Module 1: Using natural language queries to get value from Knowledge Bases

***Section Description (from the outline):*** This section describes a number of natural language processing (NLP) techniques to get value from a knowledge base of articles. We’ll use content from Wikipedia and the gensim and NLTK open source libraries to build an LDA model that enables rich queries to be asked of Wikipedia content.

|  |
| --- |
| **Metadata**: Spot the problem, highlight it, and design the solution in 3 core steps  (To be covered in the video) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Video Number** | **Video Title** | **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| 1.1 | Course intro | TBC | TBC | TBC | TBC |

Metadata for Section 1

Section Description (from the outline): Describe what this section intends to cover- decided during outline discussion.

|  |
| --- |
| **Metadata**: Spot the problem, highlight it, and design the solution in 3 core steps  (To be covered in the video) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Video Number** | **Video Title** | **Problem / Solution (Not more than 50 words)** | **Step 1 (Not more than 10 words)** | **Step 2(Not more than 10 words)** | **Step 3(Not more than 10 words)** |
| 1.2 | Building a dataset of content from Wikipedia | Build a local dataset for building an NLP model against. How can we do this using Wikipedia content? | Create a local database using SQLite3 | Use Wikipedia API to download articles by category | Cleanse data to remove markup and store in database |
| 1.3 | Topic Modeling | What is topical modelling and how can we use it with large document sets | NLP dictionary, corpus, bag of words, lemma, topic modelling | LDA encodes the prevalent topics within a document | Document is distrbution of topics. Topics are distribution over words. |
| 1.4 | Using gensim and NLTK | How can we use gensim and NLTK to build NLP models over a document set so that we can find relevant and related documents to a user provided query | Pre-process text removing stopwords and lemmatize | Use gensim to create a dictionary and corpus | Build a model using gensim and LDA |
| 1.5 | Using gensim and NLTK | Worked NLP example using Python to build an LDA model over our document set, explore the LDA model and associated dictionary and corpus to understand how it works, and then query it to find relevant and related documents. | Pre-process text removing stopwords and lemmatize | Use gensim to create a dictionary and corpus | Build and query LDA model using gensim |
| 1.6 | Build a webservice | How can we take our model and use it in a more production-like environment, allowing remote clients to access it? | Use Python and Flask to build a web service | Load our LDA model, dictionary and corpus | Use our LDA model to find related documents |