Exercise 6.1 – Conceptual questions

Answer the following questions:

A) Define word embeddings.

Solution: Word embeddings are dense, distributed, fixed-length word vectors, built using word co-occurrence statistics as per the distributional hypothesis.

B) What is the basic assumption of word embeddings?

Solution:-Words can be sufficiently characterized by their context. Co occurrence statistics are enough to characterize each of the word.[Distributional Hypothesis]

C) Describe the difference between training word-based (such as GloVe [1]) and

character-based (such as Fasttext [2]) embedding models. What are advantages

and disadvantages of each type?

Solution:- GloVe model treats each word as an atomic entity and generates a vector based on co-occurrence of the word with other context words in the training corpus.

adv:- Good Word analogies performance. Disadv:- Cant handle rare words

Fasttext which is essentially an extension of word2vec model, treats each word as composed of character ngrams. So the vector for a word is made of the sum of this character n grams.

adv:-Can handle rare words and out-of-Vocabulary words disadv:- lot of memory. Can even accept gibberish terms

D) Which model is Law2Vec [3] trained on? What are the steps in model training?

solution:- Law2Vec is trained on a large number of legal

corpora from various public legal sources. The list comprised the following:

• 53,000 pieces of the UK legislation.

• 62,000 pieces of the European legislation.

• 5500 pieces of the Canadian legislation.

• 1150 pieces of the Australian legislation.

• 800 pieces of the English-translated legislation from EU countries.

• 780 pieces of the English-translated legislation from Japanese.

• 68 bound volumes of the US Supreme Court decisions from 1998 to 2017.

• 54 titles of the most recently updated U.S. Code.

The corpus sums up to a total of 123,066 documents which consists of 492M individual

words (tokens), including punctuation marks and numbers.

The corpus was preprocessed to discard non-UTF8 encoded characters and treat dash-separated words due to different layout styles (e.g., text from PDF documents). The text was sentence splited using the nltk library to provide the best possible input for the

models. All words were lower-cased and all numerical digits where replaced by the character ‘D’ in order to normalize numerical values. The model was trained on word2vec skip gram.

E) What do Chalkidis and Kampas [3] do against the Out-of-Vocabulary Problem in

their embeddings?

solution:- They ignored them. They considered that the Missing words (OOV) is not of concern in most legal-related tasks, as legislators, lawyers and other legal professionals articulate in high quality standards.

F) What is the difference between word embeddings and their contextual variants?

solution:-

