivan Illich, who argued that the school as an institution is dysfunctional for self-determined learning and serves the creation of a consumer society instead. Criticisms of anarchism include moral criticisms and pragmatic criticisms. Anarchism is often evaluated as ….

… The term deschooling was popularised by Ivan Illich, who argued that the school as an institution is dysfunctional for self-determined learning and serves the creation of a consumer society instead.

Criticisms of anarchism include moral criticisms and pragmatic criticisms. Anarchism is often ...

Approach & Motivation

- Scope for supervised approaches
  - Available methods use either **Bayes Formulation** or manual **Feature Extraction** (like Cue words, LDA topic modelling, cosine similarity etc)
  - All are **unsupervised** which allows them to add more information.
- Large dataset available for training (Wikipedia, 51.8 million documents)
- Useful in document summarization, question answering etc.


K sentence encodings
Split Sentence
K sentence encodings
1 ,Split point
0 ,Otherwise

Progress.

1. Dataset processing.

- Extracting 13GB wiki dump and getting sections from the documents using wikiextractor.
- 51.8 million documents
- Selecting around 500k samples, where each sample is a consecutive set of ~5 paragraphs.
- Noise Removal.

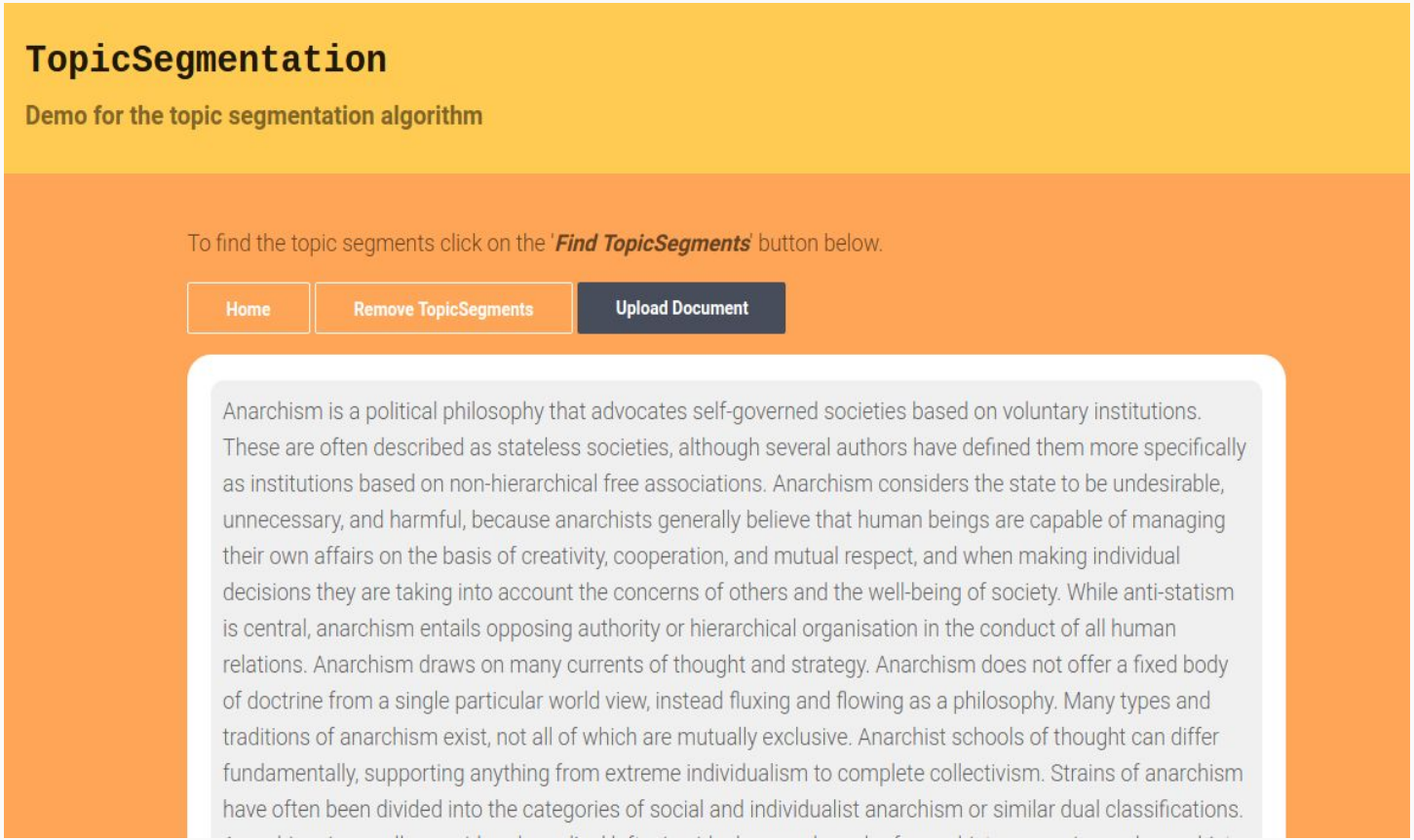
2. Sentence encoding

- A single Sample is a concatenation of sentence vectors.
- Use TF-IDF weighted normalised vectors for sentence encoding
- Move towards LSTM and memory based models.

3. Classification model

- Simple neural net using backpropagation for classification as split point.
- Switch to more non-linear model if needed.

Demo


TopicSegmentation
Demo for the topic segmentation algorithm
To find the topic segments click on the 'Find TopicSegments' button below.
Home
Remove TopicSegments
Upload Document
Anarchism is a political philosophy that advocates self-governed societies based on voluntary institutions. These are often described as stateless societies, although several authors have defined them more specifically as institutions based on non-hierarchical free associations. Anarchism considers the state to be undesirable, unnecessary, and harmful, because anarchists generally believe that human beings are capable of managing their own affairs on the basis of creativity, cooperation, and mutual respect, and when making individual decisions they are taking into account the concerns of others and the well-being of society. While anti-statism is central, anarchism entails opposing authority or hierarchical organisation in the conduct of all human relations. Anarchism draws on many currents of thought and strategy. Anarchism does not offer a fixed body of doctrine from a single particular world view, instead fluxing and flowing as a philosophy. Many types and traditions of anarchism exist, not all of which are mutually exclusive. Anarchist schools of thought can differ fundamentally, supporting anything from extreme individualism to complete collectivism. Strains of anarchism have often been divided into the categories of social and individualist anarchism or similar dual classifications.

Goal

1. Working supervised algorithm capable of segmenting a document.
2. Demonstration (publicly available)

Thank You