**Title of Talk/Workshop**

Using NPL open source Tools, Framework and Libraries to develop a Document/text Similarity Solution

**Submission type \***

45 minutes talk

**Abstract**:

The need to develop robust document/text similarity measure solutions is an essential step for building applications such as Recommendation Systems, Search Engines, Information Retrieval Systems including other ML/AI applications such as News Aggregators (that require document clustering) or automated recruitment systems used to match CVs to job specification and so on. In general, text similarity is the measure of how words/tokens, tweets, phrases, sentences, paragraphs and entire documents are lexically and semantically close to each other. Texts/words are lexically similar if they have similar character sequence or structure and, are semantically similar if they have the same meaning, describe similar concepts and they are used in the same context.

This work will demonstrate a number of strategies for feature extraction i.e. transforming documents to numeric feature vectors. This transformation step is a prerequisite for computing the similarity between documents. Typically each strategy will involve 3 steps, namely: 1) the use of standard natural language pre-processing techniques to prepare/clean the documents, 2) the transformation of the document text into a vector of vectors. 3), calculation of document feature vectors using the [Cosine Similarity](https://en.wikipedia.org/wiki/Cosine_similarity) metric.

Strategies and associated ML/NLP libraries that will be presented are:

* document pre-processing using [NLTK library](https://www.nltk.org/)
* feature extraction using Term frequency – Inverse Document Frequency (TF-IDF) with the aid of the [Scikit-Learn library](https://scikit-learn.org/stable/)
* feature extraction using a pre-trained GloVe embedding with the aid of the [Spacy NLP Library](https://spacy.io/usage/vectors-similarity)
* feature extraction using a trained [Word2Vec](https://en.wikipedia.org/wiki/Word2vec) embedding from scratch with the aid of the [Gensim](https://radimrehurek.com/gensim/) NLP library
* feature extraction using a trained Doc2Vec embedding from scratch also with the aid of the Gensim library

Models developed from these strategies will be compared against each other for the [Kaggle competition](https://www.kaggle.com/) problem of identifying duplicate questions using the [Quora dataset](https://www.kaggle.com/c/quora-question-pairs/data). These models will also be used for computing the document similarity of the popular [NPL 20 News Group problem](https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html). **Work-in-progress** details of this project including my source code can be found here.

Can your talk be classified into any of the Themes below? \*

* Tools, frameworks & libraries

Audience Level \*

* Those who are at an intermediate level of understanding and experience with ML